https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Spatially explicit seasonal modelling reveals the impact of vegetation–soil coupling on soil organic carbon prediction in bimodal-rainfall agroecosystems

Annette Achieng'[1]*, Ana Cristina Costa[1], Pedro Cabral[2,1]

[1]NOVA IMS, Universidade NOVA de Lisboa, Campus de Campolide, 1070-312 Lisbon, Portugal
[2]School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science & Technology (NUIST), Nanjing, China

*Correspondence to: Annette Achieng' (20220017@novaims.unl.pt)

Abstract. Soil organic carbon (SOC) underpins soil fertility, climate regulation, and ecosystem resilience. Yet high-resolution SOC data remain scarce in many African agroecosystems. In bimodal rainfall systems, phenological phases may imprint distinct spectral signals associated with SOC processes. However, these temporal patterns are rarely exploited in digital soil mapping. We hypothesise that SOC spatial variability is most detectable when vegetation–soil coupling is strongest and that geographically weighted machine learning enhances the ecological interpretability of SOC models in data-scarce settings. To test these hypotheses, we develop a spatially explicit, seasonally-informed framework for SOC prediction in bimodal-rainfall agroecosystems. Using environmental covariates and 11-year seasonal Landsat 7 and 8 composites to capture the full phenological cycle in Tanzania (March–May 'Long Rains', June–August 'dry season', October–December 'Short Rains'), we compare global Random Forest (RF) with Geographical Random Forest (GRF) across each seasonal window, using an annual composite as our climatological baseline.

Feature selection using Boruta reduced the covariate set by 34–50% per temporal window without compromising accuracy. Season-specific GRF models achieved R^2 values of 0.292–0.405 based on spatial cross-validation, whereas random cross-validation inflated performance to 0.444–0.471 (average overestimation = 19%). October–December covariates provided the highest predictive power using only 13 predictors. Across the temporal windows, near-infrared reflectance, green band (October–December), red band (March–May), and topsoil clay emerged as the most informative predictors. SHapley Additive exPlanations (SHAP) values extracted from the global RF were represented spatially to map the directional influence of these predictors on global predictions and subsequently assessed alongside local feature importance maps to evaluate the extent of spatial non-stationarity. This two-stage interpretability approach confirmed that the GRF models adapted dynamically to localised eco-pedological realities, such as soil texture-driven storage capacity and canopy-driven phenological masking. By demonstrating when, where and how SOC is most observable, the framework provides an operational blueprint for local organisations to design leaner, seasonally targeted soil surveys and remote sensing protocols that strengthen climate-smart land management policies and national carbon accounting systems.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

1 Introduction

Soil organic carbon (SOC) is the largest pool of organic carbon in terrestrial ecosystems (Kibblewhite et al., 2008) and plays a central role in regulating the global carbon cycle, supporting soil fertility, and sustaining ecosystem services. Small relative changes in SOC stocks can translate into substantial fluxes of carbon dioxide to or from the atmosphere, with implications for climate mitigation, land degradation neutrality, and food security agendas (Tian et al., 2015). In tropical agroecosystems experiencing pronounced agricultural expansion, land-use change, and climate variability, understanding the spatial distribution and drivers of SOC is essential for designing climate-smart land management interventions (Negassa et al., 2023; Vågen et al., 2018). This is particularly urgent in East Africa, where food security is increasingly threatened by climate extremes that trigger severe erratic rainfall and widespread crop failure (Rao Kolusu et al., 2019). Digital soil mapping (DSM) using environmental covariates and predictive models has emerged as a cost-effective approach for generating spatially explicit SOC information across large areas (Pouladi et al., 2023; Wadoux et al., 2020; Xia et al., 2022).

Leveraging existing soil datasets together with freely available satellite imagery and robust machine learning (ML) algorithms can substantially reduce monitoring costs and support evidence-based land management decisions, provided that sampling, covariate selection, and model validation strategies are explicitly designed for seasonal and spatially structured environments (FAO, 2022; Safaee et al., 2024). This optimisation raises three interlinked design questions that are addressed in this study. First, when to sample: since the efficiency of a sampling strategy depends on capturing the full feature space of the target variable (Castaldi et al., 2019a), aligning soil sampling with phenological stages that maximise the spectral variability of covariates could improve predictive performance per sample. This either increases map accuracy at a constant budget or reduces field sampling requirements for a given accuracy level. Second, what to measure: due to the highly heterogeneous nature of organic matter, SOC lacks isolated absorption peaks and instead influences reflectance across the entire visible to shortwave-infrared spectrum (Castaldi et al., 2019b). Identifying a parsimonious set of spectral and environmental predictors that consistently capture this broad signal can reduce data processing demands, simplify interpretation, and guide future sensor selection. Third, how to model: enforcing temporal consistency between soil observations and covariates, while accounting for spatial structure, is essential to ensure ecologically meaningful models and conservative performance estimates (Kim et al., 2023). By addressing these questions within a coherent analytical framework, this study proposes a spatially explicit, seasonally informed approach for SOC prediction. To illustrate its capability in resolving these complex structural and temporal dependencies, the proposed framework is applied to a bimodal rainfall agroecosystem in northern Tanzania.

Conventional bare-soil sensing is severely limited by the complex realities of persistent vegetation and crop residues in mixed-pixel scenarios such as those found in agroecosystems (Abdelbaki et al., 2025; Hosseini et al., 2026). Climatic factors such as seasonal rainfall generate distinct phenological phases (Van der Putten et al., 2013). In bimodal rainfall systems, the Long Rains (March–May) and Short Rains (October–December) drive distinct shifts in biomass, residue, and soil moisture that

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

dynamically alter surface reflectance and dictate the spectral observability of SOC (Lobell and Asner, 2002; Piao et al., 2019; Yang et al., 2015). Tanzania, like many sub-Saharan African countries, faces limited availability of harmonised, high-resolution SOC data, despite increasing demand for such information for climate policy, REDD+ reporting, and soil health monitoring (FAO, 2023; Nenkam et al., 2024). Since dense in situ sampling is unfeasible, DSM relies heavily on satellite imagery. However, standard DSM applications often use single-date or annual composites that smooth over the specific temporal windows where SOC is most detectable, obscuring the time-sensitive vegetation–soil interactions necessary for high-resolution mapping (Dvorakova et al., 2020; Yan et al., 2025). Indeed, as Hengl et al. (2021) demonstrate, while broad climatic factors often dominate coarse-scale models, high-resolution (30 m) mapping relies substantially on seasonal satellite composites.

Therefore, this study hypothesises that SOC spatial variability is most detectable when vegetation–soil coupling is strongest. By evaluating multi-temporal RS covariates across distinct seasonal windows (March–May, June–August and October–December) against an annual climatological baseline, we aim to filter out temporary surface noise and identify the optimal periods for SOC prediction.

Random Forest (RF) stands out in DSM due to its ability to capture complex nonlinear relationships, handle mixed-type environmental covariates, and provide robust feature-importance measures (Wadoux et al., 2020). However, environmental relationships are rarely uniform across large, heterogeneous landscapes. Drivers of SOC, such as soil moisture dynamics and soil parent material, often exhibit spatial non-stationarity, meaning their predictive influence shifts depending on the local environmental context (Wadoux et al., 2023). Recent developments in spatially explicit ML include geographically weighted algorithms such as Geographical Random Forest (GRF) (Georganos and Kalogirou, 2022; Kmoch et al., 2025), which account for both spatial non-stationarity and spatial autocorrelation in the data. While conventional global models assume stationary spatial relationships between the response variable and its predictors, GRF explicitly accounts for spatial heterogeneity by fitting local models that allow these relationships to vary across space (Georganos et al., 2021; Xiao et al., 2026). This makes it particularly promising for SOC prediction in data-scarce regions, where complex interactions among climate, vegetation, terrain and land use processes vary across space. Its recent implementation in Python (PyGRF) offers enhanced computational efficiency while enabling clearer interpretation of spatially varying model relationships (Sun et al., 2024).

We hypothesise that PyGRF provides a more flexible and robust framework for SOC prediction than conventional global RF models in data-scarce contexts, such as Tanzania, where sparse soil sampling and complex environmental variability present significant challenges for conventional DSM approaches. By exploiting spatial dependencies, PyGRF may better represent local soil–landscape interactions and thus enhance the reliability of SOC estimates in heterogeneous landscapes. However, assessing the true operational performance of these highly flexible models introduces a major ML paradox: the treatment of spatial autocorrelation (Jafari et al., 2026). Standard random cross-validation (RD-CV) assumes sample independence,

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

frequently yielding optimistically biased performance estimates in clustered datasets by allowing models to memorise spatial structures, rather than identify true causal relationships (de Bruin et al., 2022; Roberts et al., 2017). Conversely, spatial cross-validation (SCV) forcefully severs spatial proximity to simulate independent test data, which can impose artificial environmental extrapolation and yield overly pessimistic error estimates (Wadoux et al., 2021). Because spatially adaptive algorithms such as GRF rely heavily on nearby observations, evaluating them through a single validation strategy risks misleading conclusions. Therefore, to provide mechanistically robust diagnostics, this study implements a multi-scenario validation framework that contrasts RD-CV and SCV to rigorously evaluate both the local interpolation capacity and the regional transferability of the predictive models, thereby revealing the true drivers of SOC.

This study proposes a structured, reproducible, and transferable framework to optimise SOC mapping in complex, data-scarce environments by explicitly addressing the key design questions of when to sample, what to measure, and how to model. Our approach leverages five methodological advances. First, we stabilise the spectral baseline using 11-year median composites, ensuring that the model learns long-term soil–carbon patterns rather than transient climatic fluctuations. Second, we replace static imagery with seasonally targeted observations, evaluating three distinct phenological windows against an annual composite baseline to disentangle vegetation and soil signals, and examine how seasonal dynamics shape spectral–SOC relationships. Third, we address spatial non-stationarity by applying a GRF model, allowing predictive rules to adapt to local environmental conditions as opposed to imposing a single global relationship. Fourth, we employ a 3-tiered validation strategy: RD-CV to assess the model's true operational interpolation capability, strict SCV (k-means clustering and density-adaptive blocking) as a structural stress test to verify the stability of the target–predictor relationships, and concentric distance-based CV to bridge the scale gap between the localised interpolation of RD-CV and the extreme regional extrapolation of SCV. Furthermore, by treating spatial stress tests as a diagnostic tool to verify the authenticity of target–predictor relationships, we expect this approach to mitigate conventional mapping limitations and therefore directly inform national soil information systems and climate-smart land management policies. Finally, a key contribution is the integration of a multi-faceted interpretability analysis that couples local feature importance with global spatial attribution. This dual approach enables the identification of both the localised geographical relevance of SOC predictors and the global physical thresholds governing their seasonal variations. By linking prediction with this comprehensive interpretation, our framework provides a practical basis for targeted soil monitoring, more efficient field sampling, and improved decision-making for land management and carbon reporting.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

2 Materials and methods

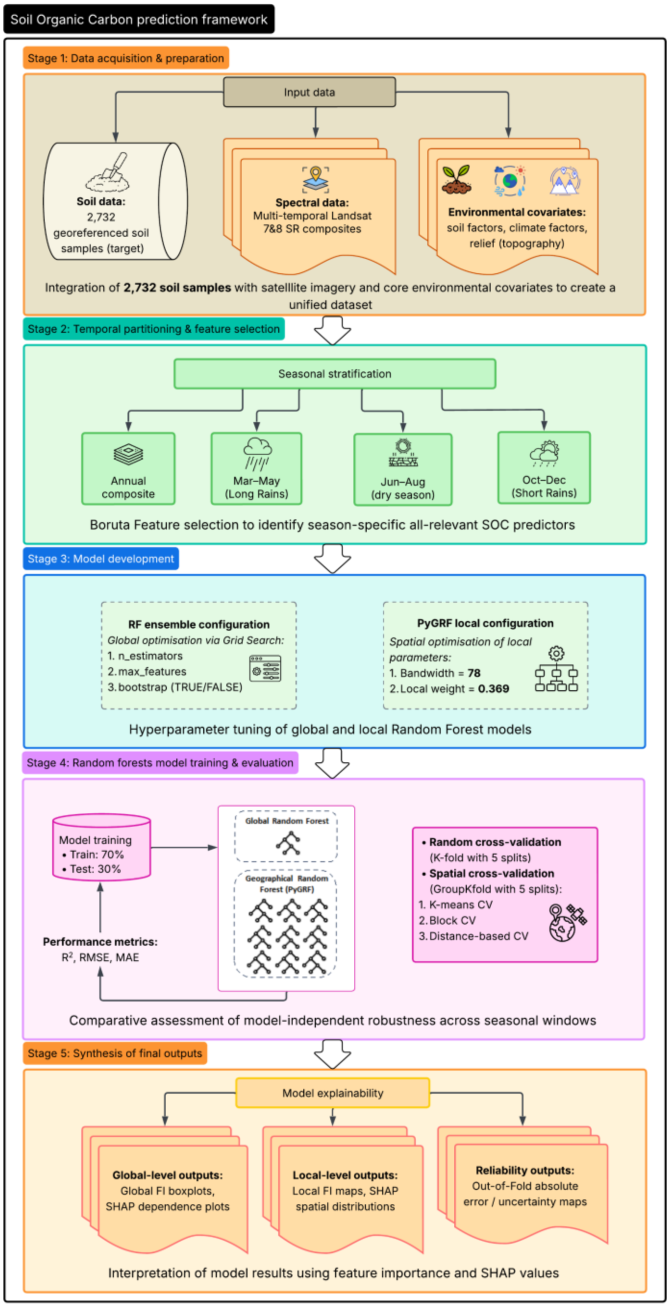


Figure 1 Five-stage conceptual framework for seasonal SOC prediction. **The framework integrates data acquisition, seasonal feature selection, global and geographical random forest modelling, robustness evaluation, and explainable spatial outputs including feature importance (FI), SHapley Additive exPlanations (SHAP) patterns, and reliability maps. Note: SR = Surface Reflectance; Mar–May = March–May; Jun–Aug = June–August; Oct–Dec = October–December; PyGRF = Geographical Random Forest.**

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

The conceptual workflow of the study in Fig. 1 illustrates how the different methodological components were linked into a single, reproducible framework for SOC prediction. The workflow begins with the integration of georeferenced soil observations, multi-temporal Landsat imagery, and ancillary environmental covariates, followed by seasonal partitioning of the predictor set to represent distinct phenological phases in the bimodal rainfall system of northern Tanzania. Feature selection was then performed independently for each temporal window using Boruta to retain only the most informative predictors for model development. These reduced predictor sets were subsequently used to train and compare a global RF model and a spatially explicit PyGRF model. Model performance and transferability were assessed through multiple cross-validation schemes, and the final stage of the workflow produced interpretable outputs, including feature importance, SHapley Additive exPlanations (SHAP) diagnostics, and spatial reliability maps, to explain both the dominant controls on SOC and the geographic stability of model predictions.

2.1 Study area

The study was conducted in the Shinyanga and Mwanza regions of northern Tanzania (Fig. 2), spanning −2.2° to −4.5° latitude and 31.5° to 34.5° longitude, covering approximately 60,000 km^2. The area experiences a semi-arid to sub-humid climate with a bimodal rainfall pattern, characterised by long rains from March to May and short rains from October to December, interspersed with a relatively dry June–August period (Yang et al., 2015). Mean annual precipitation ranges from 700 to 1400 mm, with considerable landscape and inter-annual variability. Areas around Lake Victoria such as Mwanza receive significantly higher rainfall than the semi-arid Shinyanga heartland, while the high inter-annual variability is driven by the El Niño-Southern Oscillation (ENSO) (Rao Kolusu et al., 2019).

During the study period (2008–2018), the region experienced notable extremes. These included the 2010–2011 La Niña drought, which was one of the most severe in the historical record (Lyon and Dewitt, 2012) and the 2015–2016 El Niño event, which resulted in above-normal precipitation across the Lake Victoria basin. Mean annual temperature is approximately 23°C, with limited seasonal variation (±3°C). Elevation ranges from around 1,000 to 1,700 m above sea level. Land use is dominated by mixed smallholder cropping, grazing lands, and scattered woodlands, reflecting typical agropastoral systems of northern Tanzania (Winowiecki et al., 2016). These agropastoral mosaics, together with strong rainfall gradients, create substantial spatial heterogeneity in SOC. Dominant soil types include Planosols, Cambisols, Vertisols, Phaeozems, Ferralsols and Solonetz, according to the WRB/FAO classification system, formed predominantly on weathered Precambrian basement rocks (IUSS Working Group WRB, 2022).

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

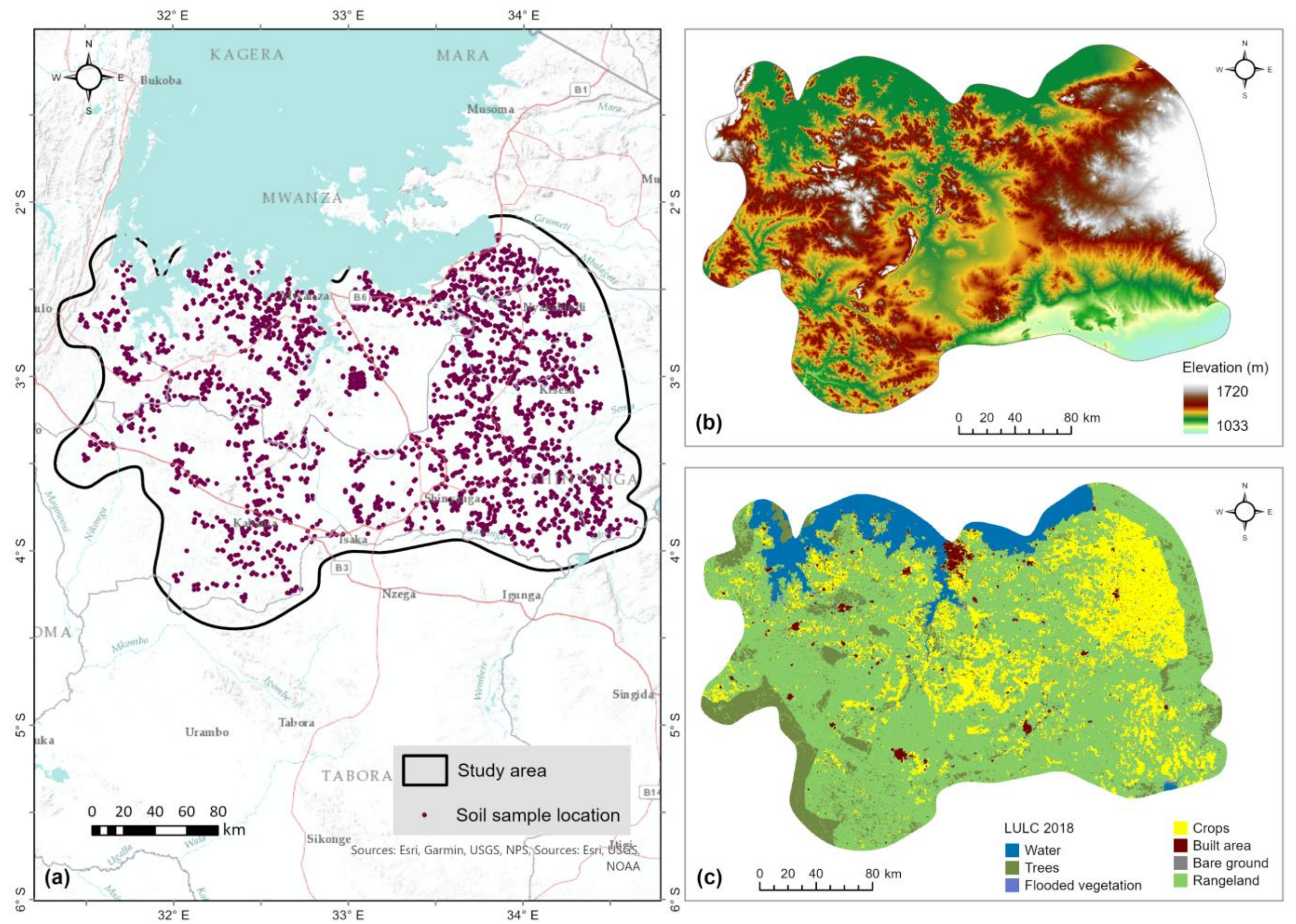


Figure 2 Spatial distribution of soil samples and environmental variables in northern Tanzania. **(a) Soil sample locations (n=2732) within Shinyanga and Mwanza regions. (b) Digital Elevation Model (DEM) showing topographic variation (range: 1,033 m to 1,720 m). (c) Land Use and Land Cover (LULC) classification for 2018, highlighting the dominance of agricultural lands (crops) and rangeland, derived from the Sentinel-2 10m global land cover dataset.** Note: **Coordinate system is set to WGS 1984 UTM Zone 36S.** Data sources: **Soil sample data from iSDA open soil dataset for Africa; Elevation data from NASA SRTM DEM 30 m; LULC map from Karra et al. (2021). Map lines delineate the study area and do not necessarily depict accepted national boundaries.**

2.2 When to sample

The initial pool of covariates was structured around the SCORPAN framework (McBratney et al., 2003) which relates soil properties to climate, organisms, relief, parent material, and spatial position. In this study, these components were represented by soil attributes (clay, sand, silt, and pH), climatic variables (annual precipitation, temperature seasonality, and soil water balance), organism-related covariates derived from multi-temporal Landsat imagery, vegetation indices, and evapotranspiration, as well as terrain variables such as elevation and slope. This structure ensured that the candidate predictor sets reflected both relatively stable soil–landscape controls and seasonally varying surface signals relevant to SOC prediction.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

2.2.1 Soil data

Soil samples were obtained from the iSDA open soil dataset for Africa (iSDA, 2023). The initial dataset comprised 5,544 georeferenced soil samples collected across the study area between 2008 and 2018. Samples were stratified into two depth intervals: 2,772 at 0–20 cm and 2,742 at 20–50 cm. SOC concentration (g kg^{-1}) was determined by laboratory analysis following standard protocols (FAO, 2023). The analysis focused on the surface soil layer (0–20 cm), as it is more responsive to recent land management, vegetation dynamics and surface spectral conditions than deeper soil layers (FAO, 2022).

To prepare the surface layer dataset for modelling, outliers were identified using a 3-standard deviation (3σ) threshold from the mean SOC concentration (mean = 10.71 g kg^{-1}, SD = 5.07 g kg^{-1}). Observations falling outside this range (below -4.50 g kg^{-1} or above 25.91 g kg^{-1}) were excluded to reduce the influence of extreme values on model calibration. This procedure removed 40 samples (1.4% of the dataset), including 22 extremely high SOC values (>26 g kg^{-1}). The remaining observations were deduplicated and spatially aligned with multi-temporal environmental covariates, resulting in a final cleaned dataset of 2,732 samples for modelling.

2.2.2 Remote sensing data

Surface reflectance (SR) data were obtained from the USGS Landsat Collection 2 Level 2 archive via GEE. To match the soil sampling period (2008–2018), a multi-sensor approach combined Landsat 7 ETM+ imagery for 2008–2013 and Landsat 8 OLI imagery for 2013–2018. To maintain spectral continuity between sensors, all images were standardised using the Collection 2 scaling factors (scale = 0.0000275; offset = −0.2). Cloud and cloud shadow contamination were removed using the QA_PIXEL bitmask (bits 3 and 5), excluding pixels affected by shadows or cirrus. This masking step was particularly essential given the persistently high cloud cover across the Lake Victoria basin and adjacent highlands. After cleaning and standardising the SR data, 11-year median composites were generated to represent long term stable-state landscape conditions. This temporal aggregation reduces the influence of interannual climatic anomalies, including ENSO-related droughts and unusually wet years, and also mitigates data gaps associated with the Landsat 7 SLC off failure. The median reducer was selected because it is robust to outliers and captures persistent spectral behaviour.

To reflect seasonal dynamics in vegetation–soil interactions, composites were produced for four temporal windows: March–May (Long Rains, characterised by peak biomass and canopy closure), June–August (dry season, characterised by senescent vegetation and maximum soil exposure), October–December (Short Rains, characterised by early-season greening and enhanced soil–moisture–carbon coupling), and an Annual composite representing the full-season climatological baseline.
For each composite, six spectral bands (blue, green, red, NIR, SWIR1, SWIR2) were extracted at 30 m resolution (Table 1). Three vegetation indices were also computed using standard formulations (Xue and Su, 2017) to characterise canopy structure and photosynthetic activity: the Normalised Difference Vegetation Index (NDVI), the Enhanced Vegetation Index (EVI), and

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

the Soil Adjusted Vegetation Index (SAVI). By using multi-year medians rather than single-date imagery, the resulting features were designed to capture persistent vegetation–soil spectral signatures while reducing the influence of transient moisture conditions, episodic disturbances, or short-term land use changes on SOC model performance.

Table 1. Selected Landsat 7&8 band parameters

Band Name	Landsat 7 (ETM+)	Landsat 8 (OLI)	Wavelength Range (μm)	Resolution	Description
Blue	SR_B1	SR_B2	0.45 – 0.52	30 m	Bathymetric mapping: soil/vegetation distinction.
Green	SR_B2	SR_B3	0.52 – 0.60	30 m	Peak vegetation reflectance; turbidity.
Red	SR_B3	SR_B4	0.63 – 0.69	30 m	Chlorophyll absorption: soil/vegetation contrast.
NIR	SR_B4	SR_B5	0.77 – 0.90	30 m	Biomass content; shorelines.
SWIR1	SR_B5	SR_B6	1.55 – 1.75	30 m	Moisture content in soil/vegetation; cloud penetration.
SWIR2	SR_B7	SR_B7	2.08 – 2.35	30 m	Improved moisture discrimination; mineralogy.

To complement the spectral data with hydrological dynamics, 10-day actual evapotranspiration and interception (ET) data from the FAO WaPOR database (version 2, decadal product L1_AETI_D, 2009–2018) were processed in a similar manner. Years with incomplete data were excluded, and the remaining time series were aggregated into multi-year seasonal and annual medians (FAO, 2020).

2.2.3 Ancillary data

Seasonal RS covariates were complemented with terrain, climate, soil property and land cover information to represent the broader environmental controls on SOC. Terrain attributes were derived from the NASA SRTM digital elevation model at 30 m resolution (Farr et al., 2007) accessed via OpenTopography. Bioclimatic variables consisting of annual precipitation, mean annual air temperature, precipitation seasonality, temperature seasonality, and isothermality, were obtained from the CHELSA Version 1.2 dataset and represent long-term climatic controls (Karger et al., 2017). Soil texture (clay, silt, sand) and pH were sourced from SoilGrids250m 2.0 (Poggio et al., 2021) at 250 m resolution and resampled to 30 m using bilinear interpolation.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Land cover was represented using a 2018 Sentinel-2-based classification, indicating broad land-use and vegetation types (Karra et al., 2021).

All raster covariates were projected to WGS84 / UTM Zone 36S (EPSG:32736) and resampled to 30 m to ensure spatial consistency across datasets. Landsat tiles were mosaicked using a median operator in overlap zones, and image pyramids were built using bilinear resampling to support efficient visualisation and analysis. Temperature variables originally stored as °C × 10 were divided by 10 to obtain degrees Celsius, whereas precipitation variables were retained in their original units (mm or mm quarter^{-1}). Evapotranspiration decadal values (mm day^{-1}) were multiplied by the scale factor 0.1 to obtain monthly equivalent values in mm month^{-1}, and soil texture variables were converted to percentage units. Soil pH values were constrained to a standard range (4–9) before extraction.

Spectral and ancillary covariates were extracted at the soil sample locations to create a unified modelling dataset. The resulting dataset contained 56 candidate predictors, including 24 seasonal spectral features (6 bands × 4 temporal windows), 12 vegetation features (3 indices × 4 temporal windows), topographical data, climate variables, and harmonised soil property covariates.

2.3 What to measure

To address the design question of what to measure, feature selection was performed using the Boruta algorithm in Python (Kursa and Rudnicki, 2010) to reduce the high dimensionality of the candidate predictor set and to improve interpretability. Boruta is a wrapper-based method built around random forests that identifies all relevant predictors by comparing the importance of observed variables with that of randomly permuted "shadow" features. Boruta was preferred over other dimensionality-reduction approaches, such as principal component analysis, because it preserves the original variables, thereby supporting eco-pedological interpretation and facilitating the translation of results into practical covariate choices for operational monitoring (Kuhn and Johnson, 2013; Kursa and Rudnicki, 2010). The method has also been applied successfully in various research fields, including in the prediction of SOC in China's croplands (Zhang et al., 2023), SOC stock estimation in agricultural lands in Vermont (USA) (Zeraatpisheh et al., 2025), and digital SOC mapping in Fengqui, China (Yan et al., 2025). Boruta was implemented using a Random Forest Regressor with 150 estimators as the base learner. To ensure a high degree of structural robustness, a strict selection threshold was applied (perc=100; maximum iterations=100), and only features that consistently outperformed their shadow counterparts at $\alpha = 0.05$ (binomial test) were retained.

The initial pool of 56 candidate predictors was grouped into four categories. The first comprised 16 core environmental covariates that remained constant across all four temporal windows: soil physical and chemical properties (clay, sand, silt and pH), bioclimatic variables (annual precipitation, mean annual air temperature, precipitation seasonality, temperature seasonality, and isothermality), relief features (elevation, slope, aspect, curvature, and Topographic Wetness Index). The rest of the variables representing organisms were stratified into the four temporal windows. The second group contained the multi-

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

temporal spectral bands (blue, green, red, NIR, SWIR1, and SWIR2) while the third consisted of vegetation indices (NDVI, EVI, and SAVI). The fourth group comprised hydrological proxies, represented here by evapotranspiration (ET), to capture additional seasonal moisture dynamics.

The selection process was carried out independently across the four distinct temporal windows. For each window, Boruta evaluated the specific covariates and classified them as either "selected" or "tentative". Only the strictly selected features were retained for subsequent RF and PyGRF model training. This parsimonious selection strategy enabled the models to focus on high-information covariates while reducing the risk of structural overfitting, which is common in models with broad, autocorrelated annual averages (Roberts et al., 2017). The full list of the selected covariates and abbreviations are provided in Table 2.

Table 2. Comprehensive list of environmental, spectral and topographic covariates according to the SCORPAN framework, representing the final selected covariates by Boruta across the four temporal windows.

SCORPAN factor	Variable	Abbreviation	Resolution (m)	Source
Soil (S)	Clay content (0–15cm)	clay_0_15cm	250	ISRIC SoilGrids250m
	Sand content (0–15cm)	Sand_0_15cm	250	ISRIC SoilGrids250m
	Silt content (0–15cm)	silt_0_15cm	250	ISRIC SoilGrids250m
	Soil pH in water (0–15cm)	pHwater_0_15cm	250	ISRIC SoilGrids250m
Climate (C)	Annual precipitation	annual_precipitation	1000	CHELSA V1.2
	Precipitation seasonality	prec_seasonality	1000	CHELSA V1.2
	Temperature seasonality	temp_seasonality	1000	CHELSA V1.2
	Soil water balance	soil_water_balance	1000	CHELSA V1.2
Organisms (O)	Blue band	blue_mar_may, blue_oct_dec, blue_annual	30	Landsat 7 & 8 (GEE)
	Green band	green_mar_may, green_oct_dec, green_annual	30	Landsat 7 & 8 (GEE)
	Red band	red_mar_may, red_oct_dec, red_annual	30	Landsat 7 & 8 (GEE)
	Near-Infrared	NIR_mar_may, NIR_june_aug, NIR_oct_dec, NIR_annual	30	Landsat 7 & 8 (GEE)

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

SCORPAN factor	Variable	Abbreviation	Resolution (m)	Source
	Shortwave Infrared 1	SWIR1_mar_may, SWIR1_june_aug	30	Landsat 7 & 8 (GEE)
	Shortwave Infrared 2	SWIR2_june_aug, SWIR2_oct_dec, SWIR2_annual	30	Landsat 7 & 8 (GEE)
	Normalised Difference Vegetation Index	NDVI_Mar_May, NDVI_Oct_Dec, NDVI_Annual	30	Calculated from Landsat composites
	Actual evapotranspiration	Mar_May_ET, Jun_Aug_ET, Oct_Dec_ET, Annual_ET	250	FAO WaPOR V2 (GEE)
Relief (R)	Elevation	elevation	30	SRTM
	Slope	slope	30	Calculated from elevation

2.4 How to model

To address the design question of how to model while accounting for spatial structure, SOC was predicted using two complementary approaches: a geographically weighted RF (PyGRF), which incorporates spatial autocorrelation through coordinate-based weighting (Sun et al., 2024), and a global RF model (Breiman, 2001) to assess algorithm-independent robustness. Model optimisation was carried out in two stages. The first stage focused on tuning the RF ensemble hyperparameters, shared by both the global RF and PyGRF models, to ensure the stability of the underlying RF learners. The second stage involved determining the spatially explicit hyperparameters unique to the PyGRF algorithm, namely bandwidth and local weight, which dictate how the model adaptively weights local observations to account for non-stationarity and spatial autocorrelation.

2.4.1 Random forests ensemble configuration

Hyperparameters for the RF base learner were tuned using a two-step approach. First, a manual sensitivity analysis was conducted to identify plausible ranges for the main parameters: the number of folds (k) used in k-fold cross-validation (n_splits: $\{3, 5, 10\}$); the number of trees ($n\_estimators$: $\{25, 40, 50, 100, 150, 200, 300\}$), and the number of maximum features at each split of a decision tree ($max\_features$: $\{1, 3, S/3\}$, where S is the total number of predictors). This exploratory step was used to assess the trade-off between computational efficiency, predictor regularisation, and model stability.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Second, an automated grid-search was used to optimise the final seasonal models within the reduced parameter space previously identified: $n\_estimators$: {100, 150, 200, 250, 300}; $max\_features$: {$\sqrt{S}$, $log_2 S$, $0.3S$, $0.5S$, $0.7S$, and integer values ranging from 1.5 to $2\sqrt{S}$}. Bootstrap sampling is used in RF to generate diverse training subsets, reducing overfitting and improving model generalisation while enabling out-of-bag error estimation (Breiman, 2001). Therefore, the use of bootstrap was also tested as a binary option (True/False). Each parameter combination was evaluated using 5-fold cross-validation, with the random state fixed at 42 to ensure reproducibility. The best performing parameter set was defined as the combination yielding the highest average R^2 score of the five folds. The grid-search was conducted independently for each temporal window, using only the Boruta-selected predictors for the corresponding period. The best performing parameter set was then used to train the corresponding global RF model and to configure the RF base learner embedded within PyGRF.

2.4.2 PyGRF local configuration

GRF combines predictions from the global model and the nearest local model through a weighting scheme that accounts for spatial proximity. The bandwidth hyperparameter defines the number of neighbours contributing to each local model. The local weight hyperparameter ranges from 0 to 1, assigning greater importance to the local model as values approach 1, and increasing the contribution of the global model as values approach 0. Sun et al. (2024) proposed using the Incremental Spatial Autocorrelation technique to determine the bandwidth and the local weight of GRF models. This theory-driven approach is not intended to maximise predictive performance, but rather to reduce the need for exhaustive hyperparameter tuning by providing an informed starting point. We could not implement that strategy because the global Moran's I index kept increasing with distance. Instead, we derived the bandwidth from Geographically Weighted Regression (GWR) diagnostics performed in ArcGIS Pro version 3.6.1 using the Golden Search optimisation algorithm. The Oct–Dec model was selected as the reference period, as the manual sensitivity analysis of hyperparameters consistently indicated superior model fit during this season. The resulting bandwidth of 78 neighbours was adopted as a standardised parameter across all temporal windows.

The local weight hyperparameter was derived from the Global Moran's I of the target variable, calculated in ArcGIS Pro version 3.6.1 using an inverse distance squared weighting scheme, where Euclidean coincident points and observations separated by less than one unit of distance were assigned a weight of 1. This procedure yielded a theory-informed local weight value of 0.369, which was integrated into the PyGRF prediction function to balance local and global prediction contributions. The final model parameters are reported in Table 3. To isolate the effect of seasonal spectral information on SOC prediction, a static bandwidth of 78 neighbours and a local weight of 0.369 was applied to all PyGRF models. This approach follows the classical GWR framework of comparing models at a common geographic scale (Brunsdon et al., 1996), ensuring that differences in predictive performance reflect seasonal variation in covariate sensitivity rather than changes in spatial model configuration, enabling a robust comparison across the bimodal rainfall cycle.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Table 3: Final hyperparameters for each model; Mar–May = March–May; Jun–Aug = June–August; Oct–Dec = October–December.

<table>
<tr><th>Model</th><th>number of predictors</th><th>n_estimators</th><th>max_features</th><th>bootstrap</th><th>band_width</th><th>local_weight</th></tr>
<tr><td>Mar–May</td><td>17</td><td>250</td><td>log2</td><td>True</td><td rowspan="4">78</td><td rowspan="4">0.369</td></tr>
<tr><td>Jun–Aug</td><td>13</td><td>300</td><td>0.3</td><td>True</td></tr>
<tr><td>Oct–Dec</td><td>13</td><td>200</td><td>0.3</td><td>True</td></tr>
<tr><td>Annual</td><td>16</td><td>300</td><td>0.3</td><td>True</td></tr>
</table>

2.4.3 Model evaluation, spatial reliability, and uncertainty assessment

An Average Nearest Neighbour (ANN) analysis confirmed a significantly clustered spatial distribution (ANN Ratio = 0.56, z = –44.24, $p < 0.001$), with an observed mean distance between sampling points of 1.47 km against an expected random distance of 2.63 km. Evaluating ML models on such clustered data using RD-CV introduces severe spatial leakage (de Bruin et al., 2022; Wang et al., 2025). When the objective moves beyond simply mapping SOC to identifying the true drivers of SOC spatial variability, RD-CV alone is insufficient. This is because it tends to mask whether the model has learned genuine biophysical rules or merely overfitted to spatial proximity (Roberts et al., 2017). To address this limitation, three distinct cross-validation (CV) strategies were implemented to evaluate model performance under different spatial dependency structures.

Scenario A – RD-CV for operational interpolation: To assess the model's practical capability to interpolate SOC within the bounds of the sampled clusters, standard k-fold RD-CV was applied. Because this method randomly assigns observations to folds regardless of their geographic coordinates, training and test samples often share similar spaces. This explicitly allows the PyGRF algorithm to fully exploit the local spatial dependencies present in the clustered dataset, providing a realistic estimate of operational map accuracy for local gap-filling where nearby observations are inherently available.

Scenario B – SCV for structural stress testing: To isolate genuine SOC drivers from spatial overfitting, the models were subjected to severe environmental extrapolation using two strict SCV partitioning methods: k-means clustering-based CV (K-means CV) and density-adaptive block CV (Block CV). First, Block CV was implemented using quantile-based discretisation applied independently to both the X (Easting) and Y (Northing) coordinate arrays. Rather than imposing a rigid, equidistant geographic grid, this approach sliced the coordinate axes into quantiles based on sample density (Fig. A1). The intersections of these X and Y quantiles formed distinct spatial blocks which were entirely withheld during testing via group k-fold, effectively maximising prediction distance and severing local spatial autocorrelation (Brenning, 2012; Roberts et al., 2017; Stock, 2025). Second, an unsupervised k-means clustering algorithm (Hartigan and Wong, 1979) was fit exclusively to the spatial coordinates (Easting and Northing) of the dataset, aggregating the samples into five distinct, spatially contiguous clusters (Fig. A2). These clusters were subsequently withheld iteratively using group k-fold. Rather than estimating operational map accuracy, these geometries were deployed purely as diagnostic stress tests to verify that the identified SOC drivers reflect genuine causal relationships rather than spatial overfitting.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Scenario C – Distance-based CV (Db-CV) for boundary extrapolation: To bridge the scale gap between the localised interpolation of RD-CV and the extreme regional extrapolation of SCV, we implemented a specific geometric configuration of distance-based CV. Rather than estimating error as a continuous function of distance as applied by Brenning (2023), the dataset was partitioned into five discrete, concentric folds radiating from the spatial centroid of the study area out to its geographical periphery (Fig. A3). This spatial arrangement was explicitly designed to evaluate the model's structural resilience across a critical operational transition: moving from multidirectional interpolation to directional boundary extrapolation, thereby simulating realistic mapping conditions. When evaluating the interior folds, the algorithm benefits from training data located in multiple surrounding directions, reflecting a stable local interpolation environment. Conversely, when evaluating the outermost peripheral folds, the model must predict outside the spatial domain of the training data, forcing it to extrapolate its learned biophysical rules into unconstrained geographic margins.

All models were evaluated using 5-fold CV and predictive performance was assessed by comparing the coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE). The spatial reliability of the final models was assessed by mapping the out-of-fold (OOF) absolute errors derived from random and distance-based cross-validation predictions for both global and geographical RF. Each observation was excluded from model training in at least one fold and subsequently predicted as part of the hold-out set. Model residuals were computed for these hold-out observations as the difference between observed and predicted SOC values, thereby representing out-of-sample prediction errors. These OOF residuals provide an unbiased estimate of model performance and were used to quantify model uncertainty. The resulting errors were then visualised geographically to assess the spatial distribution of prediction accuracy and identify areas of higher uncertainty. This approach prevents the optimistic bias associated with resubstitution (Kuhn and Johnson, 2013; Roberts et al., 2017).

2.5 Model explainability and spatial attribution of SOC drivers

A multi-faceted approach was employed to interpret the eco-pedological relationships driving SOC predictions in northern Tanzania. From PyGRF, boxplots, ranked by median importance value, were used to summarise the global consistency and variability of each feature's influence across different CV folds. Local FI values were extracted for key variables and subsequently mapped to visualise spatial non-stationarity in SOC–predictor relationships.

Due to the extreme computational demands and potential instability of deriving SHAP values for thousands of local PyGRF sub-models (Wadoux et al., 2023), we applied SHAP analysis to the global RF model. To elucidate the biophysical thresholds of the global relationships, SHAP dependence plots were generated to extract the functional forms of each covariate's relationship with predicted SOC across the entire study area.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

To assess the spatial expression of these global driver effects, we projected per-sample SHAP values onto geographic coordinates, producing a spatial representation of SHAP values that visualise both the strength and directional impact (positive or negative) of each covariate on SOC predictions. Crucially, while local FI maps quantify the absolute magnitude of a predictor's influence at a given location, spatial SHAP maps reveal the sign and geographic concentration of that influence. Together, they provide a comprehensive, spatially explicit understanding of SOC driver dynamics across different scales within the agroecosystem, ensuring the empirical functional forms match the biophysical reality of the environment.

3 Results

Confirming the central hypothesis of this study, the analysis revealed that SOC spatial variability is most effectively detected when vegetation–soil coupling is strongest. Under RD-CV, while the Annual composite yielded the highest absolute accuracy (R^2 = 0.474; 17 predictors), the October–December 'Short Rains' model was considered optimal as it achieved a highly comparable performance (R^2 = 0.471) with a more parsimonious feature set (13). Comparing validation schemes demonstrated that RD-CV systematically inflated performance estimates, with an average inflation of 19% across all models (R^2 = 0.444 – 0.471) (Fig. 3). When subjected to the demanding SCV strategies (K-means CV, Db-CV, and Block CV) to mitigate spatial overfitting, the October–December window definitively produced the most robust model, maintaining the highest predictive accuracy across all spatial stress tests (e.g., R^2 = 0.405 under Db-CV) (Table 4).

These findings are presented below in two primary phases: first, an evaluation of model consistency and validation robustness across the temporal windows, followed by the seasonal information yield of the predictors and their biophysical interpretations.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

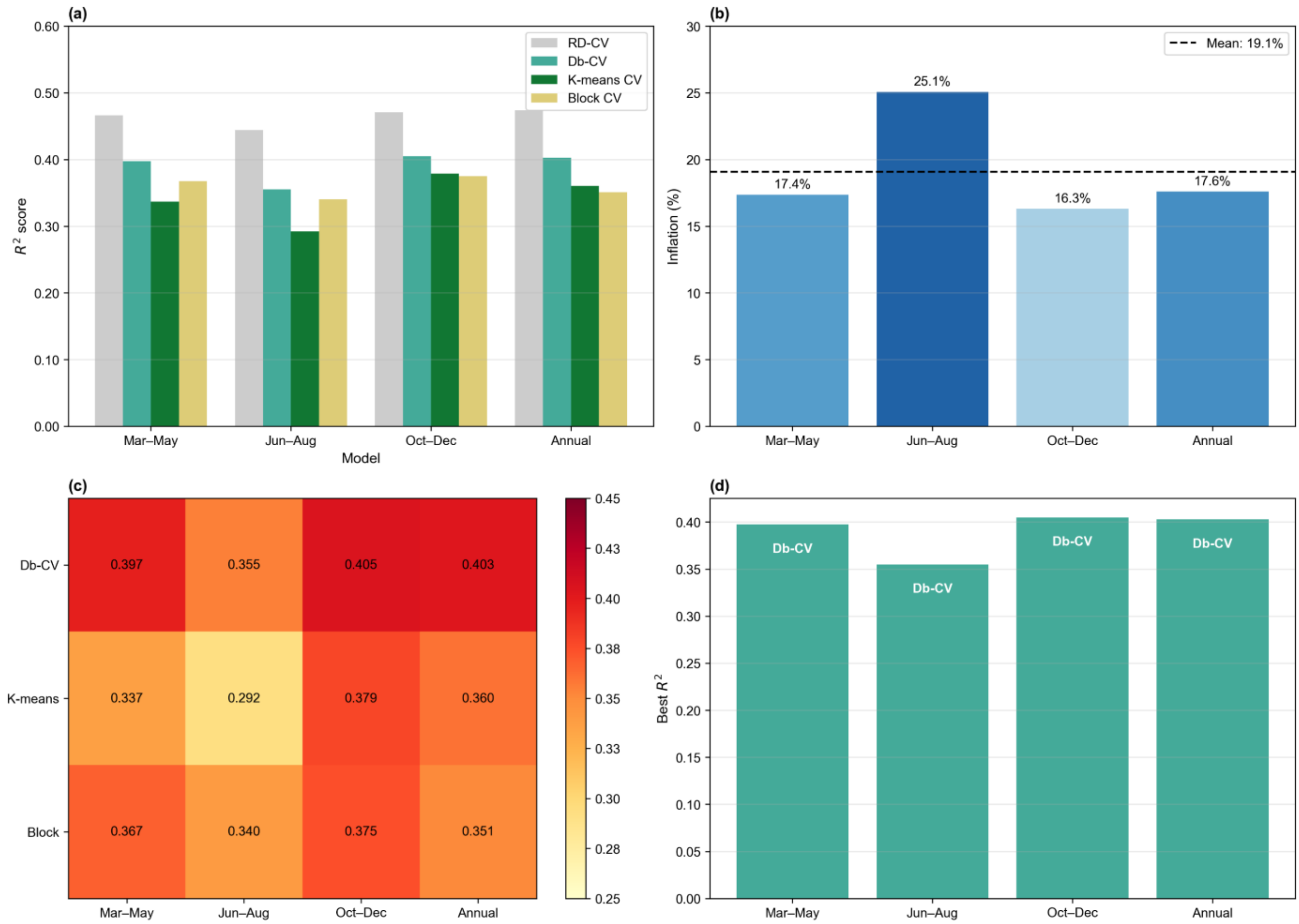


Figure 3 Visualisation of predictive performance inflation caused by spatial autocorrelation. **The charts contrast standard random cross-validation (RD-CV) metrics against spatial cross-validation methods (concentric distance-based CV [Db-CV], k-means clustering-based CV [K-means CV] and density-adaptive block CV [Block CV]), quantifying the optimistic bias that occurs when spatial proximity between training and test samples is not severed across the temporal models. (a) R² score by validation method; (b) Performance inflation from spatial autocorrelation; (c) Heatmap comparing SCV methods and (d) Best performing SCV method by model.**

3.1 Consistency of seasonal SOC observability

3.1.1 Comparison of PyGRF and global RF

To verify that seasonal differences in SOC observability reflect genuine model-independent environmental signals, the spatially explicit PyGRF model was evaluated against a global RF baseline (Table 4). While PyGRF was rigorously tested against all CV scenarios, the RF model was evaluated specifically under RD-CV and Db-CV. This comparison ensures that the seasonal trends are robust and not simply artefacts of PyGRF's local weighting structure.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Table 4. Performance metrics between global RF and PyGRF models (number of predictors in brackets) for all the cross-validation strategies.

	PyGRF models					Global RF models			
Metric	**Cross-validation strategy**	**Mar-May (17)**	**Jun-Aug (13)**	**Oct-Dec (13)**	**Annual (16)**	**Mar-May (17)**	**Jun-Aug (13)**	**Oct-Dec (13)**	**Annual (16)**
R^2	RD-CV	0.466	0.444	0.471	0.474	0.458	0.438	0.462	0.466
	Db-CV	0.397	0.355	**0.405**	0.403	0.394	0.357	0.399	0.398
	K-means CV	0.337	**0.292**	0.379	0.360	–	–	–	–
	Block CV	0.367	0.340	0.375	0.351	–	–	–	–
RMSE	RD-CV	3.270	3.338	3.256	3.248	3.295	3.355	3.284	3.272
	Db-CV	3.475	3.595	3.453	3.460	3.485	3.588	3.470	3.474
	K-means CV	3.645	3.765	3.528	3.580	–	–	–	–
	Block CV	3.561	3.636	3.539	3.607	–	–	–	–
MAE	RD-CV	2.443	2.513	2.463	2.431	2.477	2.538	2.500	2.457
	Db-CV	2.688	2.811	2.692	2.691	2.694	2.807	2.707	2.702
	K-means CV	2.828	2.973	2.745	2.783	–	–	–	–
	Block CV	2.761	2.845	2.756	2.811	–	–	–	–

Note: En dashes (–) denote SCV methods that were not used for the global RF model.

Regardless of whether spatial non-stationarity was explicitly accounted for, the October–December window consistently provided the most robust and transferable signal for SOC prediction. Although the spatially aware PyGRF algorithm demonstrated a slight accuracy advantage over the global RF baseline under both standard (RD-CV; $R^2 = 0.471$ vs. 0.462) and spatial cross-validation (Db-CV; $R^2 = 0.405$ vs. 0.399), both ML architectures confirmed the exact same seasonal hierarchy. Conversely, the June–August 'dry season' window yielded the poorest predictive performance across all tests.

3.1.2 Robustness under strict spatial cross-validation

To ensure the seasonal hierarchy was not an artefact of spatial autocorrelation, the PyGRF models were subjected to two rigorous spatial stress tests: K-means CV and Block CV. While standard RD-CV yielded higher absolute accuracies, these two spatial partitioning strategies confirmed the Short Rains season as the most robust temporal window, achieving a higher accuracy under K-means CV ($R^2 = 0.379$; RMSE = 3.528 g kg^{-1}; MAE = 2.745 g kg^{-1}) (Table 4). In contrast, enforcing spatial independence severely degraded the Jun–Aug model, whose generalisation capacity dropped significantly under the strict autocorrelation constraints of K-means CV ($R^2 = 0.292$; RMSE = 3.765 g kg^{-1}; MAE = 2.973 g kg^{-1}).

Beyond global performance, the fold-wise distributions of R^2 scores across the three SCV methods reveal important differences in how the seasonal models react to extrapolation (Fig. 4). While all models achieved relatively higher and more stable

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

accuracy under Db-CV (which preserves short prediction distances for interpolation), the strict geographic separation of K-means CV and Block CV immediately exposed the structural vulnerabilities of the Mar–May and Annual composite models.

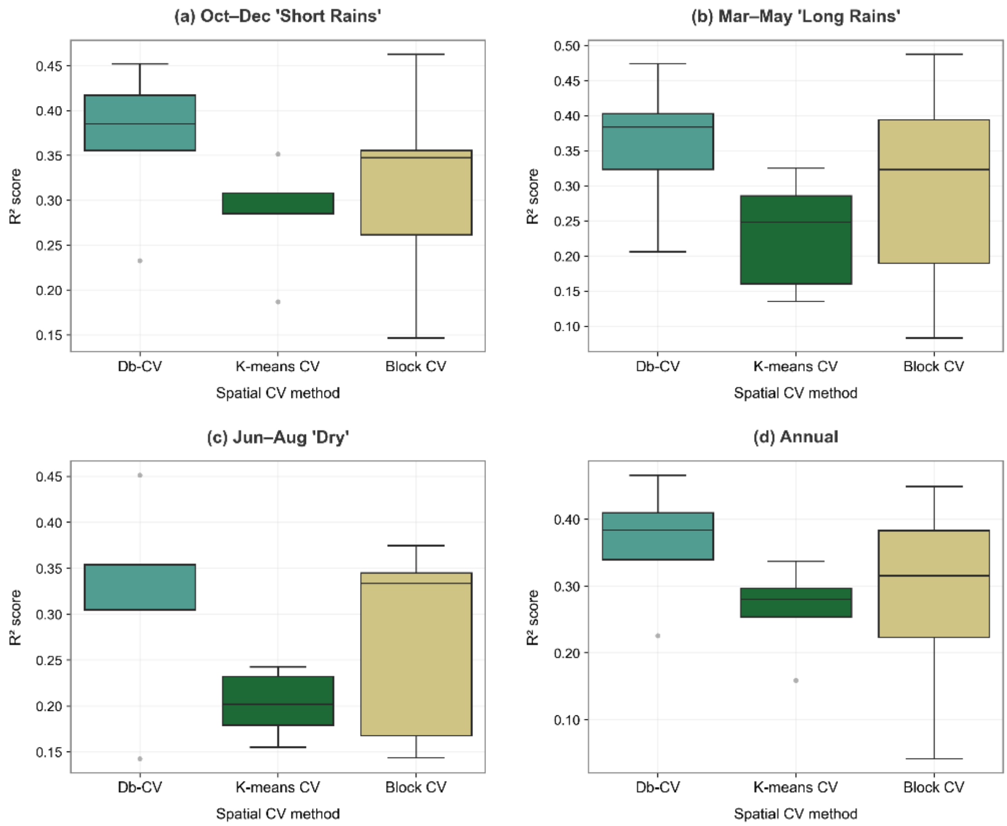


Figure 4 Fold-wise predictive variability. **Boxplots illustrating the variability and spread of R^2 scores across the different spatial cross-validation folds for each temporal window. The data reflects the absolute performance of each of the five distinct test folds, highlighting spatial variance rather than the aggregated model average.**

When phenological variance was collapsed into an annual composite, the Annual model's performance exhibited substantial volatility, especially under Block CV, with one of the folds experiencing a significant crash ($R^2 = 0.041$) as visualised in Fig. 4(d). A similarly severe crash is observed during the Long Rains (Mar–May), where the lowest fold plummets to an R^2 of 0.083 (Fig. 4b). The Oct–Dec model not only achieves the highest median performance but also exhibits one of the narrowest interquartile ranges and the fewest low-performing folds across the SCV strategies. The Jun–Aug (dry season) model avoided these near-zero crashes but exhibited highly skewed and erratic fold-wise performance. Conversely, the more stable performance under strict SCV in the Oct–Dec model suggests that the vegetation–soil coupling during the Short Rains window provides a better global signal for SOC prediction.

The model-independent nature of these results is further supported by the global FI rankings as illustrated in Fig. B1. Both PyGRF and RF models identified mostly identical primary drivers for SOC prediction in every season. For the Oct–Dec model,

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

both algorithms identified NIR_oct_dec as the single most powerful predictor, with a high relative importance of 17.1%–17.3% in PyGRF and 16.9%–17.9% in global RF. Additionally, the choice of SCV method did not significantly affect the global FI rankings (Fig. B2).

3.1.3 Spatial stability of prediction errors

While PyGRF subtly redistributes the error intensity, and Db-CV shows more pronounced darker red clusters, spatial error patterns were broadly consistent. High OOF absolute errors consistently clustered in the same geographical regions across all configurations (Fig. 5), indicating that certain locations remain inherently challenging to predict regardless of the model architecture (PyGRF vs global RF), or validation strategy (Db-CV vs RD-CV).

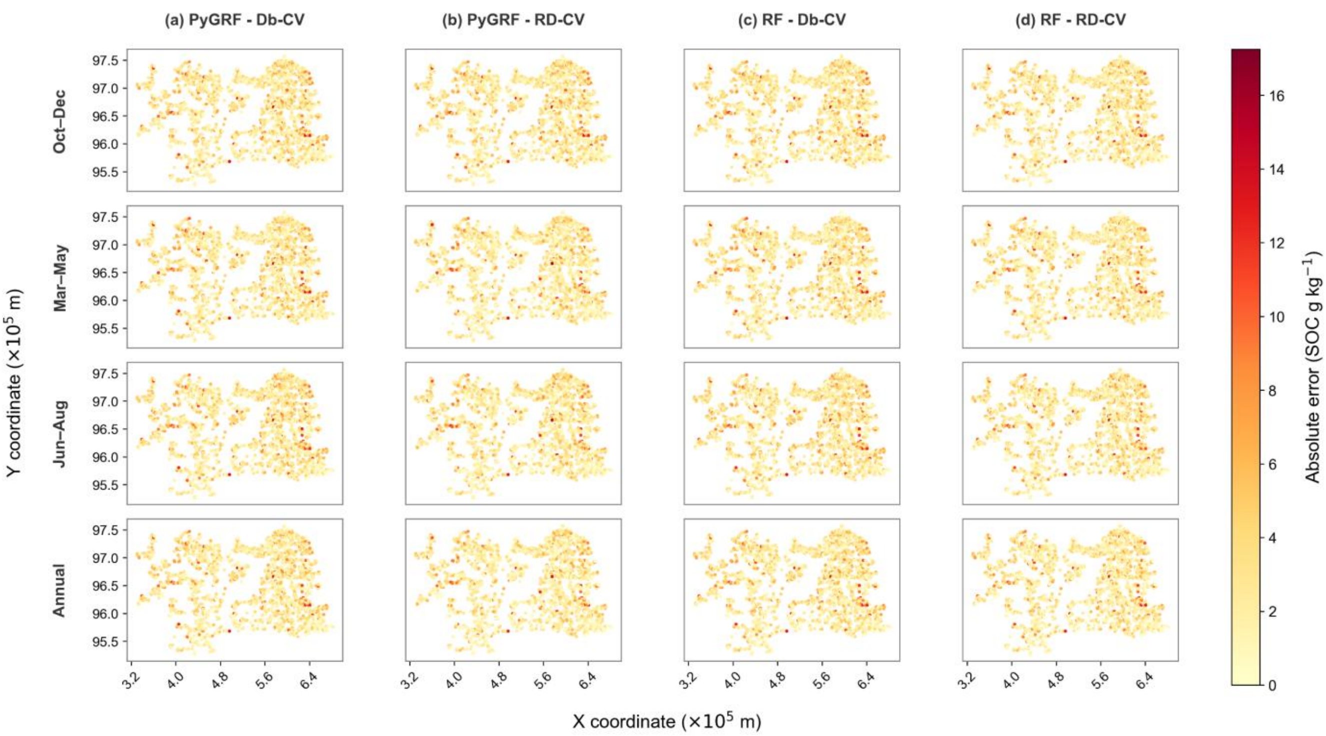


Figure 5 Spatial distribution of out-of-fold absolute errors (SOC content in grams per kilogram). **Rows represent the temporal windows; columns contrast PyGRF (geographical) and RF (global) under random (RD-CV) and spatial cross-validation (Db-CV).**

Residual analysis using the Global Moran's I further supports the conclusion that the October–December window provides the most spatially stable representation of SOC relationships in northern Tanzania (Table 5). While the residuals for the Mar–May, Jun–Aug, and Annual models all exhibited statistically significant spatial autocorrelation ($p < 0.05$, indicating dispersed error patterns), the Oct–Dec model alone approached spatial randomness (Global Moran's I = –0.016, p = 0.253).

Table 5. Residual analysis using Global Moran's I.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Season	Moran's Index	Z-score	p-value	Spatial Pattern
Mar-May	-0.029	-2.104	0.035	Dispersed
Jun-Aug	-0.036	-2.593	0.010	Dispersed
Oct-Dec	-0.016	-1.143	0.253	Random
Annual	-0.035	-2.568	0.010	Dispersed

3.2 Seasonal information yield

470 Boruta feature selection reduced the initial pool of 56 candidate predictors to 13–17 confirmed predictors per temporal window, representing a reduction of up to 50% in dimensionality without compromising predictive performance. Across all windows, vegetation indices specifically designed to suppress the soil background signal, such as EVI and SAVI, were systematically rejected, whereas NDVI was retained selectively (Table 6). For conciseness, only the top six most important variables, for each temporal window, are presented in the main text. The comprehensive visualisations for all variables are available in the Supplementary Material [Figs. S1–S12].

Table 6. Percentage contribution of individual spectral variables (raw optical bands and NDVI) to the global feature importance of the predictive models across the four temporal windows (based on Distance-based CV ranking).

Spectral variable	Mar–May (Long Rains)	Jun–Aug (dry season)	Oct–Dec (Short Rains)	Annual composite
NIR band	6.4%	14.8%	17.1%	15.0%
Red band	10.1%	–	8.0%	8.0%
Green band	8.0%	–	11.2%	8.3%
Blue band	6.4%	–	6.6%	6.0%
SWIR1 band	7.0%	9.9%	–	–
SWIR2 band	–	9.3%	5.6%	6.1%
NDVI	5.8%	–	5.3%	4.8%
Total spectral importance	43.7%	34.0%	53.8%	48.1%

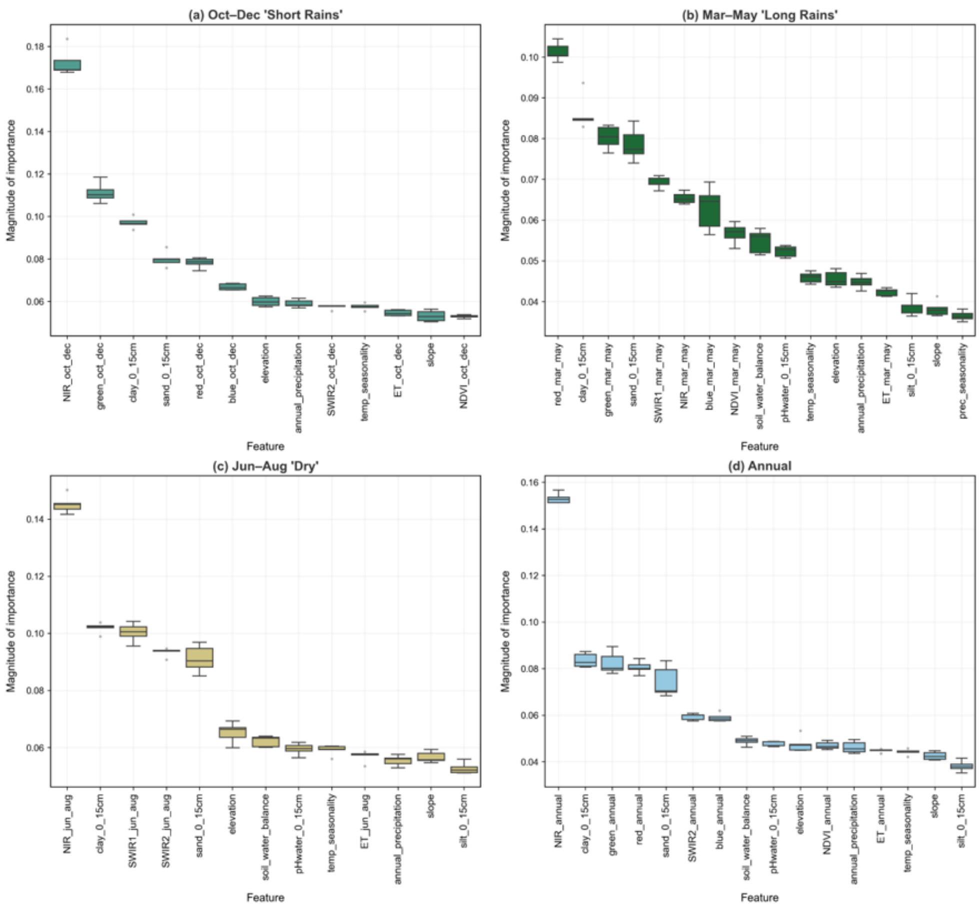
Note: En dashes (–) denote variables that were not selected by Boruta for that specific seasonal period.

Evaluating the global FI revealed a clear decoupling between stable physical covariates and highly dynamic spectral drivers as seen in Fig. 6. Static soil properties, notably topsoil clay and sand content, maintained a structurally stable baseline, consistently ranking among the top global predictors across all seasons. In contrast, the relative predictive utility of the spectral bands fluctuated dramatically depending on the phenological window (Table 6). Total spectral importance was lowest during the June–August 'dry season' (34.0%) and peaked significantly during the optimal October–December 'Short Rains' (53.8%).

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

During this October–December window, the NIR band exhibited overwhelming dominance, capturing 17.1% of the total FI. Conversely, during the March–May 'Long Rains', NIR importance dropped to 6.4%, with the red band becoming the leading spectral predictor (10.1%). During the June–August window, the algorithm increased its reliance on the short-wave infrared bands (SWIR1 at 9.9%; SWIR2 at 9.3%), alongside a NIR importance of 14.8%. When signals were temporally averaged into the Annual composite, the model prioritised the NIR (15%), green (8.3%), and red (8%) bands, while diminishing the bare-soil SWIR signals associated with the dry season (SWIR2: 6.1%).

Finally, analysing the cumulative FI for the top six covariates revealed varying degrees of signal concentration. The Oct–Dec model exhibited the highest concentration of predictive power, with the top six features accounting for 60.6% of the total global importance, predominantly driven by spectral bands (NIR, green, red, and blue; 42.9%). The Jun–Aug model showed a similarly concentrated signal (59.6%) but marked a distinct shift toward the physical landscape: static soil properties (clay, sand) and elevation rose to contribute 25.8%, while the share of spectral bands (NIR, SWIR1, SWIR2) fell to 33.9%. The Annual composite model yielded a moderate distribution (53.0%), whereas the Mar–May model displayed the most diffuse predictive signal, with the top six covariates capturing only 47.9% of the overall FI.



https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Figure 6 Global feature importance (FI) distributions across the four temporal PyGRF models. **The boxplots represent the statistical distribution and stability of global FI scores across validation iterations, highlighting the NIR band and clay content as the dominant predictors.** Note: **temp_seasonality = Temperature Seasonality; ET = evapotranspiration.**

3.2.1 Functional form of SOC–predictor relationships

To explain how the individual predictors interact with SOC globally, SHAP dependence plots were extracted for each temporal window. While these mechanistic visualisations are derived from the global RF, the variables within each six-panel figure are presented in descending order based on their PyGRF Db-CV feature importance ranking.

a) October–December (Short Rains)

During the Short Rains, the dominant spectral predictors exhibit strongly non-linear, threshold-type responses as illustrated in Fig. 7. Specifically, NIR_oct_dec exhibits a steep negative threshold effect: at lower reflectance values (below approximately 0.23), SHAP values are positive, but once reflectance exceeds approximately 0.23, contributions become negative. Similar steep negative relationships are observed for the green and red bands. In contrast, static soil properties display stable and physically interpretable relationships beneath the seasonal signal. The clay_0_15cm feature demonstrates a steady positive contribution to SOC predictions as clay content increases. Conversely, sand_0_15cm exhibits a marked negative threshold effect, with its contribution dropping once sand content exceeds approximately 60%, reflecting the diminished carbon-carrying capacity of highly porous soils.

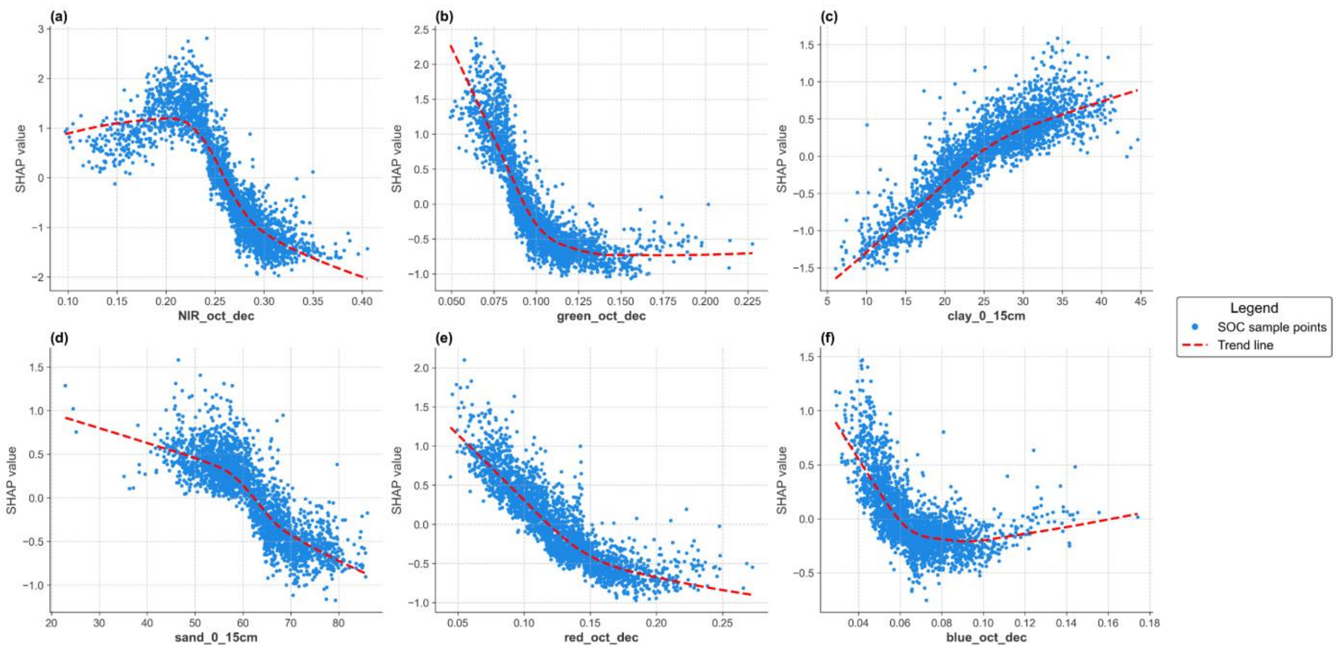


Figure 7 SHAP dependence plots for the optimal October–December window. **These plots illustrate the non-linear marginal effects of the top six environmental covariates (NIR band, green band, clay [0-15cm], sand [0-15cm], red and blue bands) on SOC predictions for the Oct–Dec model.**

b) June–August (dry)

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

During the dry season, the dependence plots (Fig. C1) show a clear shift in model behaviour. Although NIR_jun_aug still exhibits a negative threshold effect, the SWIR bands emerge as dominant predictors. SWIR1_jun_aug shows a distinct non-linear decline in its contribution once reflectance exceeds approximately 0.33, while SWIR2_jun_aug follows a more linear negative trend. Elevation also plays a prominent role during this season, with positive contributions increasing sharply above 1300 meters. Importantly, the static soil properties (clay_0_15cm and sand_0_15cm) maintain their fundamental positive and negative relationships, confirming that the underlying pedological rules remain stable even as the dominant spectral signals fluctuate.

c) March–May (Long Rains)

During the Long Rains, we observe the dominant spectral relationships shifting from NIR towards the visible bands as illustrated in Fig. C2. In this window, red_mar_may and green_mar_may exhibit the steepest non-linear, negative threshold effects, while NIR_mar_may remains important but secondary. For instance, the contribution of red_mar_may is highly positive at very low reflectance values (below approximately 0.08) but drops sharply into negative territory as reflectance increases. Although NIR_mar_may still displays a sharp negative threshold effect when reflectance exceeds approximately 0.27, its overall influence is superseded by the visible bands during this season. As in the previous seasons, the static soil variables preserve the same stable functional forms.

d) Annual composite

The Annual model further illustrates the limitations of temporal averaging. The dependence plots in Fig. C3 reveal that temporally aggregating satellite data across the year effectively dilutes the distinct bare-soil signals of the dry season, thereby forcing the model to primarily reproduce the vegetation-dominated relationships observed in the wet seasons. The NIR_annual curve closely resembles the steep negative threshold effect seen in the Oct–Dec model, with contributions dropping sharply into negative territory once reflectance exceeds approximately 0.24. The Annual model also relies on the steep negative curves of the visible red and green bands, closely mirroring the relationships found in the March–May window. Importantly, the underlying static soil properties remain unchanged.

3.2.2 Spatial distribution of the main SOC drivers

Having established the global, averaged relationship between each covariate and predicted SOC in Sect. 3.2.1, we next examined their spatial expression and local variability for each temporal window. The PyGRF local FI maps depict how the model addresses spatial non-stationarity by dynamically adapting its predictive weights to specific eco-pedological zones. Complementing this, the spatial SHAP maps reveal the geographical concentration of the global drivers, verifying that these localised adaptations remain grounded in valid eco-pedological processes.

Static soil properties (clay_0_15cm, sand_0_15cm), maintain consistent geographic footprints through time. In contrast, the spectral hotspots shift significantly across seasons: the annual and wet-season models rely primarily on vegetation-sensitive signals (NIR in the west and central regions for Oct–Dec and Annual; visible bands in the northeast for Mar–May), whereas

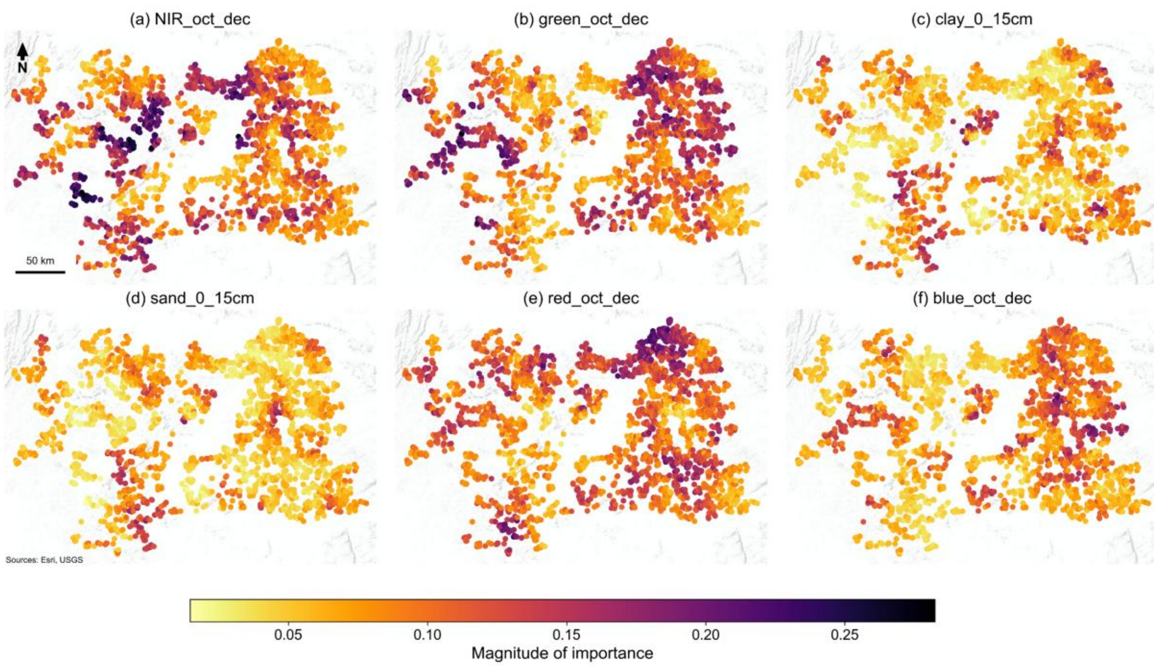
https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

the Jun–Aug model shifts toward SWIR bands and elevation. When these non-stationary relationships are compared with the OOF absolute error maps (Fig. 5), clustered error hotspots, particularly in the eastern and northeastern regions (mostly cropland; Fig. 2c), consistently coincide with areas where these primary drivers exert their most extreme local contributions. This demonstrates that the specific regions characterised by highly intense, localised biophysical relationships are also areas where model generalisation is most difficult.

a) October–December (Short Rains)

During the Short Rains, the spatial distribution of local FI and corresponding SHAP values demonstrate distinct regional dependencies. For example, the NIR_oct_dec predictor exhibits its highest local importance in the western and northwestern clusters (Fig. 8a), where it strongly drives global predicted SOC lower as visualised in Fig. 9a. However, the same variable drives predictions higher in the east where its local importance is relatively lower. The visible bands (green_oct_dec and red_oct_dec) present a different spatial pattern. They dominate model predictions in the eastern and northeastern regions (Fig. 8b, 8e) where there is more cropland (Fig. 2c). Despite this high local importance, their global positive contributions to SOC are concentrated strictly in the northeastern and northwestern regions (Fig. 9b, 9e), whereas their influence becomes slightly negative in the south.

Unlike the spectral bands, physical soil properties like clay_0_15cm maintain a more diffuse, moderate importance across the entire landscape (Fig. 8c). However, its global effect on SOC shifts geographically: clay acts as a negative influence near the southwestern boundary, but transitions into a positive driver in the eastern regions (Fig. 9c) aligning with areas of higher clay content (Fig. D1).



https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Figure 8 Spatial distribution of local feature importance for the optimal Oct–Dec model. **The maps display the absolute magnitude of local importance for the top six covariates: (a) NIR band, (b) green band, (c) clay [0–15cm], (d) sand [0–15cm], (e) red, and (f) blue band. Darker hues represent geographic regions where the specified variable exerts a stronger absolute influence on the local SOC predictions, regardless of the effect's direction.**

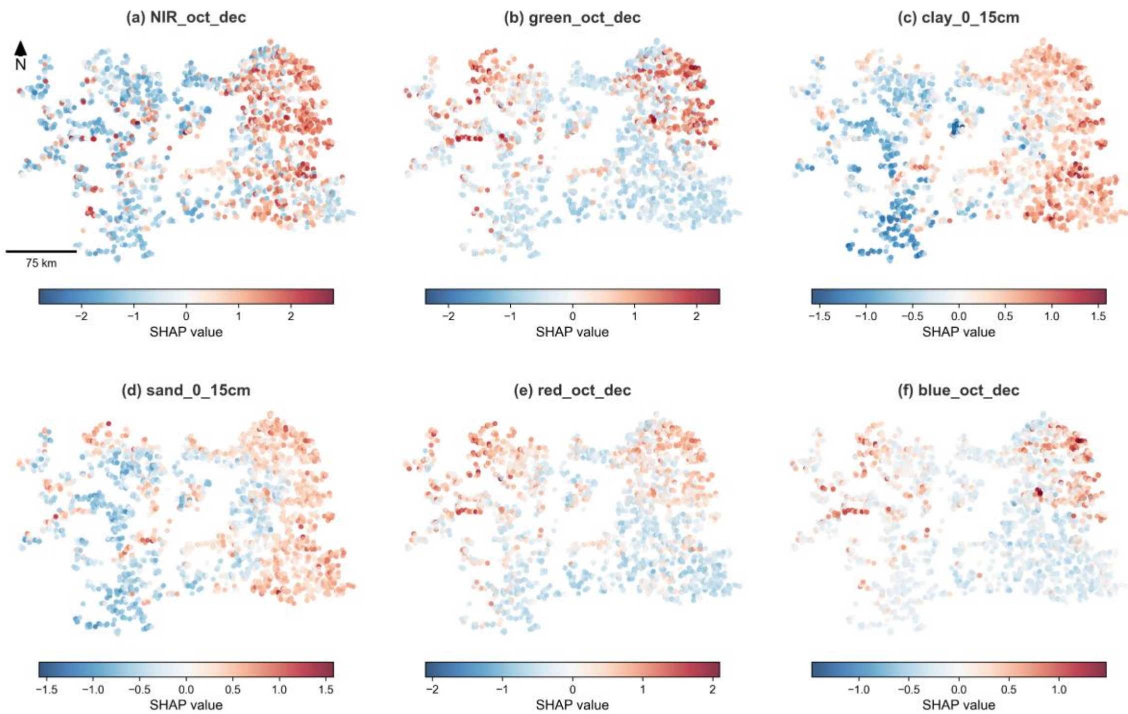


Figure 9 Spatial distribution of SHAP values for the optimal Oct–Dec model. **The diverging maps illustrate the directional impact of the top six covariates: (a) NIR band, (b) green band, (c) clay [0–15cm], (d) sand [0–15cm], (e) red, and (f) blue band on predicted SOC. The colour gradient indicates the strength and direction of a feature's effect on the model output: red dots represent positive SHAP values (the variable drives SOC prediction higher at this location), while blue dots denote negative SHAP values (the variable drives SOC prediction lower at this location).**

b) June–August (dry season)

As the land surface becomes more exposed, the local FI (Fig. C4) and spatial SHAP (Fig. C5) maps reveal a clear shift in the Jun–Aug model's primary dependencies. For the SWIR bands (SWIR1_jun_aug, SWIR2_jun_aug), global SHAP maps indicate scattered contributions with notable localised variations in the northeast. The PyGRF model actively amplifies these dry-season signals, concentrating the highest local importance for both SWIR bands in these eastern and northeastern areas. Conversely, PyGRF assigns NIR_jun_aug its highest local importance magnitudes predominantly in the western clusters, also corresponding to localised variations observed in the global SHAP map for that region.

For elevation, the global SHAP map reveals a stark, intensely positive predictive contribution isolated on the extreme northeastern edge, perfectly aligning with the region's higher altitudes (Fig. 2b). However, PyGRF does not mirror this relationship locally. Instead, its higher local importance values are spread out in several small clusters in the north, northeast,

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

southeast and southwest. Despite this seasonal shift in the spectral and topographic drivers, the static soil properties (clay_0_15cm and sand_0_15cm) remain consistent, maintaining stable geographic gradients.

c) March–May (Long Rains)

In the Mar–May model, the local FI (Fig. C6) and spatial SHAP (Fig. C7) maps highlight the dominant role of the visible bands. The influence of red_mar_may and green_mar_may is distinctly regionalised, with the highest local importance concentrated in the northeastern clusters associated with cropland. While their spatial representation remains constant between both global and local scales, the visible bands play a dual role across the landscape: they drive global SOC predictions higher in the north, while driving them significantly lower in the eastern and southeastern regions. In contrast, SWIR1_mar_may, which also shows higher global and local magnitudes in the northeast, behaves differently. It strongly drives predictions upward in the north, but it lacks the strong negative influence in the east and southeast seen with the visible bands. Meanwhile, the static soil properties (clay_0_15cm, sand_0_15cm) remain remarkably consistent, maintaining their stable geographic gradients.

d) Annual composite

The local FI (Fig. C8) and spatial SHAP (Fig. C9) distributions for the Annual composite model demonstrate a clear combination of seasonal dependencies. Global SHAP maps indicate moderate negative NIR_annual contributions heavily concentrated in the western and northwestern clusters. Correspondingly, PyGRF actively prioritises this spectral band, assigning NIR_annual its highest local importance magnitudes in this exact western region. Conversely, the visible bands (green_annual, red_annual) concentrate their highest local importance in the northeastern clusters, acting as strong positive drivers in the north while pushing global predictions downward in the eastern and southeastern regions. SWIR2_annual also emerges among the top spatial drivers with distinct local importance in the northeast (and two significant clusters in the northwest), though its directional SHAP influence remains far more diffuse compared to the visible and NIR bands. Despite the integration of these diverse spectral responses, the static physical soil properties (clay_0_15cm, sand_0_15cm) remain consistent, stabilising the model with established geographic gradients and diffuse moderate importance across the landscape.

4 Discussion

This study proposes a framework to address three critical DSM challenges in data-scarce tropical agroecosystems: identifying when SOC is most observable, which covariates provide the highest information yield, and evaluating whether spatially explicit modelling improves our understanding of SOC drivers. The discussion interprets our findings in relation to the three design questions, which can then be translated into implications for operational SOC monitoring.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

4.1 When to sample: biophysical drivers of seasonal SOC observability

In northern Tanzania, addressing the design question of *when* to sample requires re-evaluating the conventional direct-sensing paradigm. Under this paradigm, the most efficient way to map SOC is to use spectral covariates to build a sampling strategy that optimally covers the spectral feature space (Castaldi et al., 2019a). To address the geographic scarcity of bare-soil pixels, RS studies heavily prioritise techniques designed to isolate the mineral soil signal through approaches such as multi-year temporal compositing (Safanelli et al., 2020; Silvero et al., 2021) or spectral unmixing (Bartholomeus et al., 2011). To explicitly sieve and extract bare soil pixels, masking thresholds using NDVI (Loiseau et al., 2019), the Bare Soil Index (BSI) (Diek et al., 2017) or a combination of NDVI with the Normalised Burn Ratio 2 (NBR2) (Castaldi et al., 2019b) are often applied to filter out crop residues and moisture interferences. Essentially, these approaches operate on the principle that optimising sampling efficiency and prediction accuracy requires capturing the clearest, most comprehensive, spectral feature space of the target landscape. Consequently, many studies, such as Kempen et al. (2019) in Tanzania, often target the driest months to minimise the “disturbing” effect of vegetation.

However, our seasonal modelling reveals critical limitations in applying this bare-soil premise to agropastoral mosaics. While existing literature (Demattê et al., 2020; Khazaei et al., 2026) recommends targeting the driest months for satellite imagery acquisition, our results demonstrate that the Jun–Aug ‘dry season’ model actually produced the poorest predictive performance (Fig. 3, Fig. 4). In highly heterogeneous African landscapes, the dry season does not provide a clear feature space of purely exposed mineral soil (Fig. 10b). Instead, the landscape is covered by senesced grasses and non-photosynthetic vegetation (NPV) (Hosseini et al., 2026). The lignocellulose absorption features of these dry residues overlap directly with the cellulose and lignin signatures of native soil organic matter in the SWIR and NIR regions (particularly around 2100 and 2300 nm). This effectively masks the target SOC signal and results in spectral confusion (Daughtry and Hunt, 2008; Dvorakova et al., 2020), a problem further exacerbated by dominant soil types in northern Tanzania, such as Vertisols, that become extremely hard and cracked when completely dry (Somasundaram et al., 2018). As Vaudour et al. (2019) demonstrated, the compounding effects of these surface conditions and the timing of satellite imagery acquisition drive substantial variations in SOC prediction performance, particularly in croplands. Furthermore, attempting to bypass these dry-season limitations by relying on yearly averages is equally problematic. As observed from the seasonal fluctuations of the spectral bands, temporal aggregation inevitably smooths over the specific, high-contrast phenological windows where SOC observability is maximised, thereby obscuring critical eco-pedological distinctions and introducing hidden measurement variability (Surasinghe et al., 2025; Wuest, 2014). Therefore, while conventional approaches often view vegetation as spectral noise to be masked, this study leverages a vegetation–soil coupling framework that exploits phenological transitions as a predictive proxy to better understand SOC spatial variability.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Evaluating these transitions reveals highly variable dynamics across the rest of the year. During the March–May 'Long Rains' window (Fig. 10c), dense photosynthetic vegetation (PV) largely obscures the soil surface from the satellite's view, severing the structural coupling between surface reflectance and topsoil properties (Matyukira and Mhangara, 2024). Compounding this, high soil moisture content (SMC) strongly absorbs infrared light in any partially exposed areas (Lobell and Asner, 2002). Since water saturation actively masks the critical SWIR absorption features of clay minerals (particularly around 2200 nm), it alters the soil's fundamental reflectance (Abdelbaki et al., 2025; Castaldi et al., 2025). This dual effect of dense PV and water saturation confounds the spectral signatures within mixed pixels, preventing the algorithm from successfully disentangling vegetation–soil interactions. Consequently, the Mar–May model is deprived of both the active organic carbon signal and the underlying clay content required to establish a continuous baseline capacity for carbon accumulation.

Conversely, the optimal October–December window, driven by the Short Rains, establishes a delicate biophysical balance as illustrated in Fig. 10a. Unlike the arid conditions of the dry season, the moderate moisture during this window is sufficient to stimulate fresh vegetative inputs and activate the microbial carbon pump (Liang et al., 2017; Zhao et al., 2026), while simultaneously preventing the soil saturation and dense canopy obstruction that characterise the Long Rains. This allows the sensor to simultaneously detect the active vegetation dynamics and the natural physical darkening of the exposed soil surface associated with organic matter accumulation (Stenberg et al., 2010), which enables the algorithm to successfully disentangle the mixed-pixel environment. By capturing this specific phenological transition, the Oct–Dec model successfully exploits the dynamic structural coupling between vegetation and soil, as early growth stages are known to enhance spectral differentiation and strongly reflect underlying SOC content (Araya et al., 2016; He et al., 2021).

The success of this Oct–Dec model is corroborated by the recent Marmit–Leaf–Canopy model proposed by Abdelbaki et al. (2025), which establishes that accounting for the inherent complexity of real-world pixels significantly enhances predictive reliability. Specifically, Abdelbaki et al. (2025) established that pure bare-soil models consistently yield biased SOC overestimations ($R^2 = 0.71$, RMSE = 6.01 g kg^{-1}); however, integrating a comprehensive mixture of surface disturbances—including soil moisture, photosynthetic vegetation, and non-photosynthetic vegetation—achieves a highly robust and balanced predictive performance ($R^2 = 0.76$, RMSE = 5.49 g kg^{-1}). By embracing these surface disturbances rather than oversimplifying them, we demonstrate that vegetation–soil coupling is strongest during the Short Rains window, thereby securing a highly adaptable basis for reliable SOC monitoring in complex agroecosystems. As visually illustrated across the seasonal SOC-NDVI maps in Fig. 10, aligning satellite imagery acquisition with this targeted phenological window confirms that seasonal phenology fundamentally dictates SOC observability in this agroecosystem.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

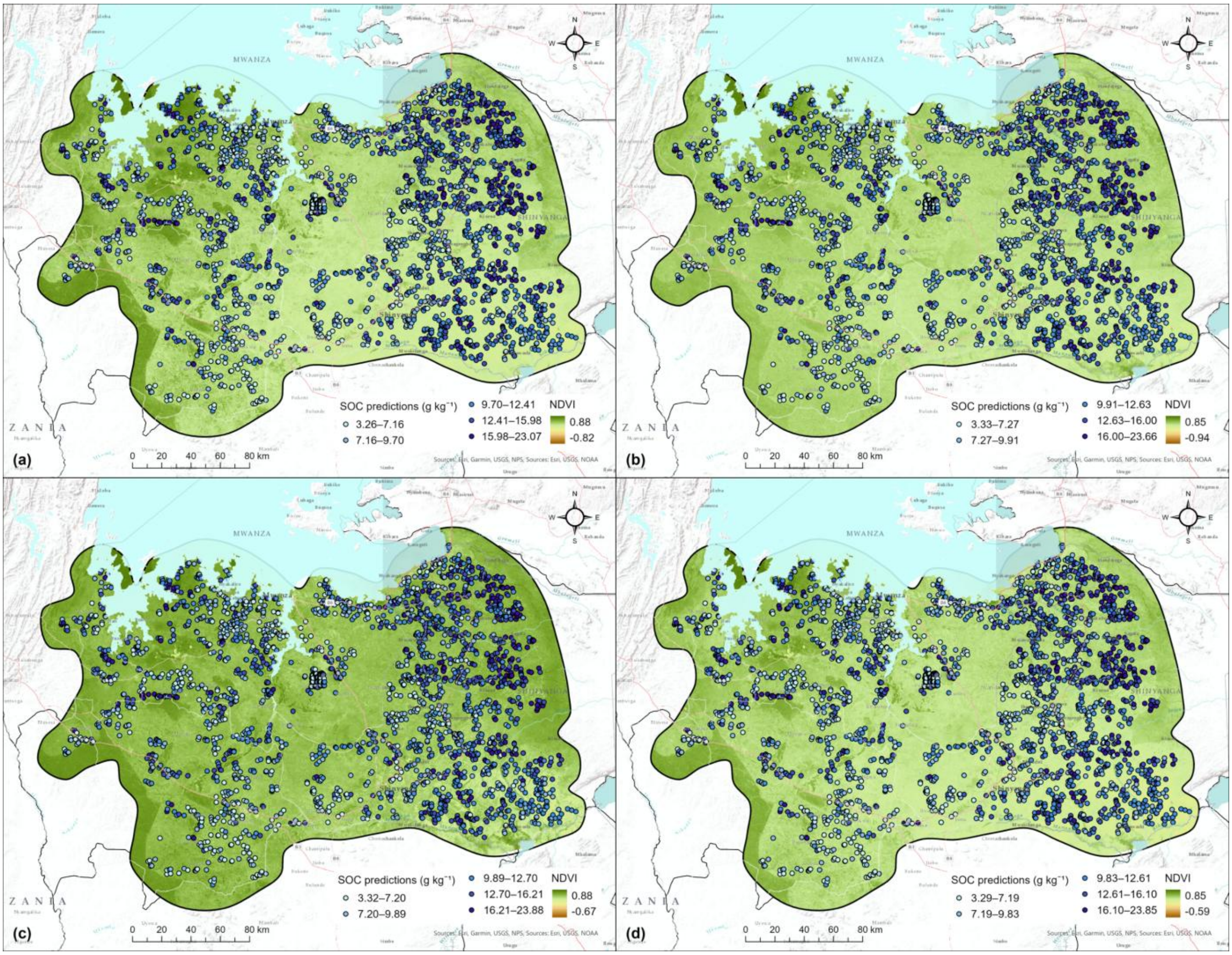


Figure 10 Spatial distribution of predicted SOC overlaid on seasonal and annual NDVI composites. **A 4-panel visualisation of the Shinyanga study area comparing localised SOC predictions (grams per kilogram) against background vegetation vigour across distinct temporal windows: (a) the optimal October–December 'Short Rains', which provides the necessary mixed-pixel environment for SOC detectability, (b) the June–August 'dry season', (c) the March–May 'Long Rains', and (d) the Annual composite. The background colour gradient represents NDVI values, transitioning from light green/yellowish tones (low vegetation vigour or bare soil) to dark green (dense canopy obstruction).** Data sources: **Background NDVI imagery derived from USGS Landsat 7&8 data accessed via Google Earth Engine; SOC predictions are based on iSDA data open soil dataset for Africa.**

4.2 What to measure: predictor utility

Once we have established the optimal time to sense SOC, we need to determine the optimal combination of environmental covariates that capture genuine SOC drivers and explain its spatial variability. We employed the Boruta algorithm to retain all ecologically relevant variables instead of just the non-redundant ones (Kursa and Rudnicki, 2010). However, since phenological patterns and carbon stabilisation rates vary significantly across ecosystems (Broeg et al., 2024; Matyukira and Mhangara, 2024), a single global model risks homogenising these distinct signals. We contend that SOC dynamics are most

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

sensitive and detectable in regions where vegetation–soil coupling is maximised, making the distinction between global and local SOC drivers essential for capturing these divergent spatial responses.

4.2.1 Global feature importance (overall contributions)

The rejection of EVI and SAVI, and the relatively low importance of NDVI in the temporal models (Fig. 6, Fig. B1, Fig. B2) highlights the limitations of dimensionally reduced indices in complex ML frameworks. While ratio-based indices are standard practice in RS, they often suffer from saturation in densely vegetated areas or lose sensitivity to subtle biomass changes, failing to distinguish between complex surface conditions within a mixed pixel (Matyukira and Mhangara, 2024). For example, NDVI frequently struggles to differentiate bare soil from crop residues (Broeg et al., 2024), while indices like NBR2 exhibit erratic, confounded responses when crop residues and high soil moisture are combined (Dvorakova et al., 2020, 2021). Feature importance ranking reported by Buthelezi et al. (2024) indicated that raw optical bands dominate the predictive feature space, accounting for up to 66% of total global importance in advanced tree-based models and consistently outperform established spectral indices. In our seasonal models, vegetation-linked spectral bands (NIR, green, red) operating in conjunction with soil texture (clay_0_15cm, sand_0_15cm) consistently emerge as the top predictors, providing strong empirical support for the vegetation–soil coupling framework.

Our global FI results (Fig. 6, Fig. B1, Fig. B2) demonstrate that robust SOC mapping requires leveraging raw spectral reflectance to access the full, multi-dimensional feature space. Because the extensive heterogeneity of organic matter components spans the entire 400–2500 nm spectrum, this full access is essential for the model to successfully decode the mixed spectral signatures of SMC, PV and NPV. Our SHAP dependence plots confirm this necessity by revealing a clear functional divergence between stable soil properties and highly dynamic seasonal spectral bands. While inherent static properties, indicated by the consistent threshold effect of clay content (Fig. 7; Fig. C1–C3), establish a continuous structural baseline for carbon storage year-round (Wiesmeier et al., 2019), the functional forms of the spectral bands shift significantly across seasons. This results from the temporal instability of surface reflectance driven by phenology (Wei et al., 2023). During the wetter phenological windows (Fig. 7, Fig. C2), vegetation-linked bands (visible and NIR) operate on aggressive, non-linear thresholds, forming steep L-shaped and cliff-edge curves where minor changes in surface brightness trigger massive penalties to predicted SOC. This sharply contrasts with the dry season, where the SWIR bands display gradual, almost linear responses to exposed crop residues and dry conditions as illustrated in Fig. C1.

This phenological variability confirms that single-date satellite imagery often yields lower accuracy as it fails to account for the dynamic interference of dense vegetation cover and crop residues throughout the year (Dvorakova et al., 2023; Zhu et al., 2024). During the Short Rains, vegetation greening in East African drylands is strongly moisture-controlled, and visible-band reflectance can be especially informative during rapid green-up and partial canopy cover (Cheng et al., 2020; Ugbaje and Bishop, 2020). In denser Long Rains canopies, standard, red-based signals tend to saturate as chlorophyll absorption in the red

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

approaches a maximum with increasing canopy closure (Gitelson et al., 1996). At that stage, increases in red reflectance can be interpreted as indicating greater soil exposure, reduced chlorophyll, or structural canopy openings (Yang et al., 2025) creating a spectral ambiguity that masks genuine variations in soil properties. This phenomenon is perfectly reflected in Fig. 6 where we observe the Mar–May (Long Rains) model preferentially selecting the red band as the leading predictor over stable soil properties.

By leveraging raw reflectance rather than simplified indices, the optimal Oct–Dec model maintained access to the full, multi-dimensional feature space required to decode the mixed spectral signatures of SMC, PV and NPV (Abdelbaki et al., 2025). Within this feature space, individual spectral bands play highly specialised roles. While the NIR band (17.3%) successfully tracks PV as a proxy for fresh carbon inputs, the green band (11.1%) drives predictions by capturing the differential, moisture-dependent phenological flush (Ugbaje and Bishop, 2020). As Gitelson et al. (1996) highlighted, unlike red-based indices that saturate rapidly under moderate canopy closure, the green channel retains high sensitivity across a wide range of chlorophyll concentrations. Lastly, the SWIR bands (SWIR2: 5.7%) allow the model to penetrate the "noise" of moisture and NPV to capture the underlying soil properties (Khazaei et al., 2026). Although moisture non-linearly decreases overall soil reflectance by darkening the surface, the SWIR region is significantly more robust for handling these variations than the visible and near-infrared (VNIR) regions; as Lobell and Asner (2002) established, it does not saturate as quickly under moist conditions and retains sensitivity to soil mineralogy. Consequently, by combining raw spectral reflectance with static soil properties (clay: 9.7%, sand: 8%), the pronounced vegetation–soil coupling observed during the October–December 'Short Rains' window reliably disentangles the mixed-pixel signals.

4.2.2 Spatially varying contributions

SOC absorption features in the VIS, NIR, and SWIR spectrums are distinct but depend heavily on local soil factors such as texture, mineralogy, and moisture (Broeg et al., 2024; Minasny et al., 2011). Understanding the true drivers of SOC in highly heterogeneous agroecosystems is fundamentally challenged by the spatial non-stationarity of environmental predictors and the inherent black-box nature of conventional ML algorithms. Purely data-driven models often learn complex empirical correlations rather than underlying physical mechanisms, which can limit interpretability and reduce generalisation beyond the training domain unless informed by valid mechanistic knowledge (Shen et al., 2023). To overcome this limitation, our study couples local feature importance values with global spatial SHAP values. Rather than conflating two different visualisations of scale, this multi-faceted interpretation operates synergistically to ensure the ecological relevance of our identified predictors: the local FI maps identify where the PyGRF algorithm adapts to spatial non-stationarity, while the global RF SHAP maps confirm whether these local adaptations are driven by genuine, biophysically valid rules.

This integration clarifies the highly distinct regional dependencies observed during the optimal Short Rains window. For example, while the PyGRF model assigns the highest local importance to the NIR band in the western and northwestern

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

clusters, the global SHAP maps reveal that this variable strongly drives SOC predictions lower in these specific zones (Fig. 8a, Fig. 9a). In the densely vegetated or forested western areas, the model identifies high NIR reflectance (dense canopy) as a physical threshold associated with distinct carbon dynamics, potentially reflecting the masking effect of persistent biomass. If the algorithm interpreted this high NIR signal as a direct soil indicator, it would incorrectly treat dense vegetation as carbon-rich soil. To avoid overestimating the SOC, it learns to apply a penalty to artificially lower the prediction in these highly vegetated zones (Wadoux et al., 2023). Conversely, in the eastern regions where the dominant land use is agriculture (Fig. 2c), the NIR local importance is notably lower. Here, NIR's SHAP contributions are strongly positive, reflecting a different ecological reality where active, early-season PV acts as a direct proxy for carbon inputs in the global RF model. This spatial inversion demonstrates how a single global covariate frequently acts as an empirical proxy for divergent local biophysical processes (Wadoux et al., 2023).

Comparing the divergence between local FI and global SHAP magnitudes for the visible bands (red, green, blue) highlights the exact mechanism by which PyGRF addresses spatial non-stationarity. For instance, in the western and northwestern regions, the high local importance of the NIR band corresponds directly to high-magnitude global SHAP values (e.g., –2.0, +2.0). In this case, the global RF model successfully captures the dominant regional relationship regardless of the direction of that influence. However, a stark contrast emerges with the visible bands in the southeastern and southwestern regions. In these southern locations, the global RF model assigns very low-magnitude SHAP values (around –0.5) to the visible bands. These low-magnitude SHAP values suggest that spatially varying biophysical surface conditions, such as differing soil mineralogy, moisture retention rates or active PV confound the global model. Since the broad absorption features of SOC overlap significantly with these properties, the SOC signal is easily masked by the heterogenous soil matrix (Castaldi et al., 2019a; Stenberg et al., 2010), making it difficult to establish a uniform physical rule. Consequently, global RF averages out the effects, essentially treating the visible bands as weak or irrelevant predictors in these zones, instead preferring NIR as the main predictor. As Broeg et al. (2024) note, this is a fundamental limitation of single global models, which often reduce complex regional interactions to a 'lowest common denominator', effectively masking out critical local drivers. By assigning high local importance where the global SHAP magnitude is low, PyGRF actively captures and utilises highly localised predictor-SOC relationships that the global RF model may completely miss. Most notably, this localised capability allows the spatially-explicit algorithm to exploit the green band's unique sensitivity to fresh, unsaturated chlorophyll in specific vegetated zones (Gitelson et al., 1996), transforming it into a powerful local predictor. This is exemplified by the green band's high local importance in locations where NIR's local importance is lower (Fig. 8b).

While phenology dictates the detectability of the soil signal, the inherent physical and chemical properties of the soil itself fundamentally alter how the PyGRF algorithm interprets that signal. Further analysis of the local FI maps reveals that the visible and SWIR/NIR bands express distinctly different spatial patterns across the study area that align with the dominant pedological zones. Although the eastern region has a relatively higher clay content (Fig. D1), the northeast is specifically

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

dominated by Planosols and Solonetz (FAO and IIASA, 2023). In their topsoils, natural weathering causes unstable minerals to dissolve while preventing new clay minerals from forming. This leaves behind highly resistant quartz (Macías and Camps-Arbestain, 2020). Pedological studies have shown that high quartz content leads to naturally high reflectance values. Against this highly reflective background, accumulated SOC acts as the primary, dominant absorbent in the visible spectrum (Stenberg et al., 2010). Consequently, the local FI maps for the visible spectrum, specifically the blue and red bands, exhibit highly concentrated hotspots of predictive importance in this northeastern region (Fig. 8b, 8e). Here, PyGRF leverages these bands to read baseline soil albedo as a physical indicator of SOC. Conversely, the rest of the study area is predominantly covered by Cambisols, where natural weathering processes actively generate new clay minerals (Macías and Camps-Arbestain, 2020). Unlike the highly reflective quartz, clay minerals express distinct absorption features (Demattê et al., 2018). In these Cambisol-dominant zones, the clay minerals fundamentally alter the spectral signal, driving the PyGRF algorithm to shift its reliance toward the SWIR and NIR regions to successfully capture the correlation between the soil's mineralogical composition and its organic carbon content.

While NIR captures the landscape's biological pulse, the SWIR2 band is deployed primarily to evaluate the soil's physico-chemical state. Widespread, faint positive SHAP contributions (+1.0) indicate the model leverages distinct SWIR clay absorption features as a consistent structural baseline for global SOC predictions (Fig. S3). However, localised pockets of negative SHAP values (-1.0) in the southeast reveal SWIR2 simultaneously acting as a diagnostic filter. Because SWIR2 is highly sensitive to surface moisture and crop residues, which artificially darken the soil and mimic high SOC (Dvorakova et al., 2020, 2023), the model applies targeted penalties here. During the Short Rains season, these southeastern pockets likely represent localised areas of temporary waterlogging or unploughed fields with heavy crop residue retention. This susceptibility to waterlogging is intrinsically linked to the high clay content in the area (Fig. D1), which restricts drainage. While clay increases the theoretical upper limit for SOC stabilisation, high-clay agricultural soils do not automatically store more carbon effectively (Schweizer et al., 2021; Soinne et al., 2023). This phenomenon is validated by the low SOC predictions found in this southeastern region. By shifting its local feature weights to overwhelmingly favour the visible bands in the northeast and southern regions, and the SWIR2 band to act as a support in the clay-rich zones, PyGRF spatially optimises its reliance on these features, capturing the true complexity of the soil matrix without being constrained by a uniform global fit.

Unlike the spectral bands, physical soil properties such as clay_0_15cm and sand_0_15cm exhibit a more diffuse, moderate local importance across the entire landscape (Fig. 8c). Global SHAP maps reveal that clay acts as a dominant positive driver in the east, whereas sand-dominated soils in the southwest trigger a negative penalty (Fig. 9c, D1, D2). Ultimately, by mapping both the local feature importance and the global directional effect of these variables, we demonstrate that the PyGRF algorithm is able to account for spatial non-stationarity in this highly heterogenous agroecosystem. Rather than implying a single uniform mechanism, the Short Rains window appears to amplify locally relevant pathways through which SOC becomes observable, including both fresh biomass dynamics and clay-mediated protection (Wiesmeier et al., 2019).

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

4.3 How to model: addressing spatial autocorrelation

By operating dynamically, models like PyGRF address spatial non-stationarity, allowing predictive rules to adapt to local environmental conditions (Georganos and Kalogirou, 2022). However, highly flexible, localised ML algorithms are particularly susceptible to overfitting these spatial dependencies (Yoo and Koo, 2025). Consequently, the true challenge in DSM is moving beyond simply maximising global prediction accuracy (Kim et al., 2023), to effectively capturing shifting environmental relationships while carefully managing underlying spatial dependence across complex landscapes. In geospatial predictive modelling, evaluating model generalisation via standard random cross-validation often violates the assumption of sample independence. This is due to the inherent spatial autocorrelation in SOC, which causes spatial leakage that masks overfitting and artificially inflates predictive accuracy (Deppner and Cajias, 2024; Roberts et al., 2017; Yoo and Koo, 2025). Aside from spatial autocorrelation, the degree of sample clustering equally alters how CV metrics should be interpreted (de Bruin et al., 2022; Wadoux et al., 2021).

The ANN analysis carried out on our target variable confirmed a statistically significant degree of spatial clustering among the sampling points. When the primary objective is mapping SOC at unsampled locations inside the same study region (Scenario A), standard RD-CV reflects operational reality reasonably well. Wang et al. (2025) demonstrate that applying RD-CV to spatially clustered data is a methodological flaw. Validation samples often occur close to training samples, violating independence assumptions. This creates spatial leakage and artificially inflates model accuracy. However, the PyGRF algorithm explicitly assumes that nearby observations are more relevant, operating by actively exploiting local target-response relationships and spatial autocorrelation (Sun et al., 2024). If local spatial dependence is the intended mathematical mechanism of the model, predicting from nearby observations is simply a reflection of the model's true interpolative capacity. The higher performance observed under RD-CV (Table 4) confirms that PyGRF successfully harnesses spatial autocorrelation to generate accurate, localised SOC predictions.

Conversely, when the deployment objective shifts from local interpolation to spatial extrapolation, spatial autocorrelation becomes a limitation rather than an advantage. In this context, the literature emphasises the importance of spatial CV as it poses a fundamental question: can the model transfer spatial knowledge to geographically distinct regions? (Brenning, 2023). By withholding contiguous spatial blocks, K-means CV and Block CV strip away the local spatial autocorrelation, forcing the algorithm to predict entirely new spatial domains based purely on the structural relationships it learned from the covariates. However, as indicated by the extreme instability observed in the fold-wise performance distribution (Fig. 4), applying strict SCV to clustered data introduces a severe penalty (de Bruin et al., 2022). Hence, in this study, we propose it as a strategy for structural stress testing.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Strict SCV (K-means CV, Block CV) yielded massive variance in predictive performance, with catastrophic drops in R^2 and spikes in RMSE occurring in specific folds (Fig. 4). Because rigid spatial partitions inherently group similar environmental characteristics, holding out massive, contiguous blocks of the landscape simultaneously masks out entire segments of the covariate feature space (Wang et al., 2025). When the algorithm is evaluated on these specific withheld blocks, it is forced to predict into environments whose specific covariate combinations were completely unrepresented during training (Roberts et al., 2017). This extreme variance corroborates the arguments of Wadoux et al. (2021) and Meyer and Pebesma (2021), who established that penalising a model for failing to extrapolate into completely unknown feature spaces yields overly punitive error estimates that obscure the model's true interpolative capacity.

When the primary goal is to isolate genuine pedological drivers from spatial artefacts without relying on the potentially destructive nature of arbitrary blocking (Scenario C) (Brenning, 2023), validation methods must enforce structural independence without inadvertently restricting the predictor space (Roberts et al., 2017). The concentric distance-based CV method achieves this balance. By expanding test folds outward from the spatial centroid, it simultaneously forces predictions beyond local spatial dependencies and neutralises the optimistic bias that is characteristic of RD-CV (Brenning, 2012). This is best observed during the Short Rains season, when early green-up provides an unmasked vegetation–soil signal. Here, the model predicts through a balanced synergy across the VNIR spectrum and static soil properties instead of heavily relying on soil texture or soil darkness as a spatial proxy. By distributing its weight between spatially autocorrelated static variables and dynamic spectral signatures, the Short Rains model likely learns fundamental physical rules (e.g., NIR as a proxy for biological carbon inputs, clay as a structural stabiliser). Because these predictive rulesets hold as the model extrapolates across shifting eco-pedological gradients toward the periphery, Db-CV yields unbiased fold-wise performance (Fig. 4), providing the most robust and ecologically meaningful validation of both model accuracy and driver observability.

The application of multiple evaluation strategies uncovers the true consequences of failing to account for phenological timing when predicting SOC. While the Annual and Mar–May models appear to perform adequately under RD-CV, their extreme fold-wise variability under Block CV (Fig. 4) indicates that this success is likely an artefact of spatial autocorrelation. Starved of a genuine SOC signal by optical barriers like dense canopy or high soil moisture, the models compensate by overfitting to non-causal, covarying spatial structures (Deppner and Cajias, 2024; Wadoux et al., 2023). This is demonstrated by the red band's dominance in the March–May 'Long Rains' model (Fig. 6). While it serves as a standard proxy for general soil darkness in bare fields (Broeg et al., 2024), red reflectance drops dramatically under dense vegetation due to intense chlorophyll absorption (Peng et al., 2017; Yang et al., 2025). Blinded by the canopy, the algorithm likely conflates these two physical processes, falsely treating the canopy-induced darkness as an indicator of carbon-rich soil. When strict SCV severs these spatially dependent proxies, the overfitted Long Rains model inevitably crashes. Crucially, as noted in the SHAP dependence plots, the Annual model reproduces these exact vegetation-dominated relationships, leading to a similar collapse. In contrast, the Jun–Aug 'dry season' model avoids these near-zero fold collapses. While the exposed mosaic of bare soil and NPV creates

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

local spatial autocorrelation that inflates its RD-CV performance (25.1% inflation; Fig. 3), the underlying biophysical signal remains sufficiently intact to prevent total failure. Ultimately, these contrasting seasonal dynamics highlight a critical limitation in conventional DSM: aggregating spectral data across a full calendar year inextricably mixes genuine phenological signals with spectral noise (Broeg et al., 2024) resulting in unreliable SOC predictions.

Lastly, our spatial uncertainty maps (Fig. 5) demonstrate the recurrence of clustered high-error spots located in the eastern region of the study area across all temporal windows, regardless of the model architecture (PyGRF vs global RF) or cross-validation method (RD-CV vs Db-CV) employed. These locations mostly coincide with croplands (Fig. 2c). Croplands are notoriously difficult to model because of their highly heterogeneous nature (Broeg et al., 2026). Localised pockets of high SOC content interspersed with depleted zones create conflicting spectral signatures easily masked by the soil matrix (Stenberg et al., 2010). In addition, active agricultural management drives extreme spatiotemporal variability via rapid phenological transitions and surface disturbances like fluctuating moisture and crop residues (Suleymanov et al., 2026). While our framework successfully exploits periods of strong vegetation–soil coupling, the dynamic, mixed-pixel nature of these actively managed zones introduces an unavoidable baseline of spectral noise, limiting absolute predictive certainty in this complex landscape.

Therefore, our 3-tiered validation strategy is not only essential for assessing model generalisability, but also for conclusively proving that SOC spatial variability is most detectable when vegetation–soil coupling is strongest, and that spatially-explicit seasonal modelling improves our understanding of SOC drivers in northern Tanzania. By aligning SOC detection with the Short Rains window, we turn the active biological signals often dismissed by conventional DSM as "noise" into predictive information that enables us to optimise SOC prediction in these highly heterogeneous agropastoral mosaics.

4.4 Implications for operational SOC monitoring in data-scarce tropical agroecosystems

In answering the three design questions of *when* to sample, *what* to measure and *how* to model, our vegetation–soil coupling framework asserts that SOC prediction in bimodal-rainfall agroecosystems can be made both leaner and more reliable by aligning sampling and covariate acquisition with the highest information yield seasonal window. National-scale carbon accounting and land management programmes in developing countries often cannot support dense and repeated field sampling, yet they still require spatially explicit and defensible estimates of SOC. Our results suggest that such efforts can be improved by prioritising Short Rains imagery for model calibration and by directing limited in situ sampling toward the spatial blind spots identified by OOF error mapping.

Detecting true, long term changes in SOC requires strict phenological consistency to ensure that transient surface conditions are not mistaken for underlying shifts in soil carbon status. While standardising the monitoring window to October–December mitigates this risk, East Africa remains highly vulnerable to the effects of ENSO-driven climate extremes. As our spatial

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

uncertainty maps demonstrate, the highest OOF absolute errors are overwhelmingly concentrated within croplands. This spatial distribution of error highlights a pressing challenge for current Monitoring, Reporting, and Verification (MRV) frameworks: the landscapes most vital for regional food security are the most difficult to monitor for soil health.

Equally important, the interpretability framework bridges the gap between predictive modelling and practical decision making. Because the global SHAP and local FI maps indicate where fresh biomass dynamics dominate and where clay-mediated stabilisation is likely more important, the outputs can support more context-specific monitoring strategies. To translate these eco-pedological dynamics into a robust predictive tool, our methodological framework addressed spatial autocorrelation by evaluating the PyGRF algorithm through a rigorous, multi-scenario spatial validation approach specifically designed to disentangle genuine SOC drivers from spatial artefacts. In that sense, the framework contributes not only to methodological advances in DSM but also to the operational design of climate-smart soil information systems.

5 Conclusion

To address the challenges of digital soil mapping in highly complex, data-scarce areas, this study developed and tested a spatially explicit, seasonally-informed framework in northern Tanzania. The results directly confirm our dual hypotheses. First, by demonstrating that SOC spatial variability is most reliably detected during the Short Rains (October–December) rather than under conventional annual composites or dry-season assumptions, we confirmed that SOC detectability peaks when vegetation–soil coupling is strongest. Second, we validated that geographically weighted ML improves our understanding of SOC in data-scarce, heterogeneous landscapes. The NIR band, green band and topsoil clay controlled SOC predictions globally, while the red, blue and SWIR2 bands controlled them locally. Utilising a parsimonious predictor set of 13 features, PyGRF successfully accounted for the spatial non-stationarity of these eco-pedological relationships.

Conventional RS approaches have historically relied on dry-season imagery to eliminate seasonal "bias" and obtain bare-soil pixels, effectively treating moderate soil moisture and partial canopy emergence as "noise". In contrast, this study demonstrates that accounting for the inherent complexity of real-world pixels significantly enhances predictive reliability in a bimodal-rainfall agroecosystem. By leveraging raw reflectance rather than simplified indices, our best model maintained access to the full, multi-dimensional feature space required to decode the mixed spectral signatures of soil moisture, photosynthetic vegetation, and non-photosynthetic vegetation. We therefore confirm that the early green-up phase of the October–December 'Short Rains' window serves as the optimal functional proxy for predicting SOC in northern Tanzania, actively harvesting this eco-pedological coupling as crucial predictive information.

Future work could extend this framework to other agroecological zones, incorporate additional sensors (e.g. Sentinel-1), and evaluate its performance for monitoring SOC stock change over time.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Appendix A: Distribution of spatial cross-validation test folds

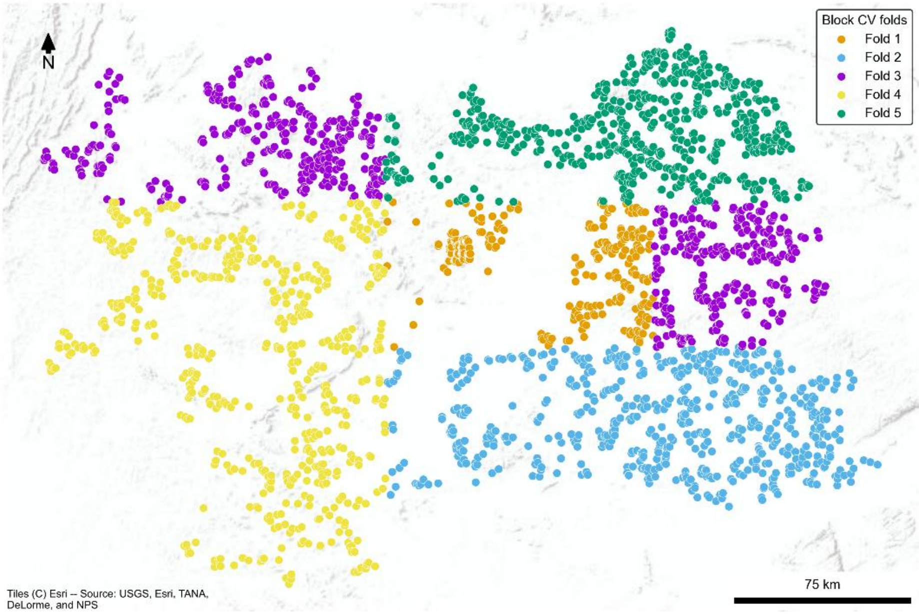


Figure A1 Spatial distribution of test folds partitioned using density-adaptive block cross-validation. **Five distinct spatial blocks were formed by applying quantile-based discretisation independently to the X and Y coordinates. Colours represent the individual test folds (Folds 1 through 5) mapped across the study area in northern Tanzania.**

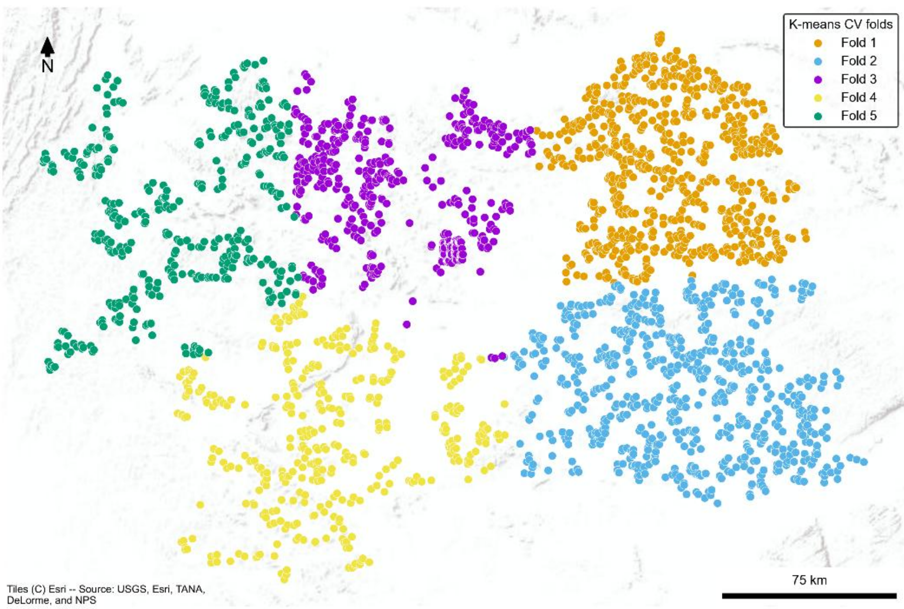
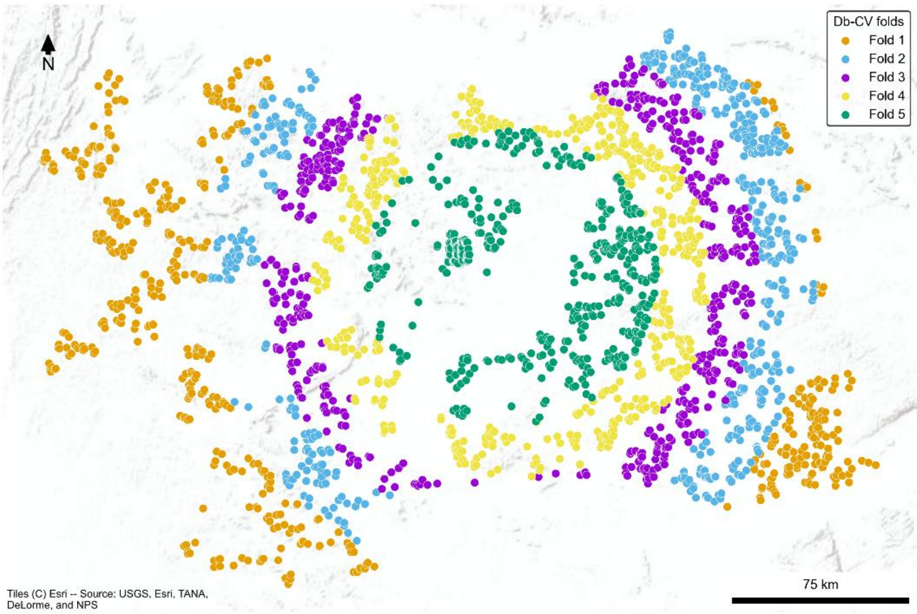


Figure A2 Spatial distribution of test folds partitioned using k-means clustering cross-validation. **Samples were aggregated into five spatially contiguous clusters using the k-means clustering algorithm. Colours represent the individual test folds (Folds 1 through 5) mapped across the study area in northern Tanzania.**

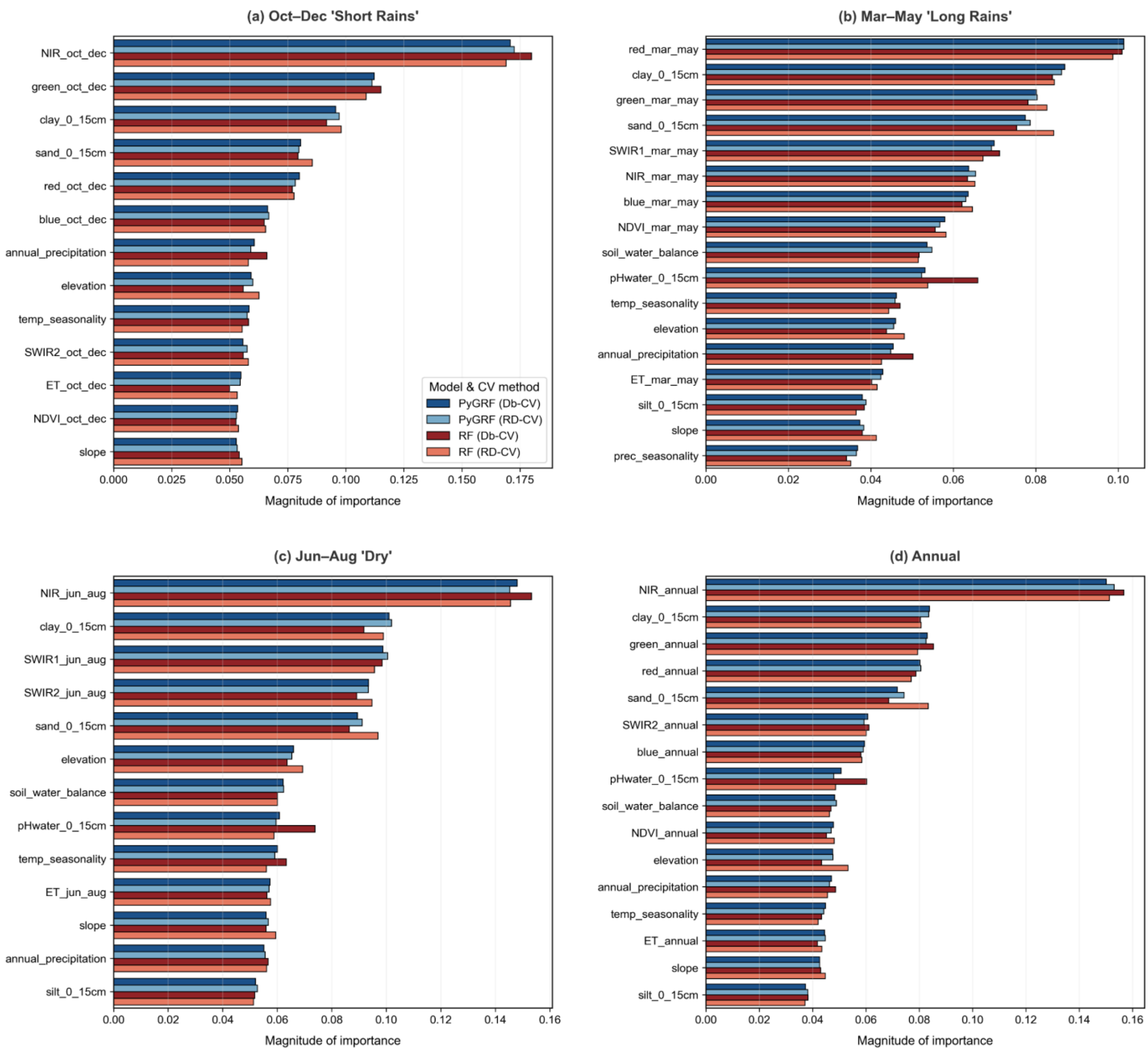
https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026



980 **Figure A3** Spatial distribution of test folds partitioned using concentric distance-based cross-validation. **The map illustrates the assignment of sample locations into five discrete, concentric folds radiating from the spatial centroid. Colours represent the individual test folds (Folds 1 through 5) mapped across the study area in northern Tanzania.**

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Appendix B: Consistency of seasonal SOC observability



985 **Figure B1** Global feature importance across modelling architectures. **Comparative analysis of predictor rankings for the PyGRF and Random Forest (RF) algorithms evaluated under distance-based (Db-CV) and random cross-validation (RD-CV). Features within each seasonal panel are sorted in descending order based on the optimal PyGRF (Db-CV) baseline to illustrate how algorithmic choice and spatial splitting strategy influence variable importance ranking.**

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

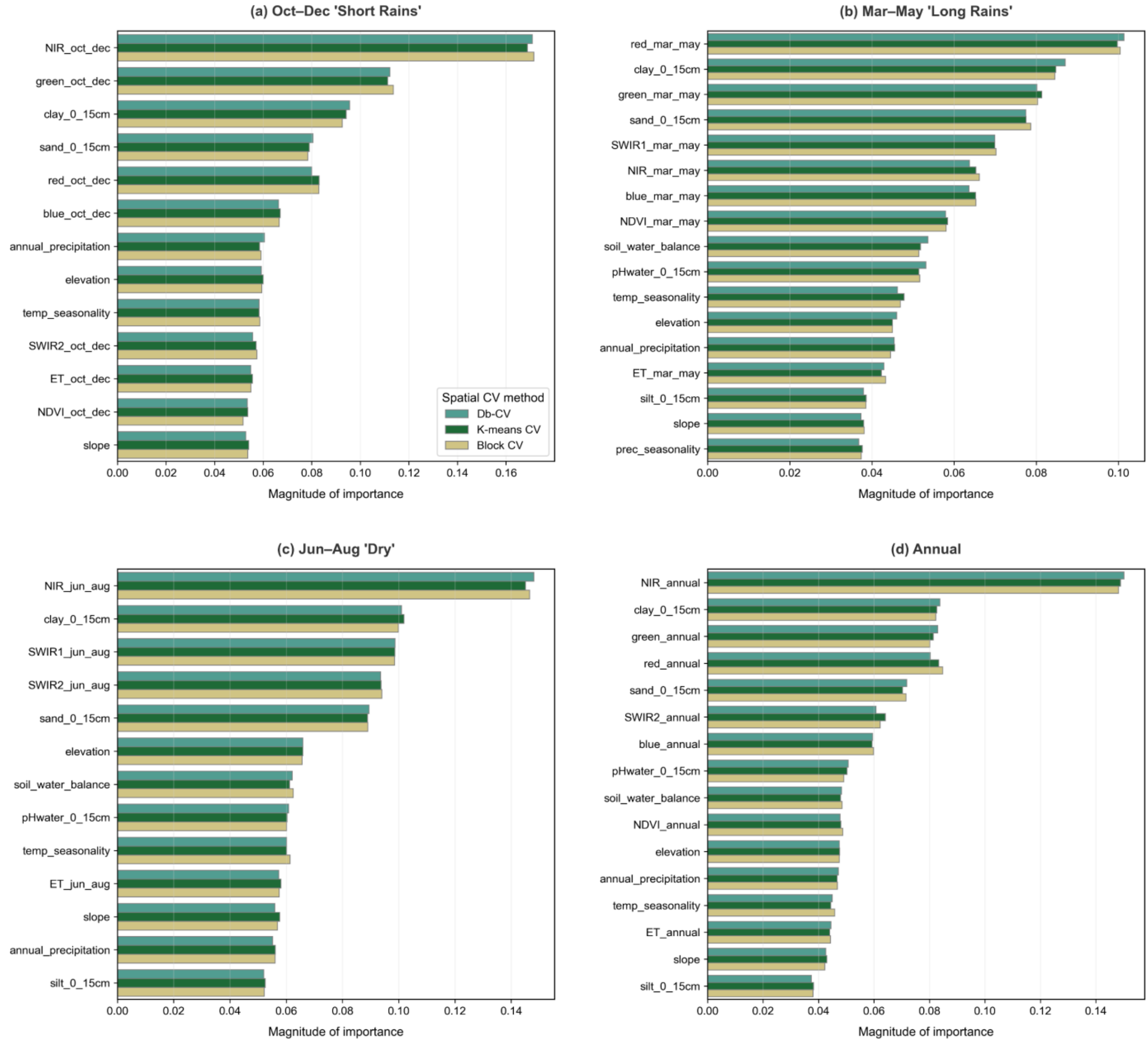


Figure B2 Sensitivity of PyGRF feature importance to spatial cross-validation. **Grouped bar charts displaying the relative importance of primary drivers across three SCV methods (Db-CV, K-means CV, and Block CV). Sorting is strictly anchored to the Db-CV (distance-based CV) method to visually demonstrate the stability and variation of SOC predictors when subjected to different spatial partitioning methods.**

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Appendix C: Seasonal information yield

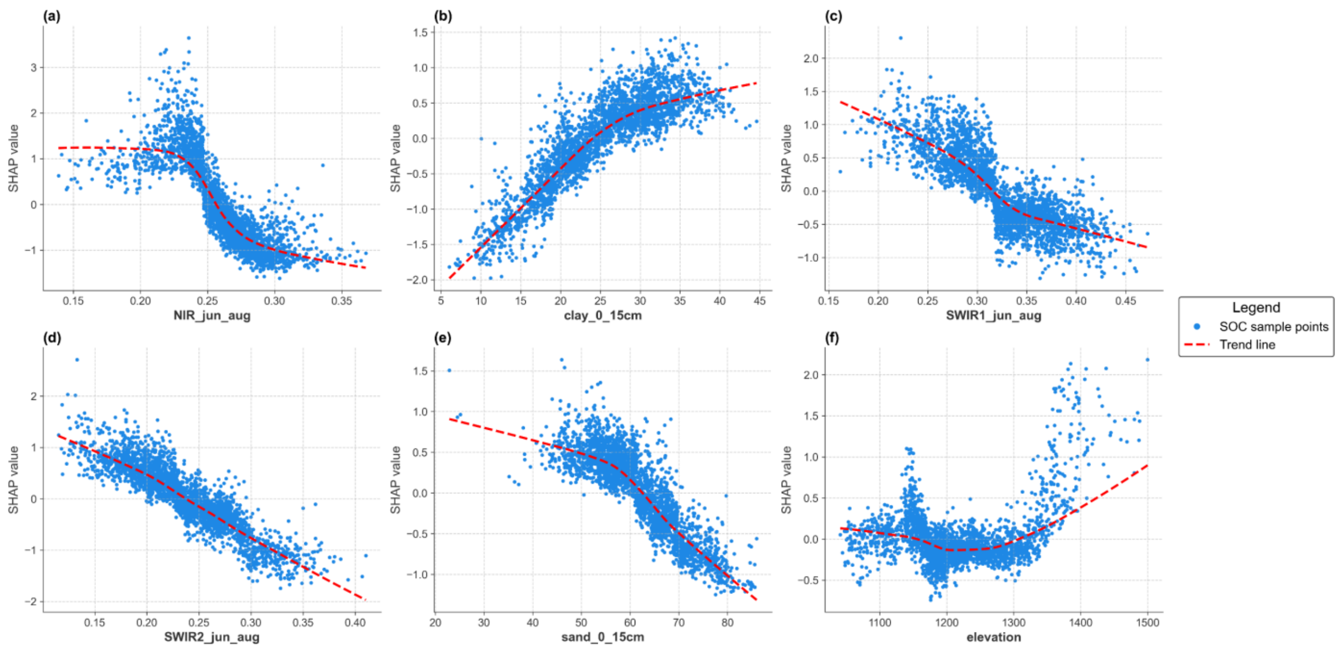


Figure C1 SHAP dependence plots **illustrating the non-linear marginal effects of the top six environmental covariates (NIR band, clay [0-15cm], SWIR1 and SWIR2 bands, sand [0-15cm], and elevation) on SOC predictions for the Jun-Aug model.**

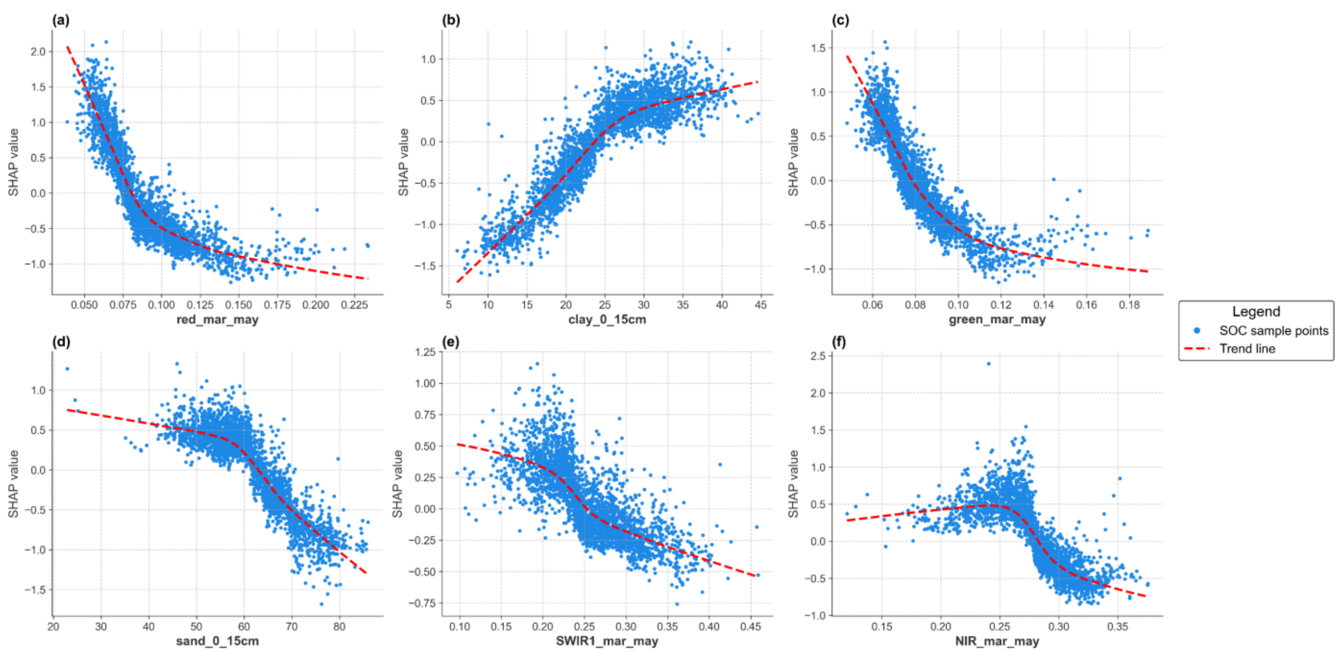


Figure C2 SHAP dependence plots **illustrating the non-linear marginal effects of the top six environmental covariates (red band, clay [0-15cm], green band, sand [0-15cm], SWIR1 and NIR bands) on SOC predictions for the Mar–May model.**

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

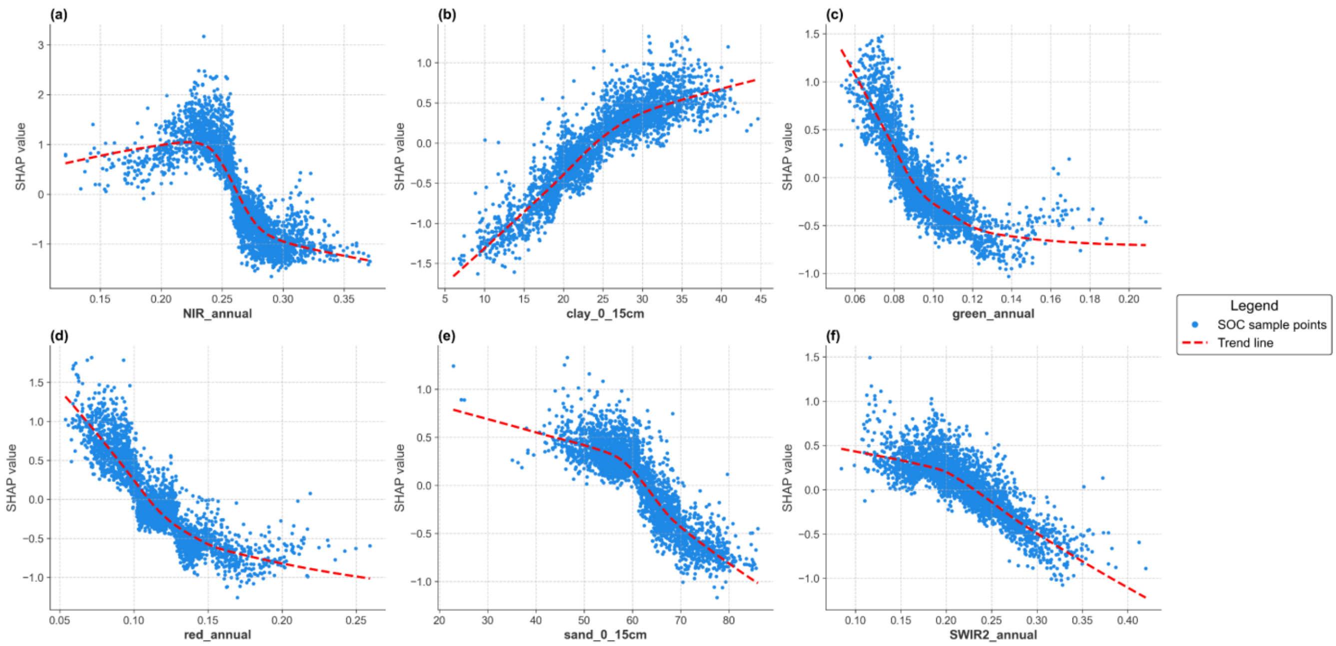


Figure C3 SHAP dependence plots **illustrating the non-linear marginal effects of the top six environmental covariates (NIR band, clay [0-15cm], green and red bands, sand [0-15cm] and SWIR2 band) on SOC predictions for the Annual composite model.**

1005

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

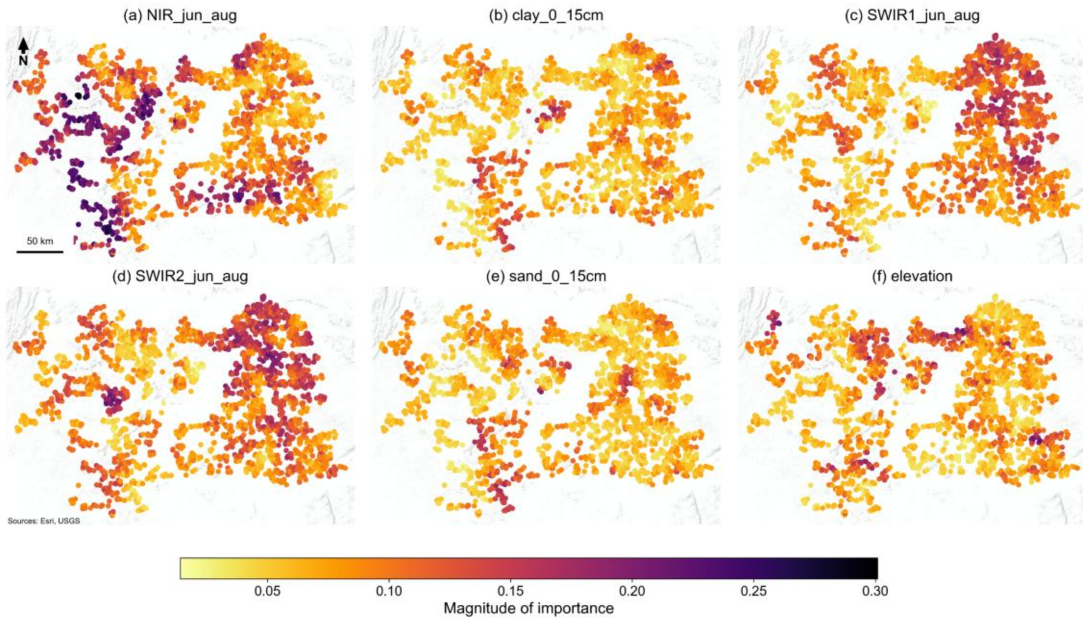


Figure C4 Spatial distribution of local feature importance for the Jun–Aug 'dry season' model. **The maps display the absolute magnitude of local importance for the top six covariates: (a) NIR band, (b) clay [0–15cm], (c) SWIR1 band, (d) SWIR2 band, (e) sand [0–15cm], and (f) elevation. Darker hues represent geographic regions where the specified variable exerts a stronger absolute influence on the local SOC predictions, regardless of the effect's direction.**

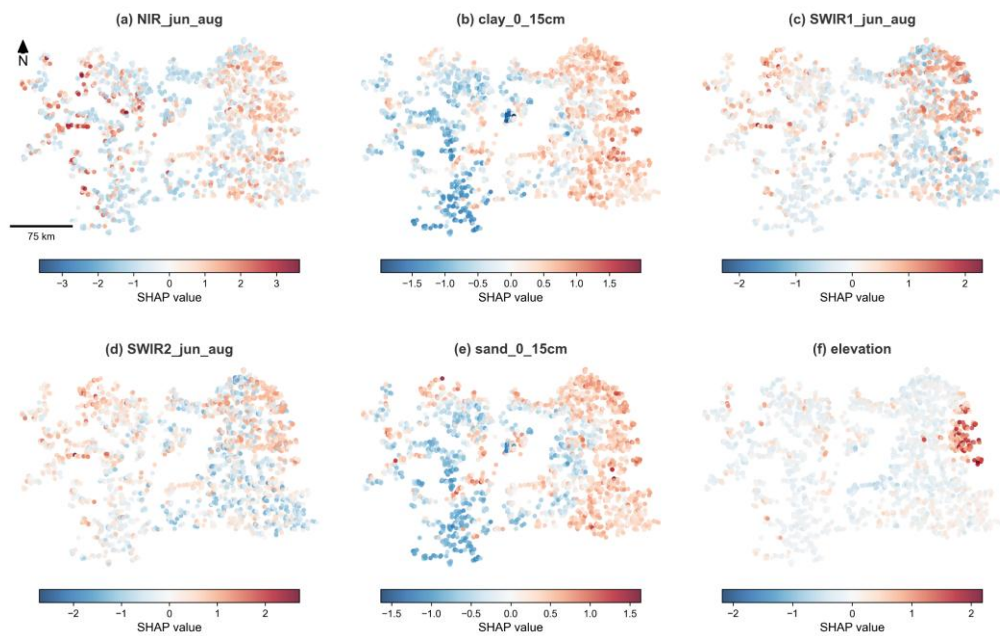


Figure C5 Spatial distribution of SHAP values for the Jun–Aug 'dry season' model. **The diverging maps illustrate the directional impact of the top six covariates: (a) NIR, (b) clay [0–15cm], (c) SWIR1, (d) SWIR2, (e) sand [0–15cm], and (f) elevation on predicted SOC. Red values indicate a local positive contribution that drives the predicted value higher, while blue values indicate a negative contribution that drives the predicted value lower.**

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

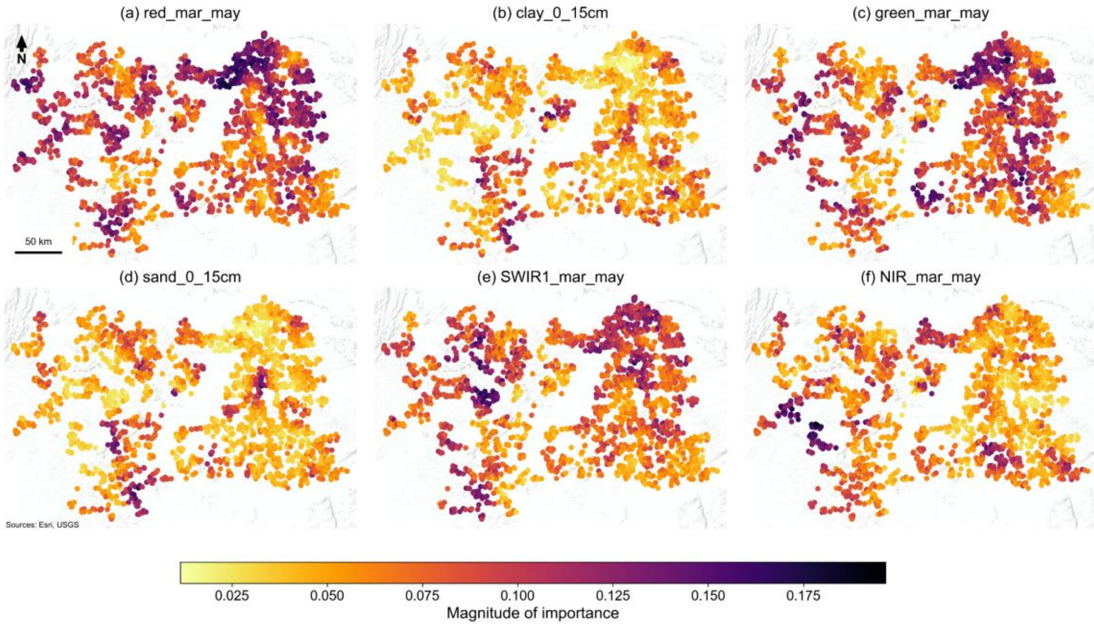


Figure C6 Spatial distribution of local feature importance for the Mar–May 'Long Rains' model. **The maps display the absolute magnitude of predictive reliance for the top six covariates: (a) red band, (b) clay [0–15cm], (c) green band, (d) sand [0–15cm], (e) SWIR1 and (f) NIR bands. Darker hues represent geographic regions where the specified variable exerts a stronger absolute influence on the local SOC predictions, regardless of the effect's direction.**

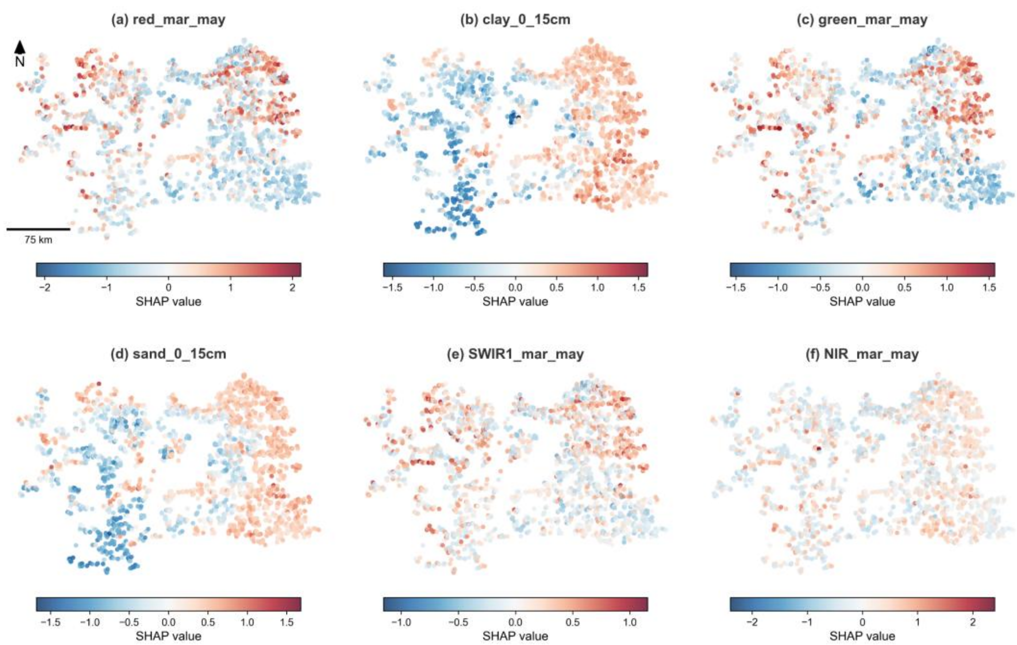


Figure C7 Spatial distribution of SHAP values for the optimal Mar–May 'Long Rains' model. **The diverging maps illustrate the directional impact of the top six covariates: (a) red, (b) clay [0–15cm], (c) green, (d) sand [0–15cm], (e) SWIR1 and (f) NIR on predicted SOC. Red values indicate a local positive contribution that drives the predicted value higher, while blue values indicate a negative contribution that drives the predicted value lower.**

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

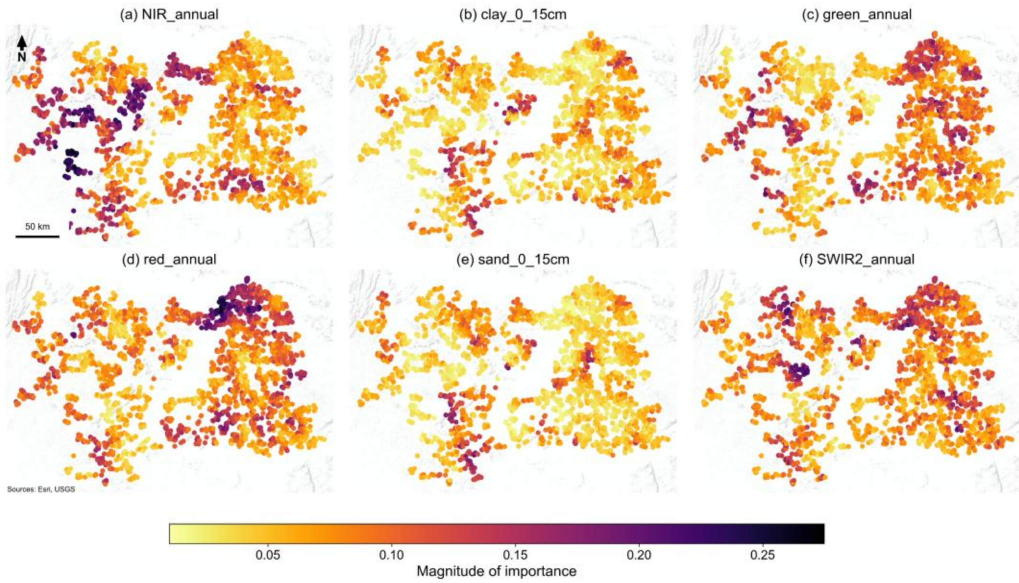


Figure C8 Spatial distribution of local feature importance for the Annual composite model. **The maps display the absolute magnitude of local importance for the top six covariates: (a) NIR band, (b) clay [0–15cm], (c) green band, (d) red band, (e) sand [0–15cm], and (f) SWIR2 band. Darker hues represent geographic regions where the specified variable exerts a stronger absolute influence on the local SOC predictions, regardless of the effect's direction.**

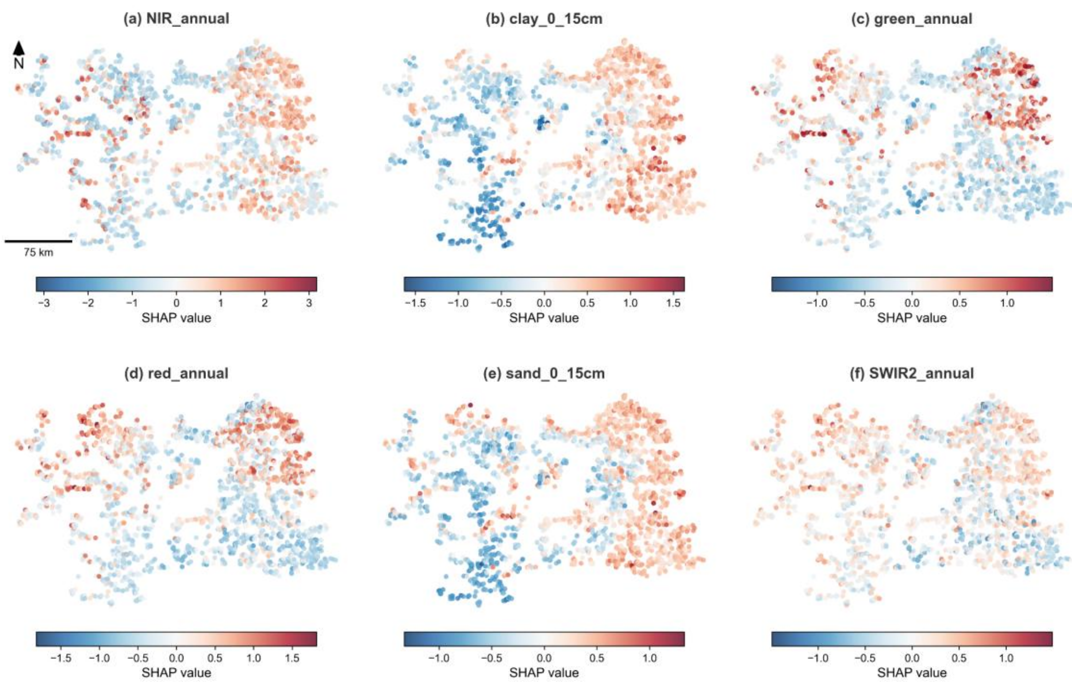


Figure C9 Spatial distribution of SHAP values for the optimal Annual PyGRF model. **The diverging maps illustrate the directional impact of the top six covariates: (a) NIR, (b) clay [0–15cm], (c) green, (d) red, (e) sand [0–15cm], and (f) SWIR2 on predicted SOC. Red values indicate a local positive contribution that drives the predicted value higher, while blue values indicate a negative contribution that drives the predicted value lower.**

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

1040 **Appendix D: Supporting figures**

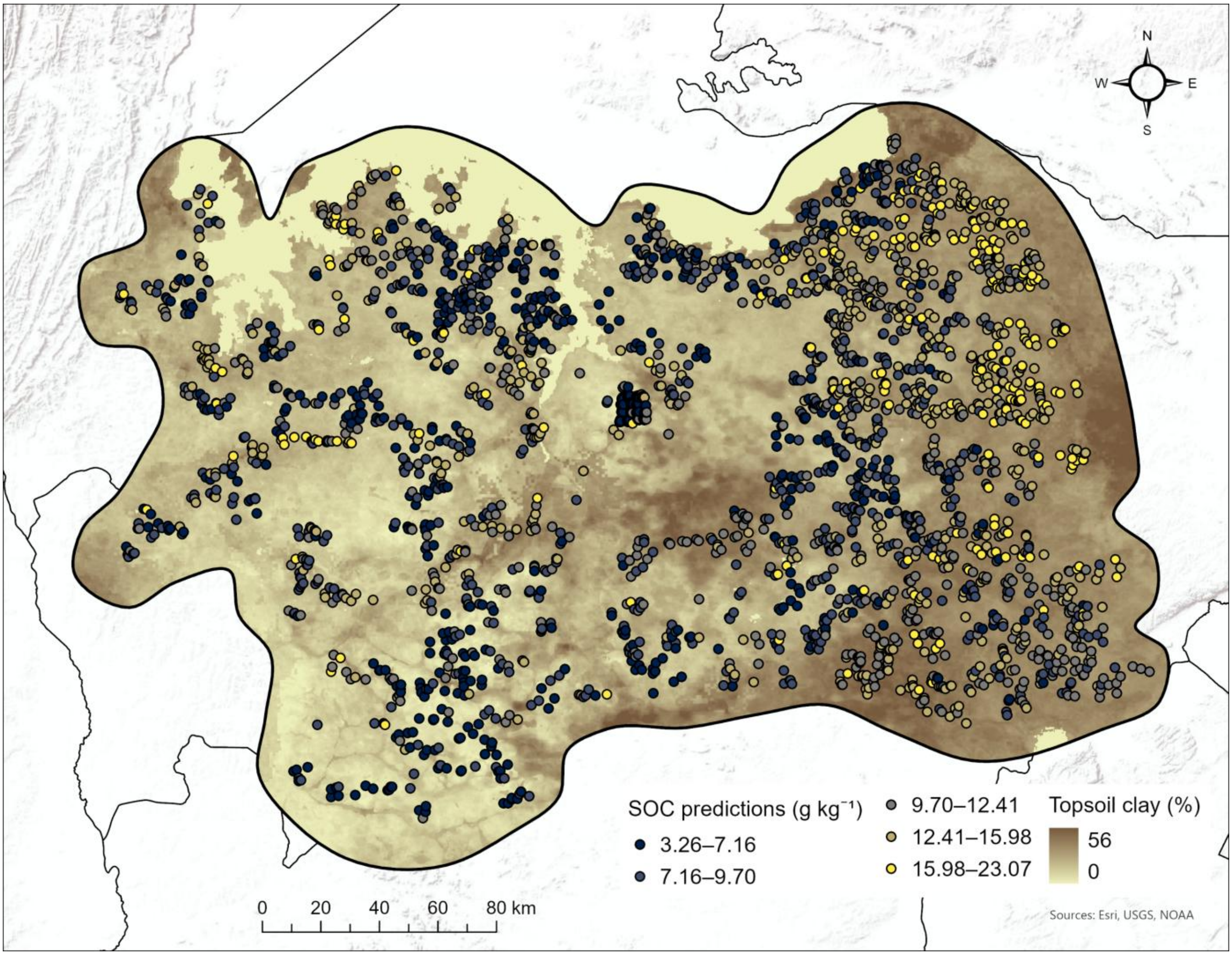


Figure D1 Spatial distribution of topsoil clay content (%) and SOC predictions (grams per kilogram) for the optimal October–December 'Short Rains' model. **The point predictions are overlaid on a continuous background of topsoil clay content (0–15 cm), highlighting how predictions align with the underlying soil texture across the northern Tanzanian study area. D**ata sources: **Background clay content map derived from SoilGrids250m 2.0, provided by ISRIC – World Soil information; SOC predictions are based on iSDA data open soil dataset for Africa.**

1045

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

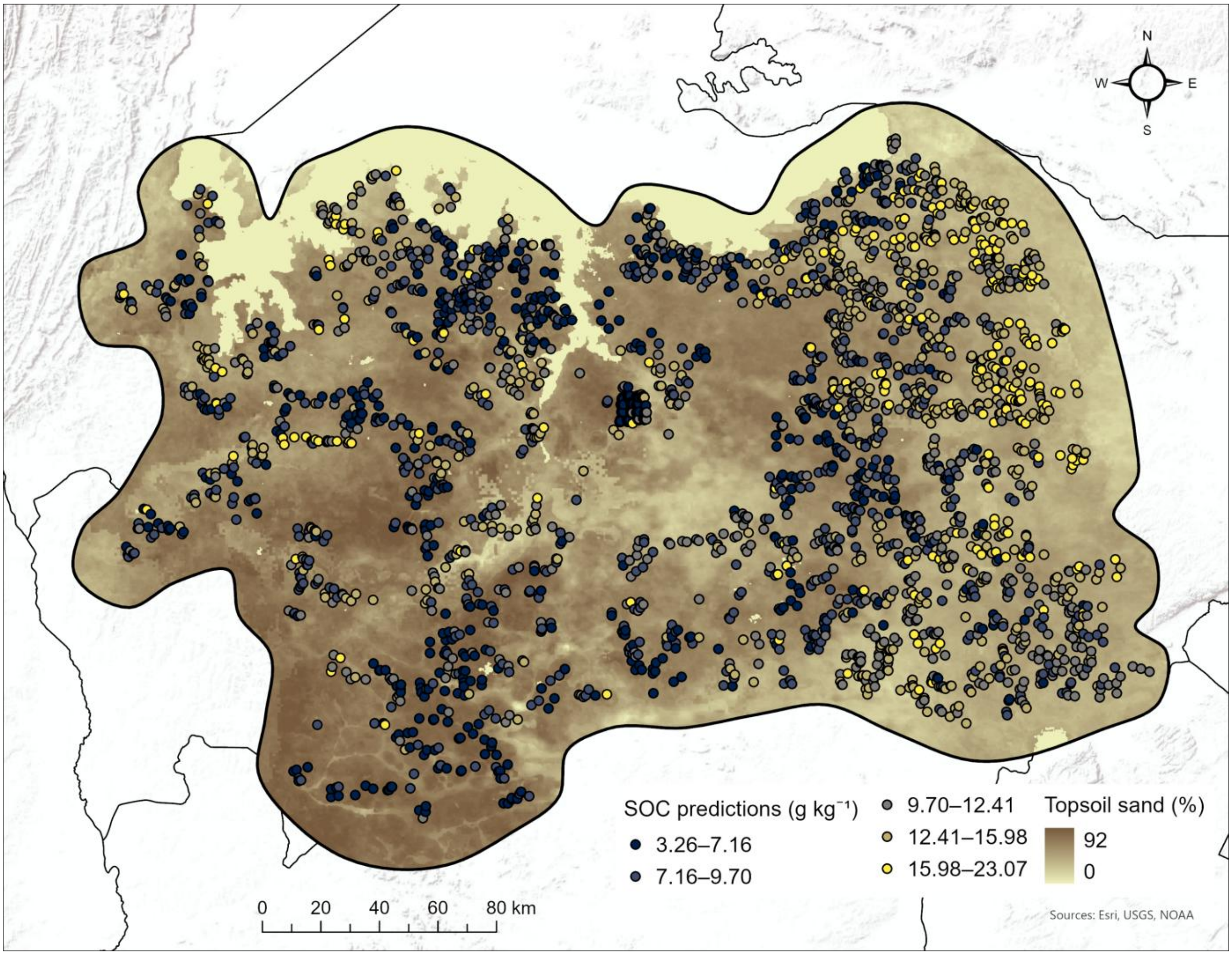


Figure D2 Spatial distribution of topsoil sand content (%) and SOC predictions (grams per kilogram) for the optimal October–December 'Short Rains' model. **The point predictions are overlaid on a continuous background of topsoil sand content (0–15 cm), highlighting how predictions align with the underlying soil texture across the northern Tanzanian study area.** Data sources: **Background sand content map derived from SoilGrids250m 2.0, provided by ISRIC – World Soil information; SOC predictions are based on iSDA data open soil dataset for Africa.**

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Code and data availability

The soil sample data and environmental covariates supporting this research were derived entirely from publicly available repositories. Soil point data were accessed via the iSDA Africa repository (https://github.com/iSDA-Africa/open-soil-data). Multispectral satellite imagery (Landsat 7 ETM+ and Landsat 8 OLI) was accessed and processed using the Google Earth Engine data catalogue. Topographic data (SRTM DEM) were acquired through OpenTopography. Bioclimatic variables (CHELSA V1.2) were sourced from the Dryad Digital Repository (https://doi.org/10.5061/dryad.kd1d4). Baseline soil properties, including texture and pH, were obtained from https://soilgrids.org/. The 2018 Esri LULC map was accessed via Sentinel-2 10m Land Use/Land Cover Time Series and evapotranspiration data were retrieved from the FAO WaPOR portal. The codes and specific dataset used in this study will be made available by the corresponding author upon reasonable request. The methodologies and formal citations for the generation of these datasets are detailed within the main text.

Author contributions

Annette Achieng': Conceptualisation, Data Curation, Formal analysis, Investigation, Methodology, Validation, Visualisation, Writing - Original Draft, Writing - Review & Editing. **Pedro Cabral:** Funding acquisition, Methodology, Supervision, Writing - Review & Editing. **Ana Cristina Costa:** Funding acquisition, Methodology, Supervision, Writing - Review & Editing.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

This work was supported by national funds through FCT,I.P. (Fundação para a Ciência e a Tecnologia,I.P.), under the project - UID/04152/2025 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS (DOI: 10.54499/UID/04152/2025), and by the European Union – NextGenerationEU under the project UID/PRR/04152/2025 (DOI: 10.54499/UID/PRR/04152/2025).

References

Abdelbaki, A., Milewski, R., Saberioon, M., Berger, K., Demattê, J. A. M., and Chabrillat, S.: Radiative Transfer Model-Integrated Approach for Hyperspectral Simulation of Mixed Soil-Vegetation Scenarios and Soil Organic Carbon Estimation, Remote Sens. (Basel)., 17, https://doi.org/10.3390/rs17142355, 2025.

Araya, S., Lyle, G., Lewis, M., and Ostendorf, B.: Phenologic metrics derived from MODIS NDVI as indicators for Plant Available Water-holding Capacity, Ecol. Indic., 60, 1263–1272, https://doi.org/10.1016/J.ECOLIND.2015.09.012, 2016.

Bartholomeus, H., Kooistra, L., Stevens, A., van Leeuwen, M., van Wesemael, B., Ben-Dor, E., and Tychon, B.: Soil Organic Carbon mapping of partially vegetated agricultural fields with imaging spectroscopy, International Journal of Applied Earth Observation and Geoinformation, 13, 81–88, https://doi.org/10.1016/J.JAG.2010.06.009, 2011.

Breiman, L.: Random Forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Brenning, A.: Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest, in: 2012 IEEE International Geoscience and Remote Sensing Symposium, 5372–5375, https://doi.org/10.1109/IGARSS.2012.6352393, 2012.

Brenning, A.: Spatial machine-learning model diagnostics: a model-agnostic distance-based approach, International Journal of Geographical Information Science, 37, 584–606, https://doi.org/10.1080/13658816.2022.2131789, 2023.

Broeg, T., Don, A., Gocht, A., Scholten, T., Taghizadeh-Mehrjardi, R., and Erasmi, S.: Using local ensemble models and Landsat bare soil composites for large-scale soil organic carbon maps in cropland, Geoderma, 444, https://doi.org/10.1016/j.geoderma.2024.116850, 2024.

Broeg, T., Don, A., Scholten, T., and Erasmi, S.: Reducing bias in cropland soil organic carbon and clay predictions using Sentinel-2 composites and data balancing, Remote Sens. Environ., 333, 115109, https://doi.org/10.1016/J.RSE.2025.115109, 2026.

de Bruin, S., Brus, D. J., Heuvelink, G. B. M., van Ebbenhorst Tengbergen, T., and Wadoux, A. M. J. C.: Dealing with clustered samples for assessing map accuracy by cross-validation, Ecol. Inform., 69, https://doi.org/10.1016/j.ecoinf.2022.101665, 2022.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E.: Geographically weighted regression: a method for exploring spatial nonstationarity, Geogr. Anal., 28, 281–298, https://doi.org/10.1111/j.1538-4632.1996.tb00936.x, 1996.

Buthelezi, M. N. M., Lottering, R., Peerbhay, K., and Mutanga, O.: Assessing the extent of land degradation in the eThekwini municipality using land cover change and soil organic carbon, Int. J. Remote Sens., 45, 1339–1367, https://doi.org/10.1080/01431161.2024.2307945, 2024.

Castaldi, F., Chabrillat, S., and van Wesemael, B.: Sampling strategies for soil property mapping using multispectral Sentinel-2 and hyperspectral EnMAP satellite data, Remote Sens. (Basel)., 11, https://doi.org/10.3390/rs11030309, 2019a.

Castaldi, F., Chabrillat, S., Don, A., and van Wesemael, B.: Soil organic carbon mapping using LUCAS topsoil database and Sentinel-2 data: An approach to reduce soil moisture and crop residue effects, Remote Sens. (Basel)., 11, https://doi.org/10.3390/rs11182121, 2019b.

Castaldi, F., Stenberg, B., Liebisch, F., Metzger, K., Ben-Dor, E., Knadel, M., Koganti, T., Wetterlind, J., Barbetti, R., Debaene, G., Klumpp, K., Lippl, M., Lorenzetti, R., Lozano Fondon, C., Sanden, T., Schaumberger, A., and Stajnko, D.: Estimating soil organic carbon using field VNIR-SWIR spectroscopy and existing soil spectral libraries: Mitigating heterogeneity, roughness and moisture effects, Smart Agricultural Technology, 12, 101353, https://doi.org/10.1016/J.ATECH.2025.101353, 2025.

Cheng, Y., Vrieling, A., Fava, F., Meroni, M., Marshall, M., and Gachoki, S.: Phenology of short vegetation cycles in a Kenyan rangeland from PlanetScope and Sentinel-2, Remote Sens. Environ., 248, 112004, https://doi.org/10.1016/J.RSE.2020.112004, 2020.

Daughtry, C. S. T. and Hunt, E. R.: Mitigating the effects of soil and residue water contents on remotely sensed estimates of crop residue cover, Remote Sens. Environ., 112, 1647–1657, https://doi.org/10.1016/J.RSE.2007.08.006, 2008.

Demattê, J. A. M., Fongaro, C. T., Rizzo, R., and Safanelli, J. L.: Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images, Remote Sens. Environ., 212, 161–175, https://doi.org/10.1016/J.RSE.2018.04.047, 2018.

Demattê, J. A. M., Safanelli, J. L., Poppiel, R. R., Rizzo, R., Silvero, N. E. Q., Mendes, W. de S., Bonfatti, B. R., Dotto, A. C., Salazar, D. F. U., Mello, F. A. de O., Paiva, A. F. da S., Souza, A. B., Santos, N. V. dos, Maria Nascimento, C., Mello, D. C. de, Bellinaso, H., Gonzaga Neto, L., Amorim, M. T. A., Resende, M. E. B. de, Vieira, J. da S., Queiroz, L. G. de, Gallo, B. C., Sayão, V. M., and Lisboa, C. J. da S.: Bare Earth's Surface Spectra as a Proxy for Soil Resource Monitoring, Sci. Rep., 10, https://doi.org/10.1038/s41598-020-61408-1, 2020.

Deppner, J. and Cajias, M.: Accounting for Spatial Autocorrelation in Algorithm-Driven Hedonic Models: A Spatial Cross-Validation Approach, Journal of Real Estate Finance and Economics, 68, 235–273, https://doi.org/10.1007/s11146-022-09915-y, 2024.

Diek, S., Fornallaz, F., Schaepman, M. E., and de Jong, R.: Barest Pixel Composite for agricultural areas using landsat time series, Remote Sens. (Basel)., 9, https://doi.org/10.3390/rs9121245, 2017.

Dvorakova, K., Shi, P., Limbourg, Q., and van Wesemael, B.: Soil organic carbon mapping from remote sensing: The effect of crop residues, Remote Sens. (Basel)., 12, https://doi.org/10.3390/rs12121913, 2020.

Dvorakova, K., Heiden, U., and Van Wesemael, B.: Sentinel-2 exposed soil composite for soil organic carbon prediction, Remote Sens. (Basel)., 13, https://doi.org/10.3390/rs13091791, 2021.

Dvorakova, K., Heiden, U., Pepers, K., Staats, G., van Os, G., and van Wesemael, B.: Improving soil organic carbon predictions from a Sentinel–2 soil composite by assessing surface conditions and uncertainties, Geoderma, 429, 116128, https://doi.org/10.1016/J.GEODERMA.2022.116128, 2023.

FAO: WaPOR database methodology, FAO, Rome, https://doi.org/10.4060/ca9894en, 2020.

FAO: Global Soil Organic Carbon Sequestration Potential Map – GSOCseq v.1.1, Global Soil Organic Carbon Sequestration Potential Map – GSOCseq v.1.1, FAO, Rome, https://doi.org/10.4060/cb9002en, 2022.

FAO and IIASA: Harmonized World Soil Database version 2.0, Harmonized World Soil Database version 2.0, FAO; International Institute for Applied Systems Analysis (IIASA);, Rome and Laxenburg, https://doi.org/10.4060/cc3823en, 2023.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D.: The Shuttle Radar Topography Mission, Reviews of Geophysics, 45, https://doi.org/10.1029/2005RG000183, 2007.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Georganos, S. and Kalogirou, S.: A Forest of Forests: A Spatially Weighted and Computationally Efficient Formulation of Geographical Random Forests, ISPRS Int. J. Geoinf., 11, https://doi.org/10.3390/ijgi11090471, 2022.

Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., and Kalogirou, S.: Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling, Geocarto Int., 36, 121–136, https://doi.org/10.1080/10106049.2019.1595177, 2021.

Gitelson, A. A., Kaufman, Y. J., and Merzlyak, M. N.: Use of a green channel in remote sensing of global vegetation from EOS-MODIS, Remote Sens. Environ., 58, 289–298, https://doi.org/10.1016/S0034-4257(96)00072-7, 1996.

Hartigan, J. A. and Wong, M. A.: Algorithm AS 136: A K-Means Clustering Algorithm, Appl. Stat., 28, 100–109, https://doi.org/10.2307/2346830, 1979.

He, X., Yang, L., Li, A., Zhang, L., Shen, F., Cai, Y., and Zhou, C.: Soil organic carbon prediction using phenological parameters and remote sensing variables generated from Sentinel-2 images, Catena (Amst)., 205, 105442, https://doi.org/10.1016/J.CATENA.2021.105442, 2021.

Hengl, T., Miller, M. A. E., Križan, J., Shepherd, K. D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S. M., McGrath, S. P., Acquah, G. E., Collinson, J., Parente, L., Sheykhmousa, M., Saito, K., Johnson, J. M., Chamberlin, J., Silatsa, F. B. T., Yemefack, M., Wendt, J., MacMillan, R. A., Wheeler, I., and Crouch, J.: African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning, Sci. Rep., 11, https://doi.org/10.1038/s41598-021-85639-y, 2021.

Hosseini, H., Hensel, O., and Nasirahmadi, A.: Monitoring African rangelands with remote sensing: Sensors, platforms, and the emerging role of machine learning, Remote Sens. Appl., 41, 101957, https://doi.org/10.1016/J.RSASE.2026.101957, 2026.

iSDA: iSDA - Open Access Soil Data for Africa, https://doi.org/10.17605/OSF.IO/A69R5, 2023.

IUSS Working Group WRB: World Reference Base for Soil Resources. International soil classification system for naming soils and creating legends for soil maps, 4th edition., International Union of Soil Sciences (IUSS), Vienna, Austria, 2022.

Jafari, A., Sarmadian, F., Heidari, A., and Rasaei, Z.: Enhancing digital mapping of soil organic carbon through spatial modeling and validation, Sci. Rep., 16, https://doi.org/10.1038/s41598-026-39496-2, 2026.

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, Sci. Data, 4, https://doi.org/10.1038/sdata.2017.122, 2017.

Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J. C., Mathis, M., and Brumby, S. P.: Global land use / land cover with Sentinel 2 and deep learning, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 4704–4707, https://doi.org/10.1109/IGARSS47720.2021.9553499, 2021.

Kempen, B., Dalsgaard, S., Kaaya, A. K., Chamuya, N., Ruipérez-González, M., Pekkarinen, A., and Walsh, M. G.: Mapping topsoil organic carbon concentrations and stocks for Tanzania, Geoderma, 337, 164–180, https://doi.org/10.1016/J.GEODERMA.2018.09.011, 2019.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Khazaei, M., Sorenson, P. T., Magagi, R., and Goïta, K.: Satellite Soil Observation (SatSoil): extraction of bare soil reflectance for soil organic carbon mapping on Google Earth Engine, Canadian Journal of Remote Sensing, 52, https://doi.org/10.1080/07038992.2025.2597003, 2026.

Kibblewhite, M. G., Ritz, K., and Swift, M. J.: Soil health in agricultural systems, Philosophical Transactions of the Royal Society B: Biological Sciences, 363, 685–701, https://doi.org/10.1098/rstb.2007.2178, 2008.

Kim, D., Song, I., Miralha, L., Hirmas, D. R., McEwan, R. W., Mueller, T. G., and Šamonil, P.: Consequences of spatial structure in soil–geomorphic data on the results of machine learning models, Geocarto Int., 38, https://doi.org/10.1080/10106049.2023.2245381, 2023.

Kmoch, A., Harrison, C. T., Choi, J., and Uuemaa, E.: Spatial autocorrelation in machine learning for modelling soil organic carbon, Ecol. Inform., 86, 103057, https://doi.org/10.1016/J.ECOINF.2025.103057, 2025.

Kuhn, M. and Johnson, K.: Applied Predictive Modeling, 1st ed., Springer New York, New York, NY, 36–37 pp., https://doi.org/10.1007/978-1-4614-6849-3, 2013.

Kursa, M. B. and Rudnicki, W. R.: Feature Selection with the Boruta Package, Journal of Statistical Software, Foundation for Open Access Statistics, 36, https://doi.org/10.18637/jss.v036.i11, 2010.

Liang, C., Schimel, J. P., and Jastrow, J. D.: The importance of anabolism in microbial control over soil carbon storage, Nat. Microbiol., 2, 17105, https://doi.org/10.1038/nmicrobiol.2017.105, 2017.

Lobell, D. B. and Asner, G. P.: Moisture Effects on Soil Reflectance, Soil Science Society of America Journal, 66, 722–727, https://doi.org/10.2136/sssaj2002.7220, 2002.

Loiseau, T., Chen, S., Mulder, V. L., Román Dobarco, M., Richer-de-Forges, A. C., Lehmann, S., Bourennane, H., Saby, N. P. A., Martin, M. P., Vaudour, E., Gomez, C., Lagacherie, P., and Arrouays, D.: Satellite data integration for soil clay content modelling at a national scale, International Journal of Applied Earth Observation and Geoinformation, 82, 101905, https://doi.org/10.1016/J.JAG.2019.101905, 2019.

Lyon, B. and Dewitt, D. G.: A recent and abrupt decline in the East African long rains, Geophys. Res. Lett., 39, https://doi.org/10.1029/2011GL050337, 2012.

Macías, F. and Camps-Arbestain, M.: A biogeochemical view of the world reference base soil classification system: Homage to Ward Chesworth, Advances in Agronomy, 160, 295–342, https://doi.org/10.1016/BS.AGRON.2019.11.002, 2020.

Matyukira, C. and Mhangara, P.: Advances in vegetation mapping through remote sensing and machine learning techniques: a scientometric review, Eur. J. Remote Sens., 57, https://doi.org/10.1080/22797254.2024.2422330, 2024.

McBratney, A. B., Mendonça Santos, M. L., and Minasny, B.: On digital soil mapping, Geoderma, 117, 3–52, https://doi.org/10.1016/S0016-7061(03)00223-4, 2003.

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

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Minasny, B., McBratney, A. B., Bellon-Maurel, V., Roger, J. M., Gobrecht, A., Ferrand, L., and Joalland, S.: Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon, Geoderma, 167–168, 118–124, https://doi.org/10.1016/J.GEODERMA.2011.09.008, 2011.

Negassa, M. K., Haile, M., Feyisa, G. L., Wogi, L., and Liben, F. M.: Soil Organic Carbon Stock Prediction: Fate under 2050 Climate Scenarios, the Case of Eastern Ethiopia, Sustainability (Switzerland), 15, https://doi.org/10.3390/su15086495, 2023.

Nenkam, A. M., Wadoux, A. M. J. C., Minasny, B., Silatsa, F. B. T., Yemefack, M., Ugbaje, S. U., Akpa, S., Zijl, G. Van, Bouasria, A., Bouslihim, Y., Chabala, L. M., Ali, A., and McBratney, A. B.: Applications and challenges of digital soil mapping in Africa, Geoderma, 449, 117007, https://doi.org/10.1016/J.GEODERMA.2024.117007, 2024.

Peng, Y., Nguy-Robertson, A., Arkebauer, T., and Gitelson, A. A.: Assessment of canopy chlorophyll content retrieval in maize and soybean: Implications of hysteresis on the development of generic algorithms, Remote Sens. (Basel)., 9, https://doi.org/10.3390/rs9030226, 2017.

Piao, S., Liu, Q., Chen, A., Janssens, I. A., Fu, Y., Dai, J., Liu, L., Lian, X., Shen, M., and Zhu, X.: Plant phenology and global climate change: Current progresses and challenges, Glob. Chang. Biol., 25, 1922–1940, https://doi.org/10.1111/gcb.14619, 2019.

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

Pouladi, N., Gholizadeh, A., Khosravi, V., and Borůvka, L.: Digital mapping of soil organic carbon using remote sensing data: A systematic review, Catena (Amst)., 232, 107409, https://doi.org/10.1016/J.CATENA.2023.107409, 2023.

Van der Putten, W. H., Bardgett, R. D., Bever, J. D., Bezemer, T. M., Casper, B. B., Fukami, T., Kardol, P., Klironomos, J. N., Kulmatiski, A., Schweitzer, J. A., Suding, K. N., Van de Voorde, T. F. J., and Wardle, D. A.: Plant-soil feedbacks: The past, the present and future challenges, Journal of Ecology, 101, 265–276, https://doi.org/10.1111/1365-2745.12054, 2013.

Rao Kolusu, S., Shamsudduha, M., Todd, M. C., Taylor, R. G., Seddon, D., Kashaigili, J. J., Ebrahim, G. Y., Cuthbert, M. O., Sorensen, J. P. R., Villholth, K. G., Macdonald, A. M., and Macleod, D. A.: The El Niño event of 2015-2016: Climate anomalies and their impact on groundwater resources in East and Southern Africa, Hydrol. Earth Syst. Sci., 23, 1751–1762, https://doi.org/10.5194/hess-23-1751-2019, 2019.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, https://doi.org/10.1111/ecog.02881, 1 August 2017.

Safaee, S., Libohova, Z., Kladivko, E. J., Brown, A., Winzeler, E., Read, Q., Rahmani, S., and Adhikari, K.: Influence of sample size, model selection, and land use on prediction accuracy of soil properties, Geoderma Regional, 36, e00766, https://doi.org/10.1016/J.GEODRS.2024.E00766, 2024.

Safanelli, J. L., Chabrillat, S., Ben-Dor, E., and Demattê, J. A. M.: Multispectral models from bare soil composites for mapping topsoil properties over Europe, Remote Sens. (Basel)., 12, https://doi.org/10.3390/RS12091369, 2020.

Schweizer, S. A., Mueller, C. W., Höschen, C., Ivanov, P., and Kögel-Knabner, I.: The role of clay content and mineral surface area for soil organic carbon storage in an arable toposequence, Biogeochemistry, 156, 401–420, https://doi.org/10.1007/s10533-021-00850-3, 2021.

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., Liu, X., Ren, W., Zheng, Y., Harman, C. J., Clark, M., Farthing, M., Feng, D., Kumar, P., Aboelyazeed, D., Rahmani, F., Song, Y., Beck, H. E., Bindas, T., Dwivedi, D., Fang, K., Höge, M., Rackauckas, C., Mohanty, B., Roy, T., Xu, C., and Lawson, K.: Differentiable modelling to unify machine learning and physical models for geosciences, Nat. Rev. Earth Environ., 4, 552–567, https://doi.org/10.1038/s43017-023-00450-9, 2023.

Silvero, N. E. Q., Demattê, J. A. M., Amorim, M. T. A., Santos, N. V. dos, Rizzo, R., Safanelli, J. L., Poppiel, R. R., Mendes, W. de S., and Bonfatti, B. R.: Soil variability and quantification based on Sentinel-2 and Landsat-8 bare soil images: A comparison, Remote Sens. Environ., 252, 112117, https://doi.org/10.1016/J.RSE.2020.112117, 2021.

Soinne, H., Keskinen, R., Tähtikarhu, M., Kuva, J., and Hyväluoma, J.: Effects of organic carbon and clay contents on structure-related properties of arable soils with high clay content, Eur. J. Soil Sci., 74, https://doi.org/10.1111/ejss.13424, 2023.

Somasundaram, J., Lal, R., Sinha, N. K., Dalal, R., Chitralekha, A., Chaudhary, R. S., and Patra, A. K.: Cracks and Potholes in Vertisols: Characteristics, Occurrence, and Management, Advances in Agronomy, 149, 93–159, https://doi.org/10.1016/BS.AGRON.2018.01.001, 2018.

Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M., and Wetterlind, J.: Visible and Near Infrared Spectroscopy in Soil Science, Advances in Agronomy, 107, 163–215, https://doi.org/10.1016/S0065-2113(10)07005-7, 2010.

Stock, A.: Choosing blocks for spatial cross-validation: lessons from a marine remote sensing case study, Frontiers in Remote Sensing, 6, https://doi.org/10.3389/frsen.2025.1531097, 2025.

Suleymanov, A., Telyagissov, S., Asylbaev, I., Mirsayapov, R., Suleymanov, R., Keshavarzi, A., Tuktarova, I., and Belan, L.: Multi-scenario modeling of soil organic carbon in semi-arid croplands with uncertainty quantification and model interpretation, Remote Sens. Appl., 41, 101903, https://doi.org/10.1016/j.rsase.2026.101903, 2026.

Sun, K., Zhou, R. Z., Kim, J., and Hu, Y.: PyGRF: An Improved Python Geographical Random Forest Model and Case Studies in Public Health and Natural Disasters, Transactions in GIS, 28, 2476–2491, https://doi.org/10.1111/tgis.13248, 2024.

Surasinghe, T. D., Singh, K. K., and Smart, L. S.: Leveraging Phenology to Assess Seasonal Variations of Plant Communities for Mapping Dynamic Ecosystems, Remote Sens. (Basel)., 17, https://doi.org/10.3390/rs17101778, 2025.

Tian, H., Lu, C., Yang, J., Banger, K., Huntzinger, D. N., Schwalm, C. R., Michalak, A. M., Cook, R., Ciais, P., Hayes, D., Huang, M., Ito, A., Jain, A. K., Lei, H., Mao, J., Pan, S., Post, W. M., Peng, S., Poulter, B., Ren, W., Ricciuto, D., Schaefer, K., Shi, X., Tao, B., Wang, W., Wei, Y., Yang, Q., Zhang, B., and Zeng, N.: Global patterns and controls of soil organic carbon dynamics as simulated by multiple terrestrial biosphere models: Current status and future directions, Global Biogeochem. Cycles, 29, 775–792, https://doi.org/10.1002/2014GB005021, 2015.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Ugbaje, S. U. and Bishop, T. F. A.: Hydrological control of vegetation greenness dynamics in Africa: A multivariate analysis using satellite observed soil moisture, terrestrialwater storage and precipitation, Land (Basel)., 9, https://doi.org/10.3390/land9010015, 2020.

Vågen, T., Winowiecki, L. A., Twine, W., and Vaughan, K.: Spatial Gradients of Ecosystem Health Indicators across a Human-Impacted Semiarid Savanna, J. Environ. Qual., 47, 746–757, https://doi.org/10.2134/jeq2017.07.0300, 2018.

Vaudour, E., Gomez, C., Loiseau, T., Baghdadi, N., Loubet, B., Arrouays, D., Ali, L., and Lagacherie, P.: The impact of acquisition date on the prediction performance of topsoil organic carbon from Sentinel-2 for croplands, Remote Sens. (Basel)., 11, https://doi.org/10.3390/rs11182143, 2019.

Wadoux, A. M. J. C., Minasny, B., and McBratney, A. B.: Machine learning for digital soil mapping: Applications, challenges and suggested solutions, Earth. Sci. Rev., 210, 103359, https://doi.org/10.1016/J.EARSCIREV.2020.103359, 2020.

Wadoux, A. M. J. C., Heuvelink, G. B. M., de Bruin, S., and Brus, D. J.: Spatial cross-validation is not the right way to evaluate map accuracy, Ecol. Modell., 457, https://doi.org/10.1016/j.ecolmodel.2021.109692, 2021.

Wadoux, A. M. J. C., Saby, N. P. A., and Martin, M. P.: Shapley values reveal the drivers of soil organic carbon stock prediction, SOIL, 9, 21–38, https://doi.org/10.5194/soil-9-21-2023, 2023.

Wang, Y., Khodadadzadeh, M., and Zurita-Milla, R.: A dissimilarity-adaptive cross-validation method for evaluating geospatial machine learning predictions with clustered samples, Ecol. Inform., 90, https://doi.org/10.1016/j.ecoinf.2025.103287, 2025.

Wei, H., Jia, K., Wang, Q., Cao, B., Qi, J., Zhao, W., and Yang, J.: Real-time remote sensing detection framework of the earth's surface anomalies based on a priori knowledge base, International Journal of Applied Earth Observation and Geoinformation, 122, 103429, https://doi.org/10.1016/J.JAG.2023.103429, 2023.

Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., von Lützow, M., Marin-Spiotta, E., van Wesemael, B., Rabot, E., Ließ, M., Garcia-Franco, N., Wollschläger, U., Vogel, H. J., and Kögel-Knabner, I.: Soil organic carbon storage as a key function of soils - A review of drivers and indicators at various scales, Geoderma, 333, 149–162, https://doi.org/10.1016/J.GEODERMA.2018.07.026, 2019.

Winowiecki, L., Vågen, T. G., and Huising, J.: Effects of land cover on ecosystem services in Tanzania: A spatial assessment of soil organic carbon, Geoderma, 263, 274–283, https://doi.org/10.1016/j.geoderma.2015.03.010, 2016.

Wuest, S.: Seasonal Variation in Soil Organic Carbon, Soil Science Society of America Journal, 78, 1442–1447, https://doi.org/10.2136/sssaj2013.10.0447, 2014.

Xia, Y., McSweeney, K., and Wander, M. M.: Digital Mapping of Agricultural Soil Organic Carbon Using Soil Forming Factors: A Review of Current Efforts at the Regional and National Scales, Frontiers in Soil Science, 2, https://doi.org/10.3389/fsoil.2022.890437, 2022.

Xiao, W., Xu, Y., Shi, Y., and Lei, S.: A Monte Carlo-augmented geographical random forest approach for spatial biodiversity modeling: managing bird communities in small urban wetlands, Ecol. Inform., 94, 103631, https://doi.org/10.1016/j.ecoinf.2026.103631, 2026.

https://doi.org/10.5194/egusphere-2026-4622
Preprint. Discussion started: 18 August 2026

Xue, J. and Su, B.: Significant remote sensing vegetation indices: A review of developments and applications, https://doi.org/10.1155/2017/1353691, 2017.

Yan, K., Wang, D., Feng, Y., Hou, S., Zhang, Y., and Yang, H.: Digital mapping of soil organic carbon in a plain area based on time-series features, Ecol. Indic., 171, 113215, https://doi.org/10.1016/j.ecolind.2025.113215, 2025.

Yang, P., van der Tol, C., Liu, J., and Liu, Z.: Separation of the direct reflection of soil from canopy spectral reflectance, Remote Sens. Environ., 316, 114500, https://doi.org/10.1016/J.RSE.2024.114500, 2025.

Yang, W., Seager, R., Cane, M. A., and Lyon, B.: The annual cycle of East African precipitation, J. Clim., 28, 2385–2404, https://doi.org/10.1175/JCLI-D-14-00484.1, 2015.

Yoo, M. and Koo, H.: Exploring the impact of spatial autocorrelation on optimistic bias in cross-validation and assessing the effectiveness of spatial cross-validation, Cartogr. Geogr. Inf. Sci., 52, 596–609, https://doi.org/10.1080/15230406.2024.2422593, 2025.

Zeraatpisheh, M., White, A., Darby, H., Neher, D. A., Faulkner, J., Hancock, L., Kretzler, B., El-Naboulsi, N., Turner, H. C., von Wettberg, E. J., Wagner, C. H., and Galford, G. L.: Spatial mapping and predictive modeling of soil organic carbon stocks in Vermont agricultural lands using machine learning and environmental variables, Comput. Electron. Agric., 237, 110727, https://doi.org/10.1016/j.compag.2025.110727, 2025.

Zhang, X., Xue, J., Chen, S., Wang, N., Xie, T., Xiao, Y., Chen, X., Shi, Z., Huang, Y., and Zhuo, Z.: Fine Resolution Mapping of Soil Organic Carbon in Croplands with Feature Selection and Machine Learning in Northeast Plain China, Remote Sens. (Basel)., 15, https://doi.org/10.3390/rs15205033, 2023.

Zhao, M., Yang, F., Li, W., Li, Z., Guo, Z., Chen, X., Zhang, G., Zhang, W., and Chen, T.: The pathways, mechanisms and influencing factors of cover crops in regulating soil organic carbon, Agric. Ecosyst. Environ., 402, 110309, https://doi.org/10.1016/J.AGEE.2026.110309, 2026.

Zhu, Y., Qi, L., Wu, Z., and Shi, P.: Spectra-based predictive mapping of soil organic carbon in croplands: Single-date versus multitemporal bare soil compositing approaches, Geoderma, 449, https://doi.org/10.1016/j.geoderma.2024.116987, 2024.