

Organic carbon dominates soil porosity across environmental gradients, with organic carbon–clay metrics providing secondary insight

Amy Thomas^{a,*}, Attila Nemes^{b,c}, David Robinson^d

^a UK Centre for Ecology & Hydrology, Bangor, United Kingdom

^b Norwegian University of Life Sciences, Faculty of Environmental Sciences and Natural Resource Management

^c Texas Tech University, Department of Plant Soil Sciences I. Lebron, UK Centre for Ecology & Hydrology, Bangor, United Kingdom

^d Division of Environment and Natural Resources, Norwegian Institute of Bioeconomy Research

ARTICLE INFO

Keywords:

Bulk density

Porosity

OC:clay ratio

Dexter ratio

Soil compaction

Soil carbon

ABSTRACT

Understanding how soil porosity varies across environmental gradients is critical for interpreting land–surface processes, including water storage, flow pathways, gas exchange, and carbon cycling. While organic carbon (OC) and soil texture are important influences on soil structure, their combined influence on porosity at landscape to continental scales remains incompletely resolved. Here, we compile topsoil datasets from Great Britain, Wales, the European LUCAS survey, and the USDA-NRCS NCSS database ($n = 6335$) to quantify relationships between porosity, OC, and clay content across diverse environmental conditions. Focusing on minimally disturbed soils, we evaluate individual and combined controls using generalized additive models.

We develop and test simple empirical descriptors of OC:clay ratio, including a signed metric (NonCOCC) representing deviation from a reference ratio of 1 g OC to 10 g clay, positioning soils along a continuum from clay excess to OC excess. Across all regions, OC has the strongest relationship to porosity, explaining substantially more variation than OC:clay ratio metrics. NonCOCC provided a modest improvement in model performance when combined with OC and clay.

Across regional datasets, the steepest portions of the fitted NonCOCC partial responses generally occurred at low positive values, suggesting a region of elevated porosity sensitivity. In a subset with aggregate-size data, this range of NonCOCC values also corresponds to a shift towards larger aggregates. This may suggest a structural pathway linking organic enrichment, aggregation, and porosity, but requires testing using data with measured OC partitions.

The findings confirm importance of OC for predicting porosity and identify added value of simple metrics (NonCOCC) from widely available soil observations.

1. Introduction

Soil porosity is a key soil property controlling the behaviour of near-surface Earth systems because it governs the storage and transmission of water, gases, and solutes in soils and underpins land–surface processes (Robinson et al., 2025; Panagos et al., 2024). Total porosity is a useful first-order indicator of soil structure and associated functions, including water and gas storage, infiltration, hydraulic conductivity and aeration (Hirmas et al., 2018; Golchin et al., 1994; Rabot et al., 2018). Other aspects of soil structure such as pore size distribution and connectivity may exert more direct control, but these properties are rarely available at large spatial scales, particularly in regional and continental surveys.

As a result, characterising variation in porosity across environmental gradients provides a practical means of understanding spatial variability in soil function at landscape to continental scales. At these large scales, the calculation of porosity relies primarily on empirical calculations from routinely measured soil properties, most commonly bulk density and organic carbon (OC) content, with some calculations including texture. Understanding these patterns is particularly important for representing soil–landscape interactions in hydrological and Earth system models, where porosity is a key control on water storage and flow pathways.

Using statistical modelling approaches, OC alone has consistently been identified as a strong predictor of bulk density and porosity across a

* Corresponding author.

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

<https://doi.org/10.1016/j.catena.2026.110657>

Received 5 June 2026; Received in revised form 22 September 2026; Accepted 25 September 2026

Available online 17 October 2026

0341-8162/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

range of soil types and environments, including large-scale datasets (Thomas et al., 2024; Robinson et al., 2022; Robinson et al., 2026). Thomas et al. (2024) found that organic matter was a much better predictor of porosity than texture metrics, and Robinson et al. (2022) established the importance of including organic matter in pedotransfer functions for modelling. Robinson et al. (2026) further demonstrated strong relationships between organic matter and porosity across Europe for the LUCAS dataset. This reflects the multiple roles of organic carbon in soil systems, including its influence on aggregation, biological activity, and pore formation (Lehmann and Kleber, 2015; Six et al., 2002; Tisdall and Oades, 1982; Bronick and Lal, 2005). These multiple roles suggest that spatial variation in OC may translate directly into landscape-scale variability in soil porosity and associated hydrological behaviour. However, the relationship between OC and porosity is not uniform, and varies depending on soil mineral composition, texture, and environmental context (e.g. Thomas et al., 2024; Feeney et al., 2024). In particular, the interaction between OC and clay has been proposed as an important control on soil structural properties (e.g. Dexter et al., 2008; Johannes et al., 2017; Koop et al., 2023). Moreover, it is likely that the geometry of organic carbon constituents, particularly the extent to which they occur as elongated, fibrous or pore-bridging structures, is a more important determinant of soil porosity than the total quantity of carbon present (Robinson et al., 2025).

Work by Dexter et al. (2008) suggests that the amount of OC complexed with clay is a more important control on porosity and soil structure than OC. This has led to the development of simple OC:clay ratios as indicators of soil condition (Dexter et al., 2008; Koop et al., 2023; Johannes et al., 2017), although the usefulness of these metrics has been questioned (e.g. Poeplau and Don, 2023; Feeney et al., 2024).

Dexter et al. (2008) reported that complexed OC was positively associated with bulk volume and matrix porosity, while Koop et al. (2023) identified complexed organic carbon and clay as an important predictor of effective porosity, which was positively correlated with the proportion of complexed organic carbon and clay across a continental-scale dataset. Using independently assessed measures of structural quality, Johannes et al. (2017) proposed SOC:clay thresholds of 1:13, 1:10 and 1:8 to distinguish degraded, moderate, good and very good soil structural condition. The framework was subsequently supported by Prout et al. (2021) using a larger dataset from England and Wales. Together, these studies suggest that OC:clay metrics can capture meaningful variation in soil structure and porosity. Conversely, analysis of a large agricultural soil dataset found weak relationships between these SOC:clay classes and bulk volume, biased by clay content, with the proportion of soils classed as degraded increasing with clay content (Poeplau and Don, 2023). This was supported by Feeney et al. (2024) who found that SOC:clay classes differentiated porosity relatively poorly across European soils and likewise found coarse-textured soils disproportionately classified as structurally “very good” and clay-rich soils as “degraded”.

Conceptually, the work by Dexter et al. (2008) implies that a 1:10 reference OC:clay ratio represents OC complexed with clay, suggesting formation of mineral-associated organic matter (MAOM) by OC and clay up to this ratio, and that any OC above this ratio would be present in other forms e.g. particulate organic matter (POM). However, the 1:10 ratio was proposed as representative rather than universal and the relationship to MAOM is difficult to validate using routine data, hence this interpretation does not commonly accompany use of OC:clay ratios. For example, while Johannes et al. (2017) found OC:clay ratios useful for assessing soil structure, they did not identify a consistent optimum ratio supporting a fixed complexation threshold, and did not frame their work in terms of complexed clay and OC.

There has also been much recent debate over the existence and consistency of an OC saturation point relating to the OC:clay ratio, above which it has been suggested that there would not be further formation of MAOM (Cotrufo et al., 2023). Recent evidence suggests that MAOM accumulation is not constrained by clay content alone, but is also

influenced by carbon inputs, pedogenic oxides, hydrology and pH, indicating that stabilisation capacity cannot be inferred from texture alone (Begill et al., 2023; Poeplau, 2023). Furthermore, apparent limits to MAOM accumulation may reflect dynamic equilibria rather than fixed storage maxima, (Song et al., 2025). Additional capacity may be due to organo-organic associations that allow additional carbon accumulation beyond that expected from mineral surface area alone (Schweizer et al., 2021; Begill et al., 2023). Clay content is also only an indirect proxy for reactive mineral surfaces. Soils with similar clay contents can differ substantially in mineralogy, specific surface area and reactive Fe and Al contents, while fine silt may additionally contribute to mineral-associated carbon storage (Schjønning et al., 2012; Schweizer et al., 2021). Consequently, soils with similar OC:clay ratios may differ substantially in their degree of carbon-mineral association, and OC:clay metrics cannot be assumed to correspond directly to MAOM and POM fractions.

Nonetheless, OC:clay ratio metrics may still act as a proxy to capture systematic variation in soil structure, by positioning soils in the OC:clay compositional space, even if their mechanistic interpretation remains ambiguous. While the general applicability of OC:clay-based thresholds has been questioned, there remains considerable evidence that these metrics capture aspects of soil structural condition. Support comes from observed associations between complexed organic carbon and clay and macroporosity (e.g. Dexter et al., 2008; Koop et al., 2023), and from the continued use of OC:clay ratios as indicators of soil structural quality and vulnerability (e.g. Johannes et al., 2017; Prout et al., 2021). On this basis, their applicability across wider geographical contexts requires further assessment (Sauzet et al., 2024).

In this study, we address this question using a compilation of soil datasets spanning Great Britain, Wales, Europe (LUCAS), and the United States (USDA-NRCS NCSS). In soils subject to tillage, management disrupts soil structure and aggregation processes (Six et al., 2000). Hence by focusing on minimally disturbed topsoils, we aim to examine structural variation driven primarily by natural and semi-natural processes. We evaluate the relationships between porosity, OC, and clay content, alongside two derived metrics representing the balance between OC and clay: a complexed OC–clay metric (from Koop et al., 2023) and a signed metric describing deviation from a reference OC:clay ratio (NonCOCC) (combining metrics from Dexter et al., 2008). We describe these metrics herein as OC–clay balance metrics, and both should be treated as empirical index metrics, given the variability in OC–mineral association in practice. For NonCOCC, positive values indicate relative OC excess and negative values indicate clay excess, without assuming a direct correspondence to specific organic matter fractions. NonCOCC should be regarded as an empirical descriptor rather than a direct measure of any specific SOM fraction. Nonetheless, it can identify systematic variation in the position of soils within OC–clay compositional space using variables that are routinely measured in large-scale monitoring programmes. Consequently, it may help identify consistent structural patterns across environmental gradients and generate hypotheses regarding soil processes that can subsequently be tested using datasets containing direct measurements of organic matter fractions and mineral associations.

Our objectives are to (i) quantify the relative importance of OC and OC–clay balance metrics in explaining porosity across large environmental gradients, (ii) assess whether derived OC–clay indices provide additional explanatory value beyond OC alone, and (iii) explore whether these gradients are associated with changes in aggregate-size distribution in subsets of the data where such information is available. By focusing on empirical patterns across diverse datasets, we aim to provide a data-driven framework for interpreting large-scale variation in soil porosity and its controls, and to assess how these controls may influence soil physical behaviour relevant to water storage and movement across landscapes. Identifying simple indicators and predictors of porosity in available data has direct relevance for Earth system and hydrological modelling, where representation of soil physical properties remains a

major source of uncertainty. By clarifying the relative roles of OC and OC–clay balance, this work could thus provide a basis for improving the representation of soil physical properties in large-scale models. Findings could also help further inform the use of OC:clay metrics as soil health and structure indicators.

2. Materials and methods

All analyses were conducted in R; scripts used for data processing, model fitting and figure generation have been provided as supplementary material.

2.1. Data

We compiled soil datasets from Great Britain, Wales, Europe, and the United States to examine relationships between porosity, organic carbon (OC), and clay content across a wide range of environmental conditions. The combined dataset spans climatic and edaphic gradients representative of temperate landscapes.

UK data were obtained from the UKCEH Countryside Survey 2019 (CS 2019) (Bentley et al., 2023) and the Glastir Monitoring and Evaluation Programme (GMEP, 2013–2016) (Robinson et al., 2019), both based on stratified random sampling designs capturing variation in climate, parent material, and land cover. European data were drawn from the LUCAS 2018 topsoil survey (Orgiazzi et al., 2018), providing harmonized soil measurements across the European Union. United States data were extracted from the USDA-NRCS National Cooperative Soil Survey (NCSS) Characterization Database (<https://ncsslabsdatamart.sc.egov.usda.gov>).

To focus on soil structural variation developed under natural or semi-natural conditions, cropland and arable sites were excluded from the European and UK datasets. For the US dataset, only non-ploughed topsoil horizons were retained. Samples were further filtered to include only those with measurements of bulk density, organic carbon, and clay content. For all datasets, topsoil samples were selected (0–15 cm for UK and Wales; 0–10 or 10–20 cm for LUCAS; A horizons within 25 cm of the surface for the US).

After filtering, the compiled dataset consisted of 6335 samples across all regions. A subset of the GMEP data representing peat soils was retained separately to illustrate patterns at high organic carbon contents. A subset of the Countryside Survey dataset ($N = 293$) containing aggregate-size distribution data was used for additional analysis of aggregation effects. These data were not available in the GMEP, LUCAS 2018 or USA data, and are only available for a subset of CS 2019 sites, a large proportion of which were arable (34%). We therefore retained arable soils to avoid further reducing the sample size.

A summary of datasets, filtering criteria, and sample sizes is provided in Table 1. Information on the distribution of variables in each subset of the data are available in Table S2 and Figs. S1–S4 in the supplementary material.

2.2. Porosity determination

Total porosity (ϕ , $\text{cm}^3 \text{ cm}^{-3}$) was calculated from measured bulk density (ρ_b , g cm^{-3}) and assumed particle density using a two-component mixture approach for organic and mineral material (Eq. (1)) (Robinson et al., 2026). Eq. (1) uses a particle density for SOM (ρ_{SOM}) ($\sim 1.4 \text{ g cm}^{-3}$), and a mineral particle density (ρ_{SM}) ($\sim 2.7 \text{ g cm}^{-3}$) (Robinson et al., 2026).

$$\phi = 1 - \left[\frac{\text{SOM} \cdot \rho_b}{\rho_{\text{SOM}}} + (1 - \text{SOM}) \cdot \frac{\rho_b}{\rho_{\text{SM}}} \right] \quad (1)$$

This approach accounts for the lower particle density of organic matter relative to mineral material and allows consistent comparison of porosity across soils with varying organic matter contents. However, this

Table 1

Summary of the datasets used in the study, how and why they were subset for the analysis, and the number of observations used (n).

Dataset	subset	reason	n
CS 2019 doi: https://doi.org/10.5285/821325f3-b353-4a51-8db2-6b2200d82aca	non arable soils, with clay measurements	remove effects of soil disturbance; clay needed for OC:clay metrics	190
	samples with aggregate-size distribution measurements, including arable	explore relationship to aggregation; data used independently so low n if removing arable soils	293
GMEP doi: https://doi.org/10.5285/85/0fa51dc6-1537-4ad6-9d06-e476c137ed09 doi: https://doi.org/10.5285/d6c3cc3c-a7b7-48b2-9e61-d07454639656	non arable soils, with clay measurements	remove effects of soil disturbance; clay needed for OC:clay metrics	673
	peat only	illustrate relationships for high OC soils	284
LUCAS 2018 https://esdac.jrc.ec.europa.eu/content/luca-s-2018-topsoil-data	non-arable soils	remove effects of soil disturbance	1979
USA http://ncsslabsdatamart.sc.egov.usda.gov/	A horizons	remove effects of soil disturbance	3493

approach does mean that SOM is included in the porosity estimation equation, and it is therefore necessary to test that this does not affect our findings on the drivers of porosity. Therefore, we assessed sensitivity of our models to the inclusion of SOM in the porosity calculation. To do this, we recalculated porosity from bulk density using an average particle density per Eq. (1b) without accounting for SOM content. When applying Eq. (1b) we used an averaged particle density of $\sim 2.53 \text{ g cm}^{-3}$ taken from analysis of the LUCAS data (Robinson et al., 2026). All models were fit separately using porosity estimates from Eq. (1) and then Eq. (1b).

$$\phi = 1 - \left[\frac{\rho_b}{\rho_s} \right] \quad (1b)$$

2.3. Derived variables: OC–clay balance metrics

To explore combined effects of organic carbon (OC) and clay content on soil porosity, we derived a set of indices describing the balance between OC and clay based on commonly used OC:clay ratio concepts (Dexter et al., 2008; Koop et al., 2023). These metrics as implemented here provide a simple way to position soils along a continuum from relative clay excess to relative OC excess.

Following previous work by Dexter et al. (2008), we first calculated theoretical “complexed” fractions of clay and OC assuming a reference ratio of 1 g OC associated with 10 g clay. This ratio has been widely used in the literature as a heuristic threshold for OC–clay interactions, although it is known to vary with mineralogy and environmental conditions (Feeney et al., 2024). While the framing of this ratio by Dexter et al. (2008) implies a relationship to specific organic matter fractions, in this study it is used solely as a pragmatic scaling approach. We acknowledge that the ratio of OC and clay complexation will vary with factors such as clay mineral type and surface area. The reference ratio also disregards other forms of OC mineral association and is therefore used only as an illustrative scaling relationship.

From this framework, we calculated:

- the “complexed OC–clay fraction” (COCC), representing the proportion of OC and clay within the reference balance following Koop et al. (2023);
- a signed index describing deviation from this balance, termed NonCOCC, which combines separate metrics for non-complexed clay and non-complexed OC from Dexter et al. (2008).

The NonCOCC metric is defined such that:

- positive values indicate relative OC excess;
- negative values indicate relative clay excess;
- values range from -1 to $+1$, representing the magnitude and direction of deviation from the reference OC–clay balance.

To aid understanding, a schematic to visualize the NonCOCC metric is supplied (Fig. 1). This shows the metric and where different soils sit in relation to the reference OC:clay ratio of 1:10. Representative examples are shown to illustrate the following soils: Balanced (1:10, NonCOCC = 0); Moderate OC excess (1:5.7, NonCOCC = 0.063); Strong OC excess (1:1, NonCOCC = 0.409); Moderate clay excess (1:15, NonCOCC = -0.313); Strong clay excess (1:26, NonCOCC = -0.595); Organic-rich (1:1.3, NonCOCC = 0.394).

Eqs. (2)–(7) explain the steps for calculation of the COCC and NonCOCC indices, and example calculations are provided in the supplementary material (Table S3). These indices represent normalized indicators of OC–clay balance which we use here to summarise variation in OC–clay balance across datasets and to test whether this variation provides explanatory power for soil.

The metrics for complexed clay and complexed carbon were proposed by Dexter et al. (2008) based on evidence that the amount of complexed carbon is a better predictor of structure than SOM in mineral

soils. Following Dexter, we use Eq. (2) to calculate complexed fraction of clay (CC) and Eq. (3) to calculate complexed fraction of carbon (COC).

$$CC = IF \left[OC < \frac{clay}{n} \right] THEN [OC * n] ELSE [clay] \quad (2)$$

$$COC = IF \left[OC < \frac{clay}{n} \right] THEN [OC] ELSE \left[\frac{clay}{n} \right] \quad (3)$$

Where OC is organic carbon mass fraction of soil ($\text{g } 100 \text{ g}^{-1}$), and clay is the clay mass fraction of soil ($\text{g } 100 \text{ g}^{-1}$) and n is 10. These are equivalent to percentage mass and relate to oven dry soil. Both CC and COC metrics calculated in this way are therefore in units of $\text{g } 100 \text{ g}^{-1}$ soil. Clay and OC must first be expressed in consistent units relative to total soil mass. Clay values should therefore be converted if they are reported relative to the mass of the mineral fraction.

Koop et al. (2023) combined these into a metric of complexed clay and OC (COCC) as a fraction of total clay and OC using Eq. (4). This gives values in g CC and COC per g total C and OC (g g^{-1} or dimensionless).

$$COCC = \frac{CC + COC}{Clay + OC} \quad (4)$$

Given that soils with differing amounts of non-complexed carbon show an inconsistent relationship between COCC and effective porosity (as identified by Koop et al., 2023) we attempt to construct a metric of the combined non-complexed fraction which captures this effect. The non-complexed fractions of clay (NCC) calculated following Dexter, using Eq. (5) and OC (NCOC) calculated using Eq. (6) are based on the same assumed ratio of 1 g of OC to 10 g of clay. However, it is inverted to give the proportion not within the complexed ratio.

$$NCC = IF[(clay - CC) > 0] THEN [clay - CC] ELSE [0] \quad (5)$$

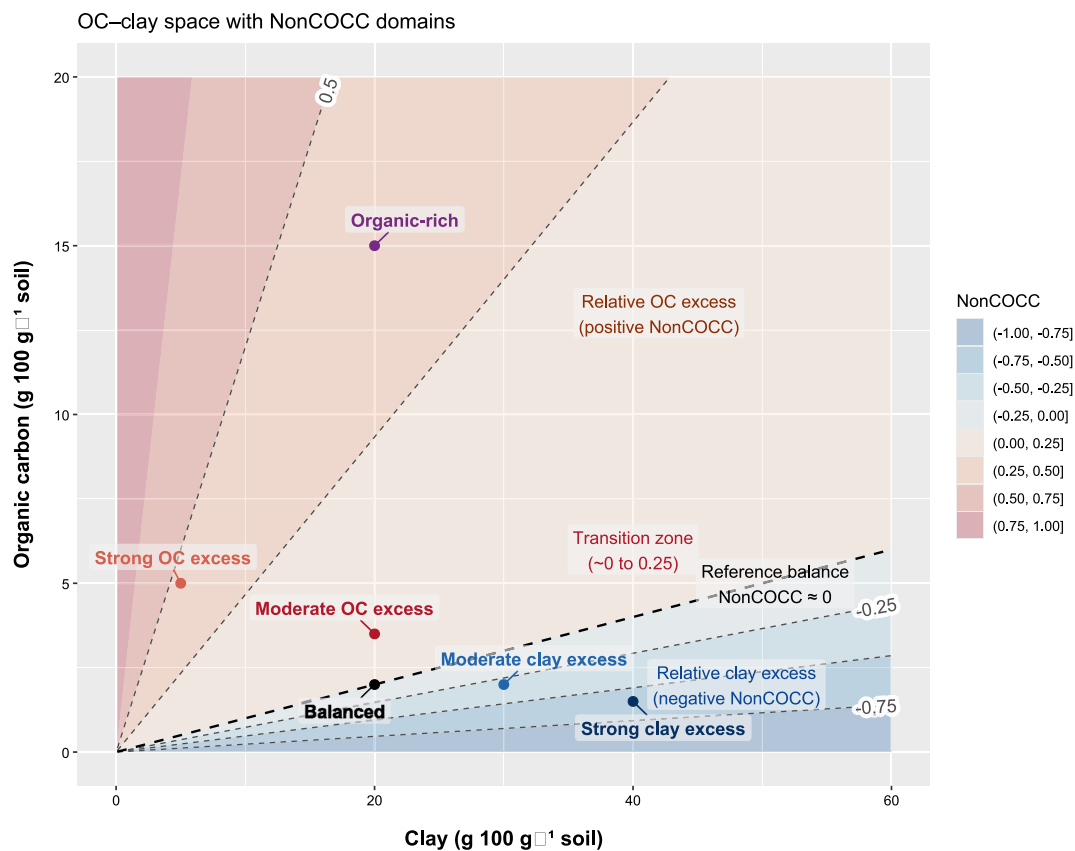


Fig. 1. Schematic of the OC–clay space showing the reference OC:clay ratio of 1:10 and the corresponding NonCOCC domains. The dashed line indicates the reference balance between organic carbon (OC) and clay. Background shading and contour lines show the signed NonCOCC metric, with negative values representing relative clay excess and positive values representing relative OC excess. Points indicate representative soil types and arrows indicate trends.

$$NCOC = IF[(OC - COC) > 0] THEN [OC - COC] ELSE [0] \quad (6)$$

To look for a potential trend across soils with non-complexed clay and soils with non-complexed OC, we have combined these metrics using Eq. (7), in a similar approach to the construction of the COCC metric. The combined NonCOCC metric allows us to look at soils which are over or under saturated with OC on a continuum, using a single metric.

$$NonCOCC = \frac{NCOC - NCC}{Clay + OC} \quad (7)$$

Eq. (7) produces a positive value when there is more than 1 g OC per 10 g clay indicating excess OC relative to the reference ratio. It produces a negative value when clay is in excess because the excess-clay term is subtracted. Values are in units representing g excess OC (positive) or excess clay (negative) per g of OC + clay, on a continuum from -1 to +1. Although the numerical values may be somewhat counterintuitive, this allows separate interpretation of behaviour of soils with excess clay or OC. Assuming that a mineral association of OC and clay represented by the 1:10 ratio captures an important driver of soil structure and porosity both in soils with excess clay and soils with excess OC, we might expect to see a porosity trend relating to this metric across the data. The expected trend is also embedded in concepts such as the OC:clay ratio metric for soil structure defined by Johannes et al. (2017). However, the indices presented are used here strictly as empirical descriptors of OC-clay balance and should not be interpreted as direct representations of soil organic matter fractions or underlying mechanistic processes.

Histograms showing distribution of the calculated metrics of COCC (Fig. S1) and NonCOCC (Fig. S2) for each dataset are presented in the supplementary material, along with further distribution plots of OC (Fig. S3) and bulk density (Fig. S4) and summary statistics (Table S2). Table S3 in the supplementary material steps through example calculations for Eqs. (2),(3),(4) and (5), and compares these values to a conventional OC:clay ratio.

2.4. Statistical modelling

To evaluate the relative contributions of organic carbon (OC), clay content, and derived OC-clay balance metrics, a series of generalized additive models (GAMs) were fitted in R (R Core Team, 2025), using the mgcv package (Wood, 2011). Porosity models used a Gaussian error distribution with an identity link, so fitted effects were additive on the original porosity scale. Aggregate-fraction models used a Tweedie distribution with a log link to accommodate non-negative, heteroscedastic responses. The objective was to infer general relationships across datasets. Dataset was therefore included as a fixed effect to account for systematic differences. Dataset-specific models and factor-smooth interaction models were used to evaluate whether predictor effects were dataset dependent. We allowed nonlinear relationships because relationships between OC, clay, and porosity are known to vary across soil types and OC ranges, and are unlikely to be adequately captured by linear formulations. GAMs were used to flexibly represent these nonlinear relationships between variables without imposing a pre-defined functional form. Models were fitted using restricted maximum likelihood (REML). Continuous predictors were represented using shrinkage cubic regression splines:

$$y \sim Dataset + s(x, bs = "cs") \quad (8)$$

where y is the response variable, x is the continuous predictor and $s(x)$ is a smooth function estimated using a shrinkage cubic regression spline ($bs = "cs"$). The Gaussian porosity models assumed normally distributed residual errors, ϵ , around the fitted mean. The spline basis allows nonlinear relationships to be represented while penalisation controls

unnecessary complexity. Smoothness parameters were estimated using REML, balancing model fit against smooth complexity. The shrinkage basis, together with `select = TRUE`, allows weakly supported smooth terms to be penalised towards zero to reduce the risk of overfitting (Wood, 2003; Wood, 2017). The models constructed directly using the mgcv package allow comparison of model parameters (Table 2). Where required for visualization in figures, fitted relationships and approximate 95% pointwise confidence bands were generated using `geom_smooth()` in the "ggplot2" package (Wickham, 2016), for plotting.

2.5. Hierarchical model-comparison framework

A hierarchical model-comparison framework was implemented to construct models 1–20 for total porosity calculated using Eq. (1), these are shown in Table 2. A parallel set of models (Table S6) were constructed to predict total porosity calculated using Eq. (1b) without accounting for the differing particle density of organic and mineral material, and thus test for influence of including OM in the porosity calculation.

Initial models (1–4) evaluated the explanatory ability of OC, clay content, NonCOCC, and COCC individually. Subsequent models (5–8) explicitly included separate smooth terms for clay and OC alongside OC-clay balance metrics to test whether these derived metrics provided additional explanatory value beyond the constituent variables from which they were calculated. Smooth-term statistics were used to describe conditional support for individual terms within each fitted model, while changes in deviance explained and AIC were used to assess whether adding predictors improved overall model fit. These variables should not be interpreted as independent predictors, since COCC and NonCOCC are alternative representations of position relative to the 1:10 reference balance. Models containing both metrics were used to assess whether their alternative representations captured different aspects of the fitted relationship, rather than to partition independent contributions between them.

To evaluate the influence of vegetation and management effects, land-cover category was incorporated into additional GAMs using two approaches. To test for systematic differences, land-cover class was included as a parametric factor. Alongside this, factor-smooth interactions ($bs = "fs"$) were used to allow predictor-response relationships to vary among land-cover categories. These models were fitted separately for OC, clay, NonCOCC and COCC (Models 9–12). Harmonized land-cover information was unavailable for the US dataset. All US observations were therefore assigned to a separate category labelled "US A horizons" in the land-cover variable, allowing them to be retained in models containing land-cover class. This category represents dataset membership and was not interpreted as a land-cover effect. Equivalent factor-smooth models were also fitted for dataset identity (13–16), enabling the predictor-porosity relationship to vary among datasets. Model performance was compared to determine whether accounting for land-cover or dataset-specific responses improved model fit. Land-cover class was also included as a factor in a model with OC, clay content, NonCOCC, and COCC together (Model 17); and as both factor and OC percent smooth (Model 18). We similarly tested using dataset as an OC percent smooth in a combined model (Model 19). Finally, we tested including OC ratio in the combined model (20) to test whether the derived COCC and NonCOCC metrics provide different information than a simple OC:clay ratio.

To examine potential differences in behaviour across soil types, analyses were also stratified by clay content (using a pragmatic threshold to split $<25\%$ and $\geq 25\%$). On these subsets, we constructed models for NonCOCC and COCC individually, and a combined model with OC, clay content, NonCOCC, and COCC together in which we can compare conditional fitted contributions (Table S8).

Table 2

Comparison of model parameters for GAMs of the relationships for OC, clay, COCC and NonCOCC to porosity.

Model	Purpose	Formula	RMSE	Deviance Explained %	AIC
Model 0	Reference model for predictor comparisons	Dataset	0.09	36.28	−12,684
Model 1		Dataset + OC (%)	0.07	64.05	−16,297
Model 2		Dataset + Clay (%)	0.09	36.43	−12,696
Model 3	Compare individual parameter models	Dataset + NonCOCC	0.08	53.18	−14,623
Model 4		Dataset + COCC	0.09	37.41	−12,788
Model 5		Dataset + OC (%) + Clay (%)	0.07	64.56	−16,376
Model 6	Test the additional value of OC–clay balance metrics	Dataset + OC (%) + Clay (%) + NonCOCC	0.07	65.27	−16,498
Model 7	when included with OC and clay	Dataset + OC (%) + Clay (%) + COCC	0.07	65.09	−16,465
Model 8		Dataset + OC (%) + Clay (%) + NonCOCC + COCC	0.07	65.26	−16,495
Model 9		Dataset + Land-cover class + (OC (%) Smoothed by land-cover class)	0.06	66.65	−16,736
Model 10	Test variation in individual predictor relationships	Dataset + Land-cover class + (Clay (%) Smoothed by land-cover class)	0.08	42.76	−13,326
Model 11	between land-cover classes.	Dataset + Land-cover class + (NonCOCC Smoothed by land-cover class)	0.07	54.63	−14,790
Model 12		Dataset + Land-cover class + (COCC Smoothed by land-cover class)	0.08	49.12	−14,067
Model 13		Dataset + Land-cover class + (OC (%) Smoothed by Dataset)	0.06	66.71	−16,754
Model 14	Test variation in individual predictor relationships	Dataset + Land-cover class + (Clay (%) Smoothed by Dataset)	0.08	42.83	−13,334
Model 15	between datasets.	Dataset + Land-cover class + (NonCOCC Smoothed by Dataset)	0.07	55.01	−14,844
Model 16		Dataset + Land-cover class + (COCC Smoothed by Dataset)	0.08	49.34	−14,095
Model 17	Combined-model baseline	Dataset + Land-cover class + OC (%) + Clay (%) + NonCOCC + COCC	0.06	66.64	−16,747
Model 18	Test land-cover variation in the OC relationship in combined model.	Dataset + OC (%) + (OC (%) Smoothed by land-cover class) + Clay (%) + NonCOCC + COCC	0.06	67.30	−16,840
Model 19	Test dataset variation in the OC relationship in combined model.	Dataset + OC (%) + (OC (%) Smoothed by Dataset) + Clay (%) + NonCOCC + COCC	0.06	66.01	−16,612
Model 20	Test the additional value of the OC:clay ratio in combined model	Dataset + OC (%) + Clay OC ratio + Clay (%) + NonCOCC + COCC	0.07	65.26	−16,495

The full suite of models is in supplementary Table S4 with additional diagnostic parameters. Models here are for porosity calculated accounting for OC fraction, which can be contrasted with models in supplementary Table S5 where porosity is calculated using assumed average particle density. See section 5.4 for definitions of statistical terms.

2.6. Model diagnostics and derivative analysis

The purpose of these models was comparative rather than predictive: to assess the relative explanatory contributions of OC and derived OC–clay indices, and to determine whether these indices provide additional information beyond OC and clay alone. Model performance and contribution of terms was evaluated using deviance explained, Akaike's information criterion (AIC), and approximate significance tests for model terms (F-statistics). These statistics were calculated using the mgcv package in R and interpreted following Wood (2017) and the package documentation (Wood, 2025). We also calculated RMSE, expressed in the units of the response variable, which therefore provides a directly interpretable measure of model error (Chai and Draxler, 2014).

Deviance explained ($D_{\text{explained}}$) was calculated using Eq. (9):

$$D_{\text{explained}} = 1 - \frac{D_{\text{residual}}}{D_{\text{null}}} \quad (9)$$

Where D_{residual} is the deviance of the fitted model and D_{null} is the null deviance. Values are reported as percentages.

RMSE was calculated on the response scale using Eq. (10):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (10)$$

Where y_i and $\hat{y}_i$ are the observed and fitted values, respectively, and n is the number of observations included in the fitted model. RMSE is therefore expressed in the units of the response and represents in-sample error rather than performance on independent data.

AIC was calculated using Eq. (11):

$$\text{AIC} = -2\log(L) + 2k \quad (11)$$

where L is the maximised model likelihood and k is the effective model degrees of freedom. AIC was compared only among models fitted to the same response observations using the same distributional assumptions. Lower values indicate a more favourable balance between model fit and complexity.

For each smooth term, the estimated degrees of freedom (edf) describe the complexity of the fitted relationship. An edf close to 1 indicates an approximately linear relationship, values greater than 1 indicate increasing non-linearity, and values approaching 0 indicate that the term has been strongly penalised towards zero. The reference degrees of freedom (Ref.df) are used in the approximate significance test and may differ from the edf.

The F-statistic and associated approximate p -value assess the conditional contribution of each smooth term, given the other terms in the model. Within a given fitted model, larger approximate F-statistics can indicate stronger evidence for, or a larger model-specific contribution from, a smooth term. However, F-statistics depend on the fitted smooth, its penalty and reference degrees of freedom, and are not standardised measures of effect size or variable importance. F-statistics were compared only among smooth terms within the same model and were interpreted alongside edf, fitted partial effects and concavity. The associated approximate p -value is indicative of statistical significance. All significance tests were interpreted alongside effect magnitude, fitted relationships and overall model fit.

Models of aggregate-size distribution were fitted using a Tweedie distribution with a log link. The Tweedie power parameter, p , was estimated during model fitting and is reported as Tweedie p . It defines

the modelled mean–variance relationship using Eq. (12):

$$\text{Var}(Y) = \phi\mu^p \quad (12)$$

Where μ is the fitted mean and Φ is the dispersion parameter. In this implementation in *mgcv*, Tweedie p is estimated between 1 and 2; values approaching 1 indicate a more Poisson-like mean–variance relationship, while values approaching 2 indicate a more gamma-like relationship.

To interpret the partial relationship between NonCOCC and porosity in the GAM including OC and clay (Model 6) we also constructed this model separately for each dataset and land-cover class, and compared the partial response curves (Table S7, Fig. S8). To investigate whether regions of rapid porosity response could be identified objectively, derivatives of the fitted GAM partial response curves were calculated using the *gratia* package (Simpson, 2024) and plotted (Fig. S9). Peak response locations were identified as the maximum derivative of the NonCOCC smooth. In addition, the range encompassing the upper 10% of derivative values was used to characterise regions of maximum porosity sensitivity. Derivative analyses were conducted for both the pooled dataset and individual regional datasets.

Because NonCOCC is derived from OC and clay content, some of the multivariable models exhibited substantial concurvity. Concurvity describes the extent to which one smooth term can be approximated by one or more other terms in the model and is analogous to multicollinearity in linear models. Values range from 0, indicating no concurvity, to 1, indicating that a term is fully represented by other model terms. Derivative analyses were therefore interpreted as describing conditional partial relationships within the fitted GAM rather than independent effects of NonCOCC.

2.7. Aggregate-size distribution analysis

For the subset of samples with available aggregate-size data (CS 2019, $N = 293$), we examined relationships between porosity, NonCOCC, and aggregate-size fractions. Aggregate-size classes were grouped into three categories (2–20 μm , 20–256 μm , and 256–2000 μm), and their relative contributions as a proportion of all size classes (including $<2 \mu\text{m}$) were analysed as a function of NonCOCC. These analyses were used to explore whether observed OC–clay balance gradients correspond to systematic changes in aggregation, which provides a potential structural pathway linking OC enrichment to porosity. Because this subset included both arable and non-arable sites, land-use type was retained as a categorical variable for interpretation but not used as a filtering criterion, given the limited sample size. This analysis provides an independent line of evidence to test for relationship of OC–clay balance to structural organisation through aggregation.

The same basic model construction approach as described in 2.4 for porosity was repeated to test statistical relationships with proportion macro-aggregates as the response variable. Here, separate models were constructed using:

- OC as a single predictor
- NonCOCC as a single predictor
- OC + NonCOCC

Each of these models was also constructed with land-cover class included as a parametric factor to allow for systematic variation. Land-cover class was then also included as a smooth to allow the relationship to the continuous predictor variable to vary between land-cover types. All models are shown in Table 4. Finally, a model with porosity as the response variable was constructed on this dataset. This was to allow inclusion of the aggregates term, and assess predictive power of proportion macro-aggregates relative to the OC percent and NonCOCC terms also included in the model (Table 5).

3. Results

Table 2 shows the performance of GAMs constructed to identify relationships between predictor variables and total porosity. These are not intended to be used as predictive models outside our datasets, but rather to test and compare the relationships in the data.

3.1. Porosity as a function of organic carbon

Comparing deviance explained by the GAMs in Table 2, OC (Model 1) explains most deviance of the individual parameter models (1–4) and models containing OC consistently perform better. This remains true for statistical models of porosity where the response variable was calculated assuming an average particle density per Eq. (1b) (see Table S5), suggesting that this finding is not an artefact of the inclusion of OM in the porosity calculations. Variation in modelled relationships between land-cover classes and datasets is discussed in Section 3.3.

Relationships between organic carbon (OC) and total porosity are shown in Fig. 2. In Fig. 2a, data are grouped by region, highlighting differences in the distribution of OC across datasets (see also Fig. S3 and Table S2). Following the filtering applied in this study, the CS 2019 and GMEP mineral soil subsets do not exceed approximately 25% OC because clay measurements were unavailable for organic soils. To provide context for higher OC contents, the GMEP peat dataset is presented separately in Fig. 2a. The LUCAS 2018 dataset includes a small number of samples with higher OC values (up to ~55%), whereas the US dataset is largely restricted to low OC soils.

Fig. 2b presents the UK and European datasets separated by broad land-use classes (grassland, shrubland, wetland, and woodland). Grassland, shrubland, and wetland categories are dominated by mineral soils, whereas woodland sites include a greater proportion of higher OC soils, particularly within the LUCAS 2018 dataset. Variability is explored statistically in Section 3.3.

Across all datasets, porosity increases with increasing OC, although the distribution of values varies among regions and land-use types. The GAM smooths capture a consistent positive relationship between OC and porosity, with variability around the trend reflecting differences in soil properties, environmental conditions and site factors such as management. The conceptual reference curves included in Fig. 2 provide context for the expected upper response of porosity to increasing OC (Robinson et al., 2025). As expected the observed data generally fall below these curves.

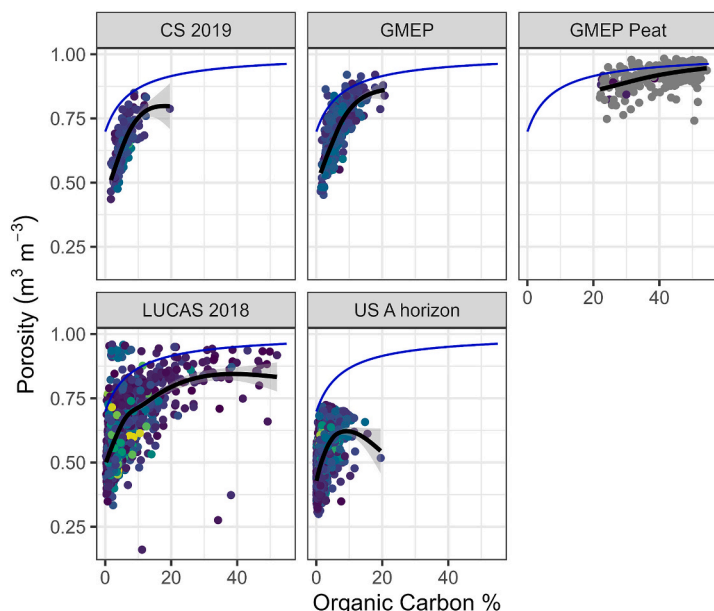
3.2. Complexed and non-complexed carbon and porosity

The relationships between clay, organic carbon (OC), and the derived OC–clay balance metrics are illustrated in Fig. 3. Fig. 3b shows the distribution of our data in the context of the schematic Fig. 1. To further aid interpretation, the NonCOCC metric is also shown in relation to a basic OC:clay ratio in Fig. S5.

Both metrics are derived from the same OC–clay compositional space but provide contrasting representations of deviation from the 1:10 reference ratio. Fig. 3a shows that low COCC values occur where either clay or OC is present in excess relative to the reference ratio, whereas high COCC values indicate conditions closer to the 1:10 reference ratio. Most samples fall within the upper range of COCC values (approximately 0.75 to 1; see also Fig. S1), reflecting the predominance of soils with relatively high OC relative to clay across the datasets. In contrast, the NonCOCC metric provides a continuous representation of deviation from the reference OC–clay balance (Fig. 3b) with negative values representing excess clay and positive values representing excess OC. The distribution of samples shows a substantial proportion clustered between 0 and 0.25 (Fig. S2), indicating that many soils fall within a relatively narrow range of moderate OC enrichment.

The two metrics exhibit markedly different relationships with porosity (Fig. 4). COCC (Fig. 4a) shows weak and highly scattered

a) Total porosity over OC percent by dataset



b) Total porosity over OC percent by land use

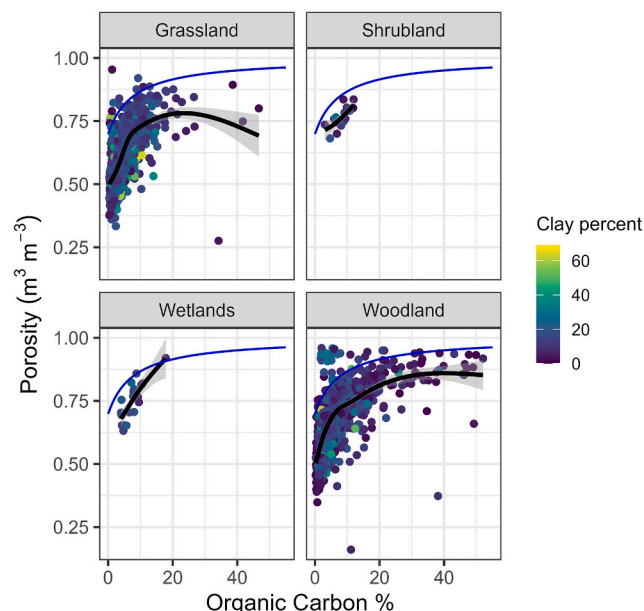
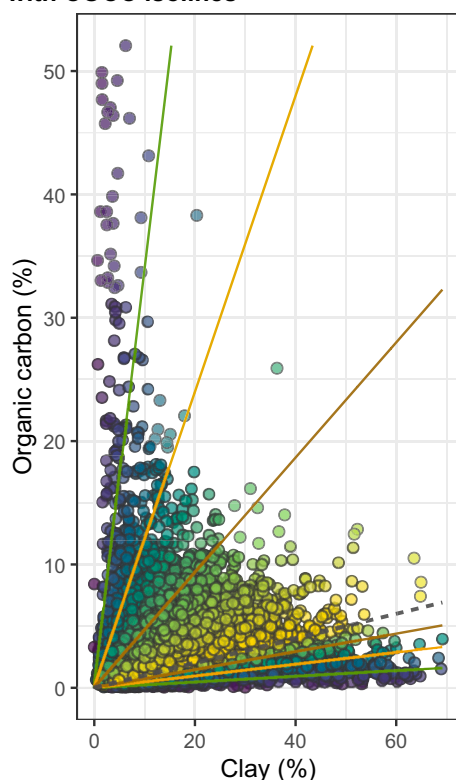


Fig. 2. The relationship between organic carbon (OC) percent and total porosity a) split by region showing UK (CS 2019, GMEP and GMEP Peat), and Europe (LUCAS 2018) and USA (US A horizon). b) The UK and Europe datasets combined and split by land use. The blue line is proposed as an upper level conceptual reference curve for the expected increase in porosity with increasing OC (per Robinson et al., 2025). This curve is illustrative and does not explicitly represent the full range of structural processes operating in soils. The black line is a GAM smooth of the relationship between porosity and OC in the data. Omission of organic soils from the CS 2019 and GMEP data subset used here explains the low OC content and small sample size for wetland habitats in Fig. 2b. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

a) OC as a function of clay with COCC isolines



b) OC as a function of clay with NonCOCC isolines

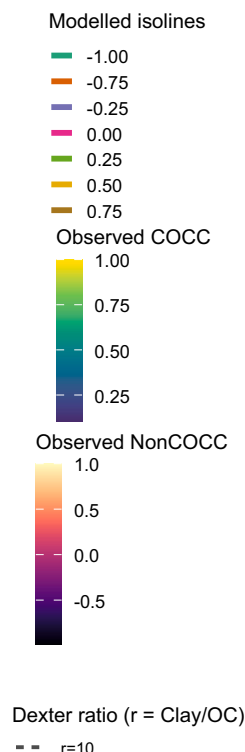
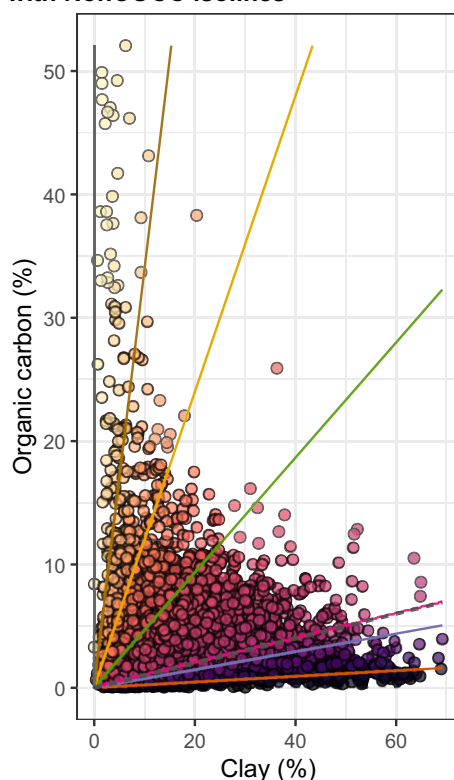


Fig. 3. COCC and NonCOCC metrics in relation to clay and OC percent. Both plots show OC as a function of clay. In Fig. 3a points are colored by the calculated COCC metric, with isolines capturing the trend of COCC values with changing clay and OC. In Fig. 3b points are colored by the calculated NonCOCC metric, with isolines capturing the trend of NonCOCC values with changing clay and OC.

associations with porosity. Conversely, NonCOCC (Fig. 4b) shows a clearer and more consistent pattern than COCC. Across the dataset, porosity tends to increase with increasing NonCOCC, although variability remains within all ranges of the metric. Relationships were weakest and most variable under relative clay excess (negative NonCOCC values), while a steeper increase in porosity was observed at low positive NonCOCC values, particularly in soils with $\geq 25\%$ clay. The concurrent influence of total OC likely contributes to scatter in Fig. 4b, with higher porosity values for a given NonCOCC value observed at higher clay contents (with associated higher OC).

When the relationship for COCC is plotted separately for the different datasets or habitats, differing trends can be seen within these subsets, but there remains a significant scatter (Fig. S10). Variation in modelled relationships between land-cover classes and datasets is explored in Section 3.3.

Model performance for COCC was poor (Model 4, Table 2), with only slightly more deviance explained than the clay only model (37.41% compared to 36.43%). Whereas the NonCOCC metric (Model 3) performed better, with 53.18% deviance explained. The NonCOCC metric also performed better for the high and low clay subsets (Table S8).

Both COCC and NonCOCC improved model performance and remained statistically significant when added to models already containing OC and clay. However, the additional explanatory value of both metrics was modest, increasing deviance explained from 64.56% (Model 5) to 65.09% by adding COCC (Model 7) and 65.27% by adding NonCOCC (Model 6). In Model 6 NonCOCC had larger F statistic than clay but smaller than OC ($F = 18$ for NonCOCC, $F = 1$ for clay, $F = 56$ for OC). Conversely, the COCC smooth term in Model 7, had an F-statistic substantially lower than OC and only slightly greater than clay (Table S6). Because the predictors exhibited high concavity, these rankings should be interpreted as model-specific rather than as a stable partitioning of independent importance. Consistent with this, and as expected from the relationship between the variables, inclusion of COCC alongside OC, clay and NonCOCC (Model 8) provided no further improvement in model performance, slightly reducing both deviance explained and AIC. The same patterns were observed for models where the porosity response variable was calculated assuming an average particle density per Eq. (1b). This was evident in both relative model performance (see Table S5) and relative smooth-term F-statistics within the individual models (see Table S6).

The partial response of NonCOCC in Model 6 (Fig. S8) closely resembled the overall relationship shown in Fig. 4b, with the greatest increase in porosity occurring between approximately 0.06 and 0.24 NonCOCC. Again, this was similar for the models of porosity response

variable calculated assuming average particle density per Eq. (1b). Derivative analysis of the modelled NonCOCC-porosity relationship can help to quantify the region of maximum response, and results in Section 3.3 indicate the range and explore consistency across our datasets.

Collectively, these results indicate that while both OC-clay balance metrics capture some variation beyond OC and clay individually, NonCOCC provides the more informative representation of OC-clay compositional variation. Both explain substantially less variance in the data than OC, which remained the dominant predictor of porosity across all models.

3.3. Land-cover and regional context

The trend with OC appears to vary more between categories in the land-cover groupings than between datasets. This difference cannot be attributed solely to land cover because the categories included an unclassified US dataset category. Allowing the OC-porosity relationship to vary among land-cover classes or datasets improved model performance (Models 9 and 13 versus Model 1), with higher deviance explained and lower RMSE and AIC. Although the land-cover- and dataset-specific smooths were significant (Table S6), they explained substantially less variation than the principal OC smooth. The dominant positive association between OC and porosity was retained across land-cover classes and datasets, although the shape of the relationship varied among habitats. When Model 6 was fit to the regional datasets, OC consistently had the largest approximate smooth-term F-statistic and showed a strong fitted relationship with porosity, and trends were visually similar (Fig. S8). Analysis of the OC derivative for the regional datasets (Table 3) did not identify an internal maximum for rate of change. In all datasets, maximum derivative values occurred at the minimum observed OC content, suggesting greater porosity sensitivity at low OC. For the combined data and the regional datasets, the marginal response of porosity to OC decreases as OC increases, suggesting diminishing marginal gains in porosity with increasing carbon content, rather than the existence of a distinct optimal OC threshold. For the land-cover class subsets, OC again had the largest approximate F-statistic, but edf varies (Table S7) reflecting different trends between habitats (Fig. S8). The relationship appears more linear within wetlands and shrublands (Fig. S8), with edf close to 1 (Table S7). Alongside, Fig. 2 these models demonstrate that OC is consistently a strong predictor of porosity across diverse soil types and environmental settings. This also highlights substantial variation in relationship to porosity associated with land use.

Allowing relationships to vary between land-cover classes or datasets for NonCOCC (Models 11 and 15 versus Model 3) and COCC (Models 12

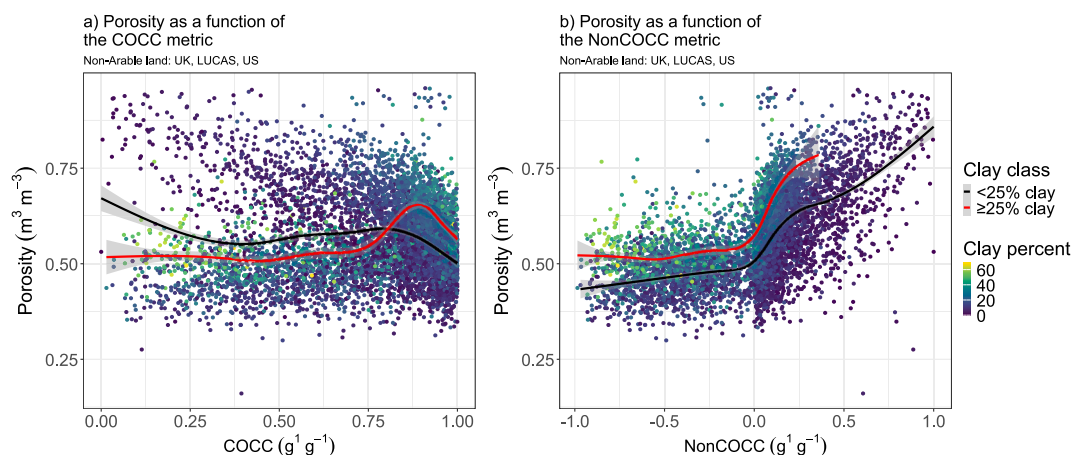


Fig. 4. Relationships between OC-clay balance metrics and porosity. a) COCC metric calculated per Eq. (4); b) NonCOCC metric calculated per Eq. (7) where negative values indicate “excess” clay and positive values indicate “excess” OC, relative to the reference balance. Trendlines for the best fit smooth of porosity predicted from COCC (a) and NonCOCC (b) using a GAM are constructed separately for datapoints with Clay <25% (black line) and for datapoints with clay $\geq 25\%$ (red line). Colour ramps indicate the clay content. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and 16 versus Model 4) also improved model performance. Both metrics remained weaker predictors of porosity than OC when comparing deviance explained to models 9 and 13 (Table 2). Improvement was greater for COCC as shown by the greater increase in deviance explained when comparing models, and (the higher approximate F-statistics for the factor-smooth terms in models 12 and 16 (Table S6). This suggests that the relationship of porosity to COCC may be more context dependent and vary more strongly among land-cover classes and regions. In contrast, improvements for NonCOCC were modest, and the land-cover class factor smooth had weaker conditional contribution in the model, although for dataset it had higher conditional contribution ($F = 4.42$ for NonCOCC Smoothed by Dataset compared to $F = 3$ for the principal NonCOCC smooth; see Table S6).

Comparing influence of metrics across regional versions of Model 6, the order of variable conditional contribution was consistently OC, NonCOCC then clay (based on F statistic, see Table S7). Across the land-cover class subset versions of Model 6 however, NonCOCC made very little contribution except in grasslands (low F statistic and relationships were penalised towards 0). When COCC was also included in regional and land-cover class subset versions of Model 8, there was generally marginal improvement in performance for the dataset or land-cover class subsets based on deviance explained and RMSE when compared to the corresponding subset version of model 6. However examination of smooth parameters shows variation in relative contribution of the COCC term in these subset models based on F statistics. For models constructed on the CS regional dataset and the woodland land-cover class subset, COCC had greater conditional contribution than clay or NonCOCC, which contribute little or nothing. In the other regional models, COCC sometimes had higher F-statistic than clay, but NonCOCC always had higher F-statistic than both. The land-cover subset models also showed variable contributions from NonCOCC and COCC. In the grassland land-cover subset model NonCOCC had higher F-statistic than clay, and COCC had lowest F-statistic. In the wetland and shrubland subset models, both NonCOCC and COCC contributed little or nothing. Across the regional and land-cover class subset versions of Model 8, OC consistently had greatest conditional contribution in the models, with consistently highest F value (Table S7).

Overall, these results suggest that the NonCOCC-porosity relationship appears more reproducible across regional datasets than within individual habitats. This indicates that the additional information

captured by NonCOCC may be most evident when considered across broad environmental gradients spanning multiple habitat types, rather than within individual habitat types.

We applied derivative analysis independently to each regional version of Model 6 and identified that the greatest rates of porosity change generally occurred at low-positive NonCOCC values (Table 3). To describe where the fitted partial response changed most rapidly, we identified the range of predictor values associated with the highest decile of estimated derivative values. The maximum estimated derivative occurred between NonCOCC values of 0.04 and 0.16 across regional dataset models, while the predictor ranges associated with the highest derivative decile were generally concentrated around the transition between relative clay-excess and OC-excess. An exception was the GMEP dataset, for which this range was broadly centred around 0, extending from -0.07 to 0.06 . The range identified for the combined dataset is most representative of the US dataset, reflecting the greater representation of US observations in the combined dataset. The precise form and range of the response varied among datasets, as shown by the derivative plots in Fig. S9. Nonetheless, all regional analyses identified a broadly consistent region of elevated sensitivity.

3.4. Carbon in excess, aggregate generation

Fig. 5 presents the subset of the Countryside Survey (CS 2019) data for which aggregate-size distributions (<2 mm) are available ($N = 293$). Arable sites (34% of the subset) were not excluded due to the limited sample size, however major differences in porosity, aggregate-size distribution, and NonCOCC values are evident between arable and non-arable samples in Fig. 5. Arable sites generally had lower porosity and a higher proportion of intermediate-sized aggregates (20–256 μm). They also had lower NonCOCC values (typically <0.25), including the small number of samples with negative NonCOCC values. These differences in the distribution of NonCOCC and aggregate fractions are reflected in the fitted trends for arable and non-arable sites. Although aggregate-size distributions differed substantially between arable and non-arable soils, statistically significant relationships for both OC and NonCOCC with aggregate abundance remained evident within each subset. This indicates that the observed patterns are not solely an artefact of differences in disturbance history or land-use category.

Across the full dataset, higher porosity values (> 0.6) are associated

Table 3

Derivative analysis for regional and combined models to identify region of maximum (top 10%) porosity sensitivity for key predictor variables in Model 6.

Derivative metrics for NonCOCC						
Dataset	Peak Derivative (rate of increase in porosity)	Predictor (NonCOCC) value at maximum estimated derivative	Peak Derivative Lower CI	Peak Derivative Upper CI	Lower predictor bound highest decile	Upper predictor bound highest decile
Combined	0.19	0.14	0.13	0.25	0.06	0.24
CS 2019	0.31	0.16	0.12	0.50	0.13	0.20
GMEP	0.17	0.04	-0.01	0.35	-0.07	0.06
LUCAS						
2018	0.14	0.08	0.08	0.21	0.01	0.18
US A						
horizon	0.25	0.13	0.17	0.33	0.07	0.24
Derivative metrics for OC						
Dataset	Peak Derivative (rate of increase in porosity)	Predictor (OC) value at maximum estimated derivative	Peak Derivative Lower CI	Peak Derivative Upper CI	Lower predictor bound highest decile	Upper predictor bound highest decile
Combined	0.05	0.04	0.04	0.06	0.04	5.29
CS 2019	0.03	1.72	0.02	0.05	1.72	3.34
GMEP	0.03	1.32	0.02	0.05	1.32	3.08
LUCAS						
2018	0.03	0.24	0.02	0.04	0.24	4.95
US A						
horizon	0.03	0.04	0.01	0.05	0.04	1.81

Confidence intervals refer to the estimated derivative at the identified maximum and not to the predictor value at which the maximum occurred. The highest derivative decile represents the range of predictor values associated with derivative estimates at or above their 90th percentile and is a descriptive interval rather than a confidence interval or threshold estimate.

with an increased proportion of macro-aggregates (256–2000 μm) and lower contributions from micro-aggregates (Fig. 5 a,e). These higher porosity conditions are generally observed at higher NonCOCC values (Fig. 5 b,f). In contrast, lower porosity values are associated with higher contributions from smaller aggregate-size classes and relatively low proportions of macro-aggregates.

The relationship between aggregate-size distribution and NonCOCC is shown explicitly in Fig. 5 b,d,f. The proportion of smaller aggregate-size classes declines with increasing NonCOCC, while the proportion of larger aggregates increases. In particular, the fraction of macro-aggregates ($>256 \mu\text{m}$) increases across the range of NonCOCC values between 0 and approximately 0.25 (Fig. 5f), with lower values observed for arable sites compared to non-arable sites. In this same range, the proportion of meso-aggregates (20–256 μm) decreases (Fig. 5d).

At higher NonCOCC values (>0.25), only non-arable sites are represented, and samples are characterised by a high proportion ($>75\%$) of macro-aggregates. These conditions are also associated with the highest porosity values, as indicated by the colour scale in Fig. 5f.

It is important to note that aggregate size class also shows a strong relationship to OC, and this can be seen in Fig. S11, with higher OC values associated with an increased proportion of macro-aggregates (Fig. S11c) and lower contributions from smaller aggregate-size classes (Fig. S11a,b). We therefore again use GAMs to test the relative contributions of OC and NonCOCC.

Table 4 shows the performance of GAMs constructed to identify relationships between predictor variables and proportion macro-aggregates response variable. Comparing deviance explained by these GAMs, OC (Model 21) individually explains more deviance than the NonCOCC metric (Model 22). However, since both showed a strong relationship with proportion macro-aggregates, we evaluated whether NonCOCC provided additional explanatory value in a combined model with OC. As in the porosity models, shrinkage penalisation allowed weakly supported smooth terms to be reduced towards zero. In Model 23, NonCOCC remains statistically supported after accounting for OC. Again, its relative contribution is smaller, as reflected in the lower F-statistic compared to OC ($F \sim 3$ for NonCOCC versus $F \sim 23$ for OC). The increase in deviance explained from 61.32% to 64.79% and reduction in AIC from 2478 to 2456 indicate that the NonCOCC transformation captured residual structure not represented by the OC smooth alone.

Accounting for land-cover class as a factor (Model 24) and as a smooth to allow different OC-aggregate proportion relationship (Model 27) both improved the basic OC-aggregate proportion model (Model 21) with higher % deviance explained and lower AIC. The increase in deviance explained was greater than achieved for Model 23 by adding NonCOCC to the model without accounting for land-cover class. However adding NonCOCC to the models accounting for land-cover class further improved performance. Deviance explained increased from 70.11% in Model 27 with OC smoothed by land-cover class, to 72.01% in Model 29 which included both OC and NonCOCC smoothed by land-cover class. AIC was also reduced indicating that the improvement in performance justifies additional complexity. Smooth terms for Model 29 indicate that the contribution from NonCOCC in this model is smaller than OC as would be expected. Overall, the relative performance of all models in Table 4 indicates that land-cover class explained a substantial component of variation in aggregate-size distribution. This confirms the importance of management and disturbance effects. However, inclusion of land-cover class did not eliminate the contribution of NonCOCC, which remained a significant predictor after accounting for land-use differences. This indicates that the observed association between NonCOCC and aggregation cannot be explained solely by differences between arable and non-arable soils.

The statistical model of porosity on this subset which includes proportion macro-aggregates as a predictor (Table 5) shows that proportion macro-aggregates, whilst significant, had lower conditional contribution in the model than both OC and NonCOCC.

4. Discussion

4.1. Porosity as a function of OC

Across all datasets analysed, organic carbon (OC) is the strongest and most consistent statistical predictor of soil porosity (Fig. 2; Table 2), supporting previous findings for GB (Thomas et al., 2024) and EU (Robinson et al., 2026) datasets, and extending them to the US. OC explains substantially more variation than NonCOCC or COCC. This finding aligns with previous work e.g. Johannes et al. (2017) which similarly identified OC as the dominant predictor over OC–clay metrics.

However, this contrasts with earlier work by Dexter et al. (2008) where clay–carbon interaction metrics, in particular COCC, showed stronger relationships with soil structural properties. A key difference is that Dexter et al. (2008) considered agricultural soils spanning a relatively narrow geographical range with low OC, whereas the datasets analysed here encompass much broader variation in soil types, climates, and carbon contents. These differences in the datasets likely underpin the divergence in findings.

Organic carbon influences porosity through multiple interacting pathways, including, geometry, aggregate formation, biological activity, and the stabilisation of pore networks (Lehmann and Kleber, 2015; Six et al., 2002; Tisdall and Oades, 1982). At regional to continental scales, these combined effects appear to dominate, making total OC a more robust predictor of porosity than derived OC–clay metrics. While indices such as NonCOCC and COCC may have value in predicting soil physical behaviour as proposed by Dexter et al. (2008), their explanatory power is lower across our heterogeneous datasets. The OC–porosity relationship is also not uniform and shows variation with dataset and land-cover class as discussed in Section 4.3.

From an applied perspective, this reinforces the use of organic carbon as a primary indicator in large-scale assessments of soil physical condition and hydrological function, particularly where more detailed structural measurements are unavailable. It also suggests that strategies aimed at increasing or maintaining soil carbon are likely to have broadly consistent benefits for porosity and associated soil functions across diverse environments.

4.2. Complexed and non-complexed carbon and porosity

The OC–clay analysis by Dexter et al. (2008) suggests that soil physical behaviour should be controlled primarily by complexed organic carbon. While our results indicate contribution from OC–clay relationships, we found that COCC was consistently a weaker predictor of porosity than either OC or NonCOCC. The 1:10 ratio established by Dexter for their data, and also applied here, was not proposed as a universal relationship. Soils with similar clay contents can differ substantially in mineralogy, reactive surface area and the abundance of Fe and Al oxides, leading to large differences in carbon stabilisation capacity (Singh et al., 2017; Begill et al., 2023). Consequently, a given OC:clay balance may correspond to very different degrees of organo-mineral association among soils, reducing the ability of OC–clay balance metrics to consistently reflect complexation status or predict structural properties.

NonCOCC nevertheless consistently explains a small but statistically detectable additional component of variation beyond that represented by separate smooths for OC and clay, and its inclusion modestly improved model performance. This suggests that NonCOCC captured some additional information in our models on soil structural behaviour across broad environmental gradients. Because NonCOCC is calculated directly from OC and clay, statistical significance alone should not be interpreted as evidence of an independent process. Rather, the result indicates that the mathematical combination represented by NonCOCC describes variation not fully captured by separate smooths for OC and clay.

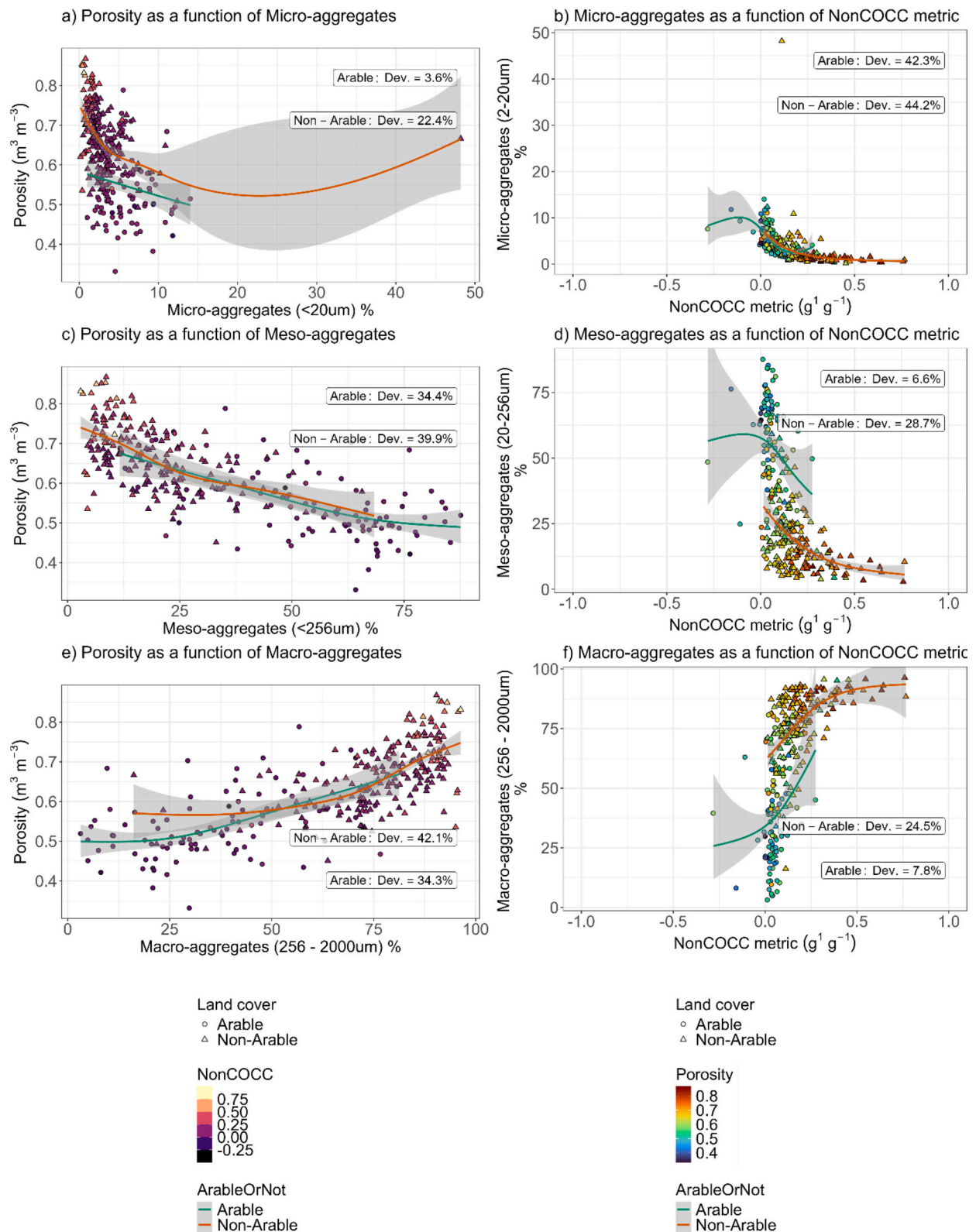


Fig. 5. Relationships to aggregate classes in 3 groups: micro-aggregates 2–20 μm ; meso-aggregates 20–256 μm , and macro-aggregates 256–2000 μm . Plots a,c,e show porosity against fraction of aggregates in that class, with colour indicating NonCOCC. Plots b,d,f show fraction of aggregates in that class against the NonCOCC metric with colour indicating porosity. In all plots, shape indicates which sites are arable and therefore likely subject to tillage. Data from CS 2019, $N = 293$. Note the differences in the axes for aggregate %, since these differ in order of magnitude with the aggregate size (i.e. micro-aggregates tend to make up a smaller cumulative contribution as they are individually smaller). In each plot, the trendlines are constructed separately for arable and non-arable sites, and deviance explained (Dev.) is reported for each trend.

Table 4

Comparison of model parameters for GAMs of the relationships for OC and NonCOCC to proportion macro-aggregates response variable.

Model	Purpose	Formula	Deviance Explained	RMSE	AIC	tweedie <i>p</i> value
Model 21	Compare individual parameter models	OC (%)	61.32	14.00	2478	1.042
Model 22		NonCOCC	43.7	17.27	2589	1.048
Model 23	Test added value of NonCOCC with OC	OC (%) + NonCOCC	64.79	13.44	2456	1.04
Model 24		Land-cover class + OC (%)	67.13	12.90	2437	1.038
Model 25	Test added value of land-cover class parametric term	Land-cover class + NonCOCC	53.98	15.29	2532	1.036
Model 26		Land-cover class + OC (%) + NonCOCC	68.61	12.59	2427	1.037
Model 27	Test added value of land-cover class smooth to allow relationships to vary between land-cover classes	Land-cover class + (OC (%) Smoothed by land-cover class)	70.11	12.12	2417	1.037
Model 28		Land-cover class + (NonCOCC Smoothed by land-cover class)	54.83	15.10	2535	1.035
Model 29		Land-cover class + (OC (%) Smoothed by land-cover class) + (NonCOCC Smoothed by land-cover class)	72.01	11.74	2404	1.014

Smooth terms for statistical models for fraction macro-aggregates					
Model	Smooth Term	edf	Ref.df	F	<i>p</i> -value
Model 23	OC (%)	4.83	9	22.98	<2e-16
Model 23	NonCOCC	2.3	9	2.98	1.30E-06
Model 29	OC (%) by Land-cover class	9.25	31	7.3	<2e-16
Model 29	NonCOCC by Land-cover class	2.72	31	0.53	2.79e-04

Model 29 tests relative contribution of parameters. The significance of each term is identified by *p*-value in the parameters for variables included in the models. Relative F values can be used to identify model-specific conditional support for smooth terms but are not standardised measures of effect size or independently explained variation. See section 5.4 for definitions of statistical terms.

Table 5

statistical model of porosity on aggregates data subset to test relative contribution of aggregate and NonCOCC metrics.

Statistical model of total porosity. Where total porosity is calculated accounting for proportion OC. Model was constructed on the subset of data with aggregate information.			
Formula	Deviance Explained	RMSE	
Land-cover class + OC (%) + NonCOCC + Fraction macro-aggregates	72.18	0.05	

Smooth terms parameters for model constructed on the subset of data with aggregate information.				
Smooth Term	edf	Ref.df	F	<i>p</i> -value
OC (%)	3.12	9	10.57	<2e-16
NonCOCC	1.74	9	1.55	0.0003
Fraction macro-aggregates	1.24	9	0.93	0.002

The relationship between NonCOCC and porosity is asymmetric across its range as might be expected based on mechanistic modelling (Robinson et al., 2025). At negative values of NonCOCC, representing relative clay excess, the relationship with porosity is weak and highly scattered. A clearer trend may have been expected based on proposed uses of OC:clay ratios as indicators. For example, the OC:clay ratio metric for soil structure defined by Johannes et al. (2017), splits soils with excess clay (OC:clay <1:10) into moderate (less excess clay) and degraded (more excess clay) implying an expected trend over this region of the data. However, Feeney et al. (2024) reported limited correspondence between porosity and OC:clay ratio classes derived from Johannes et al. (2017), except for soils in the highest ratio class. Similarly, our results indicate that soils with excess clay do not follow a simple, uniform porosity response at large spatial scales, and therefore that OC:clay ratios and metrics may be less useful at low OC:clay. The broad range of porosity values within the negative NonCOCC values corresponding to Johannes et al. (2017) lower ratio classes suggests that fixed thresholds are unlikely to provide a reliable indicator of porosity in clay-rich or OC-

poor soils. This suggests that any mechanisms associated with increasing NonCOCC under OC-excess conditions may not operate in the same way when the OC:clay ratio is below 1:10. In this part of the gradient, porosity may be more influenced by other factors, including clay mineralogy and geometry, total OC and clay contents, other mineral components, aggregation status, and other environmental controls (e.g. Lehmann et al., 2021; Hamza et al., 2025; Thomas et al., 2024). This reinforces the view that OC:clay ratio concepts should not be considered as direct proxies for porosity, and may require adjustment for geographical context (Sauzet et al., 2024).

The stronger relationship observed in the positive range of NonCOCC has steepest increase between approximately 0.06 and 0.24. This porosity response was more pronounced in higher-clay soils within the low positive NonCOCC range. This pattern is broadly consistent with Koop et al. (2023), who found that carbon-related controls on effective porosity were more evident in clay-rich soils. For a given NonCOCC value, soils with more clay also tend to have higher porosity, which may be largely due to greater total OC.

The partial effect of NonCOCC after accounting for OC and clay (Fig. S8) retained the same region of rapid porosity increase within the low positive range, indicating that the observed pattern is not explained solely by total OC concentration. The partial response also increased at strongly negative NonCOCC values once the independent effects of OC and clay were accounted for. This suggests that after controlling for OC content, deviation from the reference balance in either direction was associated with increased fitted porosity. However, the range of greatest response differed between clay-excess and OC-excess conditions. Given the high concavity among OC, clay and NonCOCC, these patterns should be interpreted as features of their joint mathematical relationship rather than evidence of a mechanism attributable to NonCOCC itself.

The range of steepest porosity increase with NonCOCC from 0.06 to 0.24 also corresponds to the most pronounced increase in macro-aggregates (Fig. 5f), suggesting that the NonCOCC metric captures a component of structural variation. This aligns with previous findings on the importance of OC:clay ratio for aggregate stability (Feller and Beare, 1997) and with aggregate hierarchy theory, under which relatively young and labile organic matter promotes macro-aggregate formation through transient organic binding agents (Six et al., 2000). This is

discussed in more detail in Section 4.4. Under this framework, increasing positive NonCOCC values indicating OC beyond that associated with the reference OC:clay balance, may create conditions that favour aggregate formation and structural development. However, causality and directionality cannot be resolved from the observational data used here. Because MAOM and POM fractions were not measured directly, this interpretation remains hypothetical and requires validation using independent SOM fractionation approaches. Additionally, when included in a model for this subset of soils, the proportion of macro-aggregates had lower conditional contribution than the NonCOCC term. This suggests that changes in aggregate-size distribution contribute to, but do not fully explain, the porosity response associated with increasing NonCOCC.

OC:clay metrics such as NonCOCC are intended to act as proxies for variation in the balance between mineral-associated and excess organic matter, which is rarely measured directly in large-scale monitoring programmes. However, the relationship between clay and OC-mineral association capacity is known to vary between soils due to differences in mineralogy and environmental conditions (e.g. Dexter et al., 2008; Sauzet et al., 2024). Such variation is likely to contribute to the scatter observed in our data and limits direct mechanistic interpretation. Consequently, the same NonCOCC value cannot be assumed to represent the same degree of organo-mineral association in all soils, and the metric is only a proxy. The value of NonCOCC lies in offering a simple descriptor of relative OC excess and clay excess which is readily calculated from available measurements. Unlike COCC, NonCOCC retains information on the direction of deviation from the reference balance and may help identify regions of the OC-clay gradient associated with different structural responses. More broadly, NonCOCC can act as a hypothesis-generating metric across large datasets, since it can easily be calculated from more common measurements. The present datasets cannot determine whether the observed relationships arise from differences in MAOM, POM, aggregate organisation, or other mechanisms. However, the broad reproducibility of the response across independent regional datasets suggests that the NonCOCC metric identifies a meaningful component of soil structural variation. This should be investigated using datasets containing direct measurements of OC fractions. However, the weaker and less consistent relationships observed within land-cover class sub models indicate that the metric may be less useful within habitat classes.

4.3. Land-cover class and regional context

Land-cover class and environmental setting showed relationships to porosity in our models, but importance differed among predictors. Improvements in model performance when land-cover and dataset-specific smooths were included indicates that porosity is shaped by a range of environmental and management factors in addition to OC and clay content. This is consistent with previous studies showing that porosity varies systematically among habitats attributed potentially to differences in vegetation, rooting patterns, biological activity, management intensity and climate (Thomas et al., 2024; Robinson et al., 2026).

Variability between datasets observed in Fig. 2 indicates modulation of the relationship between OC and porosity by soil type, climate, vegetation, and management. Land-cover- and dataset-specific smooths were significant, however their contribution was substantially smaller than the principal OC smooth. This suggests that the dominant positive association between OC and porosity is broadly conserved across contrasting soil systems. Similarly, Thomas et al. (2024) demonstrated that whilst OC was the dominant predictor of topsoil porosity across habitats, habitat also exerted an independent influence from differences in vegetation inputs, rooting activity and management intensity. Forest and semi-natural habitats exhibited greater porosity than arable systems (Thomas et al. (2024) attributed this difference partly to perennial vegetation, which supports biologically derived pore networks, and to the greater physical disturbance experienced by arable soils. Feeney

et al. (2024) showed that typical OC for a soil varies systematically among habitats and climates, with the highest values occurring in cool, wet forest and grassland systems and the lowest in Mediterranean croplands. These patterns were attributed to differences in organic matter inputs, decomposition rates, soil texture and land-use disturbance, which together influence both SOC accumulation and pore development. Their central finding was that soils with higher SOC relative to the typical SOC expected for their pedo-climatic setting had significantly greater porosity and carbon stocks. Consistent with these studies, the present analysis found that allowing OC-porosity relationships to vary among land-cover classes and datasets improved model performance. However, in our analysis the principal OC smooth consistently explained substantially more variation than the land-cover- or dataset-specific deviations, suggesting a broadly consistent positive relationship between OC and porosity across environments. This finding aligns with previous large-scale studies showing strong links between OC, bulk density and porosity which persist across diverse habitats and environmental settings (Thomas et al., 2024; Robinson et al., 2022, 2026). Overall, the persistence of the OC effect across independent UK, European and US datasets analysed here supports the interpretation that OC is a fundamental correlate of soil porosity, while the influence of dataset and land-cover class emphasises the importance of environmental context in driving OC levels and shaping the relationship to porosity.

Both OC-clay balance metrics showed some sensitivity to environmental context, with model performance improving when land-cover- and dataset-specific responses were included. This is unsurprising since the relationship between clay content and OC:mineral association varies between soils due to differences in mineralogy, reactive oxides, hydrology and climate (Begill et al., 2023; Sauzet et al., 2024). As discussed in Section 4.2, this means that similar OC:clay balances may represent different degrees of organo-mineral association in different soil systems, limiting the extent to which an OC-clay metric can be expected to perform consistently across broad environmental gradients.

However, NonCOCC showed a more consistent relationship with porosity across regions than across habitats. Whilst factor-smooth interactions improved the NonCOCC models, the magnitude of this improvement was smaller than for COCC. This may be because COCC collapses both clay-excess and OC-excess conditions into a single measure of proximity to the reference balance, whereas NonCOCC retains information on both the magnitude and direction of deviation. Across all regional datasets, NonCOCC was consistently retained as a secondary predictor after OC and had greater conditional contribution than clay. In contrast, within habitat-specific models NonCOCC contributed only in grasslands and was considerably weaker in woodland, wetland and shrubland systems. This suggests that NonCOCC may be most informative when describing broad compositional gradients across contrasting soil systems, rather than as a universal predictor within individual habitat types.

Notably, derivative analyses identified a broadly consistent region in which the fitted partial response of porosity to NonCOCC changed most rapidly, close to the transition between clay-excess and OC-excess conditions. Across the independent UK, European and US datasets, peak responses occurred between NonCOCC values of approximately 0.04 and 0.16, while the highest-response regions were generally concentrated between 0.06 and 0.24. Although the precise shape and magnitude of the response varied, the broadly similar occurrence of elevated sensitivity within the low positive NonCOCC range suggests that this region captures a recurring feature of the porosity response across diverse soil systems. However, the exact position and magnitude of the response varied among datasets, as might be expected given the variability in degrees of organo-mineral association for a given OC:clay balance as discussed above.

Climate-related structural processes may be important for variability in our observed relationships. Previous work has shown that macroporosity responds to climatic controls such as precipitation regime,

drying intensity and freeze-thaw frequency, with measurable effects occurring over decadal timescales (Hirmas et al., 2018). Similarly, freeze-thaw cycles can reorganise pore networks, alter aggregate distributions and modify hydraulic properties without necessarily changing total OC contents (Leuther and Schlüter, 2021). These processes provide potential mechanisms through which climate and management may influence porosity independently of OC and OC-clay balance.

The performance of factor smooths in our models suggests that environmental setting influences the magnitude and form of porosity responses to our predictor variables, but the high contribution from OC was consistent. For NonCOCC, the broadly consistent region of elevated sensitivity identified by the derivative analyses was reproduced across independent regional datasets despite substantial variation in climate, land use and soil properties. However, the reduced conditional contribution from NonCOCC within individual habitat classes suggests that part of its explanatory value may derive from systematic differences in OC-clay balance among habitats. Thus, the metric may be particularly useful as a descriptor of broad compositional gradients spanning multiple soil systems, while within individual habitats other controls on structure may become comparatively more important.

4.4. Carbon in excess, aggregate generation

Lehmann and Kleber (2015) proposed a soil continuum model describing the progressive transformation of organic matter in soils, from fresh plant and animal residues to increasingly processed forms that interact with mineral surfaces and contribute to soil structure. Within this framework, less processed organic inputs—often associated with particulate organic matter—are gradually decomposed and redistributed, while more assimilable forms become increasingly mineral-associated. Throughout this continuum, organic materials are involved in aggregate formation and destruction, adsorbing and desorbing on mineral surfaces.

The observed pattern in our data suggests that the increase in macro-aggregates ($>250\text{ }\mu\text{m}$) with increasing NonCOCC is most pronounced when OC is moderately in excess relative to clay (steep trend in the 0.06–0.24 range of NonCOCC, Figs. 4b and 5f). A similar pattern is observed of increasing macro-aggregates with OC which is steepest up to ~6% OC (Fig. S7e). These trends are evident in both arable and non-arable subsets, although arable soils show a lower proportion of macro-aggregates at comparable OC or NonCOCC values, likely reflecting the effects of tillage and disturbance on aggregate stability (Six et al., 2000). The models in Table 4 indicate that the relationship to OC is stronger, however NonCOCC has additional explanatory power in a combined model (Models 23 and 29) suggesting some added influence.

This interpretation aligns with established hierarchical models for soil aggregation (Edwards and Bremner, 1967; Tisdall and Oades, 1982; Oades and Waters, 1991; Sullivan et al., 2022). In these models, microaggregates (2–20 μm) are stabilised by mineral–organic interactions and inorganic binding agents, while larger aggregates (20–250 μm and $>256\text{ }\mu\text{m}$) are increasingly influenced by organic inputs such as roots, fungal hyphae, and partially decomposed residues. The association of higher OC and NonCOCC values with increased macro-aggregate abundance in our data is consistent with these models, suggesting that organic enrichment plays a key role in structural organisation.

Aggregation and changes in aggregate size distribution affect pore size distribution and connectivity, thereby influencing wider soil structural characteristics and associated functions such as water flow, gas exchange and microbial activity (Lado et al., 2004; Bronick and Lal, 2005). The observed relationships between OC, NonCOCC and macro-aggregate abundance therefore may have important implications for soil structure and function. However, although macro-aggregate abundance was a significant predictor of porosity when included in the model of porosity on this subset (Table 5), it had less conditional contribution in the model than both OC and NonCOCC. This suggests that aggregate

development contributes to, but does not fully explain, the porosity response associated with increasing NonCOCC. The lower explanatory power of aggregate abundance relative to NonCOCC suggests that additional mechanisms may also contribute to this relationship. One possibility is that carbon accumulating beyond the reference balance (1:10 OC:mineral) is increasingly stabilised through organo-organic associations within MAOM. Recent nanoscale imaging and molecular simulations suggest that MAOM commonly occurs as heterogeneous, patchy and multilayer organo-mineral assemblages, in which organo-organic interfaces may be more abundant than direct organo-mineral interfaces, while water retains access to mineral surfaces (Possinger et al., 2020; Underwood and Bourg, 2024). Such structures may alter the effective geometry of the mineral-organic matrix by increasing its irregularity, surface roughness and internal heterogeneity. This interpretation aligns with the analytical porosity framework of Robinson et al. (2022, 2025), which identifies particle shape and particle geometry as important controls on porosity emergence beyond composition alone. Under this view, the NonCOCC metric may supply information on variation in both the quantity of carbon present, and its physical organisation within the soil matrix once mineral loading expectations are exceeded. The observed association between NonCOCC and porosity may therefore be consistent with aggregate development, organo-organic associations, and changes in mineral-organic architecture that modify particle packing and pore formation. The datasets analysed here cannot confirm these potential mechanisms, which would require further investigation on more appropriate data.

Additionally, these relationships should not be interpreted as evidence of a simple causal pathway whereby increasing OC necessarily leads to greater aggregation. Aggregate formation can itself contribute to OC accumulation, because organic matter occluded within aggregates, particularly micro-aggregates formed within macro-aggregates, is physically protected from microbial decomposition and may therefore persist for longer periods in soil (Six et al., 2000). Consequently, feedbacks between OC accumulation and aggregate formation are likely to contribute to the observed associations.

It is important to note that many other variables, not directly considered in this analysis, are also known to influence aggregation. For example, increased aggregate size and stability can be associated with higher alkalinity and carbonate concentrations, or with hydrophobic substances that reduce rapid wetting of aggregates (Bronick and Lal, 2005). There is also variation associated with vegetation type due to the role of plant roots in aggregate development (Hamza et al., 2026) in addition to habitat variation relating to land management such as tillage (Six et al., 2000). Consistent with this, our analysis found that land use exerted a larger influence on aggregate-size distribution than NonCOCC, highlighting the importance of disturbance-related processes. Nevertheless, NonCOCC retained explanatory power after accounting for land-use effects, indicating that OC-clay balance captures variation in aggregate organisation that is not solely attributable to land-cover class. These patterns may be less detectable within individual land-cover classes. However, the small sample size of the aggregate dataset prevented reliable land-cover-specific analyses.

4.5. Limitations

This study is based on observational datasets and therefore identifies statistical associations rather than causal relationships. The analyses compare soils across environmental gradients and cannot determine whether equivalent structural responses would occur following temporal changes in OC within individual soils. In addition, the datasets differed in sampling design, depth and environmental context. Although dataset and land-cover class were accounted for statistically, we omit factors such as climate, mineralogy, management history and pore-size distribution which also influence porosity. Land-cover composition was strongly unbalanced among datasets, and land-cover information was unavailable for the US observations, for which a dataset-specific

category was used in place of an observed land-cover class. Land-cover composition was also unbalanced among the remaining datasets. Consequently, land-cover effects cannot be cleanly separated from dataset-specific differences and should be interpreted as broad contextual associations rather than independent management effects.

The OC-clay balance metrics are empirical descriptors derived from a literature-based reference ratio that is not expected to represent a universal stabilisation threshold across all mineralogical and environmental contexts. In addition, direct measurements of MAOM, POM, pore-size distribution, pore connectivity and aggregate stability were not available across datasets, limiting mechanistic interpretation. Consequently, the mechanisms linking NonCOCC to porosity cannot be confirmed from these datasets. Analysis using MAOM, POM and independent structural measurements would be required to determine whether the observed relationships reflect variation in organic matter beyond the mineral-associated pool.

Multivariable GAMs exhibited high concavity, particularly models containing OC, clay and derived OC-clay balance metrics simultaneously (e.g. Supplementary Table S4). This is expected because NonCOCC and COCC are mathematically derived from OC and clay content. High concavity does not substantially affect the predictive performance of the models, but reduces confidence in the unique attribution of variation to individual smooth terms. Consequently, derivative analyses of the NonCOCC smooth should be interpreted as conditional descriptions of the fitted response surface after accounting for OC and clay, rather than evidence of an independent effect of NonCOCC itself. The broad consistency of derivative patterns across regional datasets provides some support for the robustness of the observed response, but its precise shape and magnitude should be interpreted cautiously.

5. Conclusions

This study used a large compilation of soil datasets from Europe, Great Britain, and the United States to examine controls on soil porosity across a wide range of environmental conditions. The analysis demonstrates that organic carbon (OC) is the strongest predictor of porosity at these scales, explaining more variation than OC–clay balance metrics alone. This supports previous findings for GB and EU.

The derived NonCOCC metric provides additional insight by capturing variation in the balance between OC and clay content. While its contribution is modest and much smaller than that of total OC, NonCOCC consistently explains a small additional component of variation in porosity across independent regional datasets after accounting for both OC and clay. The steepest portions of the fitted NonCOCC relationships generally occurred at low positive values, where relatively small increases in OC relative to clay are associated with greater increase in porosity, although their locations and uncertainty varied among datasets. This region also corresponds to shifts towards larger aggregate-size fractions. Together, these patterns suggest that the NonCOCC gradient is associated with variation in soil structural organisation, particularly in soils close to the transition between relative clay excess and OC excess. However, fraction macro-aggregates had low conditional contribution in the combined model of porosity suggesting that aggregation does not fully explain the porosity response associated with increasing NonCOCC. The underlying mechanisms cannot be resolved directly from the available data and require validation using independent measurements of soil organic matter fractions and organo-mineral associations.

In contrast, soils with relative clay excess (negative NonCOCC values) show weak and variable relationships between OC–clay balance and porosity, reflecting the influence of additional controls such as OC, mineralogy, soil type, and environmental conditions. This finding highlights the limitations of using simple OC:clay ratio thresholds as direct indicators of porosity across heterogeneous datasets.

Overall, the results indicate that organic carbon is the strongest predictor of porosity across large environmental gradients, while OC-

clay balance metrics capture a smaller component of variation associated with soil structural organisation that is most consistently expressed across regional datasets. The NonCOCC metric provides a useful complementary descriptor to total OC by identifying regions of the compositional gradient where structural changes are most pronounced. However, the relationship was substantially more reproducible across independent regional datasets than within individual habitat classes. This suggests that NonCOCC captures a broad-scale component of variation associated with differences among soil systems and environmental settings, rather than representing a dominant or universal control on porosity within individual habitats.

These findings support the use of large-scale observational datasets to characterise soil structural variation and provide a framework for integrating commonly measured soil properties into assessments of soil physical condition across environmental gradients. In practical terms, the results support the prioritisation of OC measurements in soil monitoring programmes, while highlighting the potential of simple balance metrics such as NonCOCC to identify soils where changes in OC relative to clay may be associated with larger porosity responses. This provides a pathway for integrating observational soil data into improved predictions of hydrological and biogeochemical function at large spatial scales.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used co-pilot in order to improve the consistency of the language and clarity of the text. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRediT authorship contribution statement

Amy Thomas: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Attila Nemes:** Writing – review & editing, Conceptualization. **David Robinson:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Funding acquisition, Conceptualization.

Funding

The research was funded by, The Research Council of Norway, Climasoil, Project number: 325253. Natural Environment Research Council award to UKCEH for the National Capability for UK Challenges Programme NE/Y006208/1. The Glastir Monitoring and Evaluation Programme (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 program 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] as part of AI4SoilHealth.

Declaration of competing interest

There are no known conflicts of interest.

Acknowledgements

General: We thank all the Glastir Monitoring and Evaluation Programme (GMEP) team who contributed to collecting all the GMEP data in Wales and the UK Centre for Ecology and Hydrology UKCEH Countryside Survey team from 2007 who collected all the data from England,

Wales and Scotland. 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.catena.2026.110657>.

Data availability

The UK data sets 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.; Kobizek, 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. doi:<https://doi.org/10.5285/821325f3-b353-4a51-8db2-6b2200d82aca>

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). doi:<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). doi:<https://doi.org/10.5285/0fa51dc6-1537-4ad6-9d06-e476c137ed09>

The EU soil dataset used in this study are also publicly available through the Joint Research Centre European Soil Data Centre (ESDAC): Fernandez-Ugalde, O.; Scarpa, S.; Orgiazzi, A.; Panagos, P.; Van Liedekerke, M.; Marechal, A. & Jones, A., EUR 31144 EN, Publications Office of the European Union, Luxembourg. 2022, ISBN 978–92–76–54,832–4, doi:<https://doi.org/10.2760/215013>, JRC129926: <https://esdac.jrc.ec.europa.eu/content/lucas-2018-topsoil-data>

The US soil dataset used in this study are also publicly available through the USDA-NRCS NCSS data repository: <http://ncsslabdatamart.sc.egov.usda.gov/>.

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.

The graphical abstract uses boundary data from: www.geoBoundaries.org. Runfola, D. et al. (2020) geoBoundaries: A global database of political administrative boundaries. PLoS ONE 15(4): e0231866. doi:<https://doi.org/10.1371/journal.pone.0231866>

References

- Begill, N., Don, A., Poeplau, C., 2023. No detectable upper limit of mineral-associated organic carbon in temperate agricultural soils. *Glob. Chang. Biol.* 29, 4662–4669. <https://doi.org/10.1111/gcb.16804>.
- 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., Kobizek, 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>.
- Bronick, C.J., Lal, R., 2005. Soil structure and management: a review. *Geoderma* 124, 3–22. <https://doi.org/10.1016/j.geoderma.2004.03.005>.
- Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)?—arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* 7 (3), 1247–1250. Jun 30.
- Cotrufu, M.F., Lavallee, J.M., Six, J., Lugato, E., 2023. The robust concept of mineral-associated organic matter saturation: A letter to Begill et al., 2023. *Glob. Chang. Biol.* 29 (21).
- Dexter, A.R., Richard, G., Arruays, D., Czyż, E.A., Jolivet, C., Duval, O., 2008. Complexed organic matter controls soil physical properties. *Geoderma* 144, 620–627. <https://doi.org/10.1016/j.geoderma.2008.01.022>.
- Edwards, A.P., Bremner, J.M., 1967 Mar. Microaggregates in soils 1. *J. Soil Sci.* 18 (1), 64–73.
- Feeney, C.J., Bentley, L., De Rosa, D., Panagos, P., Emmett, B.A., Thomas, A., Robinson, D.A., 2024. Benchmarking soil organic carbon (SOC) concentration provides more robust soil health assessment than the SOC/clay ratio at European scale. *Sci. Total Environ.* 951, 175642. <https://doi.org/10.1016/j.scitotenv.2024.175642>.
- Feller, C., Beare, M.H., 1997. Physical control of soil organic matter dynamics in the tropics. *Geoderma* 79, 69–116.
- Golchin, A., Oades, J., Skjemstad, J., Clarke, P., 1994. Soil structure and carbon cycling. *Soil Res.* 32, 1043–1068. <https://doi.org/10.1071/SR9941043>.
- Hamza, A., Karcauskienė, D., Mockevičienė, I., Repšienė, R., Tahir, M.A., Manzoor, M.Z., Kousar, S., Lodhi, S.S., Rasool, N., Ullah, I., 2025. Soil aggregate dynamics and stability: natural and anthropogenic drivers. *Agriculture* 15. <https://doi.org/10.3390/agriculture15232500>.
- Hamza, A., Karcauskienė, D., Tahir, M.A., Mockevičienė, I., Noor Us, S., Repšienė, R., Skuodienė, R., 2026. Root-soil interactions a key driver of aggregate formation and stability: a trait-based and mechanistic review. *Plant Soil* 521, 1037–1066. <https://doi.org/10.1007/s11104-026-08453-3>.
- Hirmas, D.R., Gimenez, 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, 100–103. <https://doi.org/10.1038/s41586-018-0463-x>.
- Johannes, A., Matter, A., Schulz, R., Weisskopf, P., Baveye, P.C., Boivin, P., 2017. Optimal organic carbon values for soil structure quality of arable soils. Does clay content matter? *Geoderma* 302, 14–21. <https://doi.org/10.1016/j.geoderma.2017.04.021>.
- Koop, A.N., Hirmas, D.R., Billings, S.A., Li, L., Cueva, A., Zhang, X., Wen, H., Nemes, A., Souza, L.F.T., Ajami, H., Flores, A.N., Rudick, A.K., Guthrie, A., Klammer, L.M., Unruh, M., Sullivan, P.L., 2023. Is macroporosity controlled by complexed clay and soil organic carbon? *Geoderma* 437. <https://doi.org/10.1016/j.geoderma.2023.116565>.
- Lado, M., Paz, A., Ben-Hur, M., 2004. Organic matter and aggregate-size interactions in saturated hydraulic conductivity. *Soil Sci. Soc. Am. J.* 68, 234–242. <https://doi.org/10.2136/sssaj2004.2340>.
- Lehmann, J., Kleber, M., 2015. The contentious nature of soil organic matter. *Nature* 528, 60–68. <https://doi.org/10.1038/nature16069>.
- Lehmann, P., Leshchinsky, B., Gupta, S., Mirus, B.B., Bickel, S., Lu, N., Or, D., 2021. Clays are not created equal: how clay mineral type affects soil parameterization. *Geophys. Res. Lett.* 48. <https://doi.org/10.1029/2021gl095311>.
- Leuther, F., Schlüter, S., 2021 Jun 11. Impact of freeze–thaw cycles on soil structure and soil hydraulic properties. *Soil* 7 (1), 179–191.
- Oades, J.M., Waters, A.G., 1991 Dec 1. Aggregate hierarchy in soils. *Austr. J. Soil Res.* 29 (6), 815–828.
- Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., Fernández-Ugalde, O., 2018. LUCAS soil, the largest expandable soil dataset for Europe: a review. *Eur. J. Soil Sci.* 69, 140–153. <https://doi.org/10.1111/ejss.12499>.
- Panagos, P., De Rosa, D., Liakos, L., Labouyrie, M., Borrelli, P., Ballabio, C., 2024. Soil bulk density assessment in Europe. *Agric. Ecosyst. Environ.* 364. <https://doi.org/10.1016/j.agee.2024.108907>.
- Poeplau, C., 2023. Response to: “the robust concept of mineral-associated organic matter saturation: a letter to Begill et al.(2023)”. *Glob. Chang. Biol.* 29 (21), e4–e6.
- Poeplau, C., Don, A., 2023. A simple soil organic carbon level metric beyond the organic carbon-to-clay ratio. *Soil Use Manag.* 39, 1057–1067. <https://doi.org/10.1111/sum.12921>.
- Possinger, A.R., Zachman, M.J., Enders, A., Levin, B.D., Muller, D.A., Kourkoutis, L.F., Lehmann, J., 2020 Nov 30. Organo–organic and organo–mineral interfaces in soil at the nanometer scale. *Nat. Commun.* 11 (1), 6103.
- Prout, J.M., Shepherd, K.D., McGrath, S.P., Kirk, G.J., Haeefe, S.M., 2021. What is a good level of soil organic matter? An index based on organic carbon to clay ratio. *Eur. J. Soil Sci.* 72 (6), 2493–2503.
- Rabot, E., Wiesmeier, M., Schlüter, S., Vogel, H.J., 2018. Soil structure as an indicator of soil functions: A review. *Geoderma* 314, 122–137. <https://doi.org/10.1016/j.geoderma.2017.11.009>.
- Robinson, D.A., Astbury, S., Burden, G., Carter, A., Emmett, H., Garbutt, B.A., Giampieri, A., Hall, C., Henrys, J., Hughes, P.A., Hunt, S., Jarvis, A., Jones, S.G., Keenan, D.L., Lebron, P., Nunez, I., Owen, D., Patel, A., Pereira, M., Seaton, M.G., Sharps, K., Thompson, N., Williams, B., 2019. Topsoil Physico-Chemical Properties from the Glastir Monitoring and Evaluation Programme, Wales 2013–2016. <https://doi.org/10.5285/0fa51dc6-1537-4ad6-9d06-e476c137ed09>.

- 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., 2022. Analytical modelling of soil porosity and bulk density across the soil organic matter and land-use continuum. *Sci. Rep.* 12, 7085. <https://doi.org/10.1038/s41598-022-11099-7>.
- Robinson, D.A., Friedman, S.P., Thomas, A., Hirmas, D., Sullivan, P.L., Nemes, A., 2025. Soil bulk density and porosity connecting macro- and micro-scales through geometry. *Earth Sci. Rev.* 268. <https://doi.org/10.1016/j.earscirev.2025.105173>.
- Robinson, D.A., Fendrich, A., Thomas, A., Reinsch, S., Leifeld, J., Moore, T., Seward, J., Borrelli, P., Ballabio, C., Panagos, P., 2026. Soil porosity prediction across Europe with a focus on soil particle density determination. *Int. Soil Water Conserv. Res.* 14. <https://doi.org/10.1016/j.iswcr.2026.100614>.
- Sauzet, O., Johannes, A., Deluz, C., Dupla, X., Matter, A., Baveye, P.C., Boivin, P., 2024. The organic carbon-to-clay ratio as an indicator of soil structure vulnerability, a metric focused on the condition of soil structure. *Soil Use Manag.* 40. <https://doi.org/10.1111/sum.13060>.
- Christensen, B.T., Olesen, J.E., 2012 Feb 1. Clay dispersibility and soil friability—testing the soil clay-to-carbon saturation concept. *Vadose Zone Journal*. 11 (1), vzj2011-0067.
- Schweizer, S.A., Mueller, C.W., Höschen, C., Ivanov, P., Kögel-Knabner, I., 2021 Dec. The role of clay content and mineral surface area for soil organic carbon storage in an arable toposequence. *Biogeochemistry* 156 (3), 401–420.
- Simpson, G.L., 2024. Gratia: an R package for Exploring Generalized Additive Models (arXiv preprint arXiv:2406.19082).
- Singh, M., Sarkar, B., Biswas, B., Bolan, N.S., Churchman, G.J., 2017 Jun 1. Relationship between soil clay mineralogy and carbon protection capacity as influenced by temperature and moisture. *Soil Biol. Biochem.* 109, 95–106.
- 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, 681–689. <https://doi.org/10.2136/sssaj2000.642681x>.
- Six, J., Callewaert, P., Lenders, S., De Gryze, S., Morris, S.J., Gregorich, E.G., Paul, E.A., Paustian, K., 2002. Measuring and understanding carbon storage in afforested soils by physical fractionation. *Soil Sci. Soc. Am. J.* 66, 1981–1987. <https://doi.org/10.2136/sssaj2002.1981>.
- Song, X., Wu, H., Li, S., He, P., Wu, X., 2025 Mar 1. The need to update and refine concepts relating to mineral-associated organic matter saturation in soil. *Soil Biol. Biochem.* 202, 109672.
- Sullivan, P.L., Billings, S.A., Hirmas, D., Li, L., Zhang, X., Ziegler, S., Murenbeeld, K., Ajami, H., Guthrie, A., Singha, K., Gimenez, 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.
- Team, R.C., 2025. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Thomas, A., Seaton, F., Dhiedt, E., Cosby, B.J., Feeney, C., Lebron, I., Maskell, L., Wood, C., Reinsch, S., Emmett, B.A., Robinson, D.A., 2024. Topsoil porosity prediction across habitats at large scales using environmental variables. *Sci. Total Environ.* 922, 171158. <https://doi.org/10.1016/j.scitotenv.2024.171158>.
- Tisdall, J.M., Oades, J.M., 1982. Organic matter and water-stable aggregates in soils. *J. Soil Sci.* 33, 141–163. <https://doi.org/10.1111/j.1365-2389.1982.tb01755.x>.
- Underwood, T.R., Bourg, I.C., Rosso, K.M., 2024 Nov 12. Mineral-associated organic matter is heterogeneous and structured by hydrophobic, charged, and polar interactions. *Proc. Nat. Acad. Sci.* 121 (46), e2413216121.
- Wickham, H., 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer New York.
- Wood, S.N., 2003. Thin-plate regression splines. *J. Royal Stat. Soc. (B)* 65 (1), 95–114. <https://doi.org/10.1111/1467-9868.00374>.
- 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, 3–36. <https://doi.org/10.1111/j.1467-9868.2010.00749.x>.
- Wood, S., 2017. *Generalized Additive Models: An Introduction with R*. Chapman and Hall CRC.
- Wood, S.N., 2025. mgcv Package Manual. Available at <https://r-universe.dev/manuals/mgcv.html> (accessed 17 September 2026).