https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

SOIL Discussions

Continental-scale controls on soil organic carbon across sub-Saharan Africa

Sophie F. von Fromm[1,2], Alison M. Hoyt[1,3], Gifty E. Acquah[4], Ermias Aynekulu[5], Asmeret Asefaw Berhe[6], Stephan M. Haefele[4], Markus Lange[1], Steve P. McGrath[4], Keith D. Shepherd[5], Andrew M. Sila[5], Johan Six[2], Erick K. Towett[5], Susan E. Trumbore[1], Tor-G. Vågen[5], Elvis Weullow[5], Leigh A. Winowiecki[5], Sebastian Doetterl[2]

[1] Department of Biogeochemical Processes, Max-Planck Institute for Biogeochemistry, Jena, Germany
[2] Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland
[3] Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[4] Department of Sustainable Agriculture Sciences, Rothamsted Research, Harpenden, UK
[5] World Agroforestry Centre (ICRAF), Nairobi, Kenya
[6] Department of Live and Environmental Sciences, University of California Merced, Merced, CA, USA

Correspondence to: Sophie F. von Fromm, Max-Planck Institute for Biogeochemistry, Hans-Knoell-Street 10, 07745 Jena, Germany, sfromm@bgc-jena.mpg.de, phone: +49 3641-576184

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Abstract

Earlier studies have demonstrated that soil texture and geochemistry strongly affect soil organic carbon (SOC) content. However, those findings primarily rely on data from temperate regions with soil mineralogy, weathering status and climatic conditions that generally differ from tropical and sub-tropical regions.

We investigated soil properties and climate variables influencing SOC concentrations across sub-Saharan Africa. A total of 1,601 samples were analyzed, collected from two depths (0–20 cm and 20–50 cm) at 45 sentinel sites from 17 countries as part of the Africa Soil Information Service (AfSIS) project. The dataset spans climatic conditions from arid to humid and includes soils with a wide range of pH_{H20} values, weathering status, soil texture, exchangeable cations, extractable metals and a variety of important land cover types.

The most important SOC predictors were identified by linear mixed effects models, regression trees and random forest models. Our results indicate that SOC is primarily controlled by aridity index (PET/MAP), exchangeable calcium (Ca_{ex}) and oxalate-extractable aluminum (Al_{ox}); this was found across both depth intervals. Oxalate-extractable iron (Fe_{ox}) emerged as the most important predictor for both depth intervals in the regression tree and random forest analyses. However, its influence on SOC concentrations was strong only below Fe_{ox} concentrations of 0.25 wt-%. This suggests that Fe_{ox} can act as a pedogenic threshold – even on a continental scale. Across modelling approaches, clay and fine silt content (< 8 µm) and land cover were not significant SOC predictors, in contrast to common assumptions.

Our findings indicate that the key controlling factors of SOC across sub-Saharan Africa are similar to what has been reported for temperate regions – except for soil texture and vegetation cover. However, the strength and importance of the controlling factors vary across the environmental gradient we studied.

Keywords: biogeochemistry, land-use, soil organic matter, clay mineralogy, pedogenic threshold

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

1. Introduction

Soil conservation and sustainable management is crucial to address some of the main challenges that humanity is facing, such as climate change, food security, environmental degradation, and loss of soil biodiversity. Assessing the state of soils and their potential response to climate and land-use change requires complex analytical approaches with a large number of parameters at various locations (IPCC 2019). One key component is soil organic carbon (SOC). Due to the variety of sources, transformations and stabilization mechanisms, SOC is chemically very complex and spatially heterogeneous. This causes significant uncertainties in global climate models (Friedlingstein et al. 2014). It also complicates the extrapolation of SOC to a global scale, using statistical relationships to build robust global SOC products, such as SoilGrids and the Harmonized World Soil Database (Tifafi et al. 2018). Thus, to improve our understanding of global C dynamics, it is important to better understand the factors that control SOC stabilization and destabilization in soils from regional to global scales (Blankinship et al. 2018; Heimann and Reichstein 2008).

Drivers and processes that stabilize SOC have been intensively studied over the past several decades. Dokuchaev (1883) and Jenny (1941) shaped the understanding that soil properties are correlated with (independent) variables – the so-called soil-forming factors (eq. 1):

$$s = f'(cl, o, r, p, t) \qquad (1)$$

where *s* stands for any type of soil property, such as pH, carbon content, mineralogy, etc., and is determined by the function *f'* of soil-forming factors: *cl* – climate, *o* – organisms, *r* – topography, *p* – parent material, and *t* – time. This concept is still relevant and forms the basis for many experiments and research. However, the importance of the individual factors of equation (1) at different spatiotemporal scales remains unclear (Doetterl et al. 2015; Rasmussen et al. 2018; Wiesmeier et al. 2019). This uncertainty hinders the implementation of equation (1) correctly in Earth System models, resulting in a gap between the (theoretical) understanding of SOM dynamics and our ability to improve terrestrial biogeochemical projections that rely on existing models (Blankinship et al. 2018; Rasmussen et al. 2018; Schmidt et al. 2011). Despite the long history of studying SOC stabilization (Greenland 1965; Oades 1988), there still is increasing demand for data on SOC dynamics at landscape to global scales (Blankinship et al. 2018), especially from sub-tropical and tropical ecosystems.

SOC stabilization is commonly conceptualized as competition between accessibility for microorganisms versus chemical associations with minerals (Oades 1988; Schmidt et al. 2011). These aspects are usually not considered in most Earth System models (Blankinship et al. 2018; Schmidt et al. 2011). Instead, these models rely on broader variables such as clay content to describe SOC turnover rates and fluxes between soil and atmosphere (Rasmussen et al. 2018). These more generic variables integrate a variety of processes within their predictive power that are not easy to disentangle and might differ strongly across ecosystems and climate zones. Hence, improving the predictive capacity of those models requires not only a better understanding of the factors that control SOC dynamics, but also verification (or falsification) of those new findings in regions that are underrepresented in field studies and models.

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

For example, Rasmussen et al. (2018) found that exchangeable Ca was correlated with the quantity of SOC in water-limited soils, while Al_{ox} was a better predictor of SOC in wet, acidic soils. However, those findings may not be directly transferable to sub-tropical and tropical soils, since they differ greatly in climate, parent material and vegetation (Six et al. 2002b), which usually results in more weathered and older soils compared to soils from temperate regions (Feller and Beare 1997). This was illustrated recently by Quesada et al. (2020), where SOC variation in highly weathered forest soils from across the Amazon basin was best explained by clay content, whereas the best explanatory variables for less-weathered soils were Al species, pH and litter quality. Feller and Beare (1997) also found that tropical soils, dominated by low-activity clays (i.e. 1:1 clays), show a strong relationship between SOC and clay + silt content. However, the relationship for high activity clays (i.e. 2:1 clays) was less clear and contrasting trends between SOC and clay + silt content have been reported (Feller and Beare 1997; Six et al. 2002a). In terms of SOC distribution across sub-Saharan Africa, Vågen et al. (2016) showed, by using a similar dataset as in this paper, that SOC content was highest in equatorial and warm temperate climates, where sand content, sum of base concentrations and pH values were low. With regard to land cover, it has been shown for several sites in Eastern Africa that forests usually contained the highest amount of SOC, whereas differences between cropland, grassland and shrubland were marginal (Abegaz et al. 2016; Winowiecki et al. 2016a). Cropland cultivation decreased carbon content by 50% compared to forested and semi-natural plots for sites in Tanzania, regardless of sand content and topographic position (Winowiecki et al. 2016b). However, independent of vegetation cover, land degradation (i.e. erosion) resulted in SOC concentration decrease in those ecosystems (Winowiecki et al. 2016a).

To address these diverging explanations of SOC variations on regional scales, we analyzed a comprehensive soil data set collected across the African continent using the Land Degradation Surveillance Framework (Vågen et al. 2010).The dataset used in this study covers a wide range of climatic and mineralogical conditions – from very arid to humid regions, with different pH_{H2O} values, clay content, exchangeable cations and extractable metals – allowing us to test different parameters to explain the variation in SOC content in subtropical and tropical soils across sub-Saharan Africa.

The following research questions will be addressed in this study:

1. Which soil properties and climate parameters best explain SOC variation across sub-Saharan Africa?
2. Do findings from sub-Saharan Africa differ from temperate regions?
3. Do we see differences across the various climate regions and soil conditions in sub-Saharan Africa?

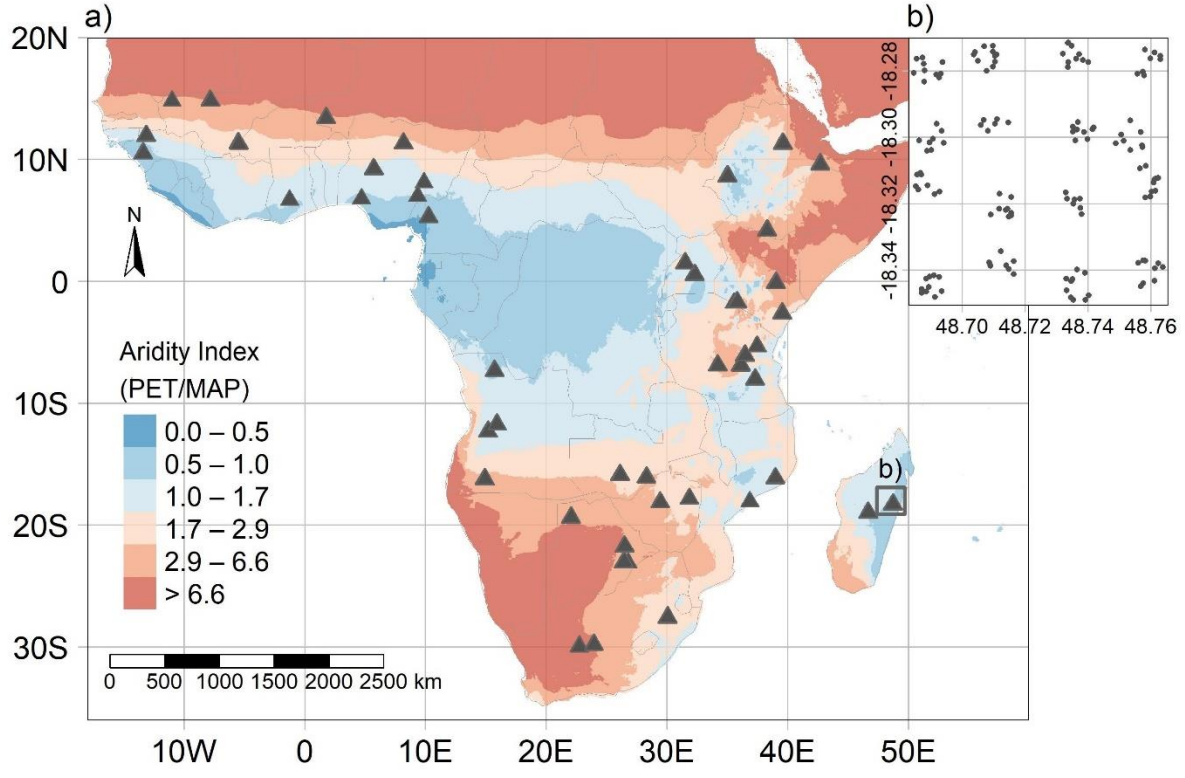
2. Methods

2.1. Study area and data collection

Soil data used in this study was collected during the AfSIS (Africa Soil Information Service) project. In total, 18,257 soil samples were taken from 60 sentinel sites and from two different depths (topsoil: 0–20 cm and subsoil: 20–50 cm). Samples

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

stem from 19 countries across sub-Saharan Africa and were collected between 2009 and 2012, following the well-established Land Degradation Surveillance Framework (Vågen et al. 2010). The 60 sentinel sites (100 km²) were stratified across sub-
100 Saharan Africa according to Koeppen-Geiger zones (Vågen et al. 2016). Ten 1000 m² plots were randomized within sixteen spatially stratified 1 km² clusters per site (Figure 1). This hierarchical sampling design allows the identification of processes at a continental scale without losing the ability to understand and quantify local heterogeneity (Vågen et al. 2010). For more details about sampling design and field survey, see Towett et al. (2015), Vågen et al. (2013a) and Winowiecki et al. (2016a).
Our analyses built upon a subset of samples (11% of total, n = 2,002) which were originally selected as reference samples for
105 lab measurements. These samples were used to calibrate mid-infrared spectroscopy data (Terhoeven-Urselmans et al. 2010) and to predict the remaining 16,255 soil samples (Vågen et al. 2016; Winowiecki et al. 2017). The calibration subset was chosen to maximize the variation of the spectral data. This selection strategy results in unequally distributed samples across 51 of the 60 sentinel sites, but still captures the variation of the original dataset.



110 **Figure 1: a) Aridity Index map and sampling scheme (n_{total} = 1,601). Grey triangles represent individual sentinel sites where clusters of samples were collected. The top-right inset (b) shows the exact sampling points within one of the sentinel sites (i.e., Didy, Madagascar) as an example.**

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

2.2. Sample and data processing

Soil material was air-dried and sieved to a particle size < 2 mm at the Soil-Plant Spectroscopy Laboratory at the World Agroforestry Centre (ICRAF) in Nairobi, Kenya. All soil properties (except for soil texture, which was measured at ICRAF), were analyzed at Rothamsted Research in Harpenden, U.K.

In order to cover a wide range of soil properties that have been identified to relate to SOC stabilization mechanisms (Oades 1988; Rasmussen et al. 2018), while maximizing the number of samples and minimizing the correlation among variables included in our analysis, the following soil parameters were selected: SOC (wt-%), pH_{H20}, amorphous oxalate-extractable aluminum (Al_{ox}, wt-%) and iron (Fe_{ox}, wt-%), exchangeable calcium (Ca_{ex}, $cmol^+/kg$), clay and fine silt content (< 8 µm, %), and total element concentrations of Al (wt-%), Ca (wt-%), K (wt-%), and Na (wt-%).

SOC was calculated from the difference of total C and inorganic C. The latter was directly measured with a Primacs AIC100 analyzer (Skalar Analytical B.V., Breda, Netherlands) by treating the sample with phosphoric acid and heating it to 135 °C in a closed system. Inorganic C in the sample was converted into CO_2 and then measured by Non-Dispersive Infrared Detection (NDIR). Total C was determined with the TruMac Total N and C combustion analyzer (Leco, St. Joseph, Michigan, USA). Soil pH_{H20} was performed in a 1:2.5 soil:water suspension. The extraction of Al and Fe with oxalic acid and ammonium oxalate solution was done by shaking the solution for 4 h at 25 °C in the dark. Carbonate-rich samples were pre-treated with ammonium acetate at pH 5.5 to remove any $CaCO_3$. Hexamine-cobalt trichloride solution was used as extractant to determine Ca_{ex}. Aqua regia acid digestion was applied for major and trace elements, including Al, Ca, K and Na. Samples were digested in tubes in time and temperature-controlled heating blocks. All extracted elements (Al_{ox}, Fe_{ox}, and Ca_{ex}) were measured with ICP-OES (Optima 7300 DV, Perkin Elmer, Waltham, Massachusetts, USA). Particle size distribution was measured using a Laser Diffraction Particle Size Analyzer (LDPSA) Model LA 950 (Horiba, Kyoto, Japan). Each sample was shaken for 4 min in a 1% sodium hexametaphosphate (calgon) solution before measuring. We used 8 µm as cut-off to capture all clay and fine silt particles. Aluminum, Ca, K and Na concentrations were used to calculate the chemical index of alteration (CIA) after Nesbit and Young (1982), using the following equation:

$$CIA = Al_2O_3 / (Al_2O_3 + CaO + K_2O + Na_2O) * 100 \quad (2)$$

where CaO is the amount incorporated in the silicate fraction. Correction is necessary for carbonates and apatite (Nesbit and Young 1982). We adopted an approach introduced by McLennan (1993): the correction assumes that Ca is typically lost more rapidly than Na during weathering. Whenever a soil sample contained inorganic C ($C_{total} - C_{org}$) and when $CaO > Na_2O$ (n = 476), the CaO concentration was set to that of Na_2O (Malick and Ishiga 2016). After applying the correction, no obvious correlation remained between CIA and inorganic C (Figure A1). The index increases (i.e. more highly-weathered soil) with the loss of Ca^{2+}, K^+, and Na^+.

Samples were removed that contained missing or negative values for one or more of the above-mentioned parameters. In addition, a single sample with extraordinarily high SOC content (> 22 wt-%) was excluded. This resulted in a total of

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

1,601 soil samples at 45 sentinel sites across 17 countries. Note that due to the sample selection, not all profiles had data from both topsoil and subsoil layers (Table B1).

The remaining soil samples (n = 1,601) were paired (based on longitude and latitude at the profile level) with mean annual temperature (MAT, °C) and mean annual precipitation (MAP, mm) from the *WorldClim* dataset at 30 sec resolution (Fick and Hijmans 2017). Potential annual evapotranspiration (PET, mm) was added from Trabucco and Zomer (2019), who calculated it after the Penman-Monteith method, based on the WorldClim data. Mean annual precipitation and PET were used to calculate an annual aridity index, defined as PET/MAP (Budyko 1974). Values > 1 indicate water-limited (dry) regions and ratios < 1 point to energy-limited (wet) regions. For the monthly aridity index, we used monthly climate data at the same spatial resolution and from the same data sources.

Land cover data was used from the collected field data. The land cover groups were re-classified into four major groups: a) Cropland (including all cultivated plots), b) Forest, c) Grassland and d) Other (including mainly woodland, shrubland, and bushland, but also samples classified as other). Ten missing values were gap-filled from a prototype high resolution Africa land-cover map at 20 m resolution based on one-year of Sentinel-2A observations from December 2015 to December 2016 (http://2016africalandcover20m.esrin.esa.int/).

2.3. Statistical analyses

Linear mixed effect modeling was performed to determine soil and climate parameters which best explain SOC variation across sub-Saharan Africa by using the *nlme* R package (Pinheiro et al. 2020). We chose linear mixed effect models because they address the non-independent nature and hierarchical structure of the clustered samples (clusters within sites) and the two paired sampling depths within one profile (Pinheiro et al. 2020). This allows the slope and intercept of the regression to vary for each site, for each cluster within the same site, and for each sample within the same profile (Harrison et al. 2018). The variance inflation factor was used to check for multi-collinearity among predictor variables with a threshold of < 3.0 (Zuur et al. 2010).

To standardize variation among variables and to meet linear mixed effect model normality assumptions, all continuous parameters were transformed to a normal distribution using Box-Cox transformation, followed by standardization to a mean of 0 and standard deviation of 1 by using the R package *bestNormalize* (Peterson and Cavanaugh 2019). The transformed and standardized data allow for direct comparison of regression-coefficient estimates among predictors – higher absolute values have a stronger relationship with the response variable (Schielzeth 2010).

In order to identify the best parameters explaining the variation of SOC, we first ran a full model, including PET/MAP, MAT, clay and fine silt content, pH_{H2O}, Al_{ox}, Fe_{ox}, Ca_{ex}, CIA, depth, land cover and the interaction term pH_{H2O}*Al_{ox} as fixed effects and site/cluster/profile as random effects. We then performed a step-wise reduction of the fixed effects based on the smallest absolute effect size (t-value) until the Akaike Information Criterion (AIC) did not improve further (Burnham and Anderson 2002; Crawley 2013). In addition, we analyzed the two sampling depths separately ($n_{Topsoil}$ = 791; $n_{Subsoil}$ = 810) to

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

determine whether the same factors were important for topsoil versus the deeper soil layer. For this model approach, we did not include profile as a random effect since each profile only contained one sample in each depth model.

In addition, we tested how SOC drivers varied across different ranges of PET/MAP, soil pH_{H2O} and weathering. We chose these three variables since they have been shown to impact other soil properties, such as Al_{ox}, Fe_{ox} and Ca_{ex} (Rasmussen et al. 2018; Quesada et al. 2020). This approach may also help to identify pedogenic thresholds that may occur under different moisture, soil pH and weathering regimes for certain soil properties. Soil pH_{H2O} and CIA data were grouped using hierarchical clustering, with the number of classes chosen to maximize the number of samples in each class and to correspond with common pH_{H2O} and weathering categories (Nesbit and Young 1982). The pH_{H2O} classes were a) strongly acid (3.89–5.22 pH_{H2O}, n = 404), b) moderately acid (5.23–6.10, n= 399), c) neutral (6.11–7.52, n = 398) and d) alkaline (7.53–9.92, n = 400), and for CIA a) moderately weathered (10.3–88.1 % CIA, n = 801) and b) strongly weathered (88.2–99.9 %, n = 800). In order to take seasonality of the sites into account, the data was divided into three groups, based on the number of wet months (i.e. months with P/PET > 1): a) no wet months (n = 572), b) 1–3 wet months (n = 367) and c) 4–7 wet months (n = 662). For each class within the three sub-categories (pH_{H2O}, CIA and number of wet months), a linear mixed effect model was built with the same fixed and random effects as for the entire dataset. Except for the two depth models, we did not include land cover in those linear mixed effects models. This is justified because the distribution of the different land cover groups was too unequal among the sub-models, and land cover was not identified as an important predictor in the full model (Table B2). Standardization and normalization of the data was performed after grouping each sub-group separately.

Regression tree (R packages: *rpart* and *rpart.plot;* Milborrow 2019; Therneau and Atkinson 2019) and random forest analyses (R packages: *ranger;* Wright and Ziegler 2017) were conducted to identify non-linear relationships between SOC and any explanatory variable. This also enabled the identification of pedogenic thresholds within the data. Each analysis was conducted with the same explanatory variables as for the linear mixed effect models. However, no data transformation was needed due to the non-linearity of the models.

Regression tree analysis was applied to obtain an easily interpretable and non-linear model for the entire dataset and for both depth layers that best describes the existing data (Breiman et al. 1984). Since regression trees are known to easily overfit data, we used a grid search to prune the model (Boehmke and Greenwell 2020) regarding the minimum number of data points required to attempt a split, and the maximum number of internal nodes between the root node and terminal nodes in order to minimize the cross-validation error (Breiman et al. 1984). The overall performance of the regression tree analyses was tested using five-fold spatial cross-validation (R package: *mlr;* Bischl et al. 2016). Spatial partitioning has been used to split the data into five disjoint subsets, using the coordinates from each sample, and repeating the partitioning 100 times (Figure A2). This results in a bias-reduced assessment of model performance (Brenning 2012; Lovelace et al. 2019).

Random forest was used to build more generalized models since it is an ensemble of multiple decorrelated trees. Tuning of the model hyperparameters was done based on spatial tuning (R package: *mlr;* Bischl et al. 2016; Lovelace et al. 2019). These hyperparameters included the number of predictors used at each split, the minimum number of observations in a

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

terminal node and the fraction of samples used in each tree (Probst et al. 2019). The best hyperparameter combination search was done for the complete dataset via a five-fold spatial cross-validation with one repetition (Lovelace et al. 2019).

Partial dependence plots were used to further explore the relationship between the predicted SOC content and the explanatory variables of the tuned random forest models (R package: *pdp;* Greenwell 2017). These plots were used to investigate the marginal effect of individual explanatory variables (such as Al_{ox}, Ca_{ex}, etc) on the predicted SOC content (Friedman 2001). This makes it possible to identify thresholds within the data. They can also provide an approximation of how important an explanatory variable is within a specific range in order to predict SOC concentration. All statistical analyses were performed within the R computing environment (Version 4.0.0; R Core Team 2020).

3. Results

3.1. Data distribution across sub-Saharan Africa

All soil and climate variables spanned at least one order of magnitude (except MAT and PET), demonstrating the diversity of this continent-wide data set. All variables, based on skewness, kurtosis, histograms, and Shapiro-Wilk-tests (data not shown for the latter two), were not normally distributed (Table 1).

Table 1: Summary statistics of all numerical soil and climate variables (n = 1,601)

Variable	Mean	SD	P0	P25	P50	P75	P100	Skewness	Kurtosis
SOC [wt-%]	1.84	1.51	0.07	0.65	1.42	2.54	9.19	1.42	2.23
MAT [°C]	21.7	3.2	13.7	19.8	21.5	23.0	29.8	0.17	-0.12
MAP [mm]	1070	487	255	648	1057	1432	2708	0.29	-0.63
PET [mm]	1810	310	1350	1571	1759	1933	2949	1.19	1.96
PET/MAP	2.35	1.73	0.71	1.2	1.54	3.16	9.54	1.46	1.31
Clay + silt [%]	55.4	22.6	0.1	37.7	57.9	74.7	100.0	-0.26	-1.00
Al_{ox} [wt-%]	0.28	0.36	0.01	0.12	0.20	0.29	3.71	4.52	25.29
Fe_{ox} [wt-%]	0.38	0.56	0.01	0.10	0.21	0.40	4.46	3.60	14.96
Ca_{ex} [$cmol^{+}/kg$]	10.29	11.01	0.03	1.34	5.86	16.49	75.66	1.28	1.32
pH_{H2O}	6.3	1.3	3.9	5.2	6.1	7.5	9.9	0.27	-1.11
CIA [%]	87.7	9.3	10.3	81.7	88.1	96.0	99.9	-1.04	3.88

SD: Standard deviation; P: Percentile; SOC: Soil organic carbon; MAT: Mean annual temperature; MAP: Mean annual precipitation; PET: Potential evapotranspiration; Al_{ox}: Oxalate-extractable Al; Fe_{ox}: Oxalate-extractable Fe; Ca_{ex}: Exchangeable Ca; CIA: Chemical Index of Alteration

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

In total, 429 samples were classified as cropland, 228 as forest, 242 as grassland and 702 as other land-covers, including mainly shrubland, bushland and woodland. The SOC content decreased among those groups in the following sequence: Forest (2.69 ± 1.15 wt-%) > Cropland (2.21 ± 1.68 wt-%) > Grassland (1.77 ± 1.55 wt-%) > Other (1.35 ± 1.28 wt-%; Figure 2a). Clay and fine silt content and SOC showed a positive relationship across the entire dataset, yet with a large spread (Figure 2b). However, individual sites showed contrasting correlations between SOC and clay and fine silt content – including, none, positive, and negative (Figure 2c; Figure A3 for all individual sites).

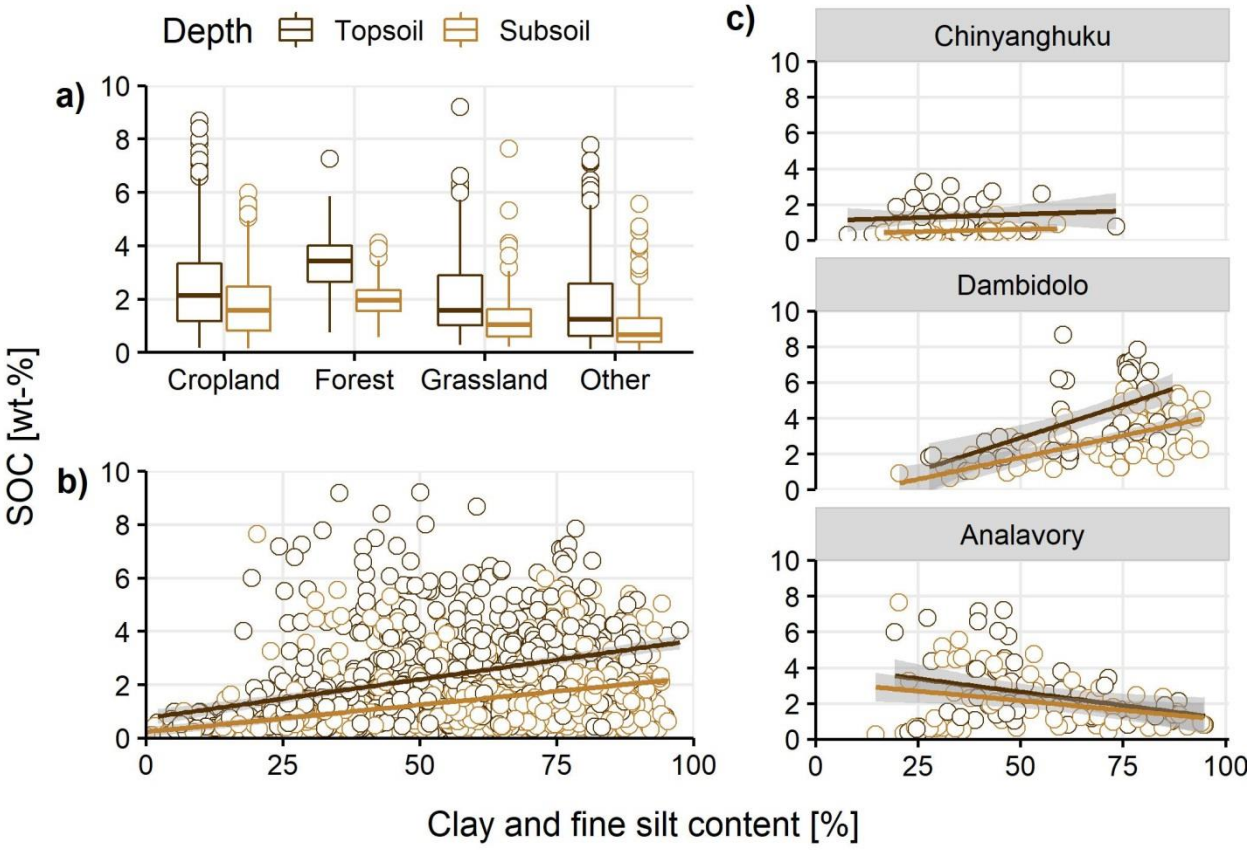


Figure 2: a) Soil organic carbon (SOC) content [wt-%] for the different land-covers (cropland, forest, grassland, other (bushland, shrubland, woodland) by depth (topsoil: 0–20 cm, subsoil: 20–50 cm); b) SOC [wt-%] and clay and fine silt content [%] by depth for the entire dataset; c) SOC [wt-%] and clay and fine silt content [%] by depth for three example sites that show contrasting trends. Gray area around fitted linear regressions in b) and c) show the 95% confidence interval. For the relationship between SOC and clay and fine silt content for all individual sites, see Figure A3.

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

3.2. Predictors of soil organic carbon

Linear mixed effect modelling

The final linear mixed effect model for the entire dataset (n = 1,601) had a marginal R^2 of 0.71. Ca_{ex} was the strongest predictor of SOC content, followed by aridity index (PET/MAP) and soil pH_{H2O} (Table 2). After accounting for those three variables, Al_{ox} was the strongest predictor, followed by depth (subsoil vs topsoil), and the weathering index (CIA). MAT and pH_{H2O}*Al_{ox} were the weakest SOC predictors. Ca_{ex} and Al_{ox} were positively correlated with SOC, whereas all other parameters showed a negative relationship with SOC concentration (Table 2). The negative coefficient for depth indicates that the SOC content in the subsoil layer is in general lower as compared with the topsoil samples. Based on the AIC of the individual models and their p-values, clay and fine silt content, Fe_{ox} and land cover were not included in the final model (Table B3). However, forest and other (shrubland, woodland, bushland) land covers differed significantly in terms of SOC content from croplands, whereas grasslands did not (Table B2).

Table 2: Final linear mixed effect model results of the standardized and normalized fixed effect parameters for the entire data set (n = 1,601)

Variable	Estimates	Lower CI	Upper CI	DF	p-value
Ca_{ex}	0.54	0.49	0.60	344	< 0.001
Aridity Index (PET/MAP)	-0.42	-0.53	-0.31	786	< 0.001
pH_{H2O}	-0.42	-0.48	-0.37	344	< 0.001
Al_{ox}	0.31	0.28	0.35	344	< 0.001
Depth (Subsoil)	-0.31	-0.34	-0.29	344	< 0.001
CIA	-0.24	-0.28	-0.20	344	< 0.001
MAT	-0.17	-0.26	-0.08	786	< 0.001
pH_{H2O}*Al_{ox}	-0.16	-0.19	-0.12	344	< 0.001
Intercept	-0.08	-0.20	0.04	786	0.196

CI: 95% confidence interval; DF: Degree of freedom; Ca_{ex}: Exchangeable Ca; PET: Potential evapotranspiration; MAP: Mean annual precipitation; Al_{ox}: Oxalate extractable Al; CIA: Chemical Index of Alteration; MAT: Mean annual temperature

Differences between the predictors were marginal for the two depth models (Figure A4). Only pH_{H2O} and the weathering index (CIA) showed significant differences between topsoil and subsoil. Clay and fine silt content only had a negative relationship with SOC in the topsoil, yet the coefficient was relatively weak (-0.07). Land cover predictors were again not significant, yet differences among the land cover groups in terms of SOC content were larger in topsoils as compared with subsoils (Table B4).

Grouping by pH_{H2O} (Figure 3a) was used to test if the explanatory power of the fixed effects differed across soil pH_{H2O}. Al_{ox} was most important in acidic soils and less or not important in neutral to alkaline soils. Ca_{ex}, however, was not only

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

important in alkaline and neutral soils, but had an even stronger influence in moderately acidic soils (Figure 3c). In neutral soils, clay and fine silt content had a positive coefficient. The interaction term pH_{H2O}*Al_{ox} was not significant in any of the pH_{H2O} models (Table B5). Overall, only four parameters were significant in alkaline soils, whereas seven parameters were
265 significant in the other three pH_{H2O} models.

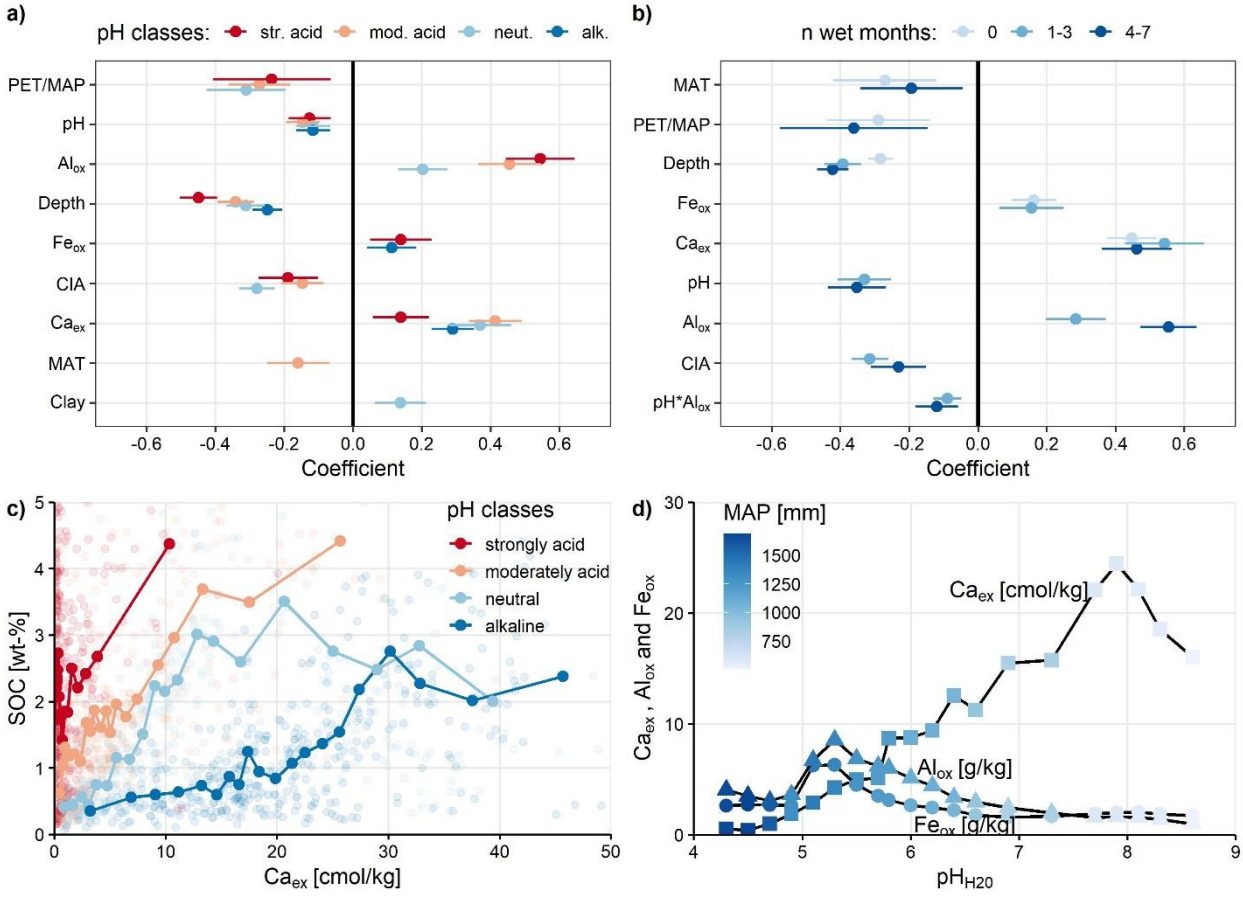


Figure 3: a) Results of the final linear mixed effect pH models (strongly acid: 3.9–5.2 pH, moderately acid: 5.2–6.1, neutral: 6.1–7.5, alkaline: 7.5–9.9); b) Results of the final linear mixed effect number of wet months models; c) SOC [wt-%] and Ca_{ex} [$cmol^+/kg$] content colored by pH classes with a moving average (bold points; n = 20). Note that x-axis is truncated for visual reasons; d) Al_{ox},
270 **Fe_{ox} [g/kg] and Ca_{ex} [$cmol^+/kg$] averaged content (n = 20) across pH_{H2O} and mean annual precipitation [mm]. Only relevant predictors (based on AIC and p-values) are shown for the two linear mixed effect models (a and b).**

Based on the number of wet months (P/PET > 1), Al_{ox}, pH_{H2O}, CIA, and pH_{H2O}*Al_{ox} were more important in wetter regions and Fe_{ox} and MAT in drier regions (Figure 3b). Again, Ca_{ex} was equally important across all three groups. Clay and fine silt content were not important in any of the three models (Table B6).

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

275 The strongly-weathered (CIA: 88–100%) model was dominated by Al_{ox}, while Ca_{ex} and Fe_{ox} were more important in less-weathered soils (Figure 4). Mean annual temperature was only important in highly-weathered soils compared to less-weathered soils. Clay and fine silt content were not important in either model (Table B7).

In summary, in the linear mixed effect models, Al_{ox} was more important in wetter regions, acid and highly weathered soils, whereas Ca_{ex} was more important in drier regions, alkaline and less weathered soils. However, Ca_{ex} also played a
280 considerable role in wetter regions and moderately acid and highly weathered soils.

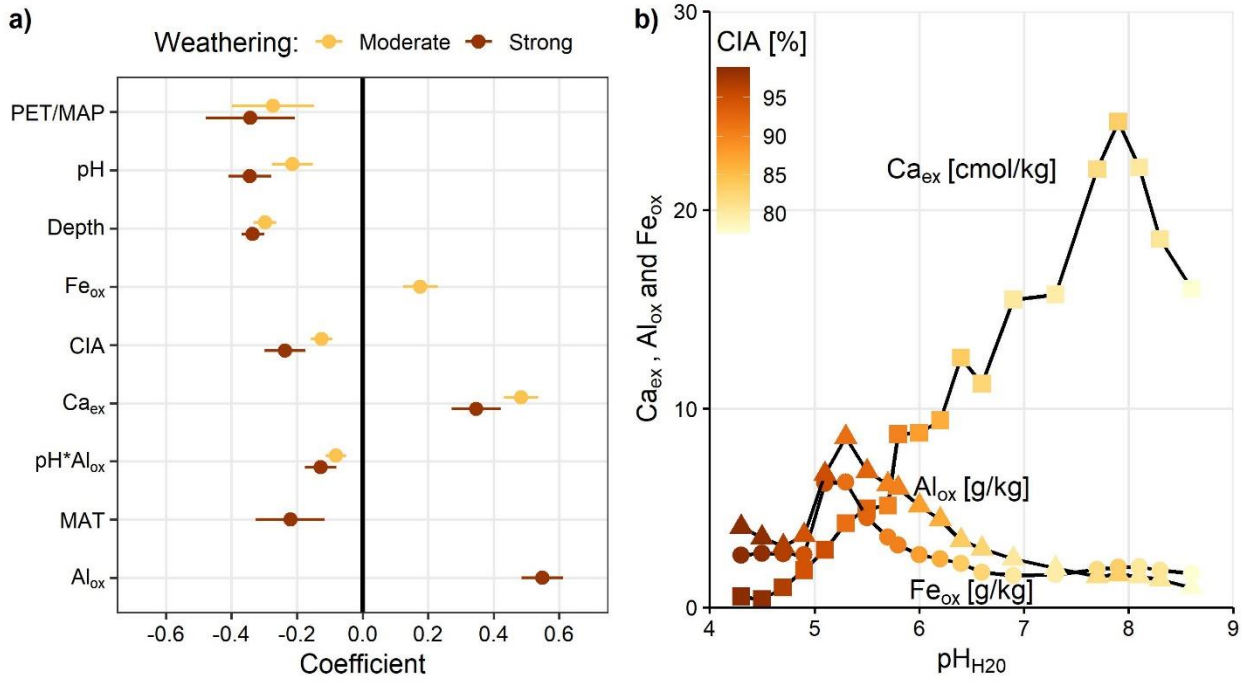


Figure 4: a) Results of the final linear mixed effect weathering (CIA) models (moderate: 10–88%, high: 88-100%); b) Ca_{ex} [$cmol^+/kg$] and Al_{ox}, Fe_{ox} [wt-%] averaged content (n = 20), respectively, colored by weathering status. Only relevant predictors (based on AIC and p-values) are shown.

285 *Regression tree and random forest*

The root median square error (RMSE) for the topsoils was 1.47 wt-% (range: 0.80–3.11 wt-%) and for the subsoils 0.67 wt-% (range: 0.44–2.26 wt-%); the relative RMSEs were 0.65% and 0.48%, respectively. In the topsoil regression tree (Figure 5b) Fe_{ox}, MAT and PET/MAP were the most important predictors to split and explain variation in SOC concentration. About 23% of the SOC variation could be explained by Fe_{ox} and MAT alone. In general, lower Fe_{ox}, Al_{ox} and
290 Ca_{ex} values resulted in lower SOC content. This was equally true for the subsoil tree (Figure A5b). While much of the SOC variation was explained by climate parameters in topsoils, the subsoil regression tree was more dominated by geochemical variables, namely Fe_{ox} and Al_{ox}. About 40% of the subsoil SOC variation could be explained by Fe_{ox} only. In both trees, clay and fine silt content and land cover poorly predicted SOC.

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

In summary, topsoil and subsoil regression trees contained the same predictors, yet the topsoil regression tree seemed to be more dominated by climate variables and the subsoil regression tree more by geochemistry. However, the results showed that the explanatory variables did not differ much between the depth intervals, while their magnitude did.

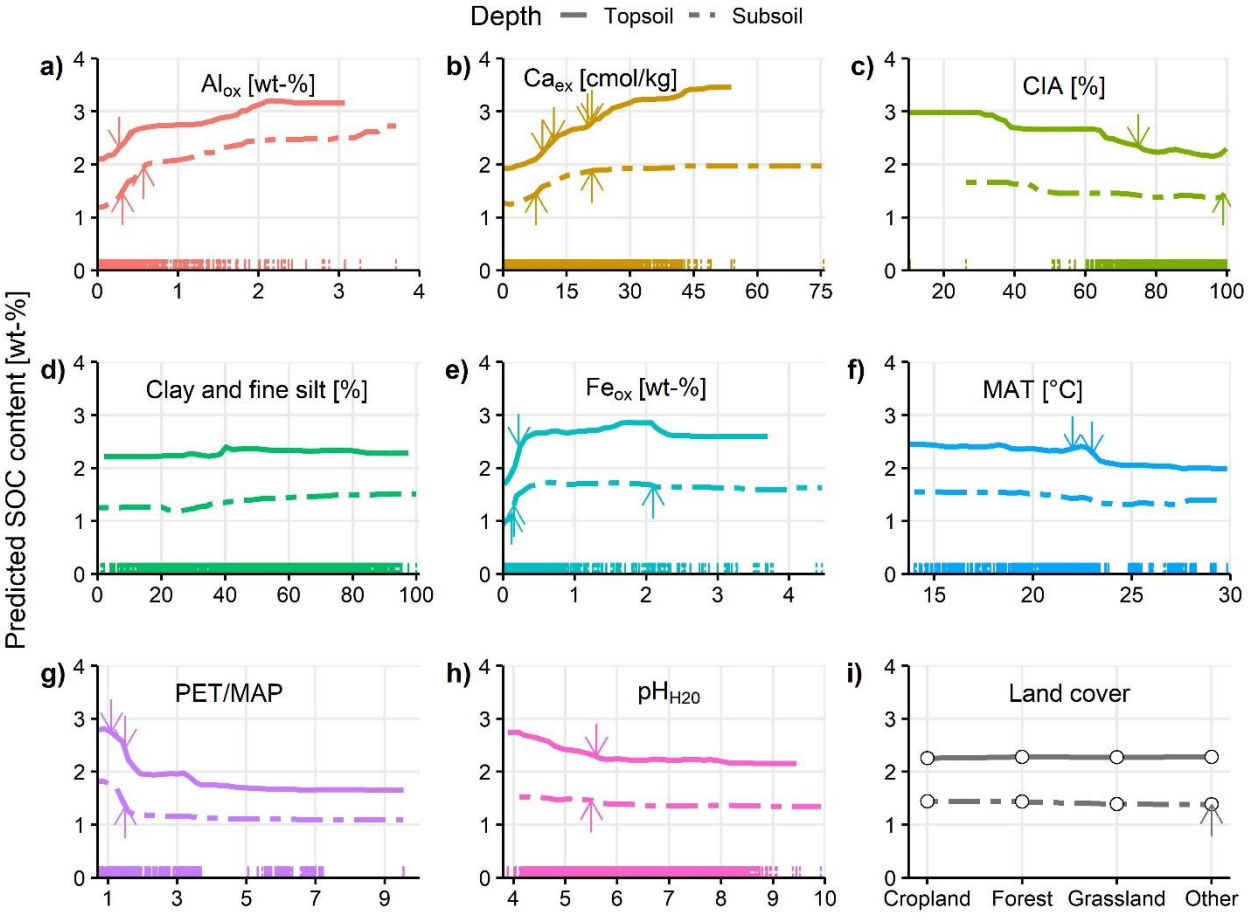


Figure 5: Partial dependence plot for each explanatory variable of the random forest models (topsoil and subsoil). X-axis always corresponds to the range of the explanatory variable, respectively. Arrows indicate splitting points in the regression tree (Figure A5). Each colored tick mark along the x-axis represents one sample.

The random forest models showed a RMSE of 1.31 wt-% and a R^2 of 0.70 for the topsoil samples, and a RMSE of 0.87 wt-% and a R^2 of 0.72 in the subsoils. Based on the partial dependence plots (Figure 5), Al_{ox} and Ca_{ex} were important in predicting SOC over the entire range of each variable (Figure 5a and b). However, in subsoils, the predictive power of Ca_{ex} is reduced – the predicted SOC content is relatively uniform above Ca_{ex} concentrations of about 20 $cmol^+/kg$ (Figure 5b). We also observed a decrease in predicted SOC with increasing soil weathering status (CIA) – this trend was relatively constant over the entire CIA range. However, due the low number of samples with CIA values below 60%, the relationship should be

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

interpreted with caution in this range (Figure 5c). Clay and fine silt content also had almost no effect on SOC, with only a weak positive trend in subsoil samples (Figure 5d). The relationship between Fe_{ox} concentration and predicted SOC content varied with Fe_{ox} concentration. At low concentrations (< 0.25 wt-%), there was a strong positive relationship between predicted SOC content and Fe_{ox}. For higher concentrations, the predicted SOC was relatively constant; in the topsoils, predicted SOC even decreased as Fe_{ox} concentrations increased above 2.2 wt-% (Figure 5e). MAT correlated negatively over the entire range with predicted SOC concentration, with a relatively sharp decrease above 22 °C in topsoils (Figure 5f). For PET/MAP, the predicted SOC declined sharply as PET/MAP increased from 1 to 2 (transition from wet to dry water regimes), and decreased again in the topsoils at about 3.5 (Figure 5g). The relationship between pH_{H2O} and predicted SOC content was not strong. The steepest decline was between pH_{H2O} 4 and 6 (Figure 5h). For land cover, there was almost no difference between the classes within the same depth layer; however, topsoils had higher SOC content (2.2 wt-%) compared with the subsoil samples across all land covers (1.5 wt-%; Figure 5i).

4. Discussion

As shown with the linear mixed effect models, the sub-groups number of wet months, pH_{H2O} and weathering classes (Figure 3 and 4) emerged as key parameters that influence how other parameters such as Ca_{ex}, Al_{ox} and Fe_{ox} explained SOC content variation across sub-Saharan Africa. In contrast, the differences in predictors were much smaller across depth classes. This may be due to the fact that we had only two depth layers (topsoil: 0–20 cm and subsoil 20–50 cm). In these two depth layers, the factors controlling SOC content were similar and any differences were probably mostly driven by the fact that the subsoil samples usually contain less SOC due to lower C inputs at depth (Jobbágy and Jackson 2000).

These findings were supported by the regression trees (Figure A5) and partial dependence plots (Figure 5), where Ca_{ex}, Al_{ox} and Fe_{ox} seemed to be more important in explaining the variation of SOC concentration compared to pH_{H2O}, PET/MAP and CIA. For example, soil pH_{H2O} was important in the full linear mixed effect model, yet it mainly influenced Ca_{ex}, Al_{ox}, and Fe_{ox} concentrations in correlation with MAP (Figure 3d); the same was true for weathering (Figure 4b). Similar relationships have been found for temperate regions, where the importance of Ca_{ex} increased with increasing pH_{H2O} and decreasing precipitation, whereas the opposite was true for Al_{ox} (Oades 1988; Rasmussen et al. 2018). In this particular study from Rasmussen et al. (2018), Fe_{ox} and clay content were not found to be important in explaining SOC variation. In contrast, Quesada et al. (2020) found that clay content and Al_{ox} both explained the variation of SOC in differently weathered soils in the Amazon basin. Soil texture, along with rainfall and soil management, was also found by Feller (1993) to be important in determining organic matter content in low activity soils clay (i.e. 1:1 clay) in various tropical soils.

In the following discussion, we focus on the variables that showed the most explanatory power in terms of SOC content across all models and compare their explanatory power with those reported in other studies for different regions. In addition, we discuss the role of clay and fine silt content and land cover since they have been found to be important in other studies.

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Exchangeable Calcium

Strong and positive relationships emerged between Ca_{ex} and SOC concentration across all models, even though Ca_{ex} concentration showed strong pH_{H20} and precipitation dependence (Figure 3). Typical Ca^{2+} sources in soils are from a) weathering of bedrock or surface rock formations, b) decomposition of Ca^{2+}-rich organic materials, c) lateral movement of Ca^{2+}-rich water, d) atmospheric dust and rain deposition or e) anthropogenic inputs (Likens et al. 1998; Rowley et al. 2018). Characteristically, Ca^{2+} is weathered easily from both primary and secondary minerals (Likens et al. 1998). This usually leads to its accumulation in semi-arid to arid environments that are characterized by low rates of water flow through the soil profile that drives slow weathering rates and high pH_{H2O} values (Figure 4). In such environments, Ca^{2+} plays an important role as a cation bridge that facilitates aggregate formation (Rimmer and Greenland 1976; Tisdall and Oades 1982) and bonding of clay minerals to organic matter functional groups because of their divalent charge, relative abundance and modest hydration radius (Likens et al. 1998; Muneer and Oades 1989). However, we found that Ca_{ex} was not only important in alkaline and less-weathered soils in dry regions, but also in acidic and more-weathered soils under wetter conditions (Figure 3). It is likely that the main Ca^{2+} source in those regions derives from atmospheric deposition (Albani et al. 2015; Goudie and Middleton 2001) and/or biological cycling by plants (Likens et al. 1998). This is supported by the fact that Ca_{ex} showed a stronger relationship with SOC in topsoil than subsoil layers (Figure 5b). Since land cover, which is a major driver of C inputs into the soil, did not show a strong relationship with SOC in the models, we speculate that biological cycling of Ca^{2+} does not play a major role in explaining the observed differences in SOC content. Yet, further analysis with better proxies for biological Ca^{2+} inputs is needed to test this hypothesis. High Ca^{2+} concentrations in acid soils can also be derived from the development of those soils from Ca^{2+}-rich parent material which are out-of-equilibrium with modern climate conditions (Slessarev et al. 2016).

In conclusion, whenever Ca_{ex} was in the soil system, it had a strong and positive relationship with SOC, independent of acidity and weathering status of the soil, as well as the available amount of water in the region. This can be linked to the bridging of clays and organic material by Ca^{2+} to form organo-mineral complexes that are physically, chemically and biologically stable (Muneer and Oades 1989; Oades 1988; Rowley et al. 2018).

Oxalate Al and Fe

Similar to Ca_{ex}, Al_{ox} showed a positive and strong correlation with SOC content across all models. The relationship was strongest in wet regions, acid and highly weathered soils (Figure 3 and 4). Hydrous oxides of Al and Fe are usually highly reactive because of their large specific areas with a high proportion of reactive sites (Parfitt and Childs 1988), which results in the adsorption of organic matter to Fe and Al oxides and the formation of stable soil aggregates (Tisdall and Oades 1982). Acid-oxalate extraction in particular dissolves short range-order minerals such as ferrihydrite (Fe), allophane and imogolite (Al), as well as other amorphous and organic Fe and Al minerals (Parfitt and Childs 1988), which are known to be important in SOC stabilization (Theng 1976; Torn et al. 1997). In humid regions, high rates of mineral weathering may release Fe, Al

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

and Si faster than crystalline minerals can precipitate (Rasmussen et al. 2018). Therefore, Fe_{ox} and Al_{ox} are usually found to be important in SOC stabilization in humid and acid soils (Eusterhues et al. 2003; Kramer and Chadwick 2018).

This is also true for Al_{ox} in sub-Saharan Africa. However, we could only find small predictive power of Fe_{ox} in acidic soils and wet regions (Figure 3a and b). Yet, Fe_{ox} was one of the most important explanatory variables in the regression tree and partial dependence plots, although only within a very narrow range and at low Fe_{ox} concentrations (Figure 5e). Inagaki et al. (2020) showed that higher amounts of soil organic matter were co-localized with Fe in drier regions compared to sites with higher rainfall, whereas the content of Al co-localized with organic matter was not affected by precipitation changes. This may be linked to the different oxidation levels of Fe. At higher precipitation levels, Fe oxides can be reduced, resulting in a release of associated SOC to the aqueous phase (Berhe et al. 2012; Chen et al. 2020; Thompson et al. 2011). This mechanism is probably responsible for the low correlation between SOC and high Fe_{ox} concentrations in our data (Figure 5e), pointing to the fact that Fe_{ox} can act as pedogenic threshold, depending on its oxidation level in the soil system. Interestingly, Al_{ox} and Fe_{ox} behave differently in SOC stabilization across sub-Saharan Africa, even though they show similar trends in their concentrations (Figure 3d and Figure 4b). However, since we only have data for acid-oxalate extraction, we cannot further speculate about their diverging behavior in the models.

Clay and fine silt content

Even though clay and fine silt content were never an important predictor of SOC concentration within our different models, we discuss them here because earlier studies indicated that total clay content explains a large proportion of SOC storage and stabilization due to the sorption of soil organic matter to surfaces of clay minerals and building of aggregates (e.g. Amelung et al. 1998; Kahle et al. 2002). Furthermore, this relationship is used in various Earth System models to describe turnover and storage of SOC. However, this simplified correlation may not account for the different stabilization mechanisms related to various clay minerals, e.g. 1:1 vs 2:1 clay minerals (Oades 1988). There are contradictory results on whether clay content explains the variation in subtropical and tropical soils or not. Bruun et al. (2010) showed for various tropical soils that clay mineralogy, Fe_{ox} and Al_{ox} are better explanatory variables for SOC content compared to clay content alone. In contrast, Quesada et al. (2020) found a strong relationship between clay and SOC content for highly weathered soils in the Amazon Basin that are dominated by 1:1 clay minerals such as kaolinite, whereas soils in the same system, dominated by 2:1 clay minerals, showed stronger relationship between SOC and Al species. In a comparison between tropical and temperate soils, Six et al. (2002b) found that less C was associated with the clay and silt fraction in tropical soils than in temperate soils.

Due to the broad spatial scale, soils in the AfSIS dataset contain different clay minerals (Butler et al. 2020). No clear relationship between clay and fine silt content and SOC concentration was observed in the models, although the raw data indicates overall a positive trend between clay and fine silt content and SOC concentration (Figure 2b). This overall positive relationship does not hold across all sites (Figure 2c) and variable relationships across individual sites may explain the low predictive power of clay and fine silt content in this dataset. Instead, variables that better capture the different behavior of clay-sized minerals, e.g. Ca_{ex}, Fe_{ox} and Al_{ox}, are likely more suitable soil parameters to explain the variation of SOC content

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

– even in highly weathered soils across sub-Saharan Africa. This is supported by the fact that a clay and fine silt-only model resulted in a very small R^2 (linear mixed effect model: 0.10; random forest: 0.12; Table B8).

Land cover

Surprisingly, land cover was not an important predictor of SOC content in our models, especially in topsoils. One possibility may be that the relatively large 0–20 cm depth interval might dilute differences that would be more marked in the top few centimeters. However, we did observe differences in SOC content across land cover classes, with forests containing the highest amount of SOC – especially in topsoils (Figure 2a). Croplands had higher SOC content than grasslands, opposite of what is commonly observed in temperate regions (Prout et al. 2020). Another possible explanation for the lack of land cover as an important predictor in our models, is that land cover covaries with other parameters (temperature, precipitation, geochemistry) to such a degree that it is not an explanatory variable. However, the land cover-only models resulted in small R^2 (linear mixed effect model: 0.10; random forest: 0.10–0.16) which suggests that land cover is a poor predictor for our SOC data in general (Table B8). This might be due to the high variation of SOC content within the different land cover classes at this large spatial scale (Figure 2a). Overall, our data for sub-Saharan Africa suggests that SOC content is better explained by stabilization potential in soils (climate, geochemistry) than by different aboveground C inputs (vegetation). More work is needed to test this hypothesis. For example, land management and land degradation are known to impact SOC stocks (Winowiecki et al. 2016a), and likely also explain some of the variation of SOC within the data.

5. Conclusions

In summary, we used a continental-scale dataset from sub-Saharan Africa to test relationships between SOC and various soil properties and climate variables. This enabled us to address our core research questions:

Which soil properties and climate parameters best explain SOC variation across sub-Saharan Africa? Do findings from sub-Saharan Africa differ from temperate regions?

We have shown for tropical and subtropical soils under various climate conditions across sub-Saharan Africa that parameters similar to temperate regions are important to explain SOC variation, namely Ca_{ex}, Al_{ox}, and PET/MAP and to some extend also pH_{H2O}. However, land cover and clay and fine silt content did not explain much of the variation in SOC content, which is opposite to some findings from other regions and studies.

Do we see differences across the various climate regions and soil conditions in sub-Saharan Africa?

In dry regions with alkaline and less weathered soils, Ca_{ex} explained most of the variation of SOC concentration, whereas Al_{ox} was more important in wetter regions with acid and highly weathered soils. Ca_{ex} was still important in acidic and more weathered soils and wetter regions. Fe_{ox} as a predictor of SOC content was only important at low concentrations in moderately weathered and wet soils. This observed trend leads to the assumption that Fe_{ox} can play an important role in pedogenic thresholds in various soils across sub-Saharan Africa.

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Overall, a combination of soil pH_{H2O}, PET/MAP, Ca_{ex} and Al_{ox} seems to be an appropriate set of variables to explain variation of SOC content on a continental scale across sub-Saharan Africa. This does not imply that other variables, such as clay and fine silt content and land cover are not good predictors on a regional scale. However, the variables identified by this study showed a consistent predictive power of SOC content across various climate regions. Future studies on SOC stabilization on large scales should consider measuring those soil properties so that they can be included in global Earth System models. This would help to improve the predictive capacity of these models and to close the gap between our (theoretical) understanding of SOC dynamics and our ability to improve terrestrial biogeochemical projections that rely on existing models.

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Code availability

As a R markdown file (pdf) in the supplement materials.

Dataset availability

The lab data used in this study is available from the corresponding author upon reasonable request. Field data (i.e. land cover) for the sampling locations can be received from Vågen et al. (2013b). The climate data used (MAT, MAP and PET) can be downloaded from the sources cited: WorldClim: Fick and Hijmans (2017) and Trabucco and Zomer (2019). Land-cover data that has been used for gap-filling, can be received from http://2016africalandcover20m.esrin.esa.int/.

Author contribution

Conceptualization of the study for this manuscript was done by SvF, AH, AAB, ST, and SD, with input from EA, SH, SMG, KS, JS, TGV and LW. Data curation, investigation and resources were done and provided by GA, EA, SH, SMG, KS, AS, ET, TGV, EW and LW. The formal analysis, methodology, and visualization for the manuscript was performed by SvF with substantial input from AH, ML, SD and ST as well as feedback from all authors. SvF wrote the initial draft and all authors were involved in the review and editing of the manuscript.

Competing interest

SD and AAB are liaison editors of the special issue *Tropical biogeochemistry of soils in the Congo Basin and the African Great Lakes region* and JS is executive editor of the SOIL journal. However, none of them was involved in the review process of this manuscript. All other authors declare that they have no conflict of interest.

Acknowledgement

SvF receives funding from the International Max-Planck Research School for Global Biogeochemical Cycles. ST and AH acknowledge support from the European Research Council (Horizon 2020 Research and Innovation Program, grant agreement 695101; 14Constraint). SD receives supportive funds through DFG Emmy Noether Group “TropSOC” (project number: 387472333). The analytical data used in the study was produced by the “Chemical and Biological Assessment of AfSIS soils” project, funded by Biotechnology and Biological Sciences Research Council (BBSRC)/Global Challenges Research Fund (GCRF) (BBS/OS/GC/000014B). SPM, SH and GA are partly funded by the Institute Strategic Program (ISP) grants, “Soils to Nutrition” (S2N; grant number BBS/E/C/000I0310). Original field surveys and sample analysis costs at ICRAF were covered by the AfSIS Phase I project funded by the Bill and Melinda Gates Foundation Grant Number 51353.

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

References

Abegaz A, Winowiecki LA, Vågen T-G, Langan S, Smith JU (2016) Spatial and temporal dynamics of soil organic carbon in landscapes of the upper Blue Nile Basin of the Ethiopian Highlands. *Agriculture, Ecosystems & Environment,* 218: pp. 190-208, doi: 10.1016/j.agee.2015.11.019.

Albani S, Mahowald NM, Winckler G, Anderson RF, Bradtmiller LI, Delmonte B, François R, Goman M, Heavens NG, Hesse PP, Hovan SA, Kang SG, Kohfeld KE, Lu H, Maggi V, Mason JA, Mayewski PA, McGee D, Miao X, Otto-Bliesner BL, Perry AT, Pourmand A, Roberts HM, Rosenbloom N, Stevens T, Sun J (2015) Twelve thousand years of dust: the Holocene global dust cycle constrained by natural archives. *Climate of the Past,* 11(6): pp. 869-903, doi: 10.5194/cp-11-869-2015.

Amelung W, Zech W, Zhang X, Follett RF, Tiessen H, Knox E, Flach K-W (1998) Carbon, Nitrogen, and Sulfur Pools in Particle-Size Fractions as Influenced by Climate. *Soil Science Society of America Journal,* 62(1): pp. 172-181, doi: 10.2136/sssaj1998.03615995006200010023x.

Berhe AA, Suttle KB, Burton SD, Banfield JF (2012) Contingency in the direction and mechanics of soil organic matter responses to increased rainfall. *Plant and Soil,* 358(1): pp. 371-383, doi: 10.1007/s11104-012-1156-0.

Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, Casalicchio G, Jones ZM (2016) mlr: Machine Learning in R. *Journal of Machine Learning Research,* 17(170): pp. 1–5, doi.

Blankinship JC, Berhe AA, Crow SE, Druhan JL, Heckman KA, Keiluweit M, Lawrence CR, Marín-Spiotta E, Plante AF, Rasmussen C, Schädel C, Schimel JP, Sierra CA, Thompson A, Wagai R, Wieder WR (2018) Improving understanding of soil organic matter dynamics by triangulating theories, measurements, and models. *Biogeochemistry,* 140(1): pp. 1-13, doi: 10.1007/s10533-018-0478-2.

Boehmke B, Greenwell BM (2020) *Hands-On Machine Learning with R.* Chapman and Hall/CRC: Boca Raton, Florida.

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) *Classification and Regression Trees.* Taylor & Francis, pp.°368.

Brenning A. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. 2012 IEEE International Geoscience and Remote Sensing Symposium, 22-27 July 2012 2012. 5372-5375.

Bruun TB, Elberling B, Christensen BT (2010) Lability of soil organic carbon in tropical soils with different clay minerals. *Soil Biology and Biochemistry,* 42(6): pp. 888-895, doi: 10.1016/j.soilbio.2010.01.009.

Budyko MI (1974) *Climate and Life.* Academic Press: New York, U.S.A., pp.°508.

Burnham KP, Anderson DR (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* Springer: New York, U.S.A., pp.°488.

Butler BM, Palarea-Albaladejo J, Shepherd KD, Nyambura KM, Towett EK, Sila AM, Hillier S (2020) Mineral–nutrient relationships in African soils assessed using cluster analysis of X-ray powder diffraction patterns and compositional methods. *Geoderma,* 375: pp. 114474, doi: 10.1016/j.geoderma.2020.114474.

Chen C, Hall SJ, Coward E, Thompson A (2020) Iron-mediated organic matter decomposition in humid soils can counteract protection. *Nature Communications,* 11(1): pp. 2255, doi: 10.1038/s41467-020-16071-5.

Crawley MJ (2013) *The R Book, 2nd Edition.* Wiley: Chichester, United Kingdom, pp.°1076.

Doetterl S, Stevens A, Six J, Merckx R, van Oost K, Casanova Pinto M, Casanova-Katny A, Muñoz C, Boudin M, Zagal Venegas E, Boeckx P (2015) Soil carbon storage controlled by interactions between geochemistry and climate. *Nature Geoscience,* 8(10): pp. 780-783, doi: 10.1038/ngeo2516.

Dokuchaev VV (1883) Russian Chernozem. Report to the Imperial Free Economic Society (Tipogr. Declerona i Evdokimova) [in Russian]. St. Petersburg, Russia, pp.

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Eusterhues K, Rumpel C, Kleber M, Kögel-Knabner I (2003) Stabilisation of soil organic matter by interactions with minerals as revealed by mineral dissolution and oxidative degradation. *Organic Geochemistry,* 34(12): pp. 1591-1600, doi: 10.1016/j.orggeochem.2003.08.007.

Feller C (1993) Organic inputs, soil organic matter and functional soil organic compartments in low-activity clay soils in tropical zones. In: Mulongoy K & Merckx R (eds.) *Soil Organic Matter and Dynamics and Sustainability of Tropical Agriculture.* John Wiley & Son: Chichester, U.K., pp. 77-88.

Feller C, Beare MH (1997) Physical control of soil organic matter dynamics in the tropics. *Geoderma,* 79(1): pp. 69-116, doi: 10.1016/S0016-7061(97)00039-6.

Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology,* 37(12): pp. 4302-4315, doi: 10.1002/joc.5086.

Friedlingstein P, Meinshausen M, Arora VK, Jones CD, Anav A, Liddicoat SK, Knutti R (2014) Uncertainties in CMIP5 Climate Projections due to Carbon Cycle Feedbacks. *Journal of Climate,* 27(2): pp. 511-526, doi: 10.1175/jcli-d-12-00579.1.

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *Annals of Statistics,* 29(5): pp. 1189-1232, doi: 10.1214/aos/1013203451.

Goudie AS, Middleton NJ (2001) Saharan dust storms: nature and consequences. *Earth-Science Reviews,* 56(1): pp. 179-204, doi: 10.1016/S0012-8252(01)00067-8.

Greenland DJ (1965) Interaction between clays and organic compounds in soils. Part II Adsorption of soil organic compounds and its effect on soil properties. *Soils and Fertilizers,* 28: pp. 415-425.

Greenwell BMW (2017) pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal,* 9(1): pp. 421-436, doi: 10.32614/RJ-2017-016.

Harrison XA, Donaldson L, Correa-Cano ME, Evans J, Fisher DN, Goodwin CED, Robinson BS, Hodgson DJ, Inger R (2018) A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ,* 6: pp. e4794-e4794, doi: 10.7717/peerj.4794.

Heimann M, Reichstein M (2008) Terrestrial ecosystem carbon dynamics and climate feedbacks. *Nature,* 451(7176): pp. 289-292, doi: 10.1038/nature06591.

Inagaki TM, Possinger AR, Grant KE, Schweizer SA, Mueller CW, Derry LA, Lehmann J, Kögel-Knabner I (2020) Subsoil organo-mineral associations under contrasting climate conditions. *Geochimica et Cosmochimica Acta,* 270: pp. 244-263, doi: 10.1016/j.gca.2019.11.030.

IPCC (2019) *Climate Change and Land, an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems.* IPCC: Geneva, Switzerland.

Jenny H (1941) *Factors of soil formation – a system of quantitative pedology.* McGraw-Hill: New York, USA.

Jobbágy EG, Jackson RB (2000) The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecological Applications,* 10(2): pp. 423-436, doi: 10.1890/1051-0761(2000)010[0423:Tvdoso]2.0.Co;2.

Kahle M, Kleber M, Jahn R (2002) Carbon storage in loess derived surface soils from Central Germany: Influence of mineral phase variables. *Journal of Plant Nutrition and Soil Science,* 165(2): pp. 141-149, doi: 10.1002/1522-2624(200204)165:2<141::Aid-jpln141>3.0.Co;2-x.

Kramer MG, Chadwick OA (2018) Climate-driven thresholds in reactive mineral retention of soil carbon at the global scale. *Nature Climate Change,* 8(12): pp. 1104-1108, doi: 10.1038/s41558-018-0341-4.

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Likens GE, Driscoll CT, Buso DC, Siccama TG, Johnson CE, Lovett GM, Fahey TJ, Reiners WA, Ryan DF, Martin CW, Bailey SW (1998) The biogeochemistry of calcium at Hubbard Brook. *Biogeochemistry*, 41(2): pp. 89-173, doi: 10.1023/A:1005984620681.

Lovelace R, Nowosad J, Muenchow J (2019) *Geocomputation with R.* Chapman and Hall/CRC: Boca Raton, Florida, USA, pp.°335.

Malick BML, Ishiga H (2016) Geochemical Classification and Determination of Maturity Source Weathering in Beach Sands of Eastern San' in Coast, Tango Peninsula, and Wakasa Bay, Japan. *Earth Science Research*, 5(1): pp. 44-56, doi: 10.5539/esr.v5n1p44.

McLennan SM (1993) Weathering and Global Denudation. *Journal of Geology*, 101(2): pp. 295-303, doi: 10.1086/648222.

Milborrow S (2019) *rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'*. https://CRAN.R-project.org/package=rpart.plot.

Muneer M, Oades J (1989) The role of Ca-organic interactions in soil aggregate stability .III. Mechanisms and models. *Soil Research*, 27(2): pp. 411-423, doi: 10.1071/SR9890411.

Nesbit HW, Young GM (1982) Early Proterozoic climates and plate motions inferred from major element chemistry of lutites. *Nature*, 299: pp. 715-717, doi: 10.1038/299715a0.

Oades JM (1988) The retention of organic matter in soils. *Biogeochemistry*, 5(1): pp. 35-70, doi: 10.1007/BF02180317.

Parfitt R, Childs C (1988) Estimation of forms of Fe and Al - a review, and analysis of contrasting soils by dissolution and Mossbauer methods. *Soil Research*, 26(1): pp. 121-144, doi: 10.1071/SR9880121.

Peterson RA, Cavanaugh JE (2019) Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics*: pp. 1-16, doi: 10.1080/02664763.2019.1630372.

Pinheiro J, Bates D, Debroy S, Sarkar D, R Core Team (2020) *nlme: Linear and Nonlinear Mixed Effects Models.* https://CRAN.R-project.org/package=nlme.

Probst P, Wright MN, Boulesteix A-L (2019) Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, 9(3): pp. e1301, doi: 10.1002/widm.1301.

Prout JM, Shepherd KD, McGrath SP, Kirk GJD, Haefele SM (2020) What is a good level of soil organic matter? An index based on organic carbon to clay ratio. *European Journal of Soil Science*: pp. 1-11, doi: 10.1111/ejss.13012.

Quesada CA, Paz C, Oblitas Mendoza E, Phillips OL, Saiz G, Lloyd J (2020) Variations in soil chemical and physical properties explain basin-wide Amazon forest soil carbon concentrations. *SOIL*, 6(1): pp. 53-88, doi: 10.5194/soil-6-53-2020.

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

Rasmussen C, Heckman K, Wieder WR, Keiluweit M, Lawrence CR, Berhe AA, Blankinship JC, Crow SE, Druhan JL, Hicks Pries CE, Marin-Spiotta E, Plante AF, Schädel C, Schimel JP, Sierra CA, Thompson A, Wagai R (2018) Beyond clay: towards an improved set of variables for predicting soil organic matter content. *Biogeochemistry*, 137(3): pp. 297-306, doi: 10.1007/s10533-018-0424-3.

Rimmer DL, Greenland DJ (1976) Effects of Calcium carbonate on the swelling behaviour of a soil clay. *Journal of Soil Science*, 27(2): pp. 129-139, doi: 10.1111/j.1365-2389.1976.tb01983.x.

Rowley MC, Grand S, Verrecchia ÉP (2018) Calcium-mediated stabilisation of soil organic carbon. *Biogeochemistry*, 137(1): pp. 27-49, doi: 10.1007/s10533-017-0410-1.

Schielzeth H (2010) Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, 1(2): pp. 103-113, doi: 10.1111/j.2041-210X.2010.00012.x.

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Schmidt MWI, Torn MS, Abiven S, Dittmar T, Guggenberger G, Janssens IA, Kleber M, Kögel-Knabner I, Lehmann J, Manning DAC, Nannipieri P, Rasse DP, Weiner S, Trumbore SE (2011) Persistence of soil organic matter as an ecosystem property. *Nature*, 478: pp. 49, doi: 10.1038/nature10386.

Six J, Conant RT, Paul EA, Paustian K (2002a) Review: Stabilization mechanisms of soil organic matter: Implications for C-saturation of soils. *Plant and Soil*, 241(2): pp. 155-176, doi.

Six J, Feller C, Denef K, Ogle SM, De Moraes Sa JC, Albrecht A (2002b) Soil organic matter, biota and aggregation in temperate and tropical soils - Effects of no-tillage. *Agronomie*, 22(7-8): pp. 755-775, doi: 10.1051/agro:2002043.

Slessarev EW, Lin Y, Bingham NL, Johnson JE, Dai Y, Schimel JP, Chadwick OA (2016) Water balance creates a threshold in soil pH at the global scale. *Nature*, 540: pp. 567, doi: 10.1038/nature20139.

Terhoeven-Urselmans T, Vagen T-G, Spaargaren O, Shepherd KD (2010) Prediction of Soil Fertility Properties from a Globally Distributed Soil Mid-Infrared Spectral Library. *Soil Science Society of America Journal*, 74(5): pp. 1792-1799, doi: 10.2136/sssaj2009.0218.

Theng BKG (1976) Interactions between montmorillonite and fulvic acid. *Geoderma*, 15(3): pp. 243-251, doi: 10.1016/0016-7061(76)90078-1.

Therneau T, Atkinson B (2019) *rpart: Recursive Partitioning and Regression Trees.* https://CRAN.R-project.org/package=rpart.

Thompson A, Rancourt DG, Chadwick OA, Chorover J (2011) Iron solid-phase differentiation along a redox gradient in basaltic soils. *Geochimica et Cosmochimica Acta*, 75(1): pp. 119-133, doi: 10.1016/j.gca.2010.10.005.

Tifafi M, Guenet B, Hatté C (2018) Large Differences in Global and Regional Total Soil Carbon Stock Estimates Based on SoilGrids, HWSD, and NCSCD: Intercomparison and Evaluation Based on Field Data From USA, England, Wales, and France. *Global Biogeochemical Cycles*, 32(1): pp. 42-56, doi: 10.1002/2017gb005678.

Tisdall JM, Oades JM (1982) Organic matter and water-stable aggregates in soils. *Journal of Soil Science*, 33(2): pp. 141-163, doi: 10.1111/j.1365-2389.1982.tb01755.x.

Torn M, Trumbore S, Chadwick O, Vitousek P, Hendricks D (1997) Mineral Control of Soil Organic Carbon Storage and Turnover. *Nature*, 389: pp. 170-173, doi: 10.1038/38260.

Towett EK, Shepherd KD, Tondoh JE, Winowiecki LA, Lulseged T, Nyambura M, Sila A, Vågen T-G, Cadisch G (2015) Total elemental composition of soils in Sub-Saharan Africa and relationship with soil forming factors. *Geoderma Regional*, 5: pp. 157-168, doi: 10.1016/j.geodrs.2015.06.002.

Trabucco A, Zomer R (2019) Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v2. Figshare, doi: 10.6084/m9.figshare.7504448.v3.

Vågen T-G, Shepherd KD, Walsh MG, Winowiecki L, Desta LT, Tondoh JE (2010) AfSIS Technical Specifications – Soil Health Surveillance [Version 1.0]. Nairobi, Kenya, pp. 69.

Vågen T-G, Winowiecki LA, Abegaz A, Hadgu KM (2013a) Landsat-based approaches for mapping of land degradation prevalence and soil functional properties in Ethiopia. *Remote Sensing of Environment*, 134: pp. 266-275, doi: 10.1016/j.rse.2013.03.006.

Vågen T-G, Winowiecki LA, Tondoh JE, Desta LT (2013b) Africa Soil Information Service (AfSIS) - Soil Health Mapping. Harvard Dataverse, doi: 10.7910/DVN/2JUBRA.

Vågen T-G, Winowiecki LA, Tondoh JE, Desta LT, Gumbricht T (2016) Mapping of soil properties and land degradation risk in Africa using MODIS reflectance. *Geoderma*, 263: pp. 216-225, doi: 10.1016/j.geoderma.2015.06.023.

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, Kögel-Knabner I (2019) Soil organic carbon storage as a key function of soils – A review of drivers and indicators at various scales. *Geoderma*, 333: pp. 149-162, doi: 10.1016/j.geoderma.2018.07.026.

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Winowiecki L, Vågen T-G, Huising J (2016a) Effects of land cover on ecosystem services in Tanzania: A spatial assessment of soil organic carbon. *Geoderma*, 263: pp. 274-283, doi: 10.1016/j.geoderma.2015.03.010.

Winowiecki L, Vågen T-G, Massawe B, Jelinski NA, Lyamchai C, Sayula G, Msoka E (2016b) Landscape-scale variability of soil health indicators: effects of cultivation on soil organic carbon in the Usambara Mountains of Tanzania. *Nutrient Cycling in Agroecosystems*, 105(3): pp. 263-274, doi: 10.1007/s10705-015-9750-1.

Winowiecki LA, Vågen T-G, Boeckx P, Dungait JAJ (2017) Landscape-scale assessments of stable carbon isotopes in soil under diverse vegetation classes in East Africa: application of near-infrared spectroscopy. *Plant and Soil*, 421(1): pp. 259-272, doi: 10.1007/s11104-017-3418-3.

Wright MN, Ziegler A (2017) ranger: A fast implementation fo random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1): pp. 1-17, doi: 10.18637/jss.v077.i01.

Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, 1(1): pp. 3-14, doi: 10.1111/j.2041-210X.2009.00001.x.

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Annex: Electronic supplement

Appendix A – Figures

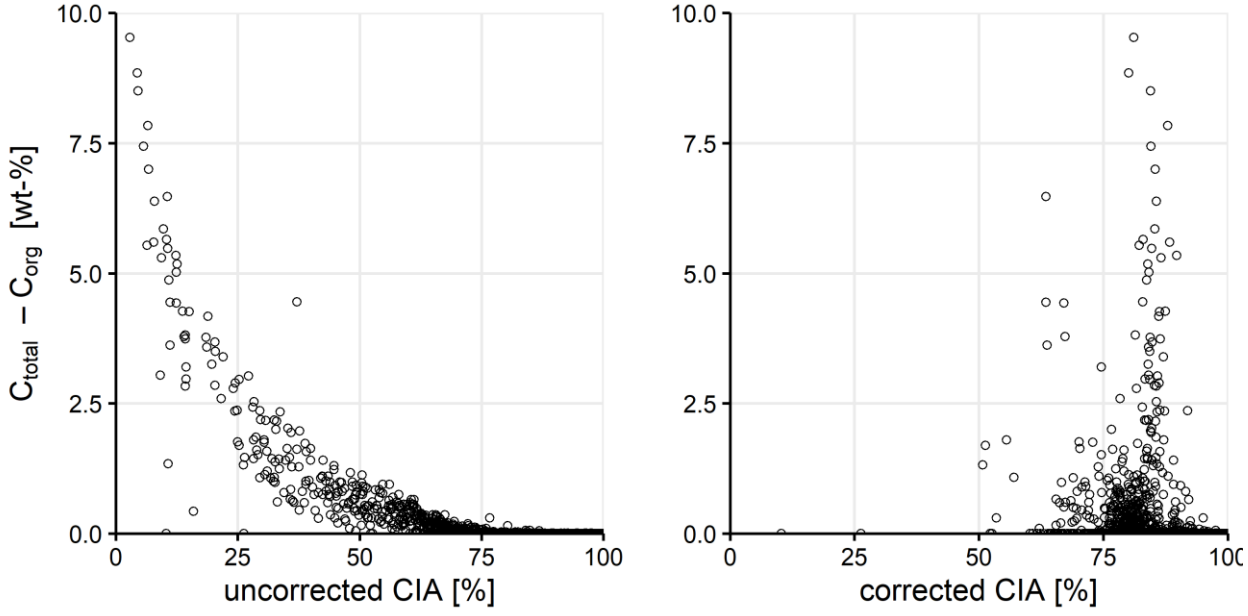


Figure A1: Scatterplot of inorganic carbon ($C_{total} - C_{org}$ [wt-%]), the uncorrected chemical index of alteration (CIA [%]; left) and the CIA [%] correct for carbonates and apatite after Nesbit and Young (1982) (right). See *methods* for more details.

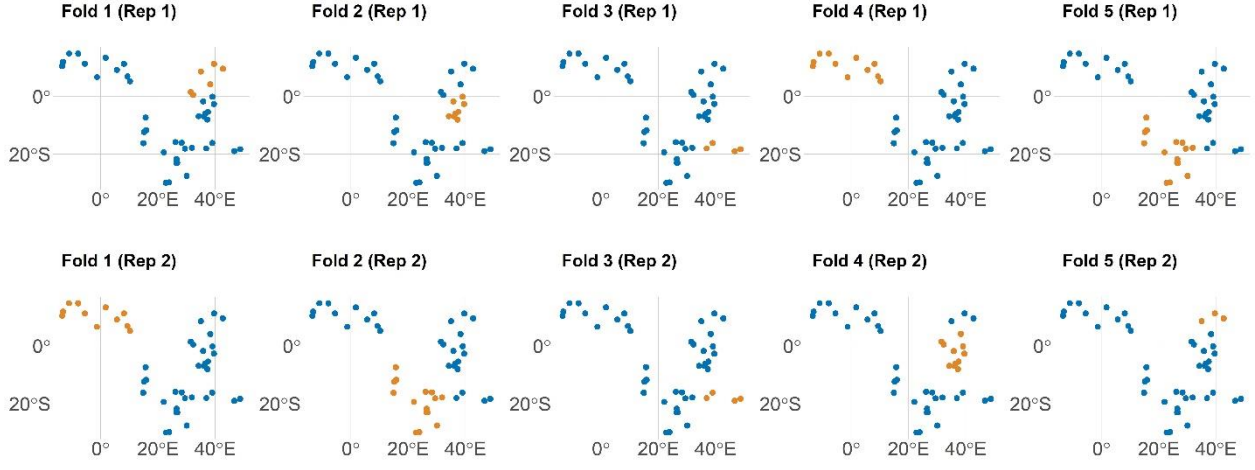


Figure A2: Spatial visualization of selected training (blue) and test (orange) observations for spatial cross-validation of two repetitions from the topsoil samples. Note: Each dot may represent multiple samples.

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

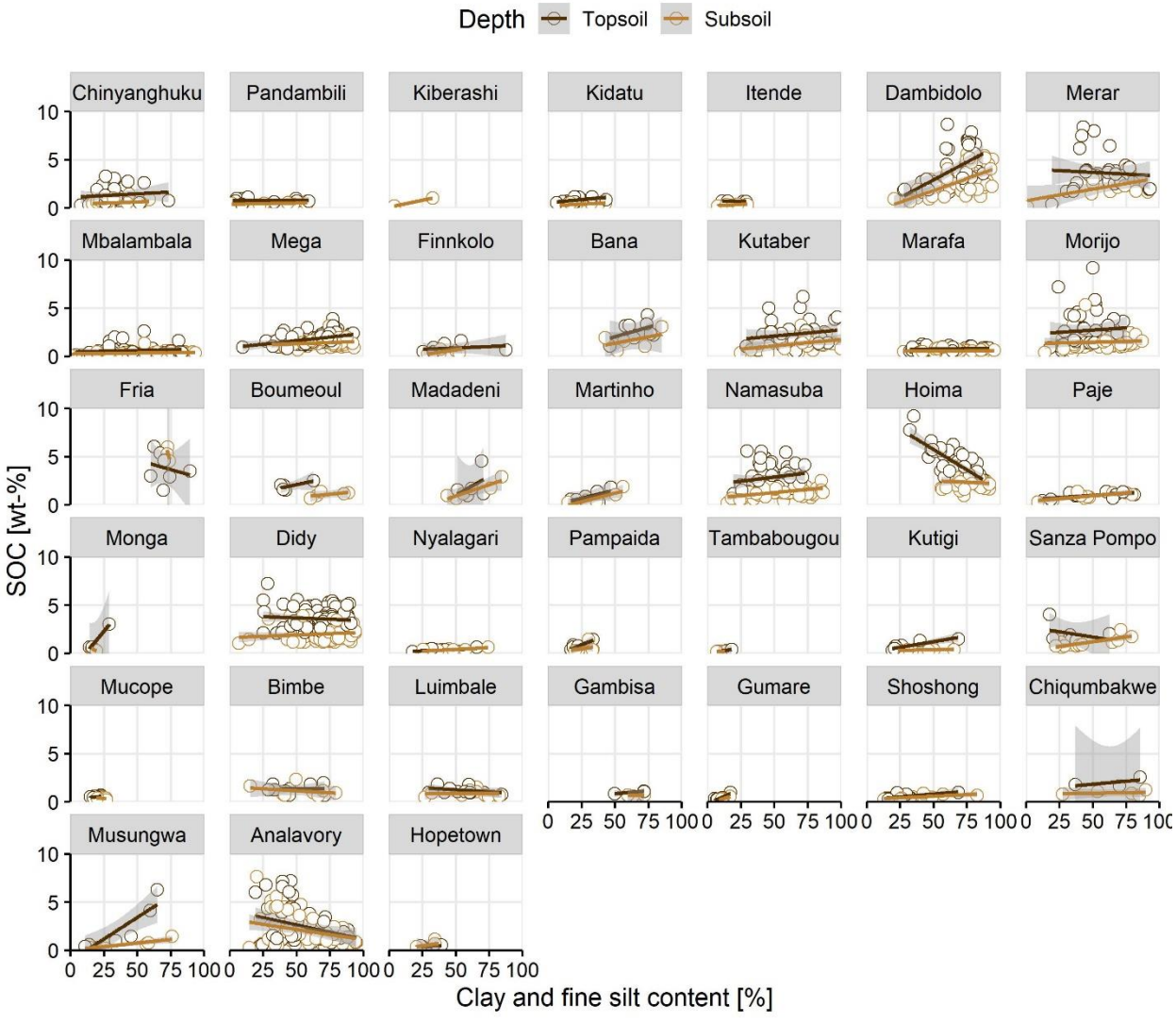
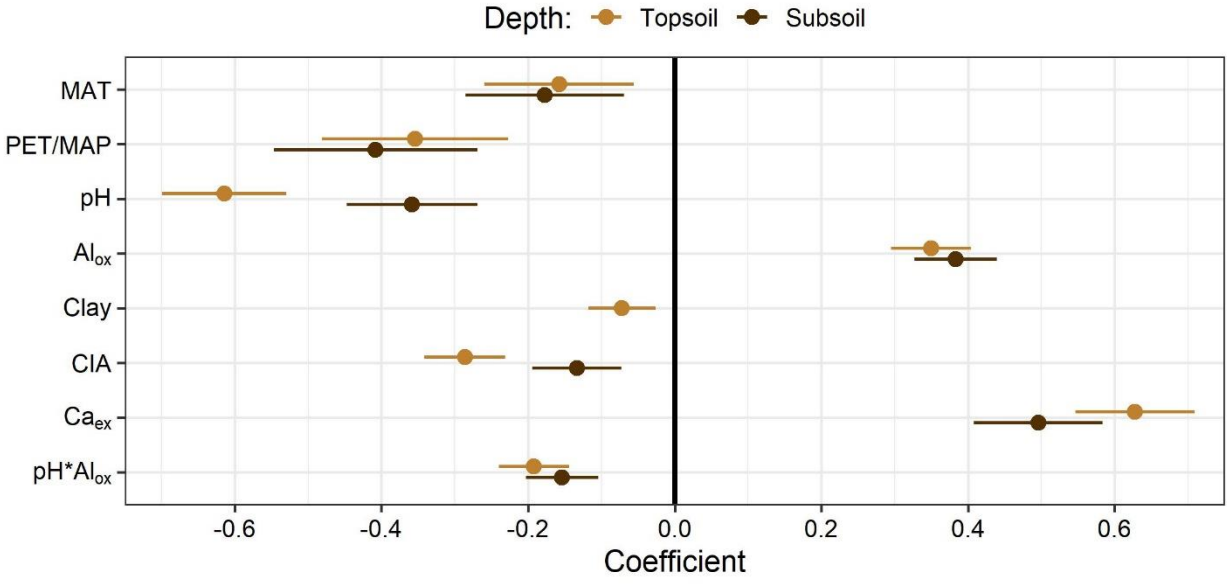


Figure A3: Soil organic carbon (SOC) [wt-%] and clay and fine silt content [%] by depth for each sampling site that contained more than one sample per depth layer (topsoil: 0-20 cm, subsoil: 20-50 cm). Gray area around fitted linear regressions represent the 95% confidence interval.

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020



660 **Figure A4: Linear mixed effect model results by depth (topsoil: 0–20 cm, subsoil: 20–50 cm) with 95% confidence intervals.**

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

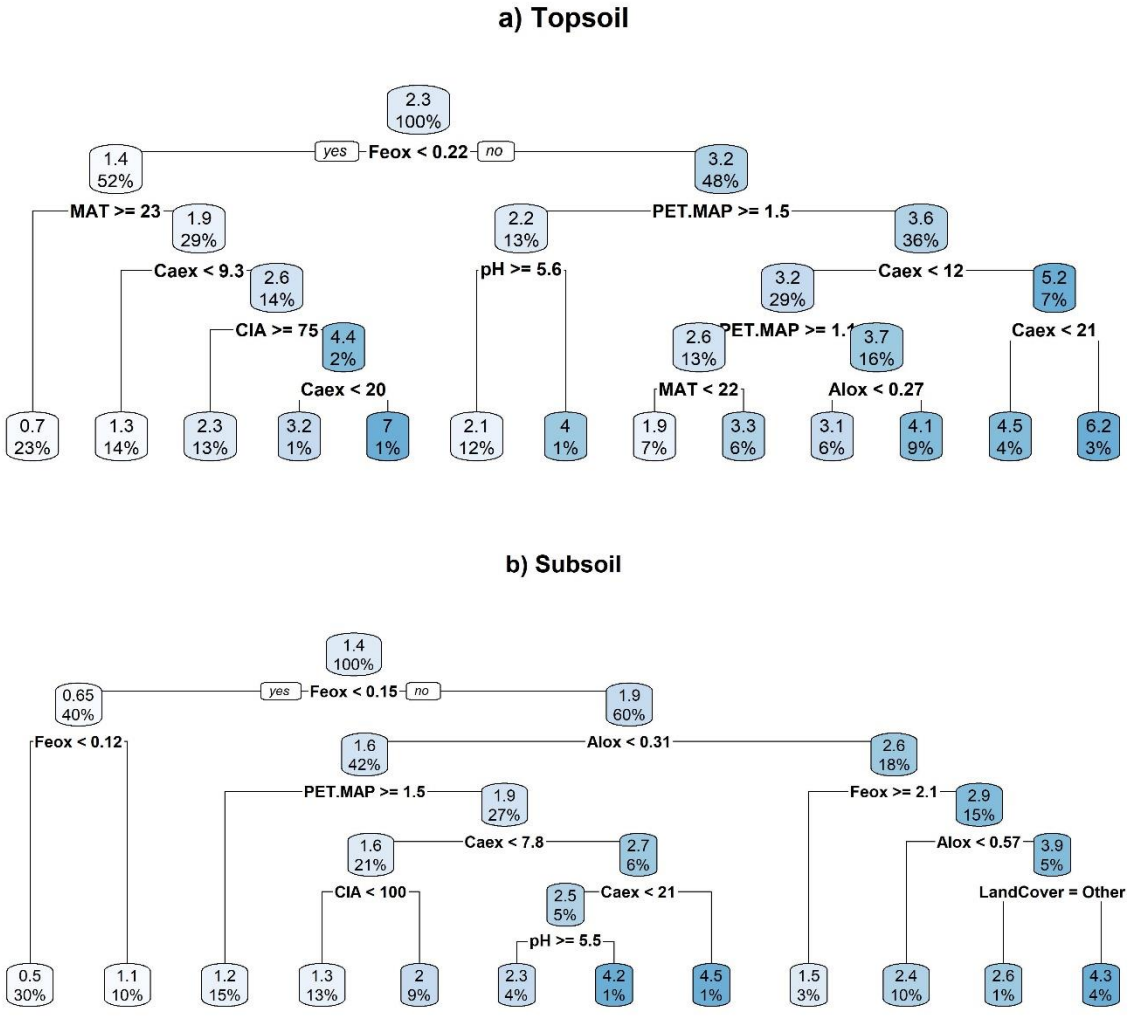


Figure A5: Regression tree for a) Topsoil (0-20 cm) and b) Subsoil (20-50 cm). Splitting values are always in the units of the parameter used for the split (for units see Table 1). Absolute values in the boxes indicate the predicted soil organic carbon (SOC) content [wt-%]. The percentage corresponds to the relative number of samples.

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Appendix B – Tables

Table B1: Overview of sample distribution used in this study across geographical region, countries, sites, depths and land cover

Region	Country	n Site	n Depth		n Land cover			
			Topsoil	Subsoil	Forest	Cropland	Grassland	Other
East	TZA	5	61	54	6	16	13	80
	ETH	4	179	165	3	153	56	132
	KEN	3	131	153	5	4	55	220
	UGA	2	99	101	0	90	29	81
	MDG	2	161	175	206	86	20	24
West	NGA	5	16	19	1	15	5	14
	MLI	3	11	14	1	9	6	9
	CMR	1	8	6	2	10	2	0
	GIN	2	12	8	1	9	1	9
	NER	1	13	11	0	12	0	12
	GHA	1	1	0	1	0	0	0
South	ZAF	3	11	11	0	0	7	15
	MOZ	2	7	6	0	4	3	6
	BWA	3	29	26	0	2	11	42
	ZMB	2	10	9	1	2	13	3
	AGO	4	36	44	1	14	17	48
	ZWE	2	6	8	0	3	4	7

TZA: Tanzania; ETH: Ethiopia; KEN: Kenya; UGA: Uganda; MDG: Madagascar; NGA: Nigeria; MLI: Mali; CMR: Cameroon; GIN: Guinea; NER: Niger; GHA: Ghana; ZAF: South Africa; MOZ: Mozambique; BWA: Botswana; ZMB: Zambia; AGO: Angola; ZWE: Zimbabwe

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Table B2: Full linear mixed effect model results of the standardized and normalized fixed effect parameters for the entire data set (n = 1,601)

Variable	Estimates	Lower CI	Upper CI	DF	p-value
Ca_{ex}	0.54	0.48	0.60	342	< 0.001
PET/MAP	-0.41	-0.52	-0.30	783	< 0.001
pH_{H2O}	-0.41	-0.47	-0.35	342	< 0.001
Depth (Subsoil)	-0.31	-0.33	-0.29	342	< 0.001
Al_{ox}	0.28	0.24	0.33	342	< 0.001
Land cover (Forest)	0.26	0.10	0.41	783	0.001
CIA	-0.23	-0.27	-0.19	342	< 0.001
pH_{H2O}*Al_{ox}	-0.16	-0.19	-0.13	342	< 0.001
MAT	-0.15	-0.24	-0.06	783	0.001
Intercept	-0.14	-0.27	-0.02	783	0.028
Land cover (Other)	0.09	0.02	0.15	783	0.014
Fe_{ox}	0.06	0.01	0.12	342	0.024
Land cover (Grassland)	0.04	-0.04	0.12	783	0.377
Clay and fine silt	-0.01	-0.04	0.03	342	0.748

CI: 95% confidence interval; DF: Degree of freedom; Ca_{ex}: Exchangeable Ca; PET: Potential evapotranspiration; MAP: Mean annual precipitation; Al_{ox}: Oxalate-extractable Al; CIA: Chemical Index of Alteration; MAT: Mean annual temperature; Fe_{ox}: Oxalate-extractable Fe

Table B3: Step-wise reduction of full linear mixed effect model based on the Akaike Information Criterion

MAT	PET/MAP	Clay	Al_{ox}	Fe_{ox}	Ca_{ex}	pH	CIA	LC	Depth	AIC
X	X	X	X	X	X	X	X	X	X	1632.49
X	X	–	X	X	X	X	X	X	X	1624.27
X	X	–	X	–	X	X	X	X	X	1621.88
X	X	–	X	–	X	X	X	–	X	1616.77

MAT: Mean annual precipitation; PET: Potential evapotranspiration; MAP: Mean annual precipitation; Al_{ox}: Oxalate-extractable Al; Fe_{ox}: Oxalate-extractable Fe; Ca_{ex}: Exchangeable Ca; CIA: Chemical Index of Alteration; LC: Land cover; AIC: Akaike Information Criterion; X – included in the model; – not included in the model

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Table B4: Full linear mixed effect model results of the standardized and normalized fixed effect parameters for the two depth models (topsoil = 791, subsoil = 810)

Topsoil (n = 791)					
Variable	**Estimates**	**Lower CI**	**Upper CI**	**DF**	**p-value**
Ca_{ex}	0.64	0.56	0.72	419	< 0.001
pH_{H2O}	-0.61	-0.70	-0.52	419	< 0.001
PET/MAP	-0.35	-0.48	-0.22	419	< 0.001
Al_{ox}	0.34	0.28	0.40	419	< 0.001
CIA	-0.28	-0.33	-0.22	419	< 0.001
Land cover (Forest)	0.27	0.07	0.46	419	0.007
Intercept	-0.21	-0.35	-0.06	419	0.004
pH_{H2O}*Al_{ox}	-0.19	-0.24	-0.14	419	< 0.001
MAT	-0.15	-0.25	-0.05	419	0.004
Land cover (Other)	0.09	-0.00	0.17	419	0.053
Clay and fine silt	-0.07	-0.12	-0.03	419	0.002
Land cover (Grassland)	0.05	-0.05	0.16	419	0.311
Fe_{ox}	0.00	-0.07	0.07	419	0.952

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Subsoil (n = 810)					
Variable	**Estimates**	**Lower CI**	**Upper CI**	**DF**	**p-value**
Ca_{ex}	0.47	0.38	0.56	440	< 0.001
PET/MAP	-0.39	-0.53	-0.25	440	< 0.001
Al_{ox}	0.34	0.28	0.41	440	< 0.001
pH_{H2O}	-0.34	-0.43	-0.25	440	< 0.001
Land cover (Forest)	0.22	-0.00	0.44	440	0.053
MAT	-0.16	-0.27	-0.06	440	0.003
pH_{H2O}*Al_{ox}	-0.15	-0.20	-0.10	440	< 0.001
CIA	-0.14	-0.20	-0.08	440	< 0.001
Intercept	-0.10	-0.25	0.05	440	0.200
Fe_{ox}	0.05	-0.02	0.13	440	0.185
Land cover (Other)	0.05	-0.05	0.15	440	0.332
Clay and fine silt	0.04	-0.01	0.09	440	0.152
Land cover (Grassland)	-0.01	-0.12	0.10	440	0.854

CI: 95% confidence interval; Degree of freedom; Ca_{ex}: Exchangeable Ca; PET: Potential evapotranspiration; MAP: Mean annual precipitation; Al_{ox}: Oxalate-extractable Al; CIA: Chemical Index of Alteration; MAT: Mean annual temperature; Fe_{ox}: Oxalate-extractable Fe

690

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Table B5: Full linear mixed effect model results of the standardized and normalized fixed effect parameters for the four pH models (strongly acid (pH: 3.9–5.2) = 404; moderate acid (pH: 5.2–6.1) = 399; neutral (pH: 6.1–7.5) = 398; neutral (pH: 7.5–9.9) = 400)

Strongly acid (n = 404)					
Variable	**Estimates**	**Lower CI**	**Upper CI**	**DF**	**p-value**
Al_{ox}	0.56	0.46	0.67	91	< 0.001
Depth	-0.42	-0.48	0.37	91	< 0.001
Aridity Index (PET/MAP)	-0.26	-0.44	-0.08	162	0.005
Ca_{ex}	0.17	0.08	0.25	91	< 0.001
pH_{H2O}	-0.14	-0.20	-0.08	91	< 0.001
Fe_{ox}	0.14	0.05	0.23	91	0.003
MAT	-0.14	-0.26	-0.02	162	0.019
CIA	-0.13	-0.22	-0.04	91	0.004
Intercept	-0.13	-0.38	0.11	162	0.284
Clay and fine silt	-0.09	-0.15	-0.02	91	0.010
pH_{H2O}*Al_{ox}	-0.03	-0.10	0.03	91	0.256

Moderately acid (n = 399)					
Variable	**Estimates**	**Lower CI**	**Upper CI**	**DF**	**p-value**
Al_{ox}	0.49	0.39	0.60	45	< 0.001
Ca_{ex}	0.46	0.37	0.55	45	< 0.001
Depth	-0.33	-0.39	-0.28	45	< 0.001
Aridity Index (PET/MAP)	-0.27	-0.36	-0.18	158	< 0.001
MAT	-0.17	-0.26	-0.08	158	< 0.001
pH_{H2O}	-0.15	-0.20	-0.10	45	< 0.001
CIA	-0.13	-0.19	-0.06	45	< 0.001
Fe_{ox}	-0.09	-0.21	0.03	45	0.139
pH_{H2O}*Al_{ox}	-0.06	-0.10	-0.01	45	0.011
Clay and fine silt	-0.03	-0.11	-0.06	45	0.533
Intercept	0.00	-0.12	0.12	158	0.999

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Neutral (n = 398)					
Variable	**Estimates**	**Lower CI**	**Upper CI**	**DF**	**p-value**
Ca_{ex}	0.36	0.27	0.45	38	< 0.001
Depth	-0.31	-0.37	0.26	38	< 0.001
Aridity Index (PET/MAP)	-0.30	-0.42	-0.18	161	< 0.001
CIA	-0.27	-0.33	-0.22	38	< 0.001
Al_{ox}	0.19	0.10	0.29	38	< 0.001
Clay and fine silt	0.13	0.05	0.20	38	0.002
pH_{H2O}	-0.11	-0.16	-0.06	38	< 0.001
Intercept	-0.05	-0.21	0.10	161	0.492
MAT	-0.04	-0.17	-0.09	161	0.543
pH_{H2O}*Al_{ox}	-0.02	-0.07	0.02	38	0.307
Fe_{ox}	0.00	-0.10	0.11	38	0.958

Alkaline (n = 400)					
Variable	**Estimates**	**Lower CI**	**Upper CI**	**DF**	**p-value**
Ca_{ex}	0.31	0.23	0.40	72	< 0.001
Depth	-0.24	-0.29	-0.20	72	< 0.001
Intercept	-0.18	-0.51	0.15	219	0.296
MAT	-0.16	-0.45	0.14	219	0.297
pH_{H2O}	-0.12	-0.17	-0.07	72	< 0.001
Fe_{ox}	0.11	0.00	0.21	72	0.042
Aridity Index (PET/MAP)	-0.07	-0.30	0.15	219	0.518
Clay and fine silt	-0.03	-0.10	0.03	72	0.331
CIA	-0.03	-0.09	0.03	72	0.311
Al_{ox}	0.01	-0.08	0.10	72	0.863
pH_{H2O}*Al_{ox}	-0.00	-0.04	0.03	72	0.825

CI: 95% confidence interval; Degree of freedom; Ca_{ex}: Exchangeable Ca; PET: Potential evapotranspiration; MAP: Mean annual precipitation; Al_{ox}: Oxalate-extractable Al; CIA: Chemical Index of Alteration; MAT: Mean annual temperature; Fe_{ox}: Oxalate-extractable Fe

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Table B6: Full linear mixed effect model results of the standardized and normalized fixed effect parameters for the three wet months (P/PET > 1) models (0 wet months = 572; 1–3 wet months = 367; 4–7 wet months = 662)

0 wet months (n = 572)					
Variable	**Estimates**	**Lower CI**	**Upper CI**	**DF**	**p-value**
Ca_{ex}	0.48	0.38	0.57	109	< 0.001
Aridity Index (PET/MAP)	-0.29	-0.44	-0.13	267	< 0.001
Depth	-0.28	-0.31	-0.24	109	< 0.001
MAT	-0.26	-0.42	-0.10	267	0.001
Fe_{ox}	0.13	0.05	0.22	109	0.003
pH_{H2O}	-0.11	-0.19	-0.04	109	0.004
CIA	-0.08	-0.14	-0.02	109	0.005
Intercept	0.02	-0.16	0.21	267	0.797
Clay and fine silt	0.02	-0.04	0.09	109	0.521
pH_{H2O}*Al_{ox}	-0.02	-0.07	0.03	109	0.388
Al_{ox}	0.01	-0.08	0.10	109	0.864

1–3 wet months (n = 367)					
Variable	**Estimates**	**Lower CI**	**Upper CI**	**DF**	**p-value**
Ca_{ex}	0.52	0.40	0.65	69	< 0.001
Depth	-0.40	-0.46	-0.35	69	< 0.001
CIA	-0.32	-0.38	-0.26	69	< 0.001
pH_{H2O}	-0.32	-0.40	-0.24	69	< 0.001
Al_{ox}	0.27	0.18	0.36	69	< 0.001
Aridity Index (PET/MAP)	-0.24	-0.46	-0.03	179	0.029
MAT	0.15	-0.08	0.38	179	0.198
Fe_{ox}	0.14	0.05	0.24	69	0.004
pH_{H2O}*Al_{ox}	-0.09	-0.13	-0.04	69	< 0.001
Clay and fine silt	0.04	-0.04	0.13	69	0.290
Intercept	-0.03	-0.35	0.29	179	0.840

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

4–7 wet months (n = 662)					
Variable	**Estimates**	**Lower CI**	**Upper CI**	**DF**	**p-value**
Al_{ox}	0.56	0.47	0.65	148	< 0.001
Ca_{ex}	0.48	0.37	0.59	148	< 0.001
Depth	-0.42	-0.47	-0.37	148	< 0.001
pH_{H2O}	-0.36	-0.45	-0.27	148	< 0.001
Aridity Index (PET/MAP)	-0.35	-0.57	-0.14	336	0.001
CIA	-0.21	-0.29	-0.12	148	< 0.001
MAT	-0.19	-0.34	-0.04	336	0.012
pH_{H2O}*Al_{ox}	-0.12	-0.18	-0.06	148	< 0.001
Intercept	-0.12	-0.42	0.19	336	0.462
Fe_{ox}	-0.03	-0.12	0.06	148	0.522
Clay and fine silt	-0.02	-0.08	0.04	148	0.452

CI: 95% confidence interval; Degree of freedom; Ca_{ex}: Exchangeable Ca; PET: Potential evapotranspiration; MAP: Mean annual precipitation; Al_{ox}: Oxalate-extractable Al; CIA: Chemical Index of Alteration; MAT: Mean annual temperature; Fe_{ox}: Oxalate-extractable Fe

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Table B7: Full linear mixed effect model results of the standardized and normalized fixed effect parameters for the two weathering (CIA) models (moderate (10–88%) = 801; high (88–100%) = 800)

Moderate CIA (n = 801)					
Variable	**Estimates**	**Lower CI**	**Upper CI**	**DF**	**p-value**
Ca_{ex}	0.45	0.39	0.51	118	< 0.001
Depth	-0.30	-0.34	-0.27	118	< 0.001
Aridity Index (PET/MAP)	-0.27	-0.39	-0.14	427	< 0.001
pH_{H2O}	-0.22	-0.28	-0.16	118	< 0.001
Fe_{ox}	0.13	0.06	0.20	118	< 0.001
CIA	-0.12	-0.16	-0.09	118	< 0.001
MAT	-0.11	-0.24	0.02	427	0.111
pH_{H2O}*Al_{ox}	-0.08	-0.11	-0.04	118	< 0.01
Al_{ox}	0.06	-0.00	0.12	118	0.055
Intercept	-0.02	-0.17	0.12	427	0.745
Clay and fine silt	0.01	-0.04	0.06	118	0.677

High CIA (n = 800)					
Variable	**Estimates**	**Lower CI**	**Upper CI**	**DF**	**p-value**
Al_{ox}	0.55	0.48	0.62	169	< 0.001
Ca_{ex}	0.36	0.29	0.44	169	< 0.001
Aridity Index (PET/MAP)	-0.36	-0.50	-0.22	320	< 0.001
pH_{H2O}	-0.34	-0.41	-0.28	169	< 0.001
Depth	-0.33	-0.36	-0.29	169	< 0.001
CIA	-0.20	-0.27	-0.14	169	< 0.001
MAT	-0.23	-0.34	-0.12	320	< 0.001
pH_{H2O}*Al_{ox}	-0.13	-0.17	-0.08	169	< 0.001
Intercept	-0.08	-0.26	0.10	320	0.377
Fe_{ox}	0.03	-0.12	0.05	169	0.411
Clay and fine silt	-0.02	-0.07	0.02	169	0.324

CI: 95% confidence interval; Degree of freedom; Ca_{ex}: Exchangeable Ca; PET: Potential evapotranspiration; MAP: Mean annual precipitation; Al_{ox}: Oxalate-extractable Al; CIA: Chemical Index of Alteration; MAT: Mean annual temperature; Fe_{ox}: Oxalate-extractable Fe

https://doi.org/10.5194/soil-2020-69
Preprint. Discussion started: 26 October 2020

Table B8: Summary table of R^2 for the different models (linear mixed effect model and random forest) with different explanatory variables (clay and fine silt, land-cover, clay and fine silt + land-cover, full) included for the entire dataset. The R^2 in brackets for the linear-mixed effect models refer to the conditional R^2 which include the variation explained by the random effects (Site/Cluster/Profile).

Model	Linear-mixed model	Random forest (topsoil)	Random forest (subsoil)
Clay and fine silt	0.01 (0.72)	0.12	0.12
Land-cover	0.01 (0.75)	0.10	0.16
Clay + Land-cover	0.02 (0.72)	0.22	0.26
full	0.71 (0.94)	0.70	0.72