Science of the Total Environment 922 (2024) 171158

Contents lists available at ScienceDirect

Science of the Total Environment

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

Topsoil porosity prediction across habitats at large scales using environmental variables

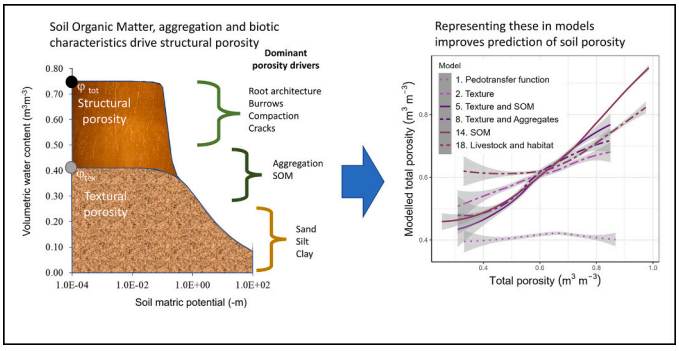
A. Thomas [a,*], F. Seaton [b], E. Dhiedt [a], B.J. Cosby [a], C. Feeney [a], I. Lebron [a], L. Maskell [b], C. Wood [b], S. Reinsch [a], B.A. Emmett [a], D.A. Robinson [a]

[a] UK Centre for Ecology and Hydrology, Environment Centre Wales, Bangor, UK
[b] UK Centre for Ecology and Hydrology, Library Ave, Bailrigg, Lancaster, UK

HIGHLIGHTS

- Soil porosity is a fundamental environmental property, often represented as static.
- We explore relative contribution of different dynamic and static predictors.
- Machine learning and statistical models were used to assess predictors of porosity.
- Habitat and soil organic matter are promising dynamic predictors.
- Dynamic estimates of soil porosity could improve feedbacks in Earth System Models.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Manuel Esteban Lucas-Borja

Keywords:
Soil porosity
Soil carbon
Earth system model
Climate change
Land use change
Soil compaction

ABSTRACT

Soil porosity and its reciprocal bulk density are important environmental state variables that enable modelers to represent hydraulic function and carbon storage. Biotic effects and their 'dynamic' influence on such state variables remain largely unknown for larger scales and may result in important, yet poorly quantified environmental feedbacks. Existing representation of hydraulic function is often invariant to environmental change and may be poor in some systems, particularly non-arable soils. Here we assess predictors of total porosity across two comprehensive national topsoil (0-15 cm) data sets, covering the full range of soil organic matter (SOM) and habitats ($n = 1385$ & $n = 2570$), using generalized additive mixed models and machine learning. Novel aspects of this work include the testing of metrics on aggregate size and livestock density alongside a range of different particle size distribution metrics. We demonstrate that porosity trends in Great Britain are dominated by biotic metrics, soil carbon and land use. Incorporating these variables into porosity prediction improves performance, paving the way for new dynamic calculation of porosity using surrogate measures with remote sensing, which may help improve prediction in data sparse regions of the world. Moreover, dynamic calculation of porosity could support representation of feedbacks in environmental and Earth System Models. Representing the

* Corresponding author.
E-mail address: athomas@ceh.ac.uk (A. Thomas).

https://doi.org/10.1016/j.scitotenv.2024.171158
Received 2 November 2023; Received in revised form 19 February 2024; Accepted 19 February 2024
Available online 20 February 2024

hydrological feedbacks from changes in structural porosity also requires data and models at appropriate spatial scales to capture conditions leading to near-saturated soil conditions.

Classification.

Environmental Sciences.

1. Introduction

Water and carbon storage and exchange at the earth's surface represent two linked environmental cycles that are central to understanding earth system dynamics and feedbacks to global environmental change. Soil structural characteristics are important mediators of these cycles, yet have been regularly overlooked in Earth System Models (ESMs) (Fatichi et al., 2020), limiting our ability to quantify human impacts on the Earth System. Soil porosity contributes to hydraulic function, and appropriate representation is important for partitioning of precipitation between run-off and infiltration (Jarvis et al., 2013). This in turn will have implications for modelling other land degradation phenomena such as soil erosion (Borrelli et al., 2021) and the magnitude and persistence of heatwaves (Lorenz et al., 2010). Similarly, accurate representation of its reciprocal, bulk density, is essential for estimating soil carbon stocks and other components of soil health (Walter et al., 2016; Panagos et al., 2024).

Pedo-transfer functions (PTFs) are commonly used in models to predict hydraulic behavior and parameters such as soil porosity. Development and implementation of PTFs are widely discussed in the literature (e.g. Pachepsky et al., 1999; Wösten et al., 2001; Rawls et al., 2004). Critically, Weihermüller et al. (2021) highlight the sensitivity of ESMs to the choice of PTF, which often represents a greater source of error than uncertainty in input parameters. Concerns of data bias when using PTFs to estimate soil hydraulic properties have led to use of alternative approaches, for example interpolation from ground-based observations was used for SoilGrids (Turek et al., 2023).

Jarvis et al. (2013) challenged current assumptions for texture based PTFs when they found that soil hydraulic conductivity was more closely related to land use, bulk density, and soil carbon than to soil texture, the key component of many PTFs. Early PTFs such as those derived in Rosetta used texture (Rosetta model versions H1 and H2) or a combination of texture and bulk density (Rosetta model versions H3-H5) (Schaap et al., 2001; Zhang and Schaap, 2017). Major improvement is observable when bulk density is incorporated. The issue is illustrated in Fig. 1 which shows the textural and structural contributions to a water retention curve, one of the fundamental descriptors of soil hydraulic function. In agricultural soils, the structural component can be much reduced by tillage for example (e.g. Bronick and Lal, 2005, Jarvis, 2007), hence, the textural component is dominant and may describe hydraulic function adequately. Conversely in undisturbed soils under grass, shrubs and trees there may be substantial development of structural porosity such as macropores from roots (e.g. Bonetti et al., 2021). Since the historical development of PTF's is based on data with a sampling bias toward agricultural, particularly arable soils that is endemic in global databases (Rahmati et al., 2018), the structural component is largely omitted. Wösten et al. (2001) call for updates to PTFs and increased availability of appropriately structured databases to support this, and Rabot et al. (2018) further advocate for open data from soil structure imaging to support improved prediction of soil function.

With regard to the textural component of porosity, Robinson et al. (2022a) discuss previous work on particle size distribution and influence on geometric packing, leading to the emergence of bulk density and porosity. Parameters representing the particle size distribution (PSD)

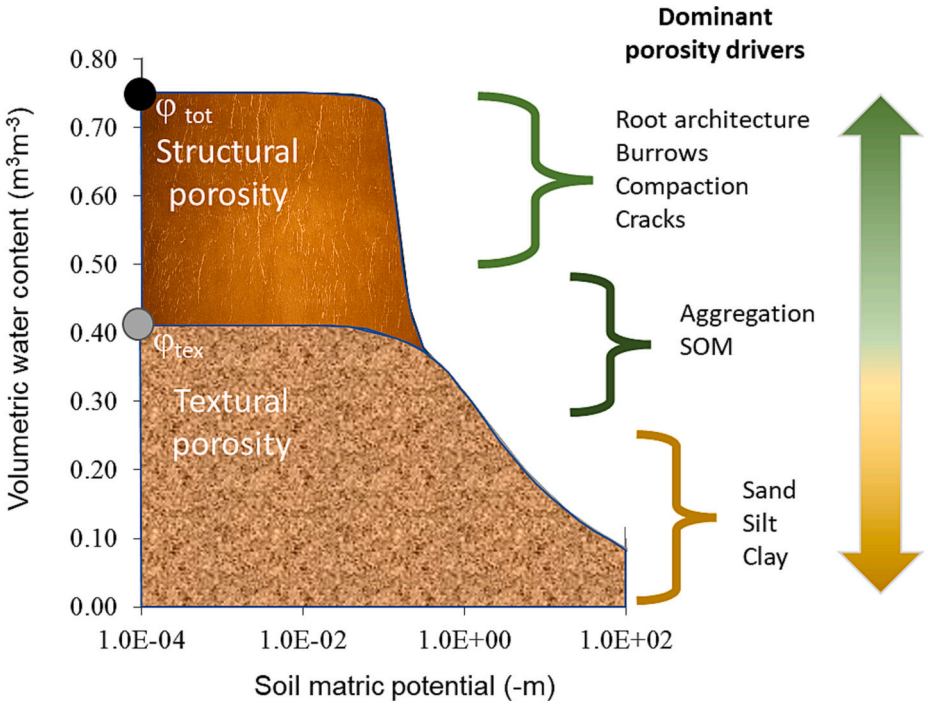


Fig. 1. A schematic diagram of a water retention curve where the volumetric water content is a function of the soil matric potential in meters (−m). The grey circle marks the textural porosity (φtex) often determined using a pedo-transfer function (PTF), whilst the black circle represents the total porosity (φtot), that of the textural and structural porosity combined. The green and yellow arrow indicates change in dominant drivers of porosity when moving between soils where porosity is primarily structural or textural. SOM = Soil organic matter content.

may capture these effects better than separate variables representing mass in different size classes, since greater heterogeneity would allow for denser packing (e.g. Martín et al., 2017). A range of mathematical approaches have been explored for representing PSD curves, including information entropy, hyperbolic and power law equations, logarithm equations and fractal approaches (summarized in Bayat et al., 2015).

For the structural component of porosity, land use and management are likely to play a key role in the pore space evolution leading to dynamic changes in porosity and environmental function of soils (Robinson et al., 2022b). Dynamic changes in porosity are also common on short temporal scales in managed soils, for example in response to tillage and subsequent precipitation events or other compacting forces (e.g. Sandin et al., 2017). Furthermore, evidence from soil data across the USA predicts changes on decadal time scales, with unknown consequences for the hydrological cycle (Hirmas et al., 2018). However, PTFs are generally modelled as a static property of the soil. Incorporating influence from biotic factors such as vegetation could support representation of a dynamic component of porosity.

In order to develop a dynamic model of soil structure to support ESMs, understanding dynamic drivers of gross measures such as porosity and its reciprocal, bulk density, is a desirable starting point. Three important challenges are thus identified that must be addressed in order to realize a step change toward a dynamic model of soil structure:

(i) Collect unbiased datasets from across habitats to assess soil structural properties and their state and change.
(ii) Narrow the uncertainty for porosity prediction by understanding the factors that contribute to variation across habitats.
(iii) Identify dynamic indicators and co-variates that can be assessed, using remote sensing for example, to track the state and change of soil structure at national scale.

Here we attempt to address these challenges using national unbiased stratified random data sets to model total porosity. These data sets constitute a novel aspect of our study, spanning a gradient of climate, parent material and vegetation across the national scale. Tifafi et al. (2018) showed that the steepest rate of change in soil carbon occurs between the latitudes of 50–60 degrees in the northern and southern hemispheres. Great Britain straddles this same latitudinal gradient and is thus an ideal study site, with soils ranging from close to 0 to 100 % soil organic matter (SOM).

We construct a series of machine learning (ML) and generalized additive mixed models (GAMMs) to determine how land use, SOM and soil aggregation contribute to topsoil (0-15 cm) total porosity at a national scale. We use the models to assess our hypothesis that key factors determining the structural porosity are SOM, aggregation and the root architecture of vegetation; we omit physical cracking in this study and focus on biotic induced structural change which will be dominant along our SOM gradient. We also include soil texture metrics where available to represent the textural component of porosity. Through this approach, we also aim to identify indicators that could lead to a dynamic model of soil structural change and impacts on structural and total porosity.

2. Materials and methods

2.1. Data

2.1.1. National scale field survey data

2.1.1.1. Study area. The research used two study areas. The Countryside Survey (CS) data was collected from across Great Britain (GB; England, Scotland and Wales) during 2007 (CS2007) and 2019 (CS2019), whilst data from the Glastir Modelling and Evaluation Programme (GMEP) was collected between 2013 and 2016 from across Wales. The GMEP data provide a semi-independent comparable data set, covering a subset of the geographical range, with which to construct comparable models or assess the predictive performance of models constructed on the full GB dataset. All methodological details for soil sample analyses can be found in the supporting information of the respective data sets described below. Only basic information is provided here. Summary statistics and number of samples in each habitat class are presented for each data set in the supplementary material (Table S26), along with maps of sampling locations (Fig. S9). The sites have assessment of vegetation, SOM and bulk density. The CS data collected in 2019 also include measurement of soil texture and aggregate size distribution using laser diffractometry to obtain spectrums of soil physical characteristics. These texture data are also available for the GMEP data. The resulting data are uniquely positioned to assess the state and change of soils along an environmental, climate, habitat and SOM gradient.

2.1.1.2. Countryside survey data. Topsoil samples (0–15 cm) were collected from across Great Britain based on a stratified random design using ITE Land Class (Bunce et al., 1996). Subsets with the variables required for our models ($n = 2570$ for 2007 and $n = 287$ for 2019) are presented in this work (Emmett et al., 2016). SOM was determined using loss on ignition and bulk density was determined on oven dry fine earth (<2 mm). All methods were based on the Countryside Survey as detailed in the supporting information in Emmett et al. (2016).

2.1.1.3. GMEP data. Topsoil samples (0–15 cm) were collected from across Wales based on a stratified random design using Land Class, a combination of parent material, climate and relief (Bunce et al., 1996), to stratify. $N = 1385$ measurements are presented in this work (Robinson et al., 2019). All methods were based on, and are compatible with, the Countryside Survey data.

Bulk density can be determined for the total soil including stones, or for the fine earth fraction removing the stones and adjusting the volume accordingly (Grossman and Reinsch, 2002; Page-Dumroese et al., 1999). Here, the bulk density was determined for the fine earth fraction following the laboratory procedure described in the supplementary material of Emmett et al. (2008), on soil cores extracted using a stainless-steel corer 5 cm diameter and 15 cm deep. Total porosity for the fine earth fraction was then determined based on the oven dry bulk density and the particle density. For this study, we define total porosity as the pore space that can be measured within the soil core contributed by the physical arrangement of individual soil particles (textural) and their aggregation (structural). Particle density was determined according to the mixing model approach of Ruehlmann (2020) using SOM proportion and assumed particle density of SOM 1.4 g cm^{-3} and mineral particle density 2.7 g cm^{-3}. We acknowledge that these soil cores cannot capture larger macropore features such as larger animal burrows, but the calculated metrics may still be considered largely representative of total porosity.

We analyzed all mineral soils (SOM fraction <0.4) using laser granulometry to obtain a full particle size spectrum. Particle size analysis was undertaken using a laser granulometer for the GMEP data and the CS2019 data. Aggregate size distribution was also determined for the CS2019 data using laser granulometry. A Beckman Coulter LS13 320 laser diffraction particle size analyser (Beckman Coulter Inc.) was used. To evaluate the accuracy of the instrument we used different size standards: nominal 500 μm glass beads (Beckman Coulter Inc.) and nominal 15 μm Garnnet (Beckman Coulter Inc.). We also used sandy soil from Gleadthorpe (Cuckney, UK), clay soil from Brimstone (Denchworth, UK) and a silty soil from Rosemaud (Bromyard, UK) as soil standards. All three soils are well-characterised farm soils from ADAS Ltd. (Helsby, WA6 0AR). In addition, we used two well-characterised internal soil standards: Bangor standard 1 (BS1) and Bangor standard 3 (BS3) soils representative of North Wales soils, they are loam and silty clay loam respectively. The soil particle size data obtained were used to calculate distribution metrics.

2.1.2. Livestock density data for the survey locations

Livestock density maps were created at a 1 km resolution. Livestock data for sheep and cattle retrieved from EDINA agCensus for years nearest to 2007 and 2019 (EDINA, 2007a, 2007b, 2010, 2016, 2018, 2019; Stuart Neil, 2022) were converted to livestock units (LU) based on conversion factors described by Nix and Redman (2021). Livestock unit data were combined to a 1-km grid for each country using country boundaries (ONS, 2020) and converted to density per ha of grassland using the Land Cover Map (Morton et al., 2014, 2022a, 2022b, 2022c). Calculations were performed in R version 4.2.1. and ArcMap version 10.7.1.

2.1.3. Particle size distribution metrics

We calculated several different multifractal metrics for the particle size distribution by implementing the moment method as described in Salat et al. (2017) and previously demonstrated for the GMEP data (Seaton et al., 2020). In our models we used D0, D1, D2, Dα and Dfα. D0, D1 and D2 represent the Rényi dimension Dq (where q = 0,1,2) of the particle size distribution, which addresses how μ (multifractal measure) varies with ε (box size or aggregation of size classes) and q (range of moment orders) defined according to Eq. (1), or Eq. (2) where q = 1.

$$D_q = \frac{1}{q-1}\lim_{\varepsilon\to 0}\frac{log\mu(q,\varepsilon)}{log\varepsilon} \tag{1}$$

$$D_q = \lim_{\varepsilon\to 0}\frac{log\mu(q,\varepsilon)}{log\varepsilon} \tag{2}$$

where $\mu(q, \varepsilon)$ is defined according to Eq. (3) and p_i is the proportion of mass in the ith box of size ε.

$$\mu(q,\varepsilon) = \sum p_i^q \tag{3}$$

These multifractal Rényi dimension parameters generated describe different aspects of the distribution. The box-counting dimension, D0 has a maximum value of 1 when all subintervals are occupied at all scales, decreasing with increasing empty subintervals. The entropy dimension, D1, quantifies how Shannon entropy scales as ε tends to 0, to represent the disorder or heterogeneity of the distribution. D1 ranges from 1 for the most heterogeneous distribution to 0 for a homogenous distribution. The correlation dimension, D2, computes the correlation of measures contained in size ε (Posadas et al., 2001). For both the GMEP and CS data, D declined with increasing q. This indicates, as expected, that the soil particle size distribution does not follow a power law distribution, hence the multifractal approach implemented here is more appropriate than a single fractal model (Posadas et al., 2001).

Dα is the spectral width for a multi-fractal spectrum, which represents the heterogeneity of particle size distribution across the full fractal structure (Wang et al., 2015). This is calculated according to Eq. (4), where α is calculated according to Eq. (5) for q from −5 to 5.

$$D\alpha = max(\alpha(q)) - min(\alpha(q)) \tag{4}$$

$$\alpha(q) = \frac{d\tau(q)}{dq} \tag{5}$$

For the moment method, τ is the mass or correlation exponent of the qth order (Salat et al., 2017), and is calculated by Eq. (6).

$$\tau(q) = (q-1)Dq \tag{6}$$

Dfα reflects the multi-fractal spectrum shape feature (Wang et al., 2015), and is calculated by Eq. (7), where $f\alpha$ is calculated by Eq. (8).

$$Df\alpha = max(f\alpha(q)) - min(f\alpha(q)) \tag{7}$$

$$f(\alpha(q)) = \alpha(q)q - \tau(q) \tag{8}$$

Further particle size distribution metrics were calculated on the texture data aggregated into 8 classes from colloid to very coarse sand using the Krumbein phi scale. The classified texture data and the aggregate data were then used to calculate distribution metrics. Weibull or Rosin-Ramler is a continuous probability distribution, here the parameters describing shape and size of the fitted distribution were calculated according to Eq. (9) using the package "fitdistrplus" in R (Delignette-Muller and Dutang, 2015).

$$P(X > x) = 100e^{\left(\frac{x}{\beta}\right)^{\alpha}} \tag{9}$$

where P(X > x) is the percentage by weight of particles (or aggregates) greater than size x, and the PSD curve is then described by α as the size parameter and β as the shape parameter describing the spread of the distribution (Keller et al., 2010). Shannon information entropy is a measure of PSD or aggregate size distribution heterogeneity, according to eq. (10), calculated using the package "entropy" (Hausser and Strimmer, 2009) in R, using the Maximum Likelihood method.

$$H = -\sum_{i=1}^{k} p_i log p_i \tag{10}$$

2.2. Modelling

Models were constructed using machine learning and GAMMs. Where possible, models were constructed on the full GB dataset (CS2007 n = 2570) and tested on the Wales dataset (GMEP n = 1385) or the corresponding texture subsets (CS2019 n = 287, GMEP n = 728). Models were used first to explore the extent to which soil texture variables, including krumlin phi size classes and the various particle size distribution metrics described in Section 2.1, are predictors of total porosity. Further models were then constructed to explore our hypotheses that soil aggregates and SOM are important drivers of total porosity. Appropriate metrics were reported for assessment of the different model types. A novel aspect of the modelling was to test inclusion of habitat, as an example of variables that are increasingly available from earth observation data, which could support upscaling for national mapping and dynamic predictions for land surface modelling. As a comparison, we also include the Rosetta H1 PTF model https://www.handbook60.org/rosetta/ which predicts porosity as the average for that USDA texture class based on the sand silt and clay content (Schaap et al., 2001).

2.2.1. Machine learning algorithm

Machine learning (ML) was performed using conditional random forests to explore relative importance of different environmental variables in predicting topsoil total porosity. ML can tell us which variables are more important and can handle large numbers of variables. Conditional random forest ML as used here also enables comparison between factor and continuous variables. Random forest is a supervised learning algorithm, which generates multiple decision trees using bootstrap samples from the data, and at the same time computes estimates of variable importance (Breiman, 2001). These were used to plot ranked variable importance according to impacts on % Mean Square Error (MSE) of randomly permuting the variable. The approach was applied primarily to explore the relative importance of different predictors in the data; in particular, to enable us to compare the impact of a large number of variables on soil particle size. Therefore, we implemented the R function "cforest" in the package "partykit" (v1.2–20; Hothorn and Zeileis, 2015) which achieves improved performance for identifying variable importance, by using conditional inference trees as base learners, and calculating the conditional permutation importance (Strobl et al., 2008; Strobl et al., 2007). This reduces issues identified with traditional random forest approaches around correlated variables and comparison of categorical and continuous variables. However, the effects of correlation between predictor variables cannot entirely be removed (Strobl et al., 2008), hence we tested a two-step approach; first

constructing the model with all available variables, and then removing the least important correlated variables.

Reported model metrics of r^2 and RMSE reflect how well out-of-bag predictions explain the target variance of the training set. We have not assessed the performance of the ML models for predictive modelling because soil particle size data were only available for a subset of samples, and because the conditional inference tree approach requires bootstrap sampling without replacement, which should only be used for the evaluation of variable importance, not for predictive modelling. Hence, these models are primarily useful for understanding the relative importance of different drivers in our data, in particular for comparing habitat with the continuous variables.

2.2.2. Statistical modelling

We separately constructed statistical models to explore the nature of nonlinear relationships in the data and look for variations in these relationships between habitats. We used a mixed model structure (Generalized Additive Mixed Model, GAMM) to account for random factors. A factor identifying the 1-km square location (each of which contained up to 5 soil sampling locations) was included as the random factor. Due to the bimodal distribution of residuals in many of these models, Gaussian distribution was not always appropriate. In these cases, a Tweedie distribution was used with p assigned in preliminary model fitting using the "gam" function in the R package "mgcv" (v1.8–42; Wood, 2011). By fitting the Tweedie distribution *p* value to the model, the distribution of residuals can be more appropriately captured.

As part of model fitting, we fitted cubic regression splines, henceforth termed "smooths", to all continuous terms in the model, to both capture nonlinearity in the relationships and test deviation away from a constant zero effect. Smooths were applied via the "gamm" function in the 'mgcv' library (Wood, 2011). These smooths allow nonlinear variation of the coefficient applied to the predictor variable. The approach here also removes spurious predictor variables, by using a double penalty smoother which allows the penalized regression routine which selects for the "wiggliness" of the smooths to also shrink covariates out of the model entirely (per Marra and Wood, 2011). We did not include interaction terms in most of the models, in order to assess the marginal influence of individual covariates, however we did separately test the interaction between SOM and habitat, to explore the influence of habitat on the relationship between SOM and total porosity. Models were constructed separately on both the CS and GMEP datasets (where required variables were available in both datasets) and the CS models were tested on the GMEP dataset to assess performance on a separate sample.

3. Results and discussion

3.1. Evaluating PTFs and the need for a structural component parameter

Our analysis framework is developed around the concept of the soil water retention curve (Fig. 1). PTFs use soil survey data such as soil texture to predict water retention characteristics including what we term the textural porosity (Fig. 1). PTFs using only soil texture work reasonably well in unstructured mineral soils (Robinson et al., 2022a). However, unless they include bulk density, PTFs are currently poor at predicting what we term the total porosity which is a combination of the structural and textural porosity. Fig. 2 illustrates this using only the sand, silt and clay data from our CS dataset (subset of the 2019 data with texture $n = 287$) with the Rosetta H1 PTF model (Table 1 Model 1) to predict the saturated water content (PTF porosity). Fig. 2 compares these estimated soil porosity values with measured porosity (see 2.1.1), in order to illustrate the missing structural porosity component to values estimated from texture alone. All data should fall on the one-to-one line, but performance is very poor due to the absence of bulk density in this model. Drivers of soil structural porosity such as SOM are also omitted from this PTF. The data clearly show that as you move away from Arable and Horticulture systems, disagreement increases, suggesting that structural porosity, hence bulk density, becomes more important. Overall performance of the PTF was very poor in predicting soil porosity (r^2–3.7, RMSE 0.215), as might be expected of a model based on texture alone.

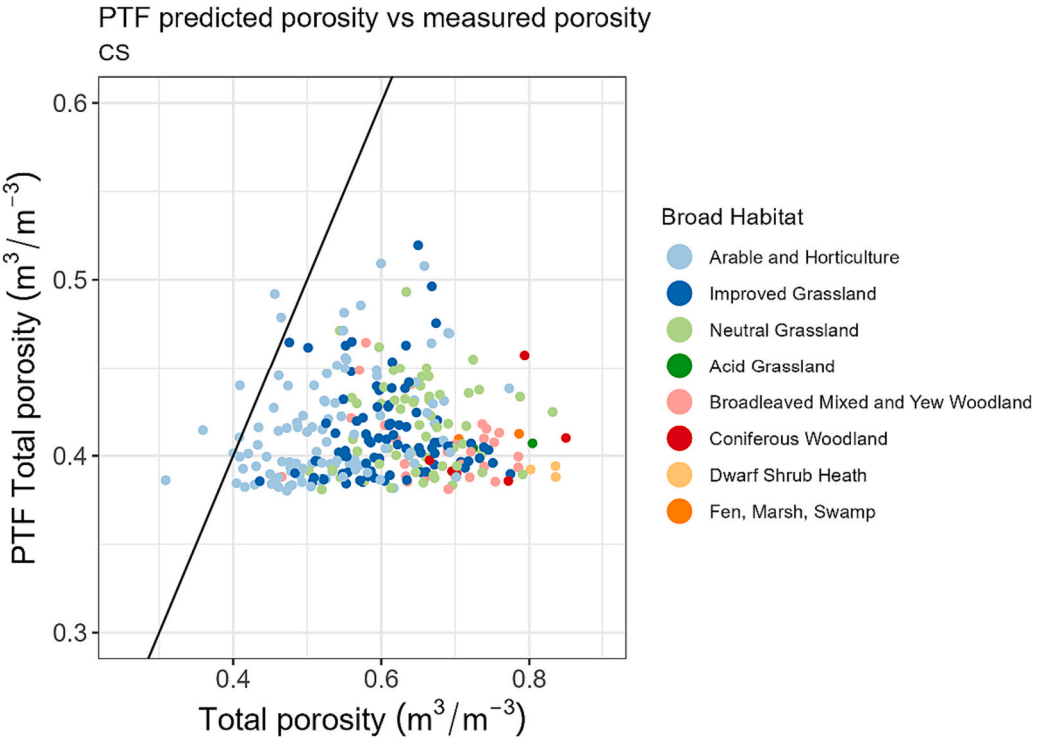


Fig. 2. Measured porosity vs the predicted porosity using a PTF applied to the CS dataset. The black line is the 1:1 line and indicates the missing structural contribution to the porosity.

Table 1
Model performance for GMEP (G) and Countryside Survey datasets (CS). (V) denotes metrics which represent the performance when testing models constructed on the CS data, to predict topsoil total porosity in the GMEP data.

Model	CS			GMEP			V
	r^2	RMSE	deviance	r^2	RMSE	deviance	RMSE
1) PTF (CS2019, G)	−3.736	0.215					
2) GAMM Texture (CS2019, G)	0.373	0.077	1.688		0.073	7.564	0.101
3) GAMM Texture distribution metrics (CS2019, G)	0.171	0.089	2.277		0.082	7.707	0.104
4) ML texture with distribution metrics, habitat (CS2019, G)	0.666	0.057		0.570	0.053		0.081
5) GAMM texture with SOM (CS2019, G)	0.747	0.049	0.680		0.047	2.372	0.066
6) GAMM texture distribution metrics with SOM (CS2019, G)	0.710	0.053	0.793		0.048	2.516	0.064
7) ML texture with distribution metrics, SOM, habitat (CS2019, G)	0.799	0.044		0.800	0.036		0.065
8) GAMM Texture and aggregates (CS2019)	0.705	0.052	0.781				
9) GAMM Texture and aggregate distribution metrics (CS2019)	0.618	0.060	1.038				
10) GAMM Texture and aggregates, SOM (CS2019)	0.764	0.047	0.625				
11) GAMM Texture and aggregate distribution metrics, SOM (CS2019)	0.709	0.053	0.796				
12) ML texture and aggregates with distribution metrics, habitat (CS2019)	0.765	0.048					
13) ML texture and aggregates with distribution metrics, SOM, habitat (CS2019)	0.804	0.044					
14) GAMM SOM (CS2007, G)		0.060	14.337		0.046	3.966	0.055
15) GAMM SOM, SOM smoothed by habitat (CS2007, G)		0.059	13.117		0.040	2.981	0.052
16) GAMM SOM, SOM smoothed by habitat and habitat (CS2007, G)		0.057	13.567		0.040	3.021	0.052
16b) GAMM SOM, SOM smoothed by habitat and habitat (CS2007, G: Grassland subset)		0.054	5.082		0.040	2.044	0.051
17)GAMM SOM, SOM and livestock smoothed by habitat and habitat (CS2007, G: Grassland subset)		0.053	4.898		0.040	2.151	0.050
18) Habitat, livestock (CS2007, G: Grassland subset)		0.089	14.101		0.071	8.492	0.078

Soil structure is commonly omitted from databases used to develop PTFs (Fatichi et al., 2020). As a consequence, early data driven PTFs were based largely on analysis of cropland soils, and their soil texture; newer functions incorporate SOM and illustrate the need for bulk density (Schaap et al., 2004; Tóth et al., 2017; Turek et al., 2023). As with other statistical models, PTFs of soil hydraulic properties should not be applied outside of the range of soils used in their development (Wösten et al., 2001). Large composite hydraulic databases are susceptible to sampling bias, over representing agricultural land uses (Rahmati et al., 2018; Robinson et al., 2022b), and under representing woodland and semi natural habitats. Given cropland accounts for ~7 % of the land surface (Ritchie and Roser, 2013), this is a substantial bias, failing to represent habitats with shrubs and trees that are likely to alter the structural porosity of soils more so than arable crops and associated management.

Fig. 3 contains data used in this study from CS2019. Fig. 3a-c are for a subset of the data with additional metrics of texture and aggregates, whilst Fig. 3d shows the full dataset. Fig. 3a shows that soil porosity does not show a strong dependence on clay content which is the existing paradigm as used in PTFs.

Fig. 3b plots very large aggregates (>256 μm) as a function of the SOM indicating a strong relationship, with aggregation increasing sharply as SOM increases between 0 and 0.2 g g^{-1}, which is the domain of mineral soils. Several reviews on aggregates using published data

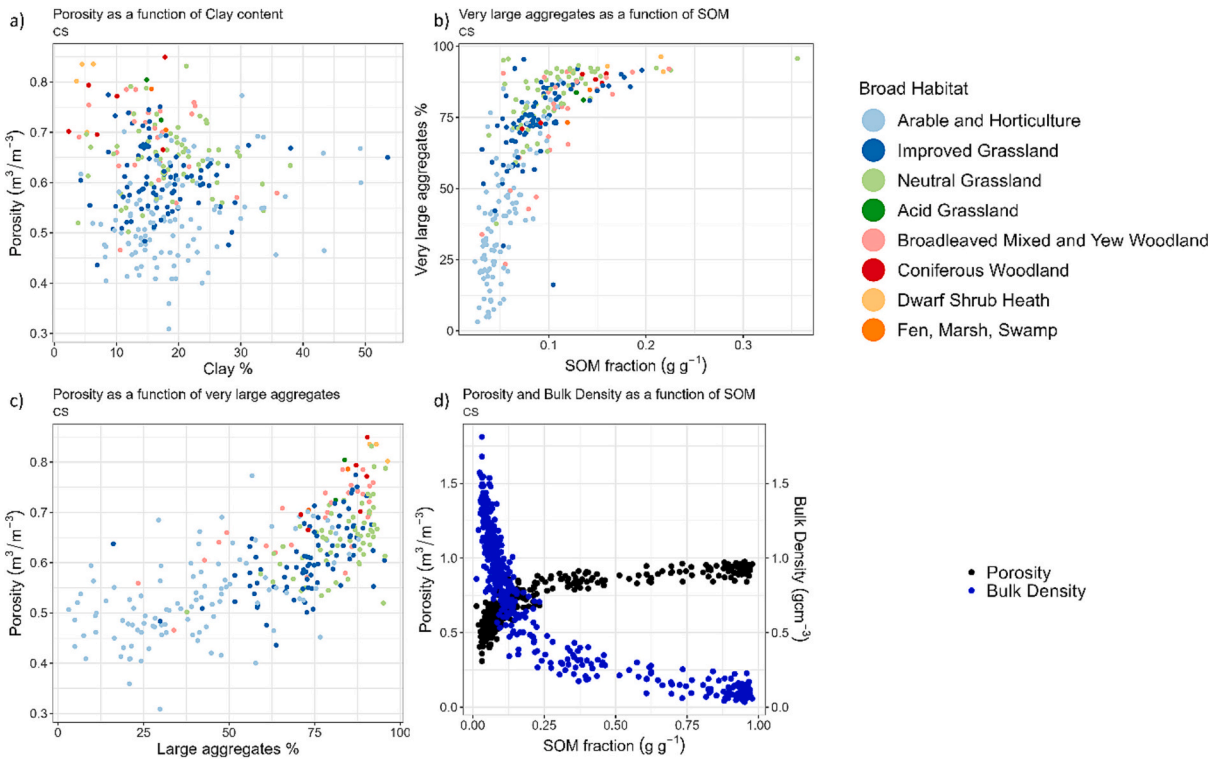


Fig. 3. Plots of relationships between percentage of clay and porosity; percentage of large aggregates and porosity; soil organic matter (SOM) fraction and large aggregates, with markers coloured by habitat. d) relationship between porosity and SOM and bulk density and SOM fraction.

have identified positive relationships between SOM and macroaggregates (e.g. Six et al., 2000, King et al., 2019, Wiesmeier et al., 2019; Sullivan et al., 2022). Relationships have been attributed to aggregate hierarchy theory, with microaggregates bound together by organic binding agents which contribute additional organic matter to soils with larger aggregates (Six et al., 2000). Wiesmeier et al. (2019) suggest this is commonly a top down process with macroaggregate (>250 μm) formation acting as the first step to long term stabilisation of soil carbon, although data from Verchot et al. (2011) suggest that this carbon stabilisation can be a bottom up process starting with the micro-aggregates. King et al. (2019) also suggest that published data may not support formation of microaggregates within macroaggregates. They found that high SOC soils had a greater proportion of microaggregates occluded in macroaggregates, and that occluded microaggregates had higher C, hence they attributed the correlation of SOC with macroaggregates to lower macroaggregate turnover. In line with this, Fig. 3b also shows generally lower values for very large aggregates and SOM for arable sites where aggregates are likely to be disrupted by tillage.

Fig. 3c shows that the increase in aggregation is also associated with an increase in soil porosity. Finally, panel 3d shows the strong relationship between SOM and soil porosity and the reciprocal bulk density. These data tell an important story, that the development of the structural and total soil porosity is a function of the aggregation in the mineral soils which in turn can be related to SOM content. Large datasets derived from agricultural soils are generally restricted to mineral soils telling only part of the story (e.g. Ramcharan et al., 2017; Rawls et al., 2004). Clearly this data from across habitats with a full range of SOM shows how porosity is related to SOM. In mineral soils the relationship is due to carbon in aggregates, but as the organic content increases the wiry shape of the SOM is likely to dominate the porosity (Robinson et al., 2022a), until the fibrous structure of peats results in soil porosities of 90 % or more.

The deviance from the 1:1 line in Fig. 2 could be due to a number of factors, e.g. omission of SOM, or largely unexplored structural factors such as the arrangement of particles that are dependent on the intrinsic characteristics of the particles. These characteristics would be better represented in the more complex ROSETTA models which include bulk density (H3, H4 and H5). Since measured values consistently exceed predicted, the deviance is unlikely to be due to packing factors such as compaction that are the result of an externally applied load, in our data-although this may be present in the datasets used to develop the PTF parameters. Additionally, it could be due to the greater development of soil structure in some habitats in our dataset, likely driven by SOM. We use particle size data and habitat as surrogate information to test these ideas, in models of total porosity constructed with and without SOM. These models (2−13) were constructed on subsets of the data, for which the texture variables were available (CS2019 $n = 287$; GMEP $n = 728$).

3.2. Testing relative contribution of soil texture to total porosity

We applied the GAMMs to the particle size data, first aggregated into texture classes on the logarithmic Krumbein phi scale, and then replacing these with PSD metrics (Table 1, models 2 and 3). Overall, model performance was poor for both texture models, and deteriorated slightly when the models constructed on the CS2019 dataset were used to predict for the GMEP sites. The poor predictive performance could suggest that our statistical models do not hold outside of the sample, or is likely a reflection of the poor performance of the models overall. The relative importance of and contribution from the various texture metrics discussed here should be considered in the context of the poor performance of these models.

By constructing models on texture classes (Table 1, model 2), we test whether certain particle sizes and their groupings might form geometrical bridges or clusters that influence the geometric arrangement of particles and associated packing density and support prediction of total porosity. The colloidal size fraction was the most important texture class in both data sets, followed by medium sand (Tables S1 and S2). Although better than model 1, performance was relatively poor (note larger deviance values for GMEP models may reflect larger sample size).

Particle size distribution (PSD) should be important for total porosity due to the influence of mixing on packing density. Since size fractions are not independent of one another, the individual fractions included in model 2 may to some extent be indicative of PSD and associated mixing effects on packing. Larger particle size fractions tended to have positive correlations with one another and negative correlations with small particle size fractions (Fig. S3). Work by García-Gutiérrez et al. (2019) linking the use of information entropy to characterize soils with the use of fractals as a PSD model demonstrates that coarse aggregations of soil texture data into triplets can be sufficient to describe the full PSD. However, they found that the triplets required vary between soil types, hence information may be lost through aggregation into inappropriate size classes, and PSD metrics may be more informative.

Exploration of PSD metrics elsewhere has identified relationships with bulk density using Shannon information entropy (Martín et al., 2017) or the Weibull equation (Keller and Håkansson, 2010) and multifractal parameters (Wang et al., 2015). We therefore tested a GAMM model using these metrics as PSD descriptors (Table 1, model 3). Model performance was poor, with RMSE and model deviance both increased compared to the model based on texture classes. Whilst PSD descriptors were important in the models, relative importance of different distribution metrics was inconsistent between the data sets (Tables S3 and S4). This inconsistency may reflect correlations between these distribution metrics (Fig. S3).

We might expect better performance of PSD metrics in the comparison studies cited, since they do not seek to explain raw field values of bulk density. Reference bulk density predicted by Keller and Håkansson (2010) could be considered analogous to the non-structural component of porosity, and likewise, in using averages of bulk density, Martín et al. (2017) remove some of the variation from structural porosity. Conversely to predict total porosity of topsoil, our models would need to represent the variation from the structural component of porosity. Nonetheless, it is important to note that our models using PSD metrics performed poorly compared to models with the individual size fractions. This suggests total porosity in our data is more strongly related to size of individual important fractions.

3.3. Testing contribution from land use to total porosity

We then introduced habitat as a surrogate for compaction or development of soil structure, with the assumption that heavily managed habitats such as arable or improved grass will experience more compaction than extensive systems and woodland, which will be subject to greater development of soil structure (Byrnes et al., 2018). We thus included habitat using a ML approach to compare relative importance of habitat with the continuous variables for soil texture. Model performance was improved (Table 1, model 4), but deteriorated when predicting for the GMEP sites using the CS model. The model fitting approach is intended to explore relative importance, hence good performance was not expected for prediction.

Habitat was most important by several orders of magnitude in the models for both data sets, whilst relative importance of the other variables was inconsistent (Figs. 4a, b and S1a, S1b). Removing the least important correlated variables led to changes in the relative importance of remaining variables, and this was also seen for our other ML models where there was a large step in relative importance. In such cases, the co-correlation between many of our related predictors (e.g. texture classes, and distribution metrics) may make them somewhat interchangeable.

Importance of habitat is supported by a recent global meta-analysis (Robinson et al., 2022b) and the analysis of Jarvis et al. (2013) identifying increased hydraulic conductivity under grassland and woodland compared to cropland given the same soil type. This reflects influence of

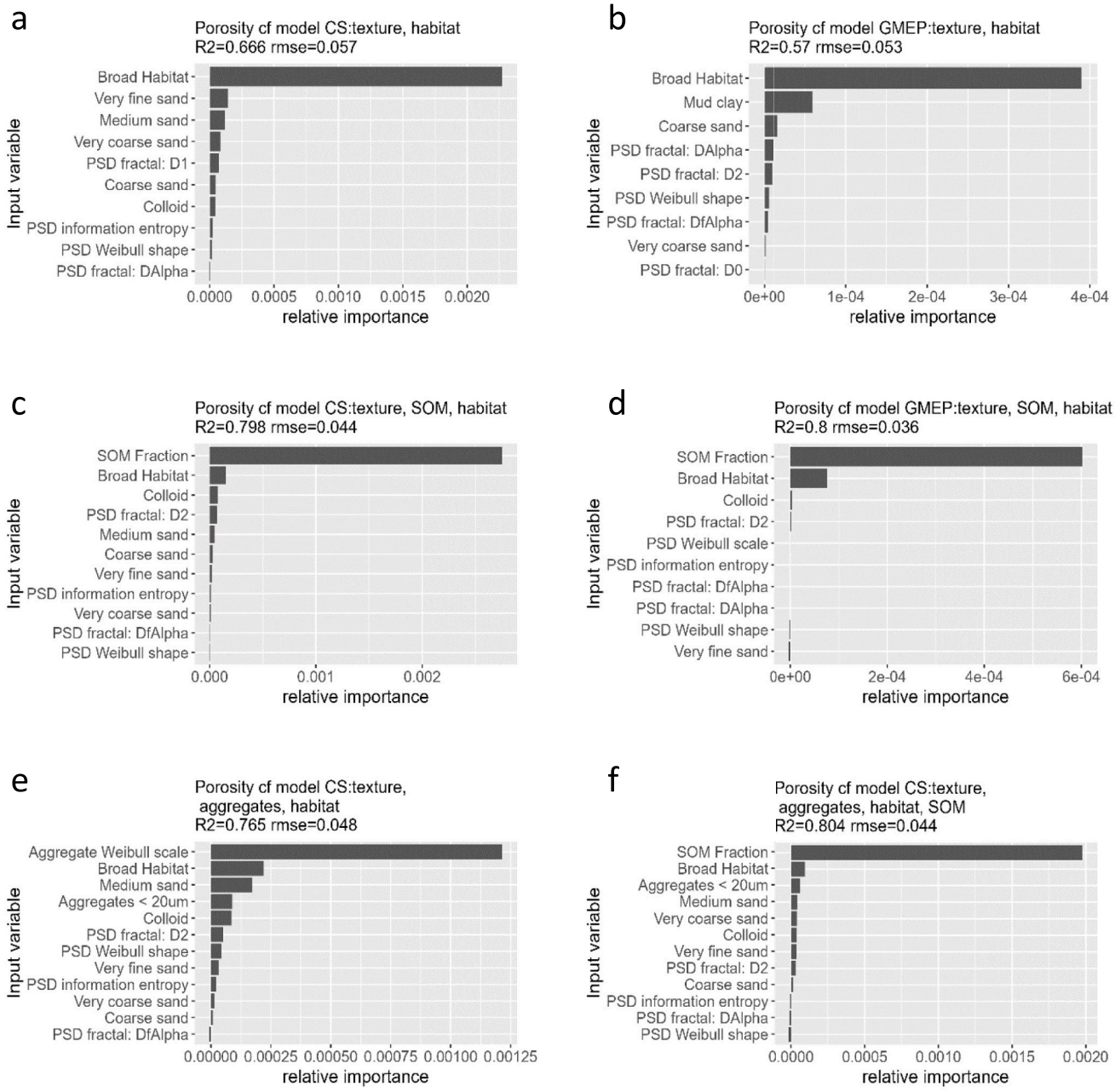


Fig. 4. Variable importance plots for conditional random forest models of total porosity. The panels relate to the following models: a) model 4 CS2019; b) model 4 GMEP; c) model 7 CS2019 d) model 7 GMEP; e) model 12 CS2019 and f) model 13 CS2019. SOM = Soil organic matter. Variables shown are those retained after removing the least important correlated variables (see methods Section 2.2.1 for description or Fig. S1 for full models with all variables).

vegetation type and management on soil hydraulic properties such as macropore development (Jarvis et al., 2013; Bonetti et al., 2021; Robinson et al., 2022b). Soil structural changes with vegetation have the most effect on hydraulic conductivity near-saturated conditions (Bonetti et al., 2021), hence the importance of this type of porosity which is seasonally variable. For much of the annual cycle the textural porosity, it's size distribution and connectivity, determines the hydraulic function of soils. However, under intense rainfall or near saturated soil conditions, the structural pores become the dominant pathways for flow and transport of water and are increasingly important for accurate representation in ESM. This was illustrated by Fatichi et al. (2020) who showed that using biotic factors in the form of proxy measures from vegetation productivity to modify soil hydraulic parameters can enhance performance of ESMs under these conditions. It must be noted that site conditions including soil properties also drive anthropogenic land use and management decisions (e.g. Smith, 1989), which will contribute to variation between habitats, complicating predictions of land use change impacts. Comparison of matrix flows with saturated conductivity in structured soils suggests that the impacts of structure are greater in finer textured soils, further complicating understanding of the impacts of vegetation (Rahmati et al., 2018).

3.4. *Exploring contribution of SOM to total porosity*

We then tested our hypothesis that SOM would be a better predictor of total porosity by including it in the GAMM models (Table 1 models 5

and 6) and the ML models (Table 1 model 7). SOM was most important in the GAMM models with texture classes (Table S5 and S6) and with PSD metrics (Table S7 and S8) and model performance was improved. Colloid and a sand class remained the most important of the texture classes, and again for the GAMM models with PSD metrics predictor importance was inconsistent between the data sets. In the ML model (Table 1 model 7), SOM was most important by several orders of magnitude, then habitat (Fig. 4c, d). Colloids were the most important textural component in both data sets, followed by D2, however these variables had low relative importance compared to SOM and habitat, and again order of importance was altered entirely when removing the least important correlated variables (Figs. S1c, S1d).

The improvements seen for inclusion of SOM in all models here contradict previous work that suggested little impact of omitting SOM from models predicting bulk density from Shannon information entropy (Martín et al., 2017). This may reflect the much lower SOM of their data compared to the GB data in our texture models (max 12 % mean 2 % compared to max 36 % mean 9 % for CS2019 and max 38 % mean 11 % for GMEP). Overall, these models indicate greater importance of SOM and habitat over texture variables when predicting total porosity. Nonetheless, the texture fractions remain important in our model, likely due to explaining the textural component of total porosity in our data (as discussed in Robinson et al., 2022a).

3.5. Testing relative contribution of soil aggregates to total porosity

We next tested our hypothesis that aggregates would be important predictors, using data for CS2019 where aggregate size fractions were recorded (n = 287). The model with aggregate fractions included showed improved performance in predicting total porosity relative to texture alone (Table 1, model 8). Very large aggregates were most important, followed by medium sand and colloid fractions (Table S9). The exponential increase in total porosity with increasing percentage of very large aggregates can be seen in Fig. 3b, whilst the partial relationship in the model more closely resembles a linear increase. Very large aggregates were correlated with the other aggregate fractions (Fig. S3C). Performance again deteriorated when the size classification data were replaced with distribution metrics (Table 1, model 9), but was greatly improved over the model with only texture PSD (Table 1, model 3). The aggregate Weibull scale parameter was most important in the model, and the only significant aggregate metric (Table S10); and is correlated with other aggregate metrics (Fig. S3c).

Including SOM in models with aggregate metrics gave less improvement than in the texture only models (Table 1, models 10 and 11). This is potentially related to the stronger correlations of SOM with the aggregate size fractions compared to weak correlations with the texture metrics (Fig. S3c). Aggregate metrics may be less useful as predictors where SOM data are available, since the relative importance of these metrics was reduced in models with SOM. However, very large aggregates remained more important than individual texture classes in model 10 (Table S11).

When habitat was also included in the ML models (Table 1 model 12), this had high importance, but the aggregate Weibull scale parameter was more important, and several orders of magnitude more important than any texture classes or distribution metrics (Fig. 4e). When SOM was also added (Table 1 model 13), this was orders of magnitude more important than habitat, texture, and aggregate variables (Fig. 4f).

SOM, habitat, and aggregate size distribution are related properties. Aggregate parameters with high importance in models without SOM (very large aggregates model 8 and aggregate Weibull scale, models 9,12) were co-correlated and were correlated with SOM (see Fig. S3c). Vegetation directly affects soil aggregates, with the rate and stability of aggregation, and the rate of aggregate turnover affected by inputs from plants to soil and influence of roots in the rhizosphere (Bronick and Lal, 2005). Soil aggregates are themselves mass fractals; i.e. their bulk density decreases with size (Anderson and Mcbratney, 1995) and aggregate size distribution will further affect the packing density of soils driving porosity change. Aggregate size distribution has also previously been shown to be correlated with pore size distribution (Lebron et al., 2002). Findings from these models (2–13) suggest that land management intensity and soil aggregation are more important than soil texture per-se, and also further support our hypothesis of the greater importance of SOM content.

3.6. Exploring variation in relationship of total porosity to SOM between habitats

Having confirmed the greater importance of SOM over texture and aggregates in predicting total porosity, GAMMs were then constructed on the full datasets (CS2007 $n = 2356$; GMEP $n = 1335$), for which those variables were not available. The increase in model deviance compared to the models on the texture subsets may be explained by the increase in n (Table 1, models 14–16; sample size was 8 times larger for the CS data and 1.8 times larger for the GMEP data).

To explore whether the relationship between SOM and total porosity varies between habitats, further GAMM models were constructed. Firstly from SOM only (Table 1 model 14), then allowing the model to adjust the gradient of the relationship for different habitats (Table 1 model 15) and then additionally allowing the intercept to differ between habitats (Table 1 model 16). Whilst habitat was important in models 15 and 16, which show small improvements in RMSE compared to the SOM only model, the gradient of the relationship between SOM and porosity did not generally vary between habitats (Supplementary table S15 to S18, and text S2). These findings suggest that impacts of habitat on total porosity are largely consistent across the SOM range sampled in GMEP and CS2007.

3.7. Contribution of land management to total porosity

Higher stocking densities are known to lead to soil compaction and thus reduced structural porosity (Byrnes et al., 2018). Robinson et al. (2022b) highlight the importance of counteracting effects in grassland of vegetation increasing structural porosity vs grazing intensity and management reducing it.

To explore this, we further tested the inclusion of data on stocking density in a model predicting total porosity for a grassland subset of the data. This resulted in a very marginal improvement in model performance (Table 1 Model 17, note improvements in model deviance reflect reduction in sample size). In both data sets partial effects from livestock density were nonlinear, and much smaller than from SOM. When models were constructed with just livestock density and habitat, model performance was reduced and cows were more important in both models (Table 1 model 18, supplementary Tables S19 and S20).

The relatively limited influence of stocking densities in models 17 and 18 may reflect the quality of the data available, and scale (1 km) in relation to the soil sampling. Because compaction effects will be localised to the fields containing stock and large-scale patterns may not be representative of stocking densities in the field or location sampled. Furthermore, livestock are commonly moved around between fields on a farm, and there will be influence from climate at the time stocking numbers are higher in the field (wet soils being more vulnerable to compaction (MAFF, 1970)). Moreover, the soil data were collected in summer and a larger stocking density effect might be observed when contrasting winter and summer, or wet and dry soils. We expect a potential seasonal change in structural porosity due to this but are yet to have suitable data to test this.

The performance for these models was nonetheless better than models based only on the texture data and PSD. Although only two broad habitats were included in the model, the habitat level differences may be sufficient to explain this, alternatively, the stocking density pattern may also follow other broad spatial drivers of climate or land

management intensity at national levels. Additionally, the opposing trends for sheep and cows in the models may reflect inverse correlation of cows with sheep in grassland areas.

3.8. Comparison of models and implications

Given the strong relationship between total porosity and SOM, it is informative to explore the performance of the models throughout the SOM range (Fig. 5b) as well as the porosity range (Fig. 5a) captured by our data. Thus comparing the models visually, the models without SOM perform poorly; total porosity is under predicted in the texture only model (2) and livestock model (18) above 0.1 g g^{-1} SOM and 0.6 porosity, whilst at lower porosity both models over-predict, with worse performance by the livestock model. Adding SOM to the texture model (5) improves performance throughout the full range, not just at higher porosity and SOM; this may be because by representing the influence of SOM the partial relationship to texture is also better captured. However, the SOM only model (14) follows a broadly similar form to the texture with SOM model (5). Conversely, the texture based PTF performs poorly (model 1), underpredicting soil porosity except at very low SOM content and porosity below 0.4.

Results confirm that SOM is key to predicting total porosity, as highlighted by the consistently greater importance in the models than any other variable. Previous work (Robinson et al., 2022a) suggests the physical reason for this is the mixture of mineral and organic geometries in the packing, pertinent to the soils studied. The mineral component is granular, and the organic component is increasingly fibrous in this data set, resulting in the emergent response curve for soil porosity to SOM. The results show the importance of incorporating a structure metric such as bulk density or SOM (which alters the bulk density) in order to predict the higher porosities. New PTFs are now doing this as data becomes more widely available (Tóth et al., 2017). Moreover, the improvement in prediction through the incorporation of even coarse habitat data indicates that the biota are contributing to structural porosity over and above surrogates such as carbon, supporting the importance of land use as a metric found in the analysis of Jarvis et al. (2013).

The incorporation of SOM and land use or habitat metrics could lead to a more dynamic representation of soil structure to aid modelling efforts such as those presented in Fatichi et al. (2020). Bonetti et al. (2021) present a framework to account for the impacts of vegetation on structural porosity and hydrological function, and this could also be used as a basis for incorporating representation of SOM. However, due to spatial

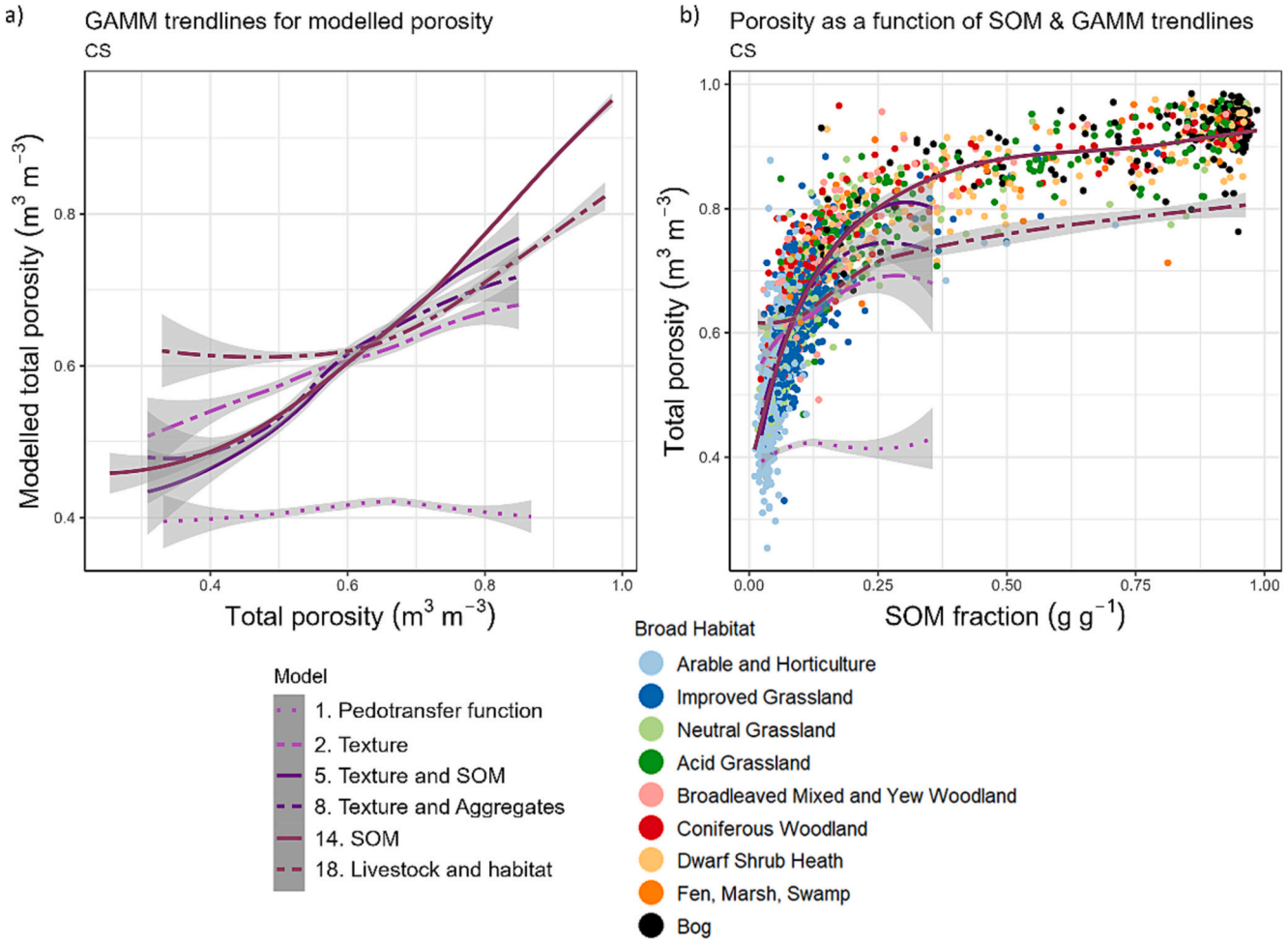


Fig. 5. a) Model trendlines for predicted total porosity as a function of measured porosity and b) Measured data points and model trendlines for total porosity as a function of soil organic matter (SOM) content. Data and models from the Countryside Survey (CS) data. Measured data point markers are color coded by Broad Habitat. Model trendlines in both plots are shown for predicted porosity on the CS data using a series of Generalized Additive Mixed Models, as listed in Table 1, alongside model 1: pedotransfer function for comparison. The lines represent LOESS mean of the predicted total porosity output from the specified model. Data were filtered to remove points missing values for predictor variables. The grey shaded area around each trendline represents the 0.95 confidence interval of the relationship to predicted total porosity for a) measured total porosity and b) SOM. Note for model 14, because porosity is predicted from only SOM the confidence interval in plot b is approximately zero and therefore too small to be visible on the plot.

variability of SOM and the nonlinear relationship between SOM and porosity, the spatial scale of both available SOM data and ESMs must be taken into account for such representation. Similarly, modelling at coarse spatial scales will average out areas of higher intensity rainfall, or areas with greater soil moisture due to topographic features, and thus may not realistically predict the saturated conditions under which structural porosity pathways are relevant (Fatichi et al., 2020). There is further complexity to representing impacts of land use change on soil hydraulic properties, since correlation between soil properties and land use decisions (Smith, 1989) contributes to observed variation between habitats, and the impacts of vegetation induced structure vary with soil type (Bonetti et al., 2021).

4. Conclusions

We show that a combination of ML and statistical approaches addresses both challenges of gaining better physical understanding of the soil porosity/SOM relationship and predictive capability. The work demonstrates the importance of representing structural porosity when predicting total porosity for unbiased datasets from across habitats. Furthermore, it suggests statistical approaches have good potential for predicting total porosity based on combining soil data and biotic remote sensing data in temperate systems. SOM still proves to be the best predictor of total porosity, with land cover providing a useful covariate. These drivers affect structural porosity, and thus total porosity and its reciprocal bulk density. Attempts to improve the prediction of total porosity by incorporating particle size metrics such as entropy and multifractal parameters did not improve performance in this analysis. The incorporation of aggregates did improve performance, indicating the importance of these emergent structures for prediction of total porosity. An attempt to incorporate animal stocking density data in the analysis may have potential, but the data are currently too coarse spatially and temporally relative to our survey data. Overall, the analysis suggests future focus should seek to better understand and incorporate biotic effects and emergent structural features such as aggregation into models. Moreover, these are more likely than texture to be affected by management and climate resulting in potentially important environmental feedbacks.

In hydrology, soil porosity provides the foundation for predicting hydraulic properties, whilst bulk density is used to estimate soil carbon stocks. The availability of annual land cover maps from remote sensing could support inclusion of dynamic soil porosity and bulk density parameters in models, helping account for feedbacks from land use or climate change. We acknowledge that this relationship between SOM and soil porosity/bulk density may only hold in temperate and northern latitudes where SOM is more plentiful, and only reflects micro-*meso* scale soil structural components. Further research and validation is necessary to develop appropriate algorithms and understand regional variation. Future work should also explore drivers of macro scale structural components such as large cracks and burrows which will further affect hydrologic processes. Nonetheless, our approach provides an important step in the assessment of drivers of soil porosity and bulk density for use in modelling at larger scales.

CRediT authorship contribution statement

A. Thomas: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **F. Seaton:** Writing – review & editing, Methodology. **E. Dhiedt:** Writing – review & editing, Formal analysis. **B.J. Cosby:** Writing – review & editing, Project administration, Conceptualization. **C. Feeney:** Writing – review & editing. **I. Lebron:** Writing – review & editing, Data curation, Conceptualization. **L. Maskell:** Writing – review & editing. **C. Wood:** Writing – review & editing, Data curation. **S. Reinsch:** Writing – review & editing, Data curation. **B.A. Emmett:** Writing – review & editing, Project administration. **D.A. Robinson:** Writing – review & editing, Writing – original draft, Project administration, Formal analysis, Conceptualization.

Declaration of competing interest

There are no known conflicts of interest.

Data availability

The main datasets used in the paper are directly available from the Environmental Information Data Centre (EIDC) through the following links:

Bentley, L.; Reinsch, S.; Alison, J.; Andrews, C.; Brentegani, M.; Chetiu, N.; Dart, S.; Dhiedt, E.; Emmett, B.A.; Fitos, E.; Garbutt, R.A.; Gray, A.; Henrys, P.A.; Hunt, A.; Keenan, P.O.; Keith, A.M.; Koblizek, E.; Lebron, I.; Millani Lopes Mazzetto, J.; Pallett, D.W.; Pereira, M.G.; Pinder, A.; Risser, H.; Rose, R.J.; Rowe, R.L.; Scarlett, P.; Seaton, F.; Smart, S.M.; Towill, J.; Wagner, M.; Williams, B.; Wood, C.M.; Robinson, D.A. (2023). Topsoil physico-chemical properties from the UKCEH Countryside Survey, Great Britain, 2018–2019. NERC EDS Environmental Information Data Centre. https://doi.org/10.5285/821325f3-b353-4a51-8db2-6b2200d82aca.

Emmett, B.A.; Reynolds, B.; Chamberlain, P.M.; Rowe, E.; Spurgeon, D.; Brittain, S.A.; Frogbrook, Z.; Hughes, S.; Lawlor, A.J.; Poskitt, J.; Potter, E.; Robinson, D.A.; Scott, A.; Wood, C.M.; Woods, C. (2016). Soil physico-chemical properties 2007 [Countryside Survey]. NERC Environmental Information Data Centre. (Dataset). https://doi.org/10.5285/79669141-cde5-49f0-b24d-f3c6a1a52db8.

Lebron, I.; Seaton, F.; Barrett, G.; Alison, J.; Burden, A.; Emmett, B.A.; Garbutt, A.; Robinson, D.A.; Williams, B.; Wood, C.M. (2020). Topsoil particle size distribution from the Glastir Monitoring and Evaluation Programme, Wales 2013–2016. NERC Environmental Information Data Centre. (Dataset). https://doi.org/10.5285/d6c3cc3c-a7b7-48b2-9e61-d07454639656.

Robinson, D.A.; Astbury, S.; Barrett, G.; Burden, A.; Carter, H.; Emmett, B.A.; Garbutt, A.; Giampieri, C.; Hall, J.; Henrys, P.; Hughes, S.; Hunt, A.; Jarvis, S.; Jones, D.L.; Keenan, P.; Lebron, I.; Nunez, D.; Owen, A.; Patel, M.; Pereira, M.G.; Seaton, F.; Sharps, K.; Tanna, B.; Thompson, N.; Williams, B.; Wood, C.M. (2019). Topsoil physico-chemical properties from the Glastir Monitoring and Evaluation Programme, Wales 2013–2016. NERC Environmental Information Data Centre. (Dataset). https://doi.org/10.5285/0fa51dc6-1537-4ad6-9d06-e476c137ed09.

The authors declare that all other data supporting the findings of this study are available within the article and its Supplementary Information files, or are available from the corresponding author upon reasonable request.

Acknowledgments

General

We thank all the Glastir Monitoring and Evaluation Program (GMEP) team who contributed to collecting all the GMEP data in Wales and the UK Centre for Ecology & Hydrology Countryside Survey team from 2007 who collected all the data from England, Wales and Scotland.

Funding

The research was funded by the Natural Environment Research Council award number NE/R016429/1 as part of the UK–ScaPE Programme Delivering National Capability. Also supported in part by the European Union's Interreg North-West Europe programme, part of the European Territorial Cooperation Programme and ERDF funding. The work was supported by grant agreement No. NWE 810, project FABulous Farmers (Functional Agro-Biodiversity in farming). The Glastir Monitoring and Evaluation Program (GMEP) was funded by the Welsh

Government as part of the Environment & Rural Affairs Monitoring and Modelling Programme (Contract reference: C147/2010/11) and NERC/Centre for Ecology & Hydrology (CEH Projects: NEC04780/NEC05371/NEC05782). This project acknowledges funding from the European Union's Horizon Europe research and innovation programme under grant agreement No.-101086179; and funding from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10053484]. The Research Council of Norway, Climasol, Project number: 325253.

Disclaimer

Work funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2024.171158.

References

Anderson, A.N., Mcbratney, A.B., 1995. Soil aggregates as mass fractals. Soil Res. 33 (5), 757–772.
Bayat, H., Rastgo, M., Zadeh, M.M., Vereecken, H., 2015. Particle size distribution models, their characteristics and fitting capability. J. Hydrol. 1 (529), 872–889. Oct.
Bonetti, S., Wei, Z., Or, D., 2021. A framework for quantifying hydrologic effects of soil structure across scales. Communications Earth & Environment 2 (1), 107. Jun 3.
Borrelli, P., Alewell, C., Alvarez, P., Anache, J.A.A., Baartman, J., Ballabio, C., et al., 2021. Soil erosion modelling: a global review and statistical analysis. Sci. Total Environ. 780, 146494.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. Oct.
Bronick, C.J., Lal, R., 2005. Soil structure and management: a review. Geoderma 124 (1–2), 3–22. Jan 1.
Bunce, R.G., Barr, C.J., Gillespie, M.K., Howard, D.C., 1996. The ITE land classification: providing an environmental stratification of Great Britain. Environ. Monit. Assess. 39, 39–46. Jan.
Byrnes, R.C., Eastburn, D.J., Tate, K.W., Roche, L.M., 2018. A global meta-analysis of grazing impacts on soil health indicators. J. Environ. Qual. 47, 758–765.
Delignette-Muller, M., Dutang, C., 2015. Fitdistrplus: an R package for fitting distributions. J. Stat. Softw. https://doi.org/10.18637/jss.v064.i04.
EDINA, 2007a. Scotland Agricultural Census 2007 [TIFF geospatial data], Scale 1:10000, Tiles: GB, Updated: 30 June 2007, Scotland Government, Using: EDINA Agcensus Digimap Service. https://digimap.edina.ac.uk. Downloaded: 2023-02-06.
EDINA, 2007b. Wales Agricultural Census 2007 [TIFF geospatial data], Scale 1:10000, Tiles: GB, Updated: 30 June 2007, Welsh Government, Using: EDINA Agcensus Digimap Service. https://digimap.edina.ac.uk. Downloaded: 2023-02-06.
EDINA, 2010. England Agricultural Census 2010 [TIFF geospatial data], Scale 1:10000, Tiles: GB, Updated: 30 June 2010, DEFRA, Using: EDINA Agcensus Digimap Service. https://digimap.edina.ac.uk. Downloaded: 2023-02-06.
EDINA, 2016. England Agricultural Census 2016 5 km [TIFF geospatial data], Scale 1: 10000, Tiles: GB, Updated: 30 June 2016, DEFRA, Using: EDINA Agcensus Digimap Service. https://digimap.edina.ac.uk. Downloaded: 2023-02-06.
EDINA, 2018. Wales Agricultural Census 2018 [TIFF geospatial data], Scale 1:10000, Tiles: GB, Updated: 30 June 2018, Welsh Government, Using: EDINA Agcensus Digimap Service. https://digimap.edina.ac.uk. Downloaded: 2023-02-06.
EDINA, 2019. Scotland Agricultural Census 2019 [TIFF geospatial data], Scale 1:10000, Tiles: GB, Updated: 30 June 2019, Scotland Government, Using: EDINA Agcensus Digimap Service. https://digimap.edina.ac.uk. Downloaded: 2023-02-06.
Emmett, B., Reynolds, B., Chamberlain, P., Rowe, E., Spurgeon, D., Brittain, S.A., Frogbrook, Z., Hughes, S., Lawlor, A., Poskitt, J., Potter, E., 2016. Soil Physico-chemical Properties 2007 [Countryside Survey].
Emmett, B.A., Frogbrook, Z.L., Chamberlain, P.M., Griffiths, R., Pickup, R., Poskitt, J., Reynolds, B., Rowe, E., Rowland, P., Wilson, J., Wood, C.M., 2008. Countryside Survey. Soils Manual.
Fatichi, S., Or, D., Walko, R., Vereecken, H., Young, M.H., Ghezzehei, T.A., Hengl, T., Kollet, S., Agam, N., Avissar, R., 2020. Soil structure is an important omission in earth system models. Nat. Commun. 11 (1), 522. Jan 27.
García-Gutiérrez, C., Martín, M.Á., Pachepsky, Y., 2019. On the information content of coarse data with respect to the particle size distribution of complex granular media: rationale approach and testing. Entropy 21 (6), 601. Jun 17.
Grossman, R.B., Reinsch, T.G., 2002. 2.1 Bulk density and linear extensibility. In: Methods of Soil Analysis: Part 4 Physical Methods, 5, pp. 201–228.
Hausser, J., Strimmer, K., 2009. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. J. Mach. Learn. Res. 10, 1469–1484. Available online from. https://jmlr.csail.mit.edu/papers/v10/hausser09a.html.
Hirmas, D.R., Giménez, D., Nemes, A., Kerry, R., Brunsell, N.A., Wilson, C.J., 2018. Climate-induced changes in continental-scale soil macroporosity may intensify water cycle. Nature 561 (7721), 100–103. Sep 6.
Hothorn, T., Zeileis, A., 2015. Partykit: a modular toolkit for recursive Partytioning in R. In. J. Mach. Learn. Res. 16, 3905–3909. https://jmlr.org/papers/v16/hothorn15a.html.
Jarvis, N., Koestel, J., Messing, I., Moeys, J., Lindahl, A., 2013. Influence of soil, land use and climatic factors on the hydraulic conductivity of soil. Hydrol. Earth Syst. Sci. 17 (12), 5185–5195. Dec 20.
Jarvis, N.J., 2007. A review of non-equilibrium water flow and solute transport in soil macropores: principles, controlling factors and consequences for water quality. Eur. J. Soil Sci. 58, 523–546.
Keller, T., Håkansson, I., 2010. Estimation of reference bulk density from soil particle size distribution and soil organic matter content. Geoderma 154 (3–4), 398–406. Jan 15.
King, A.E., Congreves, K.A., Deen, B., Dunfield, K.E., Voroney, R.P., Wagner-Riddle, C., 2019. Quantifying the relationships between soil fraction mass, fraction carbon, and total soil carbon to assess mechanisms of physical protection. Soil Biol. Biochem. 135, 95–107.
Lebron, I., Suarez, D.L., Schaap, M.G., 2002. Soil pore size and geometry as a result of aggregate-size distribution and chemical composition. Soil Sci. 167 (3), 165–172. Mar 1.
Lorenz, R., Jaeger, E.B., Seneviratne, S.I., 2010. Persistence of heat waves and its link to soil moisture memory. Geophys. Res. Lett. 37 (9). May.
MAFF, 1970. Modern farming and the soil. In: Report for the Agricultural Advisory Council on Soil Structure and Soil Fertility. London.
Marra, G., Wood, S.N., 2011. Practical variable selection for generalized additive models. Computational Statistics & Data Analysis. 55 (7), 2372–2387. Jul 1.
Martín, M.Á., Reyes, M., Taguas, F.J., 2017. Estimating soil bulk density with information metrics of soil texture. Geoderma 1 (287), 66–70. Feb.
Morton, R.D., Rowland, C.S., Wood, C.M., Meek, L., Marston, C.G., Smith, G.M., 2014. Land Cover Map 2007 (1 km percentage aggregate class, GB) v1.2, NERC Environmental Information Data Centre. https://doi.org/10.5285/289805c2-4be7-4fb5-b6ec-1539ed88c43d.
Morton, R.D., Marston, C.G., O'Neil, A.W., Rowland, C.S., 2022a. Land Cover Map 2017 (1 km summary rasters, GB and N. Ireland), NERC EDS Environmental Information Data Centre. https://doi.org/10.5285/be0bdc0e-bc2e-4f1d-b524-2c02798dd893.
Morton, R.D., Marston, C.G., O'Neil, A.W., Rowland, C.S., 2022b. Land Cover Map 2018 (1 km summary rasters, GB and N. Ireland), NERC EDS Environmental Information Data Centre. https://doi.org/10.5285/9b68ee52-8a95-41eb-8ef1-8d29e2570b00.
Morton, R.D., Marston, C.G., O'Neil, A.W., Rowland, C.S., 2022c. Land Cover Map 2019 (1 km summary rasters, GB and N. Ireland), NERC EDS Environmental Information Data Centre. https://doi.org/10.5285/e5632f1b-040c-4c39-8721-4834ada6046a.
Nix, J., Redman, G., 2021. John Nix Pocketbook for Farm Management for 2022, 52nd ed. Agro Business Consultants, Melton Mowbray.
ONS, 2020. Countries (December 2016) Boundaries GB BGC. Downloaded from. https://geoportal.statistics.gov.uk/ (web archive link, 13 February 2023). on 13/02/2023.
Pachepsky, Y.A., Rawls, W.J., Timlin, D.J., 1999. The current status of pedotransfer functions: their accuracy, reliability, and utility in field- and regional-scale modeling. In: Corwin, D.L., Loague, K., Ellsworth, T.H. (Eds.), Assessment of Non-Point Source Pollution in the Vadose Zone. American Geophysical Union, Washington, DC, pp. 223–234. Geophysical Monograph No 108.
Page-Dumroese, D.S., Brown, R.E., Jurgensen, M.F., Mroz, G.D., 1999. Comparison of methods for determining bulk densities of rocky forest soils. Soil Sci. Soc. Am. J. 63 (2), 379–383.
Panagos, P., De Rosa, D., Liakos, L., Labouyrie, M., Borrelli, P., Ballabio, C., 2024. Soil bulk density assessment in Europe. Agric. Ecosyst. Environ. 364, 108907.
Posadas, A.N., Giménez, D., Bittelli, M., Vaz, C.M., Flury, M., 2001. Multifractal characterization of soil particle-size distributions. Soil Sci. Soc. Am. J. 65 (5), 1361–1367. Sep.
Rabot, E., Wiesmeier, M., Schlüter, S., Vogel, H.J., 2018. Soil structure as an indicator of soil functions: a review. Geoderma 15 (314), 122–137. Mar.
Rahmati, M., Weihermüller, L., Vanderborght, J., Pachepsky, Y.A., Mao, L., Sadeghi, S. H., et al., 2018. Development and analysis of the soil water infiltration global database. Earth Syst. Sci. Data 10, 1237–1263.
Ramcharan, A., Hengl, T., Beaudette, D., Wills, S., 2017. A soil bulk density pedotransfer function based on machine learning: a case study with the NCSS soil characterization database. Soil Sci. Soc. Am. J. 81 (6), 1279–1287. Nov.
Rawls, W.J., Nemes, A.T., Pachepsky, Y.A., 2004. Effect of soil organic carbon on soil hydraulic properties. Dev. Soil Sci. 1 (30), 95–114. Jan.
Ritchie, H., Roser, M., 2013. Land Use. https://ourworldindata.org/land-use.
Robinson, D., Astbury, S., Barrett, G., Burden, A., Carter, H., Emmett, B., Garbutt, A., Giampieri, C., Hall, J., Henrys, P., Hughes, S., 2019. Topsoil Physico-chemical Properties From the Glastir Monitoring and Evaluation Programme, Wales 2013–2016.
Robinson, D.A., Thomas, A., Reinsch, S., Lebron, I., Feeney, C.J., Maskell, L.C., Wood, C. M., Seaton, F.M., Emmett, B.A., Cosby, B.J., 2022a. Analytical modelling of soil porosity and bulk density across the soil organic matter and land-use continuum. Sci. Rep. 12 (1), 7085. Apr 30.
Robinson, D.A., Nemes, A., Reinsch, S., Radbourne, A., Bentley, L., Keith, A.M., 2022b. Global meta-analysis of soil hydraulic properties on the same soils with differing land use. Sci. Total Environ. 15 (852), 158506. Dec.
Ruehlmann, J., 2020. Soil particle density as affected by soil texture and soil organic matter: 1. Partitioning of SOM in conceptional fractions and derivation of a variable SOC to SOM conversion factor. Geoderma 375, 114542.

Salat, H., Murcio, R., Arcaute, E., 2017. Multifractal methodology. Physica A Stat. Mech. Applic. 473, 467–487. May 1.
Sandin, M., Koestel, J., Jarvis, N., Larsbo, M., 2017. Post-tillage evolution of structural pore space and saturated and near-saturated hydraulic conductivity in a clay loam soil. Soil Tillage Res. 1 (165), 161–168. Jan.
Schaap, M.G., Leij, F.J., Van Genuchten, M.T., 2001. Rosetta: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. J. Hydrol. 251 (3–4), 163–176.
Schaap, M.G., Nemes, A., Van Genuchten, M.T., 2004. Comparison of models for indirect estimation of water retention and available water in surface soils. Vadose Zone J. 3 (4), 1455–1463.
Seaton, F.M., George, P.B., Lebron, I., Jones, D.L., Creer, S., Robinson, D.A., 2020. Soil textural heterogeneity impacts bacterial but not fungal diversity. Soil Biol. Biochem. 1 (144), 107766. May.
Six, J., Paustian, K., Elliott, E.T., Combrink, C., 2000. Soil structure and organic matter I. Distribution of aggregate-size classes and aggregate-associated carbon. Soil Sci. Soc. Am. J. 64 (2), 681–689.
Smith, B.D., 1989. Origins of agriculture in eastern North America. Science 246 (4937), 1566–1571. Dec 22.
Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics. 8 (1), 1–21. Dec.
Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinformatics. 9, 1. Dec.
Sullivan, P.L., Billings, S.A., Hirmas, D., Li, L., Zhang, X., Ziegler, S., Murenbeeld, K., Ajami, H., Guthrie, A., Singha, K., Giménez, D., 2022. Embracing the dynamic nature of soil structure: a paradigm illuminating the role of life in critical zones of the Anthropocene. Earth Sci. Rev. 225, 103873.
Tifafi, M., Guenet, B., Hatté, C., 2018. Large differences in global and regional total soil carbon stock estimates based on SoilGrids, HWSD, and NCSCD: Intercomparison and evaluation based on field data from USA, England, Wales, and France. Glob. Biogeochem. Cycles 32, 42–56.
Tóth, B., Weynants, M., Pásztor, L., Hengl, T., 2017. 3D soil hydraulic database of Europe at 250 m resolution. Hydrol. Process. 31 (14), 2662–2666.
Turek, M.E., Poggio, L., Batjes, N.H., Armindo, R.A., van Lier, Q.D.J., de Sousa, L., Heuvelink, G.B., 2023. Global mapping of volumetric water retention at 100, 330 and 15 000 cm suction using the WoSIS database. International Soil and Water Conservation Research 11 (2), 225–239.
Verchot, L.V., Dutaur, L., Shepherd, K.D., Albrecht, A., 2011. Organic matter stabilization in soil aggregates: understanding the biogeochemical mechanisms that determine the fate of carbon inputs in soils. Geoderma 161 (3–4), 182–193.
Walter, K., Don, A., Tiemeyer, B., Freibauer, A., 2016. Determining soil bulk density for carbon stock calculations: a systematic method comparison. Soil Sci. Soc. Am. J. 80 (3), 579–591.
Wang, J., Zhang, M., Bai, Z., Guo, L., 2015. Multi-fractal characteristics of the particle distribution of reconstructed soils and the relationship between soil properties and multi-fractal parameters in an opencast coal-mine dump in a loess area. Environ. Earth Sci. 73, 4749–4762. Apr.
Weihermüller, L., Lehmann, P., Herbst, M., Rahmati, M., Verhoef, A., Or, D., Jacques, D., Vereecken, H., 2021. Choice of pedotransfer functions matters when simulating soil water balance fluxes. Journal of Advances in Modeling Earth Systems. 13 (3). Mar. e2020MS002404.
Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., von Lützow, M., Marin-Spiotta, E., van Wesemael, B., Rabot, E., Ließ, M., Garcia-Franco, N., Wollschläger, U., 2019. Soil organic carbon storage as a key function of soils-a review of drivers and indicators at various scales. Geoderma 333, 149–162.
Wood, S.N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. Ser. B Stat Methodol. 73 (1), 3–6. Jan.
Wösten, J.H., Pachepsky, Y.A., Rawls, W.J., 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. J. Hydrol. 251 (3–4), 123–150. Oct 1.
Zhang, Y., Schaap, M.G., 2017. Weighted recalibration of the Rosetta pedotransfer model with improved estimates of hydraulic parameter distributions and summary statistics (Rosetta3). J. Hydrol. 547, 39–53.