https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

Soil science-informed neural networks for soil organic carbon density modelling under scarce bulk density data

Xuemeng Tian[1,2], Bernhard Ahrens[3], Leo Rossdeutscher[3], Lazaro Alonso[3], and Leandro Parente[1]

[1]OpenGeoHub, Doorwerth, The Netherlands
[2]Wageningen University and Research, Wageningen, The Netherlands
[3]Max Planck Institute for Biogeochemistry, Jena, Germany

Correspondence: Bernhard Ahrens (bahrens@bgc-jena.mpg.de)

Abstract. Soil organic carbon (SOC) density is a key variable for quantifying soil carbon stocks, yet its modelling is challenged by sparse and inconsistent measurements of bulk density and coarse fragments relative to SOC content. Conventional digital soil mapping approaches typically model SOC density as a single target variable, thereby underutilising abundant SOC content data and overlooking physical relationships among soil properties. This study evaluates a soil science-informed neural network for SOC density prediction that explicitly constrains the SOC–BD relationship, and compares it with univariate and multivariate neural network architectures. Across sparsely sampled target variables, including SOC density, bulk density, and coarse fragments, the soil science-informed model achieves comparable or slightly improved prediction accuracy relative to multivariate and univariate models. Although it yields lower accuracy for SOC content, the soil science-informed model better preserves physically plausible SOC–BD joint distributions and generates smoother, more temporally stable SOC density trajectories. Overall, the results demonstrate that incorporating soil physical constraints into machine learning models adds value beyond univariate accuracy, improving robustness, plausibility, and temporal coherence of SOC density predictions under sparse data conditions. Moreover, the latent parameters inferred by the soil science-informed model improve model interpretability and offer additional soil science relevant insights beyond predictive accuracy.

1 Introduction

Soil organic carbon (SOC) is fundamental for soil health and is increasingly recognised for its critical role in climate-change mitigation (Lal, 2004; Lehmann et al., 2020). For practical applications, stakeholders are primarily interested in SOC density ($kg\,m^{-3}$) rather than SOC content ($g\,kg^{-1}$), as SOC density directly determines SOC stocks. Digital soil mapping (DSM) provides a means to generate spatially explicit information on soil properties, including SOC. The increasing availability of SOC measurements and Earth Observation (EO) data has stimulated a rapid growth of Machine Learning (ML) applications in DSM, resulting in numerous SOC map products at global, continental and national scales (Wadoux et al., 2020).

In DSM, SOC density is typically modelled in a univariate way: SOC density is calculated from observed SOC content, bulk density (BD, $g\,cm^{-3}$), and coarse fragments (CF, unitless fraction between 0 and 1) (Hengl and MacMillan, 2019), and then learned as a single target variable. However, SOC content is usually measured far more frequently than BD and CF, and even when BD and CF are available, inconsistencies in measurement protocols can make them incompatible for direct SOC density

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

calculation (e.g., BD measured for fine earth only while CF reported as mass fraction; Poeplau et al. 2017). For example, in the LUCAS Topsoil survey, SOC content is available for 2009/2012, 2015 and 2018, with only SOC content available through all three surveys (Orgiazzi et al., 2018). This mismatch leads to a "waste" of SOC content data and limits our ability to model SOC density dynamics.

Focusing solely on SOC density may also obscure the intrinsic physical relationships among SOC content, bulk density, and coarse fragments. Different combinations of SOC content, BD and CF can yield similar SOC density values, yet correspond to distinct environmental and soil conditions represented by different regions of the feature space. Ignoring these relationships effectively discards valuable information and may limit the model's ability to learn physically meaningful patterns, ultimately undermining predictive performance and robustness.

Multivariate modelling has long been used in DSM to address such interdependencies, ranging from co-kriging (Heuvelink et al., 2016) and structural equation modelling (Angelini et al., 2017) to more recent machine-learning approaches. Multivariate ML has been applied to jointly model SOC, total nitrogen and C:N ratio (Van Der Westhuizen et al., 2023), soil properties at multiple depths (Wadoux, 2019; Taghizadeh-Mehrjardi et al., 2020), multiple soil attributes (Padarian et al., 2019; Ng et al., 2019), and the three texture components (Taghizadeh-Mehrjardi et al., 2020; Ließ and Sakhaee, 2024). However, to our knowledge, it has not yet been applied to jointly model SOC concentration, bulk density, coarse fragments and SOC density within a single modelling framework.

Beyond multivariate modelling, another way to account for relationships among soil variables is to explicitly incorporate knowledge about relationships among soil properties into machine-learning models, which has been referred to as soil science–informed ML (Minasny et al., 2024). Embedding domain knowledge can constrain model behaviour and prevent predictions that violate known soil processes. In univariate SOC dynamics modelling, purely data-driven ML approaches have been shown to produce noisy SOC time series, sometimes exhibiting implausibly large changes over short time spans (Tian et al., 2025b). Hybrid approaches that combine soil process understanding with ML have been reported to yield more realistic predictions than those derived from purely data-driven models (Zhang et al., 2023, 2024). Although the mechanistic relationship between SOC concentration and bulk density has been well documented in soil science (Stewart et al., 1970; Adams, 1973; Federer et al., 1993; Robinson et al., 2022), its integration into DSM frameworks remains limited.

The aim of this study is to compare multivariate and soil relation–informed modelling strategies for predicting SOC density under sparse data conditions while explicitly accounting for inter-variable relationships. Among ML approaches, neural networks (NNs), compared to more conventional models such as Random Forests, Cubist, and partial least squares regression, are particularly well suited for multivariate soil property mapping. This is due to ease-of-use in representing complex dependencies and their flexibility to accommodate data sparsity, where not all observations contain measurements for all target variables (Ng et al., 2019; Padarian et al., 2019). However, explicitly incorporating expert soil knowledge into ML-based DSM remains challenging (Ma et al., 2019; Wadoux, 2019), as ML models are inherently data-driven and the soil processes they represent are implicitly inferred from the training data (Wadoux et al., 2020). The `EasyHybrid.jl` package developed by Alonso et al. (2025) provides a practical framework for embedding mechanistic constraints directly into NN architectures. Therefore, in this study, all three modelling strategies are implemented using neural networks.

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

Using the LUCAS topsoil dataset, three model structures are compared: (i) a univariate neural network (UniNN), (ii) a multivariate neural network (MultiNN), and (iii) a soil relation–informed neural network (SiNN). The models are evaluated with respect to their ability to predict SOC density, as well as the soil properties from which SOC density is derived, namely SOC content, BD and CF. Using the subset of LUCAS 2018 with complete measurements for all four properties, we further examine whether MultiNN and SiNN can exploit the abundant SOC concentration data from 2009/2012, 2015, and the incomplete part of 2018 to reconstruct SOC density under sparse BD availability. Inter-variable relationships are also examined. As CF are excluded from SOC content and fine-earth BD determinations in this study, the assessment focuses primarily on the SOC–BD relationship. In addition, we assess whether the hybrid SiNN model produces more temporally stable and physically realistic SOC density trajectories. Finally, plausibility checks are performed on the latent representations of bulk density components inferred by the SiNN to evaluate the soil-science consistency of the model outputs.

2 Materials and Methods

2.1 Soil data

This analysis uses LUCAS soil data from the 2009/2012, 2015, and 2018 campaigns. Besides SOC content $g\,kg^{-1}$, the key properties are the volumetric form of CF (fraction, unitless between 0-1) and the fine-earth BD ($g\,cm^{-3}$), excluding coarse fragments larger than 2 mm, taken from the curated dataset of Pacini et al. (2023). When all three variables are available, SOC density ($kg\,m^{-3}$) is calculated as:

$$\mathrm{SOC}_{\mathrm{density}} = \mathrm{SOC}_{\mathrm{content}} \cdot \mathrm{BD}_{\mathrm{fe}} \cdot (1 - \mathrm{CF}_{\mathrm{vol}}) \tag{1}$$

This study focuses on topsoil (0–20 cm); thus, LUCAS 2018 measurements at 0–10 and 20–30 cm were excluded. For sites with three repeated measurements, we applied a quality filter based on temporal consistency of SOC content: assuming SOC changes at less than 0.5 $g\,kg^{-1}\,yr^{-1}$, we used a conservative threshold of 50 $g\,kg^{-1}\,yr^{-1}$ for the maximum absolute difference across measurements (Poeplau et al., 2011; Gubler et al., 2019). After removing time series exceeding this threshold and discarding records lacking required covariates, 56,117 measurements remained. All of them include SOC content, among them 5,194 also contain CF and BD. Regarding temporal replication, 11,690 sites have three measurements, 7,309 have two, and 6,429 have one.

2.2 Covariates

A wide range of covariates (also referred to as *predictors* or *features*) was included to represent environmental factors influencing SOC dynamics. Some covariates were available as time series (e.g., precipitation), providing multiple values for the same location across time. Others were static or long-term summaries (e.g., topography), offering a single value per location. All covariates were rescaled to the [0,1] range to ensure their suitability in the neural network, and variables lacking informative values within Europe were excluded from the analysis. This results in a covariate set with 362 individual covariates, from 15

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

Table 1. Summary statistics of soil properties: SOC content ($g\,kg^{-1}$), BD ($g\,cm^{-3}$), CF (fraction) and calculated SOC density ($kg\,m^{-3}$) per year.

Year	N sites	SOC content	BD	CF	SOC density
2009	17817	38.266 ± 72.672	-	-	-
2012	1777	19.517 ± 9.528	-	-	-
2015	19780	32.440 ± 53.596	-	-	-
2018	16743	34.836 ± 57.591	1.032 ± 0.302	0.061 ± 0.055	23.411 ± 17.764

covariate groups (Table 2). The full list of covariates is not included in the main text but is available in the supplementary materials at our Supplemetary Repository.

Table 2. Covariate groups used in this study. Note that each entry represents a group of covariate layers, which may include multiple data layers with different temporal and spatial resolutions.

Covariate group	Source	Temporal resolution	Spatial resolution
Digital terrain model and derived land surface parameters	Ho et al. (2025)	Static	30, 60, 120, 240, 480, and 960 m
Lithology type probability	Isik et al. (2024)	Static	250 m
Plant functional type	Harper et al. (2023)	Annual	300 m
Vegetation cover fraction	Sun et al. (2023)	Annual	500 m
Precipitation	Karger et al. (2021)	Annual	1 km
Land surface temperature	Wan (2006)	Annual	1 km
Water vapor	Lyapustin and Wang (2018)	Annual	1 km
Bioclimate	Karger et al. (2017)	Long-term	1 km
Landsat spectral index	Tian et al. (2025a)	Annual, long-term	30 m
Bare surface reflectance	Rogge et al. (2018)	Long-term	30 m
Sentinel backscatter	Wagner et al. (2021)	Long-term	25 m
PALSAR backscatter	Shimada and Ohtaki (2010)	Static	25 m
Peatland indicator	Widyastuti et al. (2024)	Long-term	1 km
Cropland indicator	Potapov et al. (2022)	Annual	30 m
Soil moisture	Bauer-Marschallinger et al. (2018)	Annual	1 km

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

2.3 Model architectures

Three NN architectures are evaluated and compared in this study: UniNN, MultiNN, and SiNN (Fig. 1). To improve training stability and predictive performance of the NN, all four target variables—BD, SOC content, CF, and SOC density—were transformed to reduce skewness and to constrain their values to the $[0, 1]$ range. This was achieved through log transformation and scaling using a standard scaler.

The UniNN model serves as a baseline that reflects the conventional approach in DSM, where each soil property is modelled independently using a separate NN. The MultiNN architecture advances this idea by training a single NN to simultaneously learn and predict four targets. All targets are treated as purely data-driven outputs in both UniNN and MultiNN. The SiNN architecture incorporates physical constraints directly into the learning process. In SiNN, SOC content and CF are predicted by the NN, while BD and SOC density are computed within the model using mechanistic equations. Here, BD is expressed as a function of SOC content and two latent BD parameters—organic bulk density (oBD) and mineral bulk density (mBD), following the well-established relationship of Federer et al. (1993):

$$BD = \frac{1}{\frac{SOM}{oBD} + \frac{1-SOM}{mBD}} \tag{2}$$

where the soil organic matter (SOM) could be converted from SOC content with a factor 1.724 ($SOM = 1.724 \cdot SOC_{\text{content}}$). To constrain the latent parameters oBD and mBD and to facilitate model convergence, prior ranges and recommended initial values are specified. Following Robinson et al. (2022), oBD is constrained to the range 0.05–0.40 $g\,cm^{-3}$, with an initial value of 0.20 $g\,cm^{-3}$. Mineral bulk density (mBD) is constrained to the range 0.75–2.00 $g\,cm^{-3}$, with an initial value of 1.20 $g\,cm^{-3}$. SOC density is subsequently derived from SOC content, BD, and CF using the mechanistic formulation in Eq. 2.

Across all three architectures, the supervised training targets include the four soil variables (SOC content, BD, CF, SOC density), and the loss function jointly constrains the predictions for all variables. In the SiNN model, the latent parameters oBD and mBD are also learned during training and can be output to provide additional interpretability regarding soil physical properties. The model experiments are implemented using the EasyHybrid.jl package (Alonso et al., 2025).

2.4 Model evaluation

All models were trained and evaluated using a five-fold cross-validation scheme to ensure objective assessment while maximising the use of available data. In each iteration, three folds are used for hyperparameter optimisation and model training, one fold serves as the validation set for selecting the best-performing model, and the remaining fold is held out as the test set for final evaluation. The folds are constructed through random partitioning of the dataset. By iterating across all five folds, full-coverage predictions for all samples are obtained for subsequent analysis.

Model evaluation is performed at two levels. First, the predictive accuracy of each individual target variable is assessed using all available observations for that variable. Second, joint evaluation is conducted on the complete subset of samples where all four target variables are present. Both global accuracy and stratified accuracy across land-cover classes are examined. The

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

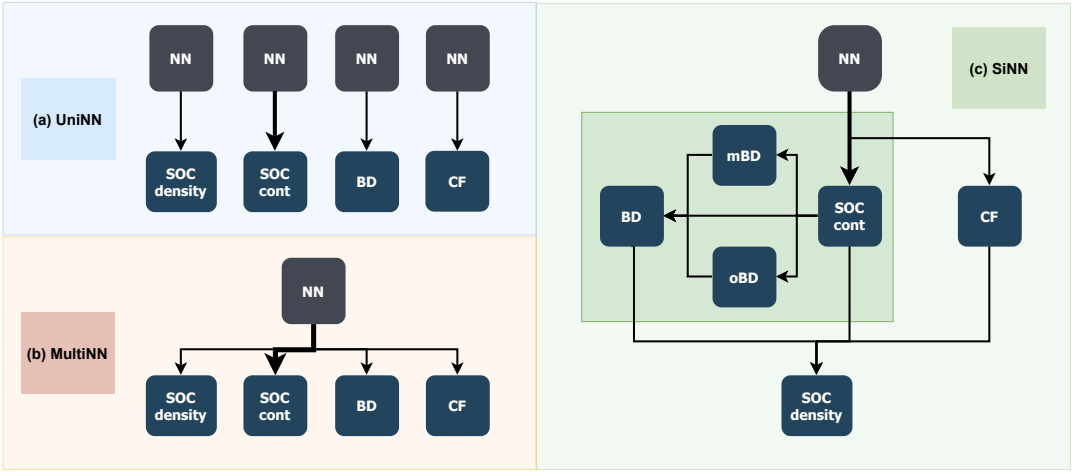


Figure 1. Schematic illustration of the three NN architectures evaluated in this study: (a) UniNN, in which a separate NN is trained for each soil variable; (b) MultiNN, in which a single NN is trained to predict all soil variables simultaneously; and (c) SiNN, in which SOC–BD relationships are explicitly embedded in the model structure. Dark grey boxes denote NN models, while dark blue boxes denote target variables and intermediate variables that are output by the model.

metrics considered include the coefficient of determination (R^2), mean squared error (MSE) and bias, defined as the mean difference between predictions and observations.

Additionally, to assess the plausibility and consistency of BD–SOC content relationships, the joint distributions of predicted SOC content and BD were compared across the three neural network architectures and against observed data. Finally, the latent parameters oBD and mBD predicted by the SiNN model are examined with respect to soil texture, land cover, SOC content, and other soil properties to assess whether their spatial patterns and magnitudes are physically reasonable.

2.5 Temporal performance

The three models are evaluated in terms of temporal performance from two complementary perspectives: temporal plausibility and temporal transferability. Temporal plausibility refers to the coherence and stability of predicted SOC density time series, and is assessed using SOC density predictions obtained from cross-validation across multiple survey years. The underlying expectation is that the SiNN produces more coherent and less noisy SOC density trajectories than purely data-driven models.

Temporal transferability, in contrast, evaluates how well a model trained on one time period generalises to other time periods. This is assessed by training the models on data from 2018 and predicting SOC density for the remaining survey years. However, as SOC density measurements in the current dataset are available for only a single survey year, direct evaluation of temporal transferability for SOC density is not feasible. Therefore, an indirect assessment is conducted using SOC content, which is consistently measured across multiple survey years, as a proxy to evaluate temporal transferability.

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

3 Results

3.1 Model accuracy

Fig 2 shows the cross-validation performance of SOC density predictions for the three models. Overall, SiNN exhibits the highest predictive performance, followed by MultiNN and UniNN, as reflected by R^2, MSE, and bias. Among the three models, SiNN produces predictions that most closely align with the 1:1 line relative to the observations. However, the performance differences among the models are relatively small.

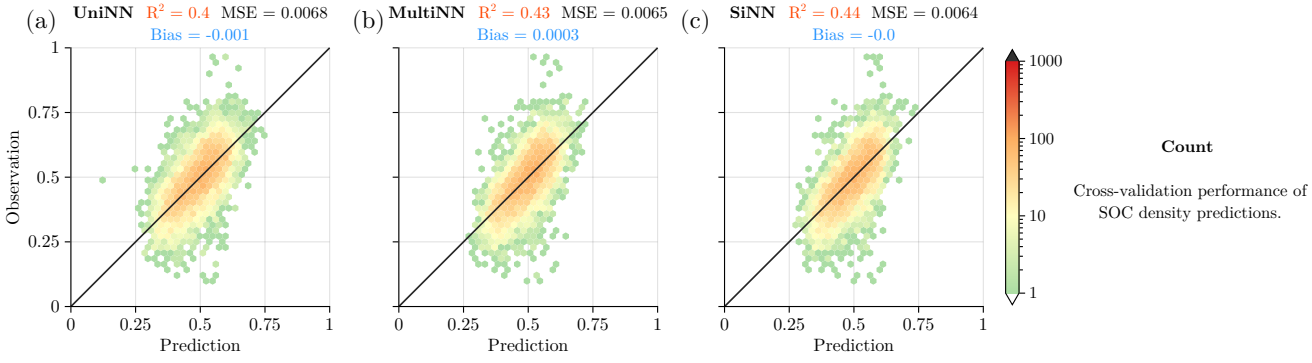


Figure 2. Accuracy of cross-validated SOC density predictions for the three models: UniNN, MultiNN and SiNN. All results are shown in the transformed target space used during model training.

These modest overall performance differences are also reflected in the land-cover-specific accuracy metrics, shown in Table 3. SiNN does not consistently outperform the other models across all land-cover classes or evaluation metrics. Based on R^2 and MSE, SiNN shows better performance in land covers typically associated with higher SOC levels, including *Grassland*, *Woodland*, and *Wetland*. Comparable performance is observed for *Cropland* and *Bareland*, whereas SiNN underperforms in *Artificial land*.

In terms of bias, UniNN tends to overestimate SOC density across most land-cover classes, with the exception of *Grassland* and *Shrubland*, resulting in an overall positive bias. In contrast, both MultiNN and SiNN exhibit overestimation in *Bareland*, *Cropland*, and *Wetland*, while showing slight underestimation in the rest land-cover classes.

3.2 Joint distribution of SOC content and BD

Figure 3 compares the observed and predicted joint distributions of SOC content and BD. The observations exhibit a clear wedge-shaped support, characterised by high SOC values occurring predominantly at low BD and an overall negative SOC–BD dependence. In the observations, BD spans a wide range (approximately 0–1.8 $g\,cm^{-3}$). In contrast, none of the three models fully reproduces the extreme ends of the BD distribution, with predicted values rarely falling below 0.2 $g\,cm^{-3}$ or exceeding 1.6 $g\,cm^{-3}$, except for two isolated cases produced by UniNN.

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

Table 3. SOC density prediction performance by land cover and model; sample sizes for each land-cover class are reported in the last row.

Metric	Model	Artificial	Bareland	Cropland	Grassland	Shrubland	Woodland	Wetland
R^2	UniNN	0.30	0.45	0.40	0.39	0.08	0.15	0.53
R^2	MultiNN	0.26	0.44	0.41	0.41	0.21	0.19	0.84
R^2	SiNN	0.18	0.44	0.41	0.43	0.17	0.21	0.90
MSE	UniNN	0.0094	0.0057	0.0050	0.0068	0.0115	0.0094	0.0239
MSE	MultiNN	0.0098	0.0058	0.0049	0.0065	0.0098	0.0090	0.0079
MSE	SiNN	0.0110	0.0058	0.0049	0.0064	0.0103	0.0088	0.0049
Bias	UniNN	0.0025	0.0149	0.0070	-0.0033	-0.0163	0.0046	0.0148
Bias	MultiNN	-0.0086	0.0138	0.0034	-0.0044	-0.0183	-0.0008	0.0197
Bias	SiNN	-0.0066	0.0164	0.0033	-0.0042	-0.0204	-0.0024	0.0360
N	–	24	203	2339	1159	161	1306	2

The negative SOC–BD relationship observed in the data is qualitatively reproduced by all three models. However, this relationship is more clearly expressed and more strongly constrained in the MultiNN and SiNN predictions, whereas UniNN exhibits a weaker coupling between SOC content and BD.

In the observations, SOC content at very low BD (<0.2 $g\,cm^{-3}$) shows a bimodal pattern, with both very high (400-500 $g\,kg^{-1}$) and low SOC values (<100 $g\,kg^{-1}$). While UniNN produces SOC predictions exceeding 200 $g\,kg^{-1}$, these occur mostly at moderate BD values rather than at low BD. MultiNN shows a reduced occurrence of such implausible combinations, while SiNN largely avoids them. Overall, the predicted SOC–BD distributions become progressively more constrained from UniNN to MultiNN to SiNN, with extreme and weakly supported combinations occurring less frequently.

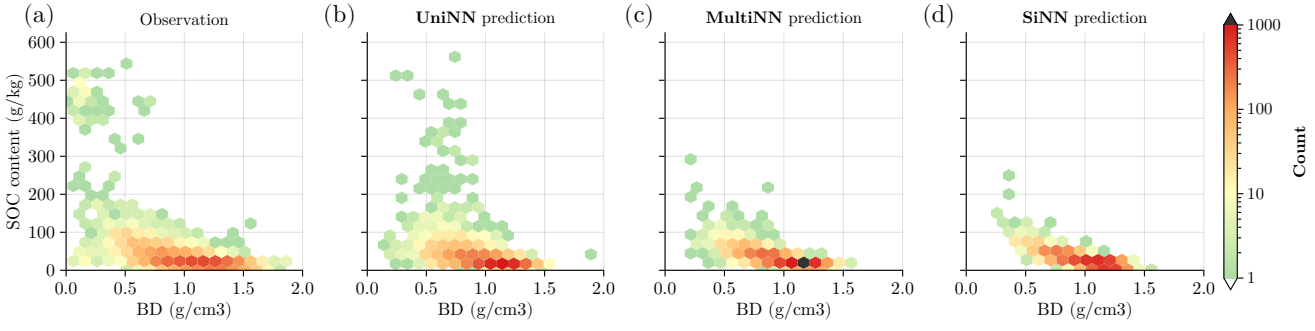


Figure 3. Joint distribution of SOC content ($g\,kg^{-1}$) and bulk density ($g\,cm^{-3}$) for observations and model predictions.

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

3.3 Temporal performance

Temporal performance is assessed from two perspectives: temporal plausibility and temporal transferability. Plausiblity is assessed based on the stability of SOC density time series derived from cross-validation predictions, whereas transferability is evaluated by training the models on 2018 data and predicting other years.

The stability of predicted SOC density time series is quantified using the temporal range, defined as the maximum absolute change in SOC density within each predicted time series across survey years. The resulting range distributions (Fig. 4) are long-tailed for all three models, with the majority of values concentrated around a change magnitude of approximately 3 $kg\,m^{-3}$. Despite similar median ranges, clear differences emerge in the upper tails of the distributions. UniNN exhibits the largest extreme changes, followed by MultiNN, whereas SiNN shows a much reduced upper outliers. Overall, the SiNN produces smoother and more temporally stable SOC density trajectories.

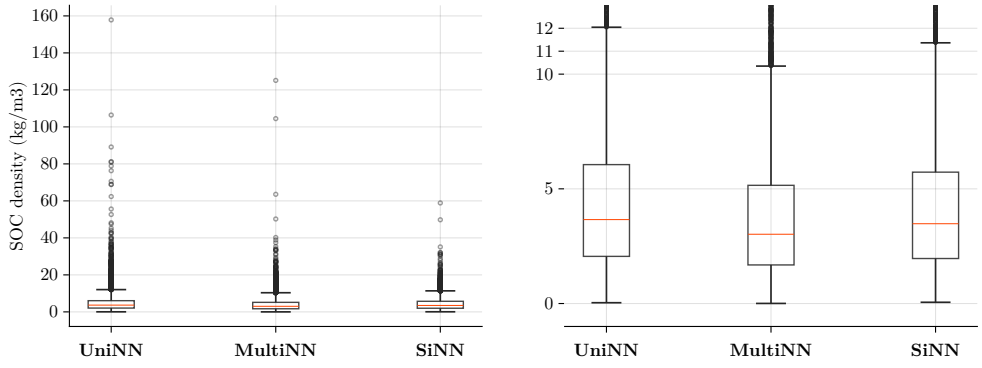


Figure 4. Temporal range of predicted SOC density time series across LUCAS survey years for the three models (left), with a zoomed-in view for detailed inspection (right). Quantile statistics of the temporal ranges are shown for UniNN (5% = 0.76, 50% = 3.66, 95% = 12.35), MultiNN (5% = 0.63, 50% = 3.02, 95% = 10.66), and SiNN (5% = 0.74, 50% = 3.48, 95% = 10.39).

Temporal transferability is assessed indirectly using SOC content. Among the three models, UniNN achieves the highest prediction accuracy when transferred to other survey years, followed by SiNN, while MultiNN shows the lowest accuracy. However, when these results are compared with the cross-validation performance reported in Appendix X, a larger performance drop is observed for UniNN (from $R^2 = 0.59$ to $R^2 = 0.49$) and MultiNN (from $R^2 = 0.49$ to $R^2 = 0.42$) than for SiNN (from $R^2 = 0.48$ to $R^2 = 0.46$). This smaller degradation in performance indicates that the SiNN model exhibits greater temporal robustness.

3.4 Plausibility of oBD and mBD

An additional advantage of SiNN is that, beyond predicting the target soil properties, it simultaneously estimates the latent parameters oBD and mBD. Figure 6 shows the distributions of the latent oBD and mBD components across seven land cover classes. Overall, predicted oBD values range from approximately 0.1 to 0.3 $g\,cm^{-3}$ and exhibit a clearer land cover relevant

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

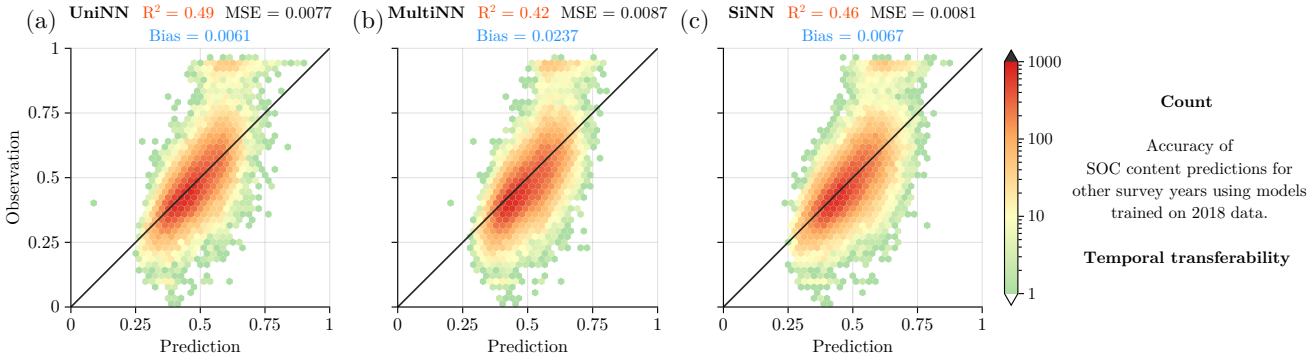


Figure 5. Accuracy of SOC content predictions for other survey years using models trained on 2018 data, assessing temporal transferability. All results are shown in the transformed target space used during model training.

pattern than mBD. From *Bareland* and *Cropland* through *Grassland* and *Shrubland* to *Woodland*, oBD decreases gradually from around 0.22 to approximately 0.15 $g\,cm^{-3}$, consistent with increasing vegetation cover. The oBD of *Artificial land* lies between that of *Cropland* and *Grassland* (around 0.2 $g\,cm^{-3}$), while *Wetland* exhibits the lowest mean oBD and largest spread among all land cover classes. Assuming peat as an organic end member, with reported peat BD values of approximately 0.1–0.2 $g\,cm^{-3}$ (Päivänen, 1969; Huat et al., 2011), and organic forest soils exhibiting oBD values around 0.11 $g\,cm^{-3}$ (Perie and Ouimet, 2008), the inferred oBD range appears physically plausible (Robinson et al., 2022).

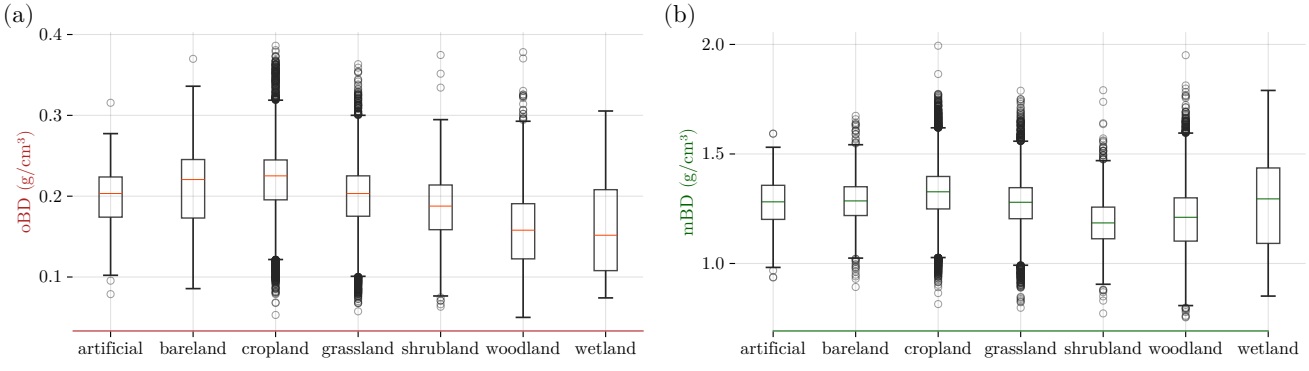


Figure 6. Distribution of latent parameters oBD (a) and mBD (b) across different land covers.

In contrast, mBD values range from approximately 1.0 to 1.5 $g\,cm^{-3}$ and show much weaker differentiation across land covers. The mBD distributions for *Artificial land*, *Bareland*, *Cropland*, *Grassland*, and *Wetland* are centered around similar values (approximately 1.3 $g\,cm^{-3}$), although *Wetland* shows a slightly larger spread. *Shrubland* and *Woodland* exhibit slightly lower mBD values, with mean values below approximately 1.25 $g\,cm^{-3}$. Compared with mBD values estimated from soil data

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

collected in Wales (1.9 $g\,cm^{-3}$), the values inferred here are lower, but are comparable to those suggested by Stewart et al. (1970) (1.4 $g\,cm^{-3}$).

Fig 7 shows the distributions of the estimated oBD and mBD across the soil texture space. Both oBD and mBD exhibit lower values in fine-textured soils, particularly where sand content is below 10 %, corresponding to silty clay and clay textures. In contrast, higher oBD and mBD values are observed in coarse-textured soils characterised by high sand content (above 70 %) and low clay content (below 40 %), corresponding to sandy clay loam and sandy loam textures. In other texture regions, the patterns are more mixed.

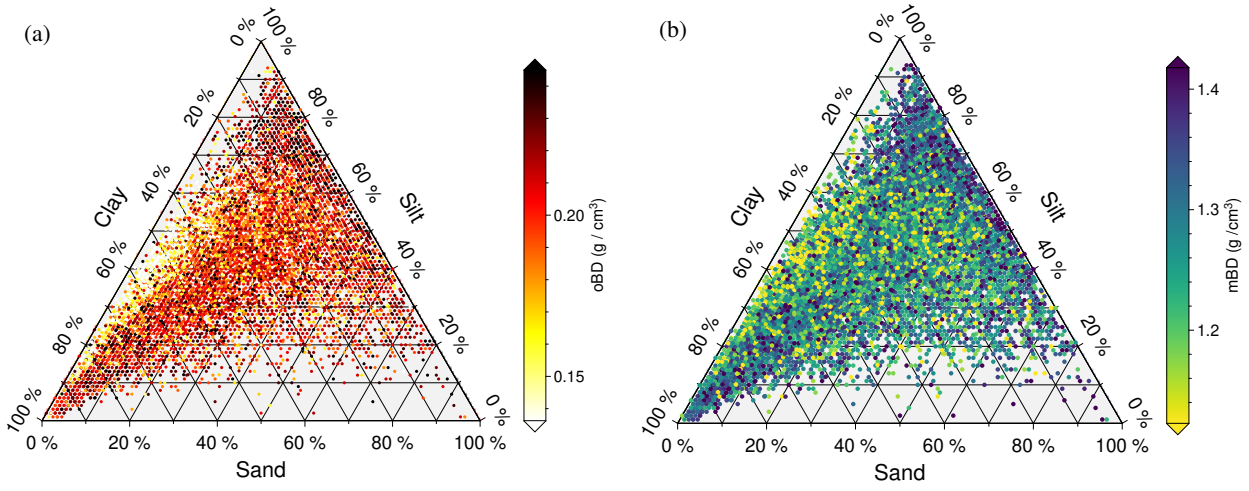


Figure 7. Distribution of latent parameters oBD (a) and mBD (b) across textures.

Fig 8 presents example maps of these latent parameters generated by SiNN. Consistent with the land-cover analysis shown in Fig. 6, clear land-cover and land-use patterns are evident: higher oBD and mBD values are observed in cropland, followed by artificial land associated with built-up areas, and lower values in woodland regions, which are indicated by dense vegetation (dark green) in the satellite imagery.

4 Discussion

4.1 Model performance under sparse data conditions

Across all target variables with sparse data, including SOC density, BD, and CF, SiNN generally exhibits comparable or slightly better performance than MultiNN, followed by UniNN, with the clearest improvements observed for SOC density (see accuracy plots for other properties in Appendix Fig. A1). These improvements, however, are not consistent across all land-cover classes.

An exception is SOC content, for which prediction accuracy decreases from UniNN ($R^2 = 0.59$) to MultiNN ($R^2 = 0.49$) and further to SiNN ($R^2 = 0.48$). UniNN achieves the highest accuracy for SOC content, which is also the most available variable, while slightly underperforming for properties with sparser observations. From a purely univariate accuracy perspective,

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

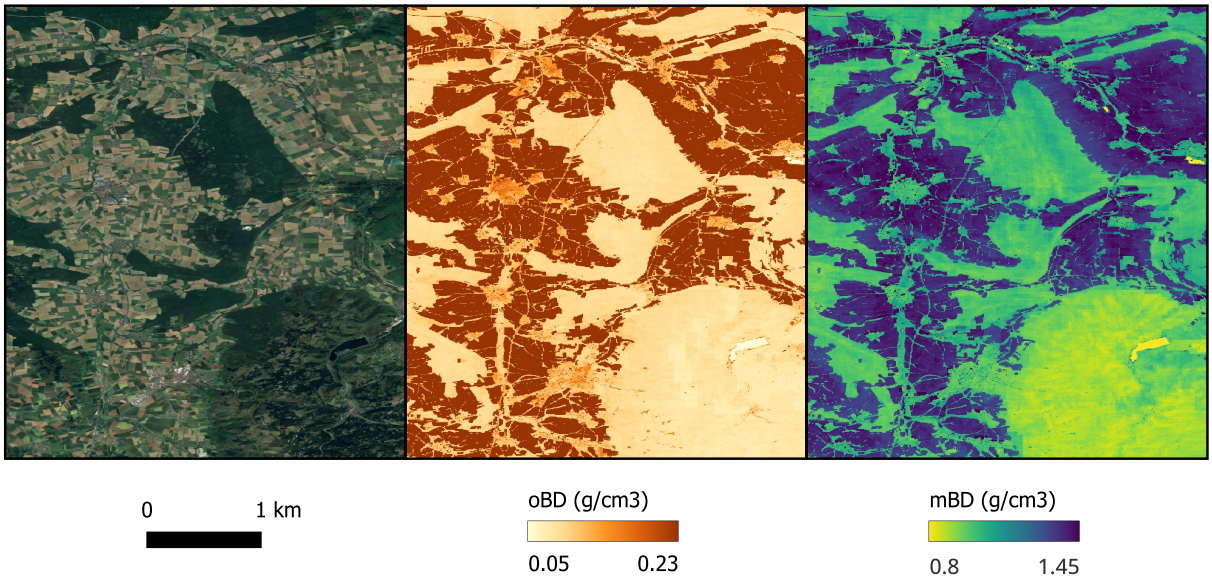


Figure 8. Maps of the predicted latent parameters oBD (middle) and mBD (right), shown alongside the corresponding satellite imagery (left). Predictions are generated at 30 m spatial resolution for the year 2018 over a region in central Germany (centre location: longitude 10.19320, latitude 51.98643). Satellite imagery: map data ©2025 GeoBasis-DE/BKG (©2009), Google.

this suggests that different model structures may be preferable depending on data availability: the soil science-informed model is advantageous for SOC density prediction under sparse auxiliary data, whereas MultiNN achieves comparable accuracy for BD and CF with lower computational cost, and the univariate approach remains competitive for well-represented variables such as SOC content.

4.2 Plausibility of predictions from the perspective of soil science

The higher SOC content accuracy achieved by UniNN appears to be associated with its ability to predict very high SOC values, whereas MultiNN and SiNN tend to underestimate the upper tail of the SOC distribution (Fig. A1-a). However, when the joint SOC–BD relationship is considered, these high SOC predictions by UniNN are often implausible. In the observations, very high SOC content typically co-occurs with very low BD values (below approximately 0.2 $g\,cm^{-3}$), whereas UniNN frequently predicts such high SOC values at moderate BD levels (approximately 0.5–1.0 $g\,cm^{-3}$), thereby placing probability mass in sparsely supported regions of the SOC–BD space.

Temporal transferability experiments further show that UniNN exhibits the largest performance degradation when models trained on one survey year are transferred to other years. This pronounced drop indicates a higher sensitivity of UniNN to temporal domain shifts and a less stable generalisation across time, compared with the more constrained multivariate and soil science-informed model structures. Consistent with this finding, temporal plausibility analyses reveal that UniNN produces noisier SOC density trajectories, with larger extreme changes exceeding 60 g cm^{-3}. Such magnitudes are highly implausible

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

over the nine-year LUCAS interval, even under strong anthropogenic intervention (Poeplau et al., 2011; Gubler et al., 2019). In contrast, MultiNN shows intermediate behaviour, while SiNN exhibits the most stable temporal dynamics. Overall, the soil science-informed model provides the most plausible and temporally coherent SOC density time series under temporal domain shifts.

4.3 Latent soil physical parameters

Beyond predicting the target soil properties, SiNN additionally yields latent estimates of oBD and mBD. Although direct ground truth data for these latent parameters are not available for validation, their inferred ranges are broadly consistent with values reported in the literature, and their plausibility is supported by systematic patterns observed across land cover classes and soil texture space. It should be noted that the learning of oBD and mBD is constrained by prior information specified in this study. In our case, these priors are knowledge-based and lead to physically reasonable and numerically stable solutions. However, fixed priors implicitly assume their validity beyond the calibration domain, and their suitability may vary. Therefore, the use of the SiNN requires selecting or adapting prior ranges based on domain knowledge and the specific environmental context of the application.

The latent parameter estimates also provide insights into the spatial variability of oBD and mBD. As illustrated in Fig. 8, clear land cover and land use associations are evident for both parameters. Such patterns are expected for oBD, as land cover and land use influence organic matter inputs and accumulation (Smith, 2008). For mBD, however, the interpretation is less straightforward. The observed spatial variation may reflect differences in soil compaction associated with land use, differences in soil texture, or a combination of both. In the absence of independent ground truth data, these hypotheses cannot be explicitly tested within this study.

Estimates of oBD and mBD can be further linked to soil porosity and compaction characteristics (Robinson et al., 2022). Soil porosity, defined as the fraction of soil volume occupied by pore space, plays a key role in regulating aeration, water movement, and nutrient transport, and is therefore closely related to soil structure and ecosystem functioning (Pagliai and De Nobili, 1993; Hao et al., 2008; Assouline and Or, 2013). Using the latent parameters inferred by SiNN, soil porosity can be derived following the analytical model developed by Robinson et al. (2022). This illustrates that the SiNN enables large-scale porosity estimation by combining legacy soil observations with EO covariates, without requiring direct porosity measurements. As shown in Appendix B, the derived porosity patterns are consistent with those reported by Robinson et al. (2022) based on soil data from Great Britain and Wales: porosity generally increases with SOC content and levels off at SOC contents of approximately 200 $\mathrm{g\,kg^{-1}}$ (Fig. B1).

Soil compaction is a major form of soil degradation (Nawaz et al., 2013; Shah et al., 2017). BD and its derivatives, such as packing density, are commonly used as proxies for soil compaction because compaction directly leads to increased BD (Stolf et al., 2011; Panagos et al., 2024). However, BD is influenced not only by compaction but also by SOC content, which complicates its interpretation. By explicitly estimating mBD, SiNN rules out the influence of SOC on BD. When further normalised by the mineral particle density (mPD), assumed to be constant, the ratio mBD/mPD can be used as a proxy indicator of soil compaction. Using this compaction indicator, Fig. B1 compares its distribution across land-cover classes. The highest

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

mean values are observed in *Cropland* and *Woodland*, with *Woodland* exhibiting the largest spread. These patterns may reflect the effects of mechanised operations associated with intensive cropping and forest silviculture (Nawaz et al., 2013). This is consistent with the findings of Panagos et al. (2024), who reported a higher susceptibility of arable land to soil compaction using packing density as an indicator.

The case of porosity and compaction illustrate how the soil science-informed model structure enables interpretation of soil physical characteristics beyond the directly predicted variables, providing added soil science value and practical relevance for soil assessment. Furthermore, these latent parameter estimates also enhance model interpretability by explicitly linking SOC content and BD within the model structure. This improves transparency and supports a more process-consistent understanding of model behaviour.

4.4 Limitations

Despite the added value introduced by the soil science-informed structure, SiNN does not fully resolve the challenge of predicting highly organic soils. Both MultiNN and SiNN fail to accurately reproduce the upper tail of SOC content associated with very low BD. This limitation can be viewed from two related perspectives: an underestimation of extreme SOC and BD values, and an incomplete reproduction of their joint relationship.

As illustrated in Fig. A2, highly organic soils—characterised by SOC content exceeding 200 $g\,kg^{-1}$ and BD below approximately 0.2 $g\,cm^{-3}$—remain at the lower edge of the predicted BD range for all models, yet their predicted BD values are generally higher than observed. While both MultiNN and SiNN largely fail to predict SOC content above 200 $g\,kg^{-1}$, they preserve the qualitative association between high SOC and low BD. In contrast, UniNN is able to generate high SOC predictions, but these are weakly constrained, often occurring at moderate BD values and failing to reproduce both the low-BD regime and the joint SOC–BD structure.

Two factors may contribute to this limitation. First, the input data may be insufficient to distinguish highly organic soils. They are intrinsically under-represented in the training set, and the available covariates may not capture the factors that differentiate these soils. Although most highly organic samples originate from *Woodland* and *Grassland*, many other samples from the same land-cover classes occupy similar covariate space while exhibiting lower SOC content and higher bulk density, thereby limiting the discriminative power of the predictors. Second, all model experiments are conducted in a transformed target space, with logarithmic scaling applied. This transformation compresses differences at the upper end of the distribution (e.g., between 200 and 600 $g\,kg^{-1}$), reducing the influence of extreme values during loss optimisation and contributing to systematic underestimation of the highest SOC contents.

A further limitation is the lack of repeated SOC density measurements across survey years, which prevents direct validation of predicted temporal changes. This constraint is intrinsic to the sparse availability of BD and CF data and necessitates indirect assessments of temporal transferability and plausibility using SOC content and relative change metrics. As future LUCAS survey rounds become available and provide more complete measurements, this limitation is expected to be alleviated.

In addition, this study focuses primarily on accuracy-based evaluation and does not include uncertainty quantification, which has become increasingly common in digital soil mapping to support map interpretation and use. Although the gains in predictive

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

accuracy introduced by structural constraints are modest, such constraints may contribute to reduced predictive uncertainty. Given the availability of established uncertainty quantification methods for neural networks (Huang et al., 2025), incorporating such approaches is identified as a priority for future work.

5 Conclusions

This study evaluated three NN architectures for reconstructing SOC density time series while preserving consistent relationships between soil properties, particularly between SOC content and BD. The results show that the soil relation informed model generally achieved the most favourable balance between prediction accuracy, temporal stability, and joint SOC–BD plausibility. While improvements in SOC density accuracy are modest, SiNN consistently produced less noisy SOC density trajectories and better preserved the physically meaningful association between SOC and BD, followed by MultiNN and UniNN. In addition, the latent parameters estimated by SiNN enable the derivation, interpretation and assessment of soil physical properties, such as porosity and compaction-related indicators.

In contrast, the univariate model performed well for abundantly sampled variables such as SOC content, but failed to reproduce plausible joint SOC–BD behaviour while exhibiting greater sensitivity to temporal domain shifts. These findings highlight that, in soil property modelling and mapping, evaluating models solely based on univariate accuracy can be misleading. Incorporating multivariate structure and soil constraints provides added value by improving robustness, plausibility, temporal coherence, and interpretability from a soil science perspective.

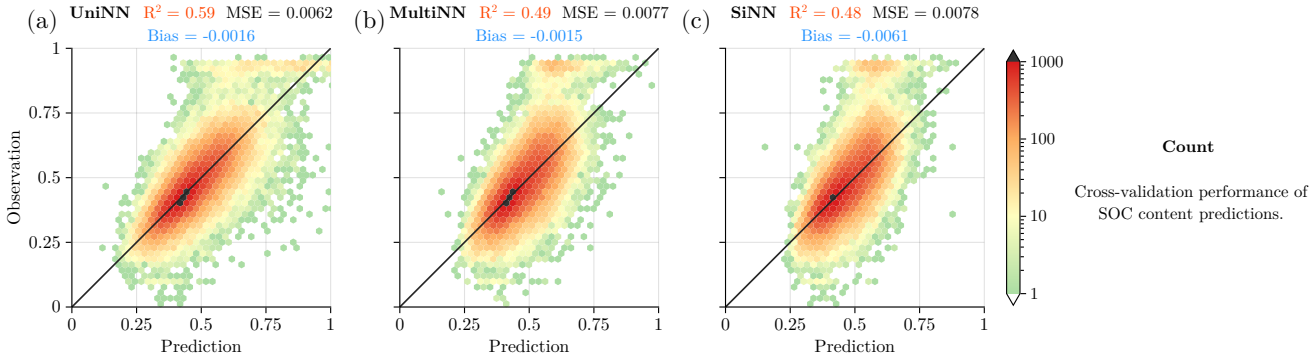
Code and data availability. The code used for the modelling experiments, analysis, and visualisation is available via a GitHub repository at https://github.com/AI4SoilHealth/EasyDensity.jl (last access: 15 January 2026). The EasyHybrid.jl package used in this study is available at https://github.com/EarthyScience/EasyHybrid.jl. The soil measurements analysed in this study are curated and maintained by the European Soil Data Centre (ESDAC) and are openly available through the LUCAS Topsoil database (https://esdac.jrc.ec.europa.eu/projects/lucas). All covariate layers used for modelling are openly accessible via the EcoDataCube platform (https://ecodatacube.eu/, last access: 24 December 2025).

Appendix A: Complementary accuracy assessment and joint distribution analysis

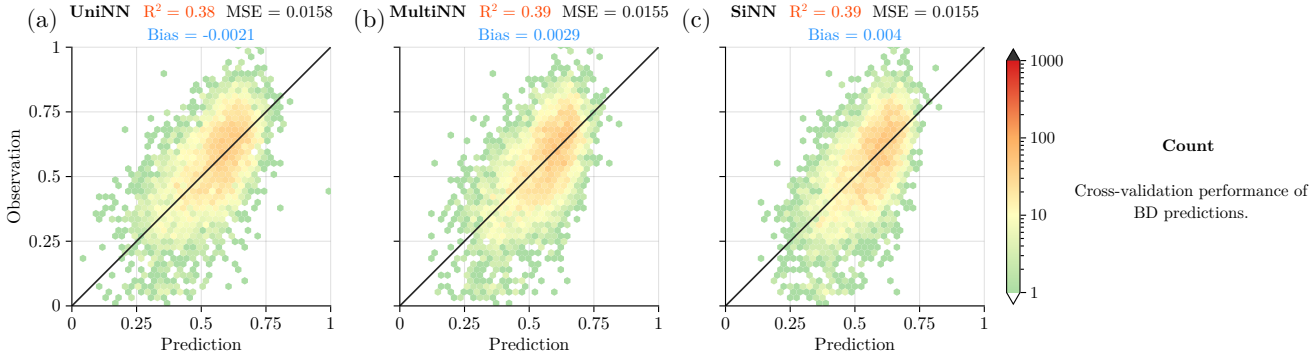
Figure A1 presents complementary cross-validation accuracy plots for the three models for the remaining target variables: SOC content (a), BD (b), and CF (c). UniNN achieves the highest prediction accuracy for SOC content ($R^2 = 0.59$), substantially outperforming MultiNN ($R^2 = 0.49$) and SiNN ($R^2 = 0.48$). In contrast, for BD and CF, UniNN shows slightly lower performance compared to the multivariate and soil science-informed models.

Figure A2 shows the joint distribution of SOC content and bulk density for highly organic soils, defined by SOC content exceeding 200 $g\,kg^{-1}$ and bulk density below approximately 0.2 $g\,cm^{-3}$. Across all models, these samples remain at the lower edge of the predicted bulk density range; however, predicted BD values are generally higher than those observed. Both

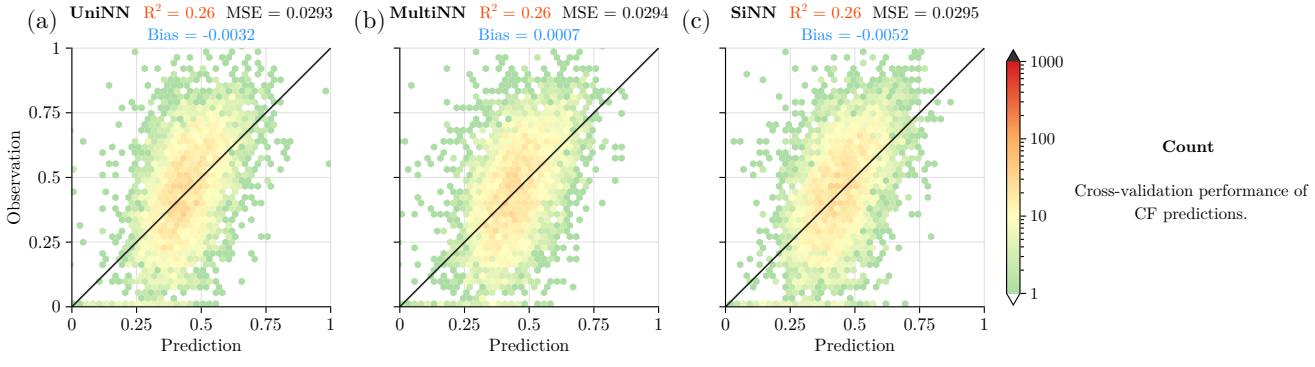
https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026



(a) SOC content prediction accuracy



(b) BD prediction accuracy



(c) CF prediction accuracy

Figure A1. Accuracy of cross-validated predictions for the three models, UniNN, MultiNN, and SiNN, across different soil properties. All results are shown in the transformed target space used during model training.

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

MultiNN and SiNN largely underpredict SOC content above 200 $g\,kg^{-1}$, although their predictions remain associated with low bulk density values. In contrast, UniNN is able to generate high SOC predictions, but these frequently occur at moderate bulk density levels.

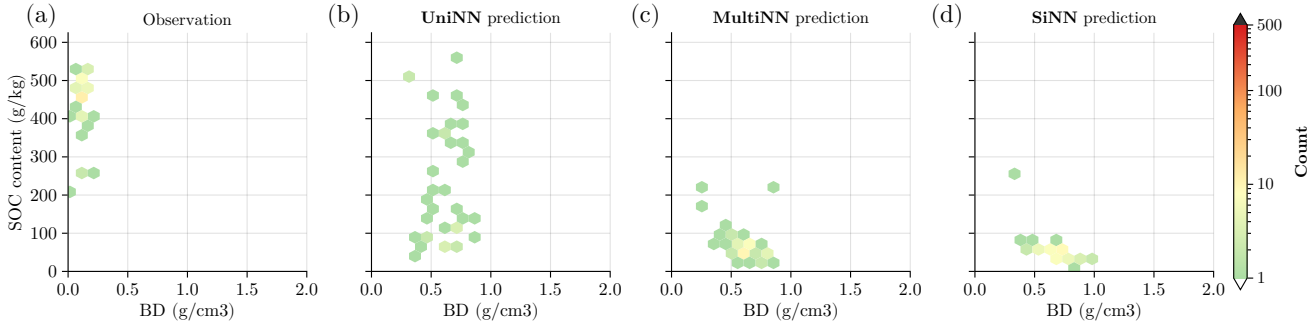


Figure A2. Joint distribution of SOC content ($g\,kg^{-1}$) and bulk density ($g\,cm^{-3}$) for observations and model predictions, restricted to samples with SOC content exceeding 200 $g\,kg^{-1}$ and bulk density below 0.2 $g\,cm^{-3}$.

Appendix B: Porosity and compaction

Once oBD and mBD are obtained from SiNN, soil porosity can be derived. Following Robinson et al. (2022), particle densities of the organic and mineral soil constituents (oPD and mPD) are assumed to be constant, with $\mathrm{oPD} = 1.4\ \mathrm{g\,cm^{-3}}$ and $\mathrm{mPD} = 2.7\ \mathrm{g\,cm^{-3}}$. Soil porosity is then calculated as:

$$\varphi = 1 - \left[\left(\frac{SOM}{oPD} + \frac{1-SOM}{mPD}\right) \div \left(\frac{SOM}{oBD} + \frac{1-SOM}{mBD}\right)\right], \tag{B1}$$

Figure B1$(a)-(g)$ shows the derived soil porosity against SOC content across land-cover classes. Porosity is lowest in *Cropland* and *Bareland*, and highest in *Wetland* and *Woodland*. Consistent with estimates reported by Robinson et al. (2022) using soil data from Great Britain and Wales, porosity generally increases with SOC content. This increase, however, levels off at SOC contents of approximately 200 $\mathrm{g\,kg^{-1}}$.

The inverse of mineral soil porosity, expressed as the ratio of mineral bulk density to mineral particle density (mBD/mPD), could be interpreted as a proxy indicator related to soil compaction. This mBD-based indicator is a modelling construct intended to reduce the influence of SOC-driven density effects. Figure B1 (e) compares the distribution of this compaction-related indicator across land-cover classes. The highest mean values are observed in *Cropland*, whereas the lowest mean values occur in *Shrubland*. The widest distributions are found in *Woodland* and *Wetland*.

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

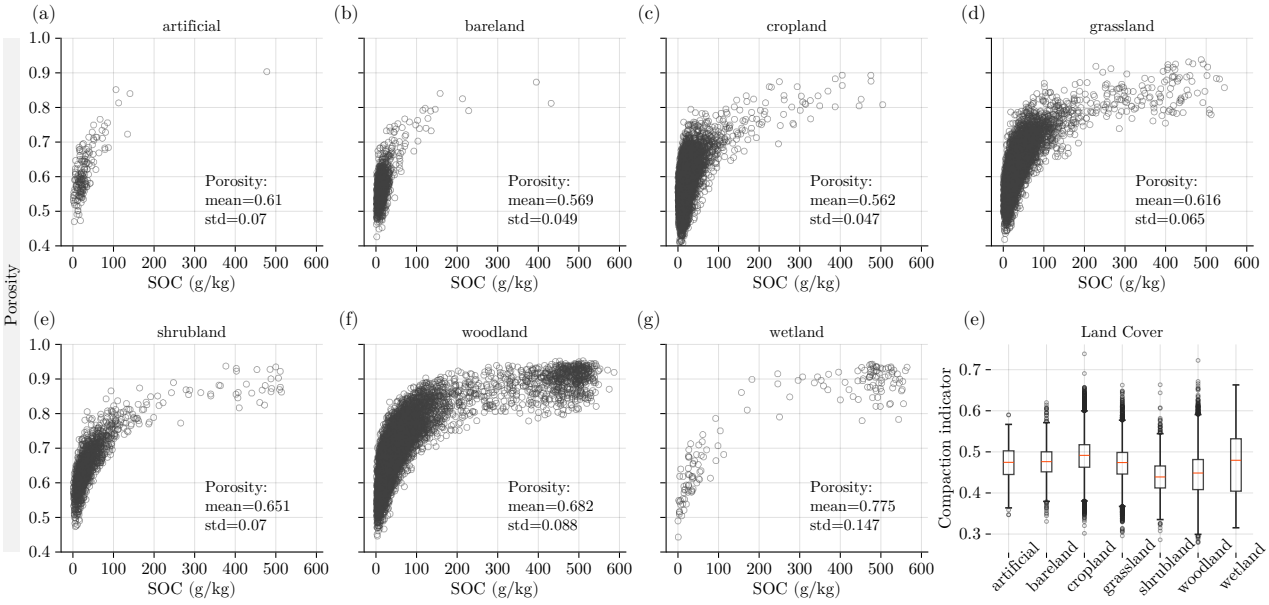


Figure B1. $(a)-(g)$. Porosity derived from predicted oBD and mBD, plotted against SOC content for each land cover class. (e) Compaction indicator calculated for each land cover.

Author contributions. Bernhard Ahrens, Xuemeng Tian, and Leo Rossdeutscher conceived the study. Xuemeng Tian, Bernhard Ahrens, and Lazaro Alonso conducted the experiments and subsequent analyses. Xuemeng Tian and Leandro Parente curated the data. Bernhard Ahrens and Lazaro Alonso developed the Julia package EasyHybrid.jl, which was used to implement the soil science–informed models in this study, and Bernhard Ahrens designed the soil science–informed models evaluated herein. Xuemeng Tian prepared the manuscript with contributions from all co-authors.

Competing interests. The authors declare that they have no competing interests.

Acknowledgements. This research was funded by the European Union under the Horizon Europe programme through AI4SoilHealth (Grant Agreement No. 101086179) and Intergenerational Open Geospatial Carbon Registry (Grant Agreement No. 101218854). The authors acknowledge the use of ChatGPT (OpenAI) to assist in improving the clarity and readability of the manuscript. The first author thank Prof. Sytze de Bruin for his valuable suggestions and insightful discussions on the experimental design and the preparation of the manuscript.

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

References

Adams, W.: The effect of organic matter on the bulk and true densities of some uncultivated podzolic soils, Journal of Soil Science, 24, 10–17, https://doi.org/10.1111/j.1365-2389.1973.tb00737.x, 1973.

Alonso, L., Ahrens, B., Reichstein, M., Tian, X., and Bi, Q.-F.: EarthyScience/EasyHybrid.jl: v0.1.7, https://doi.org/10.5281/zenodo.17794983, 2025.

Angelini, M. E., Heuvelink, G., and Kempen, B.: Multivariate mapping of soil with structural equation modelling, European Journal of Soil Science, 68, 575–591, https://doi.org/10.1111/ejss.12446, 2017.

Assouline, S. and Or, D.: Conceptual and parametric representation of soil hydraulic properties: A review, Vadose Zone Journal, 12, vzj2013–07, https://doi.org/10.2136/vzj2013.07.0121, 2013.

Bauer-Marschallinger, B., Freeman, V., Cao, S., Paulik, C., Schaufler, S., Stachl, T., Modanesi, S., Massari, C., Ciabatta, L., Brocca, L., et al.: Toward global soil moisture monitoring with Sentinel-1: Harnessing assets and overcoming obstacles, IEEE Transactions on Geoscience and Remote Sensing, 57, 520–539, https://doi.org/10.1109/TGRS.2018.2858004, 2018.

Federer, C., Turcotte, D., and Smith, C.: The organic fraction–bulk density relationship and the expression of nutrient content in forest soils, Canadian Journal of Forest Research, 23, 1026–1032, https://doi.org/10.1139/x93-131, 1993.

Gubler, A., Wächter, D., Schwab, P., Müller, M., and Keller, A.: Twenty-five years of observations of soil organic carbon in Swiss croplands showing stability overall but with some divergent trends, Environmental Monitoring and Assessment, 191, 277, https://doi.org/10.1007/s10661-019-7435-y, 2019.

Hao, X., Ball, B., Culley, J., Carter, M., and Parkin, G.: Soil density and porosity, Soil sampling and methods of analysis, 2, 743–759, 2008.

Harper, K. L., Lamarche, C., Hartley, A., Peylin, P., Ottlé, C., Bastrikov, V., San Martín, R., Bohnenstengel, S. I., Kirches, G., Boettcher, M., Shevchuk, R., Brockmann, C., , and Defourny, P.: A 29-year time series of annual 300 m resolution plant-functional-type maps for climate models, Earth System Science Data, 15, 1465–1499, https://doi.org/10.5194/essd-15-1465-2023, 2023.

Hengl, T. and MacMillan, R. A.: Predictive soil mapping with R, Wageningen: OpenGeoHub foundation, 2019.

Heuvelink, G. B., Kros, J., Reinds, G. J., and De Vries, W.: Geostatistical prediction and simulation of European soil property maps, Geoderma Regional, 7, 201–215, https://doi.org/10.1016/j.geodrs.2016.04.002, 2016.

Ho, Y.-F., Grohmann, C. H., Lindsay, J., Reuter, H. I., Parente, L., Witjes, M., and Hengl, T.: Global Ensemble Digital Terrain modeling and parametrization at 30 m resolution (GEDTM30): a data fusion approach based on ICESat-2, GEDI and multisource data, ResearchSquare, https://doi.org/10.21203/rs.3.rs-6280607/v1, 2025.

Huang, Y.-C., Padarian, J., Minasny, B., and McBratney, A. B.: Using Monte Carlo conformal prediction to evaluate the uncertainty of deep-learning soil spectral models, Soil, 11, 553–563, https://doi.org/10.5194/soil-11-553-2025, 2025.

Huat, B. B., Kazemian, S., Prasad, A., and Barghchi, M.: State of an art review of peat: General perspective, International Journal of the Physical Sciences, 6, 1988–1996, https://doi.org/10.5897/IJPS11.192, 2011.

Isik, S., Minarik, R., and Hengl, T.: Continental Europe surface lithology based on EGDI / OneGeology map at 1:1M scale, https://doi.org/10.5281/zenodo.12607973, 2024.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., and Kessler, M.: Climatologies at high resolution for the earth's land surface areas, Scientific data, 4, 1–20, https://doi.org/10.1038/sdata.2017.122, 2017.

Karger, D. N., Wilson, A. M., Mahony, C., Zimmermann, N. E., and Jetz, W.: Global daily 1 km land surface precipitation based on cloud cover-informed downscaling, Scientific Data, 8, 307, https://doi.org/https://www.nature.com/articles/s41597-021-01084-6, 2021.

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

Lal, R.: Soil carbon sequestration to mitigate climate change, Geoderma, 123, 1–22, https://doi.org/10.1016/j.geoderma.2004.01.032, 2004.

Lehmann, J., Bossio, D. A., Kögel-Knabner, I., and Rillig, M. C.: The concept and future prospects of soil health, Nature Reviews Earth & Environment, 1, 544–553, https://doi.org/10.1038/s43017-020-0080-8, 2020.

Ließ, M. and Sakhaee, A.: Deep Learning with a Multi-Task Convolutional Neural Network to Generate a National-Scale 3D Soil Data Product: The Particle Size Distribution of the German Agricultural Soil Landscape, Agriculture, 14, 1230, https://doi.org/10.3390/agriculture14081230, 2024.

Lyapustin, A. and Wang, Y.: MCD19A2 MODIS/Terra+Aqua Land Aerosol Optical Depth Daily L2G Global 1km SIN Grid V006, NASA EOSDIS Land Processes Distributed Active Archive Center, https://doi.org/10.5067/MODIS/MCD19A2.006, 2018.

Ma, Y., Minasny, B., Malone, B. P., and Mcbratney, A. B.: Pedology and digital soil mapping (DSM), European Journal of Soil Science, 70, 216–235, https://doi.org/10.1111/ejss.12790, 2019.

Minasny, B., Bandai, T., Ghezzehei, T. A., Huang, Y.-C., Ma, Y., McBratney, A. B., Ng, W., Norouzi, S., Padarian, J., Sharififar, A., et al.: Soil science-informed machine learning, Geoderma, 452, 117 094, https://doi.org/10.1016/j.geoderma.2024.117094, 2024.

Nawaz, M. F., Bourrie, G., and Trolard, F.: Soil compaction impact and modelling. A review, Agronomy for sustainable development, 33, 291–309, https://doi.org/10.1007/s13593-011-0071-8, 2013.

Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., and McBratney, A. B.: Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra, Geoderma, 352, 251–267, https://doi.org/10.1016/j.geoderma.2019.06.016, 2019.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., and Fernández-Ugalde, O.: LUCAS Soil, the largest expandable soil dataset for Europe: a review, European Journal of Soil Science, 69, 140–153, https://doi.org/10.1111/ejss.12499, 2018.

Pacini, L., Yunta, F., Jones, A., Montanarella, L., Barrè, P., Saia, S., Chen, S., and Schillaci, C.: Fine earth soil bulk density at 0.2 m depth from Land Use and Coverage Area Frame Survey (LUCAS) soil 2018, European Journal of Soil Science, 74, e13 391, https://doi.org/10.1111/ejss.13391, 2023.

Padarian, J., Minasny, B., and McBratney, A. B.: Using deep learning for digital soil mapping, Soil, 5, 79–89, https://doi.org/10.5194/soil-5-79-2019, 2019.

Pagliai, M. and De Nobili, M.: Relationships between soil porosity, root development and soil enzyme activity in cultivated soils, Geoderma, 56, 243–256, https://doi.org/10.1016/B978-0-444-81490-6.50024-8, 1993.

Päivänen, J.: The bulk density of peat and its determination, Silva Fennica, 3, https://doi.org/10.14214/sf.a14569, 1969.

Panagos, P., De Rosa, D., Liakos, L., Labouyrie, M., Borrelli, P., and Ballabio, C.: Soil bulk density assessment in Europe, Agriculture, Ecosystems & Environment, 364, 108 907, https://doi.org/10.1016/j.agee.2024.108907, 2024.

Perie, C. and Ouimet, R.: Organic carbon, organic matter and bulk density relationships in boreal forest soils, Canadian journal of soil science, 88, 315–325, https://doi.org/10.4141/CJSS06008, 2008.

Poeplau, C., Don, A., Vesterdal, L., Leifeld, J., Van Wesemael, B., Schumacher, J., and Gensior, A.: Temporal dynamics of soil organic carbon after land-use change in the temperate zone–carbon response functions as a model approach, Global change biology, 17, 2415–2427, https://doi.org/10.1111/j.1365-2486.2011.02408.x, 2011.

Poeplau, C., Vos, C., and Don, A.: Soil organic carbon stocks are systematically overestimated by misuse of the parameters bulk density and rock fragment content, Soil, 3, 61–66, https://doi.org/10.5194/soil-3-61-2017, 2017.

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

Potapov, P., Turubanova, S., Hansen, M. C., Tyukavina, A., Zalles, V., Khan, A., Song, X.-P., Pickens, A., Shen, Q., and Cortez, J.: Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century, Nature Food, 3, 19–28, https://doi.org/10.1038/s43016-021-00429-z, 2022.

Robinson, D., Thomas, A., Reinsch, S., Lebron, I., Feeney, C., Maskell, L., Wood, C., Seaton, F., Emmett, B., and Cosby, B.: Analytical modelling of soil porosity and bulk density across the soil organic matter and land-use continuum, Scientific reports, 12, 7085, https://doi.org/10.1038/s41598-022-11099-7, 2022.

Rogge, D., Bauer, A., Zeidler, J., Mueller, A., Esch, T., and Heiden, U.: Building an exposed soil composite processor (SCMaP) for mapping spatial and temporal characteristics of soils with Landsat imagery (1984–2014), Remote Sensing of Environment, 205, 1–17, https://doi.org/10.1016/j.rse.2017.11.004, 2018.

Shah, A. N., Tanveer, M., Shahzad, B., Yang, G., Fahad, S., Ali, S., Bukhari, M. A., Tung, S. A., Hafeez, A., and Souliyanonh, B.: Soil compaction effects on soil health and cropproductivity: an overview, Environmental Science and Pollution Research, 24, 10 056–10 067, https://doi.org/10.1007/s11356-017-8421-y, 2017.

Shimada, M. and Ohtaki, T.: Generating large-scale high-quality SAR mosaic datasets: Application to PALSAR data for global monitoring, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 3, 637–656, https://doi.org/JSTARS.2010.2077619, 2010.

Smith, P.: Land use change and soil organic carbon dynamics, Nutrient Cycling in Agroecosystems, 81, 169–178, https://doi.org/10.1007/s10705-007-9138-y, 2008.

Stewart, V., Adams, W., and Abdulla, H.: Quantitative pedological studies on soils derived from Silurian mudstones: II. The relationship between stone content and the apparent density of the fine earth, Journal of Soil Science, 21, 248–255, https://doi.org/10.1111/j.1365-2389.1970.tb01174.x, 1970.

Stolf, R., Thurler, Á. d. M., Bacchi, O. O. S., and Reichardt, K.: Method to estimate soil macroporosity and microporosity based on sand content and bulk density, Revista Brasileira de Ciência do Solo, 35, 447–459, 2011.

Sun, Q., Zhang, P., Jiao, X., Lin, X., Duan, W., Ma, S., Pan, Q., Chen, L., Zhang, Y., You, S., Liu, S., Hao, J., Li, H., and Sun, D.: A global estimate of monthly vegetation and soil fractions from spatio-temporally adaptive spectral mixture analysis during 2001–2022, Earth System Science Data Discussions, 2023, 1–30, https://doi.org/10.5194/essd-16-1333-2024, 2023.

Taghizadeh-Mehrjardi, R., Mahdianpari, M., Mohammadimanesh, F., Behrens, T., Toomanian, N., Scholten, T., and Schmidt, K.: Multi-task convolutional neural networks outperformed random forest for mapping soil particle size fractions in central Iran, Geoderma, 376, 114 552, https://doi.org/10.1016/j.geoderma.2020.114552, 2020.

Tian, X., Consoli, D., Witjes, M., Schneider, F., Parente, L., Şahin, M., Ho, Y.-F., Minařík, R., and Hengl, T.: Time series of Landsat-based bimonthly and annual spectral indices for continental Europe for 2000–2022, Earth System Science Data, 17, 741–772, https://doi.org/10.5194/essd-17-741-2025, 2025a.

Tian, X., de Bruin, S., Schneider, F., Herold, M., and de Beurs, K.: Modeling soil organic carbon changes using signal-to-noise analysis: a case study using European soil survey datasets, https://doi.org/10.21203/rs.3.rs-7308469/v1, 2025b.

Van Der Westhuizen, S., Heuvelink, G. B., and Hofmeyr, D. P.: Multivariate random forest for digital soil mapping, Geoderma, 431, 116 365, https://doi.org/10.1016/j.geoderma.2023.116365, 2023.

Wadoux, A. M.-C.: Using deep learning for multivariate mapping of soil with quantified uncertainty, Geoderma, 351, 59–70, https://doi.org/10.1016/j.geoderma.2019.05.012, 2019.

https://doi.org/10.5194/egusphere-2026-229
Preprint. Discussion started: 28 January 2026

Wadoux, A. M.-C., Minasny, B., and McBratney, A. B.: Machine learning for digital soil mapping: Applications, challenges and suggested solutions, Earth-Science Reviews, 210, 103 359, https://doi.org/10.1016/j.earscirev.2020.103359, 2020.

Wagner, W., Bauer-Marschallinger, B., Navacchi, C., Reuß, F., Cao, S., Reimer, C., Schramm, M., and Briese, C.: A Sentinel-1 backscatter datacube for global land monitoring applications, Remote Sensing, 13, 4622, https://doi.org/10.3390/rs13224622, 2021.

Wan, Z.: MODIS land surface temperature products users' guide, Institute for Computational Earth System Science, University of California: Santa Barbara, CA, USA, 805, 26, 2006.

Widyastuti, M. T., Minasny, B., Padarian, J., and Maggi, F.: PEATGRIDS: Mapping global peat thickness and carbon stock via digital soil mapping approach, dataset, https://doi.org/10.5281/zenodo.15070037, 2024.

Zhang, L., Heuvelink, G. B., Mulder, V. L., Chen, S., Deng, X., and Yang, L.: Using process-oriented model output to enhance machine learning-based soil organic carbon prediction in space and time, Science of the Total Environment, 922, 170 778, https://doi.org/10.1016/j.scitotenv.2024.170778, 2024.

Zhang, X., Xie, E., Chen, J., Peng, Y., Yan, G., and Zhao, Y.: Modelling the spatiotemporal dynamics of cropland soil organic carbon by integrating process-based models differing in structures with machine learning, Journal of Soils and Sediments, 23, 2816–2831, https://doi.org/10.1007/s11368-023-03516-9, 2023.