

Article

Accurate Quantification of 0–30 cm Soil Organic Carbon in Croplands over the Continental United States Using Machine Learning

Peng Fu ^{1,2,*}, Christian Clanton ¹, Kirk M. Demuth ¹, Verena Goodman ¹, Lauren Griffith ¹, Mage Khim-Young ¹, Julia Maddalena ¹, Kenny LaMarca ¹, Logan A. Wright ¹, David W. Schurman ¹  and James R. Kellner ^{1,3,4}

¹ Perennial Climate Inc., Boulder, CO 80301, USA; kdemuth@perennial.earth (K.M.D.)

² Center for Advanced Agriculture and Sustainability, Harrisburg University, Harrisburg, PA 17101, USA

³ Institute at Brown for Environment and Society, Brown University, Providence, RI 02912, USA

⁴ Department of Ecology, Evolution and Organismal Biology, Brown University, Providence, RI 02912, USA

* Correspondence: pfu@perennial.earth

Abstract: Increases in organic carbon within agricultural soils are widely recognized as a “negative emission” that removes CO₂ from the atmosphere. Accurate quantification of soil organic carbon (SOC) to a certain depth in the spatial domain is critical for the effective implementation of improved land management practices in croplands. Currently, there is a lack of understanding regarding what depth strategy should be used to estimate SOC at 0–30 cm when sample datasets come from multiple depths. Furthermore, few studies have examined depth strategies for mapping SOC at the agricultural management level (i.e., field level), opting instead for point-based analysis. Here, three types of approaches with different depth strategies were evaluated for their ability to quantify 0–30 cm SOC content based on soil samples from 0–5 (surface), 5–30 (subsurface), and 0–30 cm (full column). These approaches involved the generalized additive model and machine learning techniques, i.e., artificial neural networks, random forest, and XGBoost. The soil samples used for the model evaluation and selection consisted of the newly collected samples in 2020–2022 and the Rapid Carbon Assessment (RaCA) legacy samples collected in 2010–2011. Environmental covariates corresponding to these SOC measurements were used in model training, including long-term physical climate, short-term weather, topographic and edaphic, and remotely sensed variables. Among the models evaluated in this study, the XGB regression model with a full column depth assignment strategy yielded the best prediction performance for 0–30 cm SOC content, with an r^2 (squared Pearson correlation coefficient) of 0.48, an *RMSE* (root mean square error) of 0.29%, an *ME* (mean error) of 0.06%, an *MAE* of 0.25%, and an *MEC* (modeling efficiency coefficient) of 0.36 at the pixel level and an r^2 of 0.64, an *RMSE* of 0.32%, an *ME* of −0.20%, an *MAE* of 0.28%, and an *MEC* of 0.48 at the field level. This study highlights that machine learning models with a full column depth strategy should be used to quantify 0–30 cm SOC content in agricultural soils over the continental United States (CONUS).

Keywords: carbon accounting; depth strategy; land management; legacy soil samples; XGBoost; soil organic carbon



Citation: Fu, P.; Clanton, C.; Demuth, K.M.; Goodman, V.; Griffith, L.; Khim-Young, M.; Maddalena, J.; LaMarca, K.; Wright, L.A.; Schurman, D.W.; et al. Accurate Quantification of 0–30 cm Soil Organic Carbon in Croplands over the Continental United States Using Machine Learning. *Remote Sens.* **2024**, *16*, 2217. <https://doi.org/10.3390/rs16122217>

Academic Editors: Dominique Arrouays and Bas van Wesemael

Received: 23 April 2024

Revised: 5 June 2024

Accepted: 15 June 2024

Published: 19 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Soil organic carbon (SOC), a pool of ~1550 GT (1 GT = 1 billion metric tons), is an important component in the global carbon cycle, regulating soil fertility and structure and the functioning of terrestrial ecosystems [1,2]. In support of the Sustainable Development Goal Indicator 15.3.1, SOC has also been considered a necessary variable to assess the sensitivity of soils to land degradation [3]. It is also widely recognized that soils are a critical part of the global carbon agenda for climate change mitigation and adaptation through various high-level initiatives. For example, the “4p1000 initiative”, launched at COP21 in 2015, aspires to increase global SOC stocks in the top 0.3–0.4 m by 0.4% per year,

amounting to a carbon storage similar in magnitude to the annual fossil fuel-related CO₂ emissions [4–6]. The Koronivia Joint Work on Agriculture is a landmark decision under the United Nations Framework Convention on Climate Change (UNFCCC) [7], acknowledging soils and SOC as part of the solutions to reduce greenhouse gas emissions. The Food and Agriculture Organization (FAO) recarbonization of global agricultural soils (RECSOIL) program supports the provision of incentives for farmers to implement sustainable management practices to prevent future SOC losses and increase SOC stocks through mechanisms such as voluntary carbon credits [8]. As such, accurate estimations of SOC content and stocks to a certain depth in the spatial domain are highly sought after and hold important implications for agronomical, environmental, economic, and political issues. Specifically, in carbon accounting, it is recommended to measure SOC down to a minimum depth of 30 cm to monitor changes in SOC stocks [9,10].

Digital soil mapping (DSM) has been widely used for characterizing spatial and vertical variations in soil properties, such as SOC, based on the relationship between soil properties and environmental factors [11]. These environmental factors include soil, climate, organisms, topography, parent material, age, and space, i.e., the *scorpan* model [12]. Different from Jenny's factorial model or the Sergey–Zakharov equation [13], the *scorpan* model is used to quantify soil classes or continuous soil attributes based on empirical observations (e.g., Landsat images, digital elevation model, air temperatures, and precipitation). Regression techniques used to build the relationship between environmental covariates and soil attributes may include Cubist [14], Random Forests [15], artificial neural networks [16], XGBoost [17], and deep learning [18]. These developed regression models are applied to spatially exhaustive environmental data for large-scale mapping of soil attributes at various spatial resolutions.

Various DSM methods have been developed to estimate SOC at vertical and spatial dimensions, as described in the literature review in Section 2. The depth strategies behind these methods can be grouped into three categories. First, statistical or machine learning models are trained to predict SOC at individual depth intervals (e.g., 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm) [19]. Second, a parametric function characterizing the SOC–depth relationships, such as exponential decay [20] and power functions [21], can be used to model soil attributes by depth. The third category accounts for depth explicitly as a covariate in the geostatistical and machine learning models [22,23]. Despite the success of these existing methods to estimate SOC at depth, there is still a lack of understanding of what depth strategy is most suitable for predicting SOC at 0–30 cm at both sample and field levels. This is urgently needed in carbon accounting. In addition, in many DSM applications, soil samples may be available from different locations and times at non-standard depths (e.g., 5–30 cm). Available soil datasets can be spatially heterogeneous, non-randomly sampled, and exposed to large amounts of non-random missing observations in environmental covariates. Thus, a further understanding of the performance of the existing modeling approaches for these soil datasets in predicting SOC at 0–30 cm is highly needed.

In this study, different modeling approaches were assessed for predicting SOC content at 0–30 cm using a comprehensive dataset. This dataset included soil samples collected at 0–5 cm (surface), 5–30 cm (subsurface), and 0–30 cm (full column) depths, as well as environmental data such as long-term physical climate, short-term weather, topographic, edaphic, and remotely sensed variables. The soil samples comprised newly collected data from 2020–2022 and Rapid Carbon Assessment (RaCA) legacy samples from 2010–2011, spanning agricultural lands across various states in the continental United States. The objectives of this study were to (1) examine performances of existing soil mapping models to characterize SOC (0–30 cm) at sample and field levels and (2) explore what depth strategies should be used in the prediction model for estimating SOC.

2. Literature Review

When both the vertical and lateral dimensions, i.e., 3D, are considered in soil mapping, existing DSM methods for quantifying soil attributes generally fall into four categories,

which are further summarized in Table 1. The first category of methods refers to the spatial prediction of soil attributes for individual depth intervals, known as pseudo-3D or 2.5D mapping [24], using machine learning and/or other regression techniques with inputs of environmental variables such as remote sensing data. Depth intervals or increments are generally harmonized and standardized so that soil samples from different surveys/databases at various depth levels can be used consistently. For example, in the GlobalSoilMap project specification, soil properties were estimated at six standard depth intervals, i.e., 0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, 60–100 cm, and 100–200 cm [19]. To achieve standardization, smoothing functions, such as the equal-area spline function [25], are used to obtain vertically continuous soil profile data at each sampling location. Such a smoothing function may also be used after multilayered spatial predictions of soil attributes are made to generate continuous vertical variation of soil properties over the region of interest. Despite their wide use, these methods commonly require different combinations of environmental variables (e.g., apparent electrical conductivity) at each depth interval, which may be difficult to obtain [26]. In addition, the smooth functions to obtain soil profile data can obscure the quantification of uncertainty associated with estimations of soil attributes.

Table 1. Summary and comparisons of the four types of modeling approaches for characterizing SOC at depths, as outlined in Section 2.

Model Type	Explanation	Are Surface and Subsurface SOC Measurements Needed at the Same Location?	Is There Vertical Independence of SOC Predictions among Depth Intervals?	Is a Surface SOC Value Required to Predict at Depth?	Input Variables
A1	Models built for individual depth intervals	No	Yes	Yes	Environmental and remote sensing covariates at each interval
A2	Geostatistical model in 3D	Yes	No	Yes	Depth + Environmental/remote sensing variables
A3	Depth as a model feature in machine learning	No	No	No	Depth + Environmental/remote sensing variables
A4	A function to explain soil attributes by depth	Yes	No	Yes	Environmental and remote sensing variables

Note: SOC measurements in this table are from soil samples collected at different horizons or depth intervals, e.g., 0–5 cm, 5–15 cm, and 15–30 cm. In this study, samples from 0–5 cm are defined as surface samples, and those from deeper depths are considered subsurface samples. The questions listed in the table are for the model training phase and are used to highlight the advantages and disadvantages of each model type. These approaches can potentially be used to estimate SOC content and stocks at various depths, e.g., 0–100 cm and 0–200 cm.

The second category is based on 3D geostatistics. This approach assumes that soil properties at each location and depth can be represented by a spatial trend and a residual component. For example, a regression kriging approach was proposed in [27]. The study suggested the use of either linear regression, generalized linear models, or zero-inflated models for identifying the spatial trend component and a spline function for the vertical trend. In [22], a 3D generalized additive model (GAM) was combined with 3D kriging to quantify SOC stocks in both the vertical and lateral dimensions in Scotland. The GAM was fit to estimate the trend component of SOC using a 3D smoother with related covariates, and kriging simulations of GAM residuals were used to account for local details (i.e., the residual component). The 3D GAM has also been combined with area-to-point kriging and increment-averaged kriging [28,29] so the model can include all the soil profile data in one statistical analysis and provide explicit functions for quantifying the uncertainty associated with soil depth functions. These 3D geostatistical approaches can reduce the complexity of workflow and thus are naturally appealing. However, geostatistical models can be computationally expensive when applied at regional or global scales as a large

number of soil samples are required [27]. In addition, the geostatistical modeling may not be able to capture the non-linear relationship between soil properties and cross-correlated environmental covariates [30].

The third category points to the mapping methods incorporating soil depth along with environmental covariates as predictors for estimating 3D soil attributes [17,23,31]. In these 3D modeling approaches, soil observation from a depth interval is equal to a point observation. Similar to 3D geostatistical models, 3D depth models include all variables in one modeling analysis and thus are appealing and flexible for producing both vertical and lateral variations in soil attributes. Compared with 2.5D models for estimating soil attributes, these 3D depth models can achieve similar, if not better, prediction accuracies but with larger uncertainties in the vertical dimension at the field level [32,33].

The fourth type of mapping method involves the use of parametric soil depth functions to quantify 3D soil attributes. Different soil depth functions have been used previously, such as the exponential decay [34], spline [25], polynomial [35], and power functions [21]. The coefficients of these depth functions are modeled against environmental variables, such as terrain, land cover, and soil type, at sampling locations using regression techniques or geostatistical interpolations [36]. Once depth function coefficients are obtained using developed regression or interpolation models over the study region, soil attributes at any depth can be derived. It is assumed that these soil depth functions can be applied to the whole study region. However, such an assumption may only hold for certain soil properties that naturally follow these depth functions down the soil profile [32], or not hold at all because of anthropogenic and pedological forcing, for which abrupt variation in soil attributes can occur [37].

While 3D soil mapping methods are divided into different categories, they are not exclusive to each other and may be used in a synergistic way. For example, legacy soil profile data were standardized/harmonized into regular depth increments before applying a full 3D depth model to estimate SOC stocks in both vertical and lateral dimensions at the global scale [17]. The final predictions were also generated at seven standard depths for all numeric soil properties using a weighted average scheme (similar to a spline function). In [38], an exponential depth function of SOC was combined with 3D ordinary kriging to characterize 3D SOC in a fluvial plain in Quzhou, China. In this study, the first three types of methods, i.e., A1–A3 in Table 1, can be potentially employed to estimate SOC at 0–30 cm based on soil samples collected at 0–5 cm, 5–30 cm, and 0–30 cm and the corresponding remotely sensed and environmental data. However, further examination of the model configuration with depth strategies is necessary for understanding the performance of existing models in estimating SOC content at 0–30 cm.

3. Study Area and Data

3.1. Study Area and Soil Samples

This study focuses on mapping the top 30 cm SOC content (%) over the continental U.S. (CONUS) croplands at a 10 m spatial resolution. The interest in this mapping study is encouraged by the scientific consensus on the need to increase SOC stocks through sustainable agricultural management practices [5,39] and the availability of stimulus mechanisms, such as voluntary carbon markets [40] in the U.S. To help with this carbon mapping project, 6092 soil samples, including surface (0–5 cm; 4367 soil samples), subsurface (5–30 cm; 968 soil samples), and full column (0–30 cm; 757 soil samples) samples, were collected from agricultural lands of non-tree row crops over the CONUS in 2020, 2021, and 2022 (hereafter referred to as 2020–2022 soil samples, specific collection dates for these soil samples are listed in Appendix A Table A1). The U.S. Department of Agriculture (USDA) Cropland Data Layer [41] was used to assist in the collection of soil samples from various crop types, such as corn, soybean, wheat, alfalfa, cotton, and sunflower. The main soil types covered by these soil samples consisted of Mollisols, Alfisols, Entisols, Aridisols, and Inceptisols based on the USDA National Resource Conservation Service (NRCS) Soil Survey Geographic Database (SSURGO). Soil samples were randomly extracted from each field (441 fields in

total) and were collected either before the growing season or after harvest to minimize vegetation impacts. For each sampling location, GPS coordinates were recorded, and samples were sealed in a plastic bag. These soil samples were analyzed by a commercial laboratory with the dry combustion method [42], and SOC content measurements were provided as a percentage value.

An additional 970 soil samples, consisting of surface (962 soil samples, 0–5 cm) and full column (8 soil samples, 0–30 cm) samples from agricultural lands, were selected for this study. SOC percentage measurements of these sites are publicly available from the RaCA project initiated by the Soil Science Division of the Natural Resource Conservation Service to capture baseline SOC stocks over the CONUS [43]. Soil samples from these sites (hereafter referred to as RaCA soil samples) represent a balanced collection across different land cover types, soil regions, and ecological zones in the U.S. The selection of 970 soil samples was a result of quality assessment and filtering of soil sites available from the RaCA project. Specifically, soil sites missing spatial coordinates and SOC measurements were excluded from the analysis. Only data from central pedons in individual field plots and depths of 0–5 cm and 0–30 cm were used for analysis. These soil samples were collected in 2010 and 2011 and had a broader geographic coverage than the soil samples collected between 2020 and 2022. Figure 1 shows the number of soil samples collected from croplands across the CONUS. These soil samples were distributed across all the states except the State of Rhode Island, with the number of soil samples varying greatly among states (from less than 10 soil samples in one state to more than 1000 soil samples in another state). The soil samples do not represent a random sampling approach, and currently, this is a challenge that must be solved in carbon accounting. This unique dataset thus functions as a test set for evaluating the applicability of various digital soil mapping methods employing different depth strategies. The goal is to assess whether these methods can effectively estimate SOC content specifically within the 0–30 cm depth range.

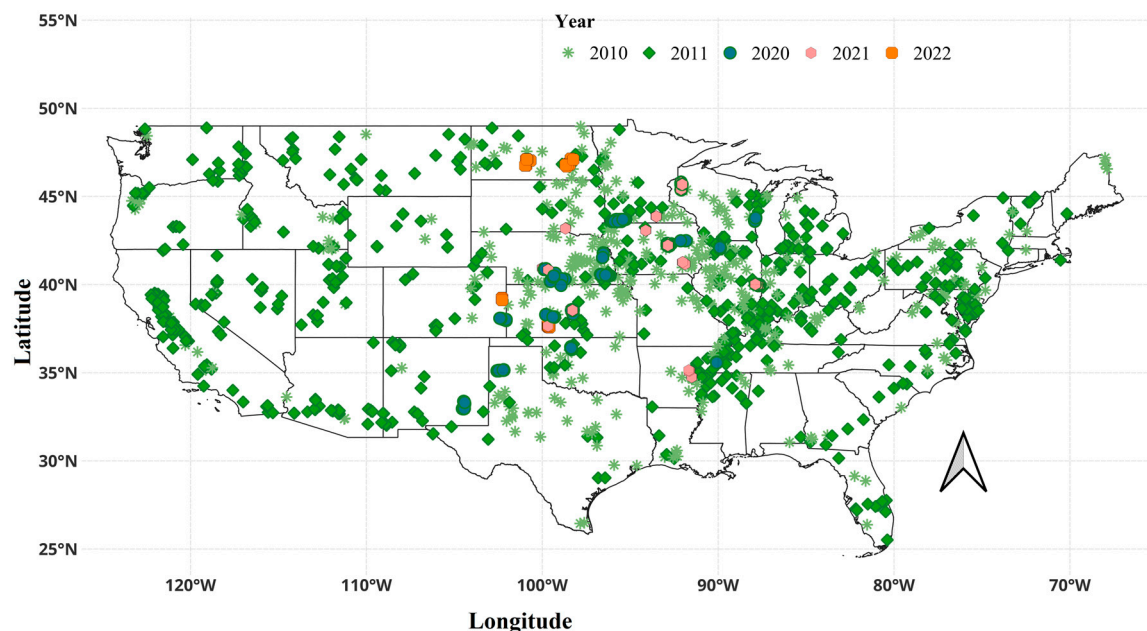


Figure 1. The spatial distribution of soil samples collected from each state across the continental United States (CONUS). The total number of soil samples used in this study was 7018. RaCA soil samples were collected in 2010 and 2011, and the rest of the soil samples were collected between 2020 and 2022. The specific dates for these soil samples are listed in Table A1 (Appendix A).

3.2. Environmental and Remote Sensing Data

Various environmental and remote sensing variables, which are predictive of the SOC content [17,44,45], were used as model predictors in this study. Specifically, fol-

lowing DSM [12], the variables used in this study, as shown in Table A2 (Appendix A), included the following: (1) long-term physical climate proxies, (2) short-term physical climate and weather data, (3) topographic and edaphic information (i.e., soil texture from SoilGrids [17]), and (4) remotely sensed data consisting of Sentinel-1 synthetic aperture radar (SAR), Sentinel-2 surface reflectance and its derived spectral indices, MODIS land surface temperatures (LSTs), and Soil Moisture Active Passive (SMAP) soil moisture. These variables were of different spatial resolutions and resampled to 10 m using a cubic spline interpolation for the final modeling analysis. In addition, location information, i.e., longitude and latitude, and depth were also used per the model selection and configuration as shown in Section 3.

The long-term climate data maps were downloaded from WorldClim with a spatial resolution of ~1 km [46]. WorldClim provides spatially interpolated monthly temperature, precipitation, solar radiation, vapor pressure, and wind speed, aggregated across the temporal interval between 1975 and 2000. It also provides separate data layers of bioclimatic variables, annual trends, seasonality, and extremes in temperature and precipitation, as derived from monthly data. These datasets have been validated using in situ observations from weather stations with cross-validation correlations larger than 0.75. In this study, the following bioclimatic variables were used to represent long-term physical proxies: BIO1 (mean annual temperature), BIO6 (precipitation of the wettest quarter), and BIO17 (precipitation of the driest quarter).

The short-term climate and weather data were obtained from the National Centers for Environmental Prediction Climate Forecast System (NCEP CFS; [47,48]). The product contains weather variables from the CFS-v2 operational 6-hourly analysis (with a spatial resolution of ~0.2°) since 1 April 2011 and CFS 6-hourly Reanalysis between 1 January 1999 and 31 March 2011. These 6-hourly weather variables, including soil moisture, mean temperature, potential evapotranspiration, transpiration, maximum temperature, minimum temperature, shortwave radiation net flux, sensible heat flux, and the sum for precipitation and water runoff, were selected and summarized into daily averages. These daily weather variables were further aggregated to averages over different seasons (spring: March–May, summer: June–August, autumn: September–November, and winter: December–February) and three years (i.e., seasonal mean and three-year mean) prior to a sampling or prediction date. Seasonal means in each of the three years prior to the sampling date were included as feature variables. For example, if soil samples were collected in September 2022, the seasonal means of weather data over each of three years, i.e., 1 September 2021–31 August 2022 (year 1), 1 September 2020–31 August 2021 (year 2), 1 September 2019–31 August 2020 (year 3), were computed and used as feature variables.

Surface elevation data were derived from the U.S. Geological Survey (USGS) 3D elevation program (3DEP) digital elevation model (DEM) at 10 m spatial resolution. The 3DEP improved the original National Elevation Dataset (NED) with better lidar point clouds and bare earth DEMs [49]. The gridded clay, sand, and silt contents from the SoilGrids250 product [17] were also used as model predictors in this study. The incorporation of existing soil model predictions into the model building in this study can be considered an ensemble prediction, and this strategy has been used in previous studies [44,45].

The remote sensing data used in this study included Sentinel-1 SAR imagery, Sentinel-2 optical surface reflectance and its derived spectral indices, MODIS LSTs, and SMAP soil moisture. Compared with optical remote sensing data, Sentinel-1 SAR images can be used regardless of weather and daylight conditions. Sentinel-1 operates on C-band and captures images in various modes at both high spatial and temporal resolutions. In this study, SAR images with dual polarizations (VH and VV) acquired in the Interferometric Wide swath (IW) mode from ascending orbits were used. These images were organized in the Level-1 Ground Range Detected (GRD) format, and multi-looked and projected to the ground range using an Earth Ellipsoid model [50]. The seasonal median values in each of the three years prior to the sampling or prediction date for individual pixels were computed and used as a covariate in the modeling analysis. In addition to SAR data, Sentinel-2 Multispectral Imager

(MSI) level-2 bottom of atmosphere (BOA) surface reflectance product [51] was also used. The twin Sentinel-2 satellites can acquire images covering 13 spectral bands at spatial resolutions of 10–60 m with a revisit cycle of 5 days. Although cloud masks in the Sentinel-2 product [52] were used to filter out bad pixels, optical images may still be subject to contamination from clouds, cloud shadow, and other poor atmospheric conditions; therefore, image composites, including seasonal median composites in each of the three years prior to the sampling or desired prediction date, were generated and used as model predictors. Specific variables used for these image composites included Sentinel-2 Blue (B2, 458–533 nm, 10 m), Green (B3, 543–578 nm, 10 m), Red (B4, 650–680 nm, 10 m), Near-Infrared (NIR, B8, 785–900 nm, 10 m), Shortwave Infrared 1 (SWIR-1, B11, 1565–1655 nm, 20 m), and Shortwave Infrared 2 (SWIR-2, B12, 2100–2280 nm, 20 m) bands, and spectral indices computed based on these spectral bands.

Table A3 (Appendix A) summarizes the spectral indices and equations used to compute these indices. The reflectance and spectral indices can help track seasonality in vegetation cover, planting and harvest time, and regenerative agricultural and management practices, such as cover cropping and tillage status, which are known to have enduring effects on soil health. Recognizing the slow and cumulative process of SOC accumulation, it is essential to consider the “legacy effects” of factors such as environmental factors, vegetation photosynthesis, and land management practices on SOC. These legacy effects embody the long-term impacts of previous seasons’ environmental conditions and agricultural practices on the current SOC levels, necessitating the inclusion of remote sensing data from various seasons to capture these comprehensive influences.

The collection 6 MODIS 8-day LST composite product (MOD11A2, 1 km; [53]) was used to obtain both daytime and nighttime LSTs. On a per pixel level, the seasonal mean LSTs within each of three years prior to the sampling or prediction date were computed and used as environmental covariates. In addition, soil moisture observations from the SMAP L3 daily 9 km product were used. Daily soil moisture data were averaged over different seasons in each of three years prior to the sampling or prediction date. Since RaCA samples were collected before the availability of Sentinel-1/2 and SMAP data, the corresponding environmental variables were treated as missing values.

All the environmental and remote sensing variables were processed in a cloud-based geospatial computing platform provided by Descartes Labs Inc. In this study, a three-year time series summary of environmental and remote sensing data was chosen to account for the influence of both historical and current variables on SOC values.

4. Methodology

Based on the soil samples available at surface (0–5 cm), subsurface (5–30 cm), and full column (0–30 cm) levels, three different types of prediction techniques were used and compared for their performances in characterizing 0–30 cm SOC content. Specifically, the pseudo-3D (or 2.5 D), 3D geostatistics, and full 3D depth-based approaches (i.e., A1–A3 in Table 1) were used and compared in this study. Since soil samples from only two depth increments, i.e., 0–5 cm and 5–30 cm, were available, the 3D geostatistical approach used in this study did not include the regression kriging and depth functions. The approach based on parametric soil depth functions was not examined in this study since these functions cannot be fitted with only two observations at each location.

Figure 2 shows the flowchart for data processing and analyses conducted in this study. First, soil samples 2020–2022 were screened and removed if their corresponding environmental covariates had missing values. After this step, 1277 surface and 410 subsurface soil samples (the training dataset) were selected for modeling analysis. A separate full column dataset (135 soil samples, roughly 1/10 number of the training data) was saved (these soil samples were randomly selected from fields that were not considered for the modeling training) for evaluating the models built for SOC predictions. Second, three types of modeling approaches, i.e., models built for each depth interval (A1), a 3D generalized additive model with depth as a feature variable (A2), and 3D machine learning models

with depth as a feature variable (A3), as shown in Sections 3.1 and 3.2, were applied to the selected datasets and compared for their modeling performances in estimating SOC percentage. In A1 and A3, three machine learning models, i.e., artificial neural networks (ANNs), random forests (RFs), and XGBoost (XGB) regression, were used and compared. In the model selection phase, the best modeling approach was identified based on a few metrics (further descriptions were provided in Section 4.4) and then applied to the dataset consisting of both *RaCA* and soil samples collected between 2020 and 2022 (7062 soil samples in total). This step aimed to evaluate whether the best model identified in the previous steps could leverage historical soil samples, which had better spatial coverage than the soil samples collected between 2020 and 2022, for estimating SOC.

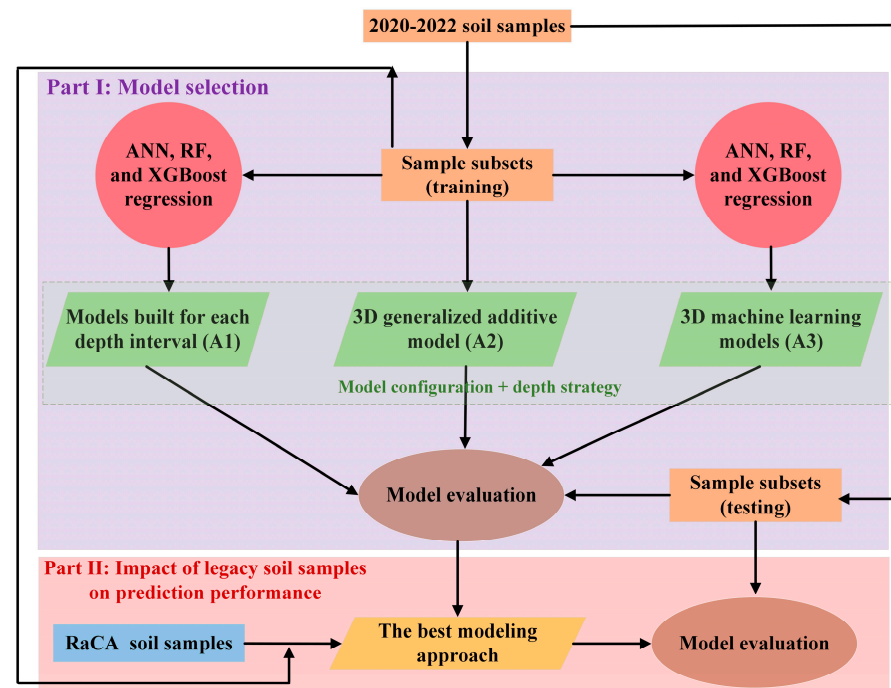


Figure 2. The flowchart showing the data analyses conducted in this study. Soil samples were divided into training and testing datasets. Three different types of approaches, i.e., machine learning models built at each depth interval (A1), the 3D generalized additive model with depth as a covariate (A2), and 3D machine learning models with depth as a covariate (A3), were compared for their performances in quantifying SOC percentage values. The machine learning models used in this study included artificial neural network (ANN), random forest (RF), and XGBoost (XGB) regression. After the best modeling approach was identified, both the *RaCA* (with missing values in the environmental covariates) and soil samples collected between 2020 and 2022 were used as inputs. This step was used to evaluate whether the identified best model can leverage the *RaCA* soil samples for estimating SOC content.

4.1. Models Built for Each Depth Interval (A1)

4.1.1. Machine Learning Models

For A1, three machine learning algorithms, i.e., XGB, RF, and ANN, were employed to characterize SOC at two individual depth intervals, i.e., 0–5 and 5–30 cm. These models were referred to as A1-XGB, A1-RF, and A1-ANN. These machine learning algorithms were selected because of their wide use and success in soil mapping [15,54,55]. The XGBoost algorithm, introduced by [56], belongs to the family of boosting algorithms. It provides resilience against overfitting and improves model generalization by weighing multiple weak predictions into single outputs, as shown in Equation (1).

$$\hat{y}_j = \phi(x) = \sum_{i=1}^N f_i(x_j), \quad f_i \in F \quad (1)$$

where x_j refers to the predictor variables corresponding to the j th response variable y_j , f denotes the individual regression tree structure and leaf weights, N represents the number of additive functions, and $\hat{y}$ is the ensemble output. To learn the additive functions (or tree structures), the following regularized objective $\mathcal{L}$ is minimized (Equations (2) and (3)).

$$\mathcal{L}(\phi) = \sum_j^K l(\hat{y}_j, y_j) + \sum_i^N \Omega(f_i) \quad (2)$$

$$\Omega(f) = \gamma T + 1/2\lambda \|w\|^2 \quad (3)$$

where l is the loss function that measures the difference between predictions and observations, K is the number of observations, the function Ω is used to penalize the model complexity, T is the number of leaves in the tree, w denotes the score of each leaf node, and γ and λ are parameters to be optimized using the stochastic gradient boosting procedure. Feature importance can also be computed in XGBoost using metric such as “gain” during its learning process. Gain measures the improvement in predictive accuracy achieved by using a specific feature to split data in decision trees. The accumulation of gains across trees reflects a feature’s overall influence on predictions.

In this study, environmental and remote sensing variables were used as inputs, and SOC was the output. To optimize the model, the following hyperparameters were tuned: the number of trees in the ensemble, the learning rate, the proportion of columns used to identify the split at each model, the proportion of samples used by individual trees, and the maximum depth of each tree. Bayesian optimization was used to optimize parameter values with a cross-validation strategy. This method can efficiently balance exploration, i.e., trying new parameter values, and exploitation, i.e., focusing on promising parameter regions, thus leading to quicker convergence and better performance compared with traditional grid and random search of hyperparameters. Further details regarding Bayesian optimization can be referred to in [57].

As a bagging ensemble learning algorithm, RF constructs and averages a large number of randomized decision trees for classification and regression [58]. In the model training phase, each tree is built based on a bootstrap sample of the original training data (with replacement). For each built tree, a subset of predictors is randomly selected to achieve the best split at each node. The randomness, introduced through the bootstrap strategy, helps the model overcome weaknesses of regression trees that are prone to overfitting. It also allows the estimation of general errors through the remaining unused subset, i.e., the out-of-bag error of predictions. Three parameters were tuned in this study for the RF regression, including the number of variables used to grow each tree, the number of trees in the forest, and the minimum number of terminal nodes. These three parameters were optimized via the minimization of root mean square error (RMSE) with a cross-validation strategy.

ANN achieves classification and regression by simulations of human learning processes through establishment and reinforcement of linkages between the input and output variables [59]. It models the relationship between input and output variables using a multi-layer network that consists of an input layer, an output layer, and one or more hidden layers [60]. Although numerous ANN algorithms have been developed, such as recurrent neural network, Hopfield neural networks, and Radial basis function (RBF)-based networks, multi-layer perceptron neural networks (MLP networks) with the back-propagation algorithm may be the most popular one. In addition, MLP neural network regression can handle non-linear relationships among data even when conflicting relationships between input and output variables exist. Thus, in this study, MLP neural network-based regression was utilized. The number of hidden layers and neurons in the MLP neural networks was optimally determined via minimizing RMSE within cross-validation for each depth interval.

4.1.2. Machine Learning Predictions Aggregated to 0–30 cm

Since the final output was to generate the SOC percentage value for the top 30 cm, predictions from the two depth intervals yielded by XGB, RF, and ANN, respectively, were integrated following the trapezoidal rule [17] as shown in Equation (4).

$$SOC = \frac{1}{b-a} \int_b^a f(x) dx \approx \frac{1}{b-a} \frac{1}{2} \sum_{k=1}^{N-1} (x_{k+1} - x_k) (f(x_k) + f(x_{k+1})) \quad (4)$$

where N is the number of depths, x_k is the k -th depth, $f(x_k)$ is the value of the soil property at depth x_k , and b and a represent the depth increment at which the soil property value is integrated. As there were only two depth intervals, Equation (2) was simplified to the following calculation (Equation (5)).

$$SOC_{30} = \frac{1}{6} * SOC_{0-5} + \frac{5}{6} * SOC_{5-30} \quad (5)$$

where SOC_{30} refers to the SOC content for the top 30 cm, SOC_{0-5} denotes the SOC at 0–5 cm, and SOC_{5-30} represents SOC at 5–30 cm.

4.2. The Generalized Additive Model for 3D SOC Mapping (A2)

4.2.1. The Generalized Additive Model

The GAM was originally developed [61] to blend generalized linear models with additive models by which the response variable is linearly dependent on nonparametric smooth functions of some predictor variables. Thus, GAMs have the interpretability advantages of linear models as well as substantially more flexibility since the relationship between predictor and response variables is not assumed to be linear. The 3D GAM, consisting both of trend and residual components, was used by [22] to characterize soil organic carbon stocks in Scotland, yielding better performance than regression kriging. Here, 3D GAM fitting (hereafter referred to as A2) was used without the residual component (Equation (6)) since data from only two depth increments were available.

$$SOC_p = f_1(N_p, E_p, D_p) + f_2(DEM_p) + f_3(Sentinel_p) + \dots + f_n(\dots) \quad (6)$$

where SOC_p refers to SOC values at location p (indicated by northing N , easting E , and depth D), and f_2 to f_n represent a one-dimensional smooth function of the covariate variables at the considered location. A scale-invariant tensor product smooth of 3D-space dimension is utilized for Equation (6), and the 3D smoother can handle the estimation of the 3D space trend and interaction [62]. The smoother can also avoid the need to make choices on the relative scaling of 3D space. In this study, the 3D GAM was fitted based on an internal cross-validation. Although the 3D GAM also used depth as a feature variable, the depth variable needed to be used with location variables (i.e., longitude and latitude, or northing and easting) in one smooth function, as shown in Equation (6). This occurred because depth variables only had a few unique values, e.g., either 5 or 30 cm, which would not be sufficient for a smooth function to work. Further technical details related to GAM and its implementation can be found in [62].

4.2.2. Strategies for Depth Assignment

The depth interval measurements were converted to point observations using several methods (referred to as depth assignment strategies, Figure 3). First, depth values were assigned as the center of observed depth intervals, i.e., 2.5 cm for 0–5 cm, 17.5 cm for 5–30 cm, and 15 for 0–30 cm (hereafter referred to as D1). Second, both the center and the upper and lower boundary depth values (e.g., at 0–5 cm, the depth values were 0, 2.5, and 5 cm) were used as inputs into the GAMs (hereafter referred to as D2). Third, the depth values used for 0–5 cm (i.e., 5 cm) and 0–30 cm (i.e., 30 cm) were used as inputs into machine learning models. In this case, SOC measurements from the 0–5 cm and 5–30 cm depth intervals at the same sampling sites (with geolocation distance less

than 10 m) were first integrated into a full column value (0–30 cm) using Equation (5) and then were used as response observations in the GAM fitting. These three ways to convert interval measurements into point observations were evaluated for their performances in characterizing 3D SOC (hereafter referred to as A2–D1, A2–D2, and A2–D3).

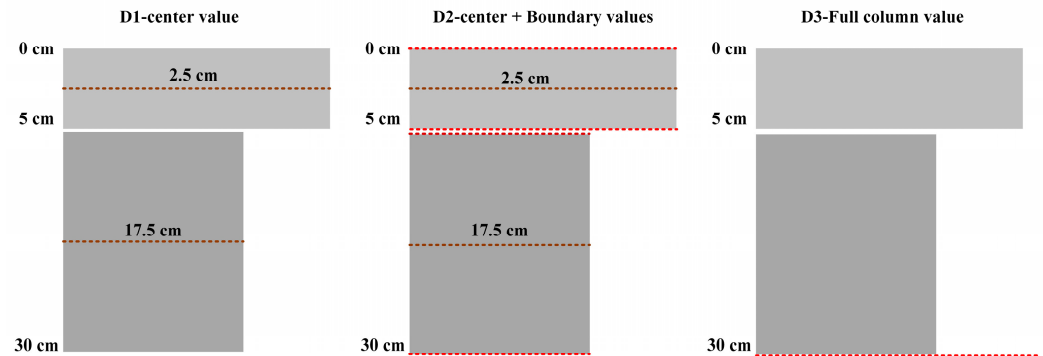


Figure 3. The different depth assignment strategies to convert depth interval-based measurement to point observations. D1-center values of depth intervals are used, D2-center + boundary values of intervals (e.g., 0, 2.5, and 5 cm for 0–5 cm surface samples) are used, and D3 full column depth value (i.e., 30 cm in this study) are used for the two depth intervals. In D3, SOC measurements from the two depth intervals need to be integrated first into one measurement following Equation (5).

4.3. Three-Dimensional Machine Learning Models for Mapping SOC (A3)

The incorporation of depth, along with other environmental covariates, in machine learning regression models for characterizing SOC in both lateral and vertical dimensions (e.g., 0–200 cm at 30 m spatial resolution) has been recently promoted [23,45]. These approaches simply assume soil samples from a depth interval as point observations. Specifically, these models assume SOC as a function of depth and a suite of spatially explicit covariate layers as shown in Equation (7).

$$SOC(x, y, d) = f(d, X_1(x, y), X_2(x, y), \dots, X_p(x, y)) \quad (7)$$

where $SOC(x, y, d)$ represents SOC content estimated at location (x, y) for easting and northing) and depth d , X_p denotes the environmental and remote sensing covariates at a specific location, and f refers to the function used to model the relationship between SOC and depth and covariates. Three ways, i.e., D1–D3, as described in Section 4.2.2 and Figure 3, were used to convert depth interval measurements into point observations. These three ways were evaluated for their performances in characterizing 3D SOC. Following Section 3.1, the XGB, RF, and ANN models, hereafter referred to as A3–XGB, A3–RF, and A3–ANN, were all used and evaluated in the full depth-based approaches for characterizing the top 30 cm SOC. Parameters in these machine learning models were optimized via minimization of RMSE using cross-validations. When combined with the ways to convert measurements from depth intervals to point observations (as shown in Section 4.2.1), there were 9 model variants evaluated, hereafter referred to as A3–XGB–D1, A3–XGB–D2, ..., A3–RF–D2, ..., A3–ANN–D2, A3–ANN–D3 (as shown in Table 2). Since the RaCA sample measurements were missing corresponding Sentinel remote sensing data, a down-weighting scheme was used to reduce the importance of RaCA samples in the model training and evaluation. Specifically, weights of 0.05–1.00 at an increment of 0.05 were used for RaCA samples and the model performance based on a full column test dataset was recorded under each weighting scenario. This down-weighting step was applied when the dataset consisting of RaCA and soil samples collected between 2020 and 2020 (weight value set at 1) was used for model training of A3–XGB–D3.

Table 2. The performance of different models in characterizing SOC percentage at 0–30 cm.

Type	Cross-Validations					Testing (0–30 cm)						
	MAE (%)	ME (%)	RMSE (%)	r^2	MEC	MAE (%)	ME (%)	RMSE (%)	r^2	MEC	b_0	b_1
A1-ANN	0.26/0.25	0.00/0.00	0.34/0.36	0.58/0.46	0.57/0.44	0.29	−0.03	0.33	0.17	0.11	−0.99	1.92
A1-RF	0.23/0.23	0.00/−0.02	0.32/0.34	0.64/0.49	0.64/0.49	0.34	−0.25	0.33	0.31	−0.11	−0.02	0.80
A1-XGB	0.24/0.24	0.00/0.00	0.33/0.34	0.62/0.49	0.62/0.49	0.25	−0.10	0.31	0.39	0.31	0.12	0.85
A2-D1	0.26	0.00	0.35	0.61	0.61	0.56	0.54	0.32	0.34	−1.91	0.76	0.49
A2-D2	0.24	0.00	0.32	0.68	0.68	0.54	0.45	0.35	0.22	−1.64	0.95	0.48
A2-D3	0.27	0.00	0.37	0.55	0.54	0.35	0.13	0.34	0.28	−0.15	0.76	0.48
A3-ANN-D1	0.25	0.01	0.33	0.64	0.63	0.55	0.52	0.36	0.16	−1.69	0.87	0.57
A3-RF-D1	0.24	0.00	0.33	0.65	0.64	0.27	−0.07	0.32	0.33	0.30	−0.05	0.47
A3-XGB-D1	0.24	0.00	0.33	0.65	0.64	0.58	0.53	0.35	0.21	−2.04	0.99	0.43
A3-ANN-D2	0.21	0.00	0.32	0.74	0.74	0.50	0.42	0.36	0.16	−1.24	0.98	0.39
A3-RF-D2	0.22	0.00	0.30	0.94	0.93	0.29	0.13	0.35	0.22	0.08	0.47	0.73
A3-XGB-D2	0.24	0.00	0.28	0.89	0.89	0.35	0.22	0.33	0.31	−0.22	0.74	0.54
A3-ANN-D3	0.24	0.00	0.33	0.65	0.66	0.46	−0.77	0.35	0.21	−6.8	0.91	0.20
A3-RF-D3	0.21	0.00	0.35	0.58	0.58	0.27	−0.09	0.33	0.29	0.22	−0.20	1.08
A3-XGB-D3	0.20	0.00	0.23	0.72	0.72	0.25	0.06	0.29	0.48	0.36	0.13	0.99

Note: A1-models built at each depth interval, A2-3D generalized additive model with depth as a feature variable, A3-3D machine learning models with depth as a feature variable, ANN-artificial neural network, RF-random forest, XGB-XGBoost, D1-the center depth value of surface or subsurface was used, D2-center + boundary depth values were used, and D3-column depth was used (further descriptions of D1–D3 can be found in Section 3.2 and Figure 3). Bold numbers indicate the best model metrics in each category (A1–A3) based on the testing dataset. In A1, cross-validation ME, RMSE, r^2 , and MEC from both surface (0–5 cm) and subsurface (5–30) levels were reported and separated by “/” as models from the two levels were built separately. b_0 and b_1 refers to the intercept and the slope of the linear regression line between predictions and observations.

4.4. Model Optimization, Evaluation, and Performance Metrics

4.4.1. Model Optimization and Evaluation

In the model selection phase (as shown in Figure 2), model hyperparameters in different types of approaches (A1–A3) used to characterize SOC at 0–30 cm were optimized via minimization of the root mean square error (RMSE) with a cross-validation strategy. Specifically, with 10-fold cross-validation, individual soil samples from each field had the chance to be the validation datasets (hereafter referred to as the sample/pixel level evaluation). In addition, the geographic cross-validation strategy, i.e., leave-one-field-out, was also used to evaluate the performance of the best model in estimating SOC at 0–30 cm (hereafter referred to as the field level evaluation). In this case, only fields with the number of soil samples larger than 5 were selected for evaluation to avoid possible errors in soil sample measurements. The mean SOC from model predictions in these fields were then validated against the mean SOC from sample measurements in each field. There were 1277 surface soil samples (0–5 cm) and 410 subsurface (5–30) soil samples involved in the model selection phase (as shown in Figure 2).

In the model testing phase, SOC percentage measurements (135 observations) from a separate full column (0–30 cm) dataset were also used to facilitate the evaluation of different types of models (A1–A3) built in this study to characterize SOC percentage values at 0–30 cm. These soil samples were collected from fields that were different from those used in the model training phase, and thus were used in evaluations at the sample level. The model performance was further evaluated when both RaCA and soil samples collected between 2020 and 2022 (7062 soil samples in total) were used.

4.4.2. Model Performance Metrics

The model performance during the training and testing phases was evaluated using the mean absolute error (MAE), mean error (ME), RMSE, squared Pearson’s correlation coefficient (r^2), and modeling efficiency coefficient (MEC) in Equations (8)–(12). In addition,

the slope and intercept values of the linear regression between predictions and observations were computed and used as supplementary metrics for evaluating the model performance.

$$ME = \frac{1}{n} \sum_{i=1}^n |(y_i - \hat{y}_i)| \quad (8)$$

$$ME = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) \quad (9)$$

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

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (11)$$

$$MEC = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (12)$$

where n is the number of soil observations, y_i refers to the measured SOC, $\hat{y}_i$ is the predicted SOC, $\bar{y}$ denotes the average measured SOC, and $\hat{\bar{y}}$ is the mean of the predicted values. The r^2 values describe the linear correlation between measured and predicted values and range between 0 (no linear correlation) and 1 (perfect linear correlation). The MEC value can be either negative or positive with 1 as the optimal value. Positive MEC values can be interpreted as an amount of variance explained by the model used to predict SOC. The optimal model would exhibit minimal ME and $RMSE$ (close to 0) and achieve a high r^2 and MEC close to 1. Parameters in machine learning and GAM models during the training phase were optimized using the cross-validations that yielded the smallest $RMSE$. To mitigate potential overfitting during model training, the recursive feature elimination (RFE) technique was applied to the top-performing model identified in the selection phase. Our analysis showed that RFE had minimal impact on model performance, indicating that overfitting was not a significant issue (details in Appendix B).

5. Results

5.1. Descriptive Statistics of SOC Percentage Measurements

Figure 4 shows the density distributions of SOC content at the surface (0–5 cm, Figure 4a,d), subsurface (5–30 cm, Figure 4b,e), and full column (0–30 cm, Figure 4c,f) levels. For the unfiltered soil samples (with missing values in corresponding environmental covariates), the mean SOC values (and standard deviations) at the surface, subsurface, and full column level (Figure 4a–c) were 1.96% (0.79%), 1.50% (0.63%), and 1.44% (0.59%), respectively. For the filtered soil samples (with no missing values in corresponding environmental covariates), the mean SOC values (standard deviations) at the surface, subsurface, and full column levels were 1.98% (0.53%), 1.51% (0.49%), and 1.34% (0.39%), respectively. SOC measurements in both unfiltered and filtered soil samples shared similar distribution statistics at the surface (Figure 4a,d), subsurface (Figure 4b,e), and full column levels (Figure 4c,f). In this study, SOC measurements from soil samples at the surface and subsurface levels (Figure 4d,e) were used as the training dataset, and at the full column level (Figure 4f) used as the testing datasets. SOC measurements from both the training and testing datasets displayed a nearly symmetric bell curve (Figure 4d–f).

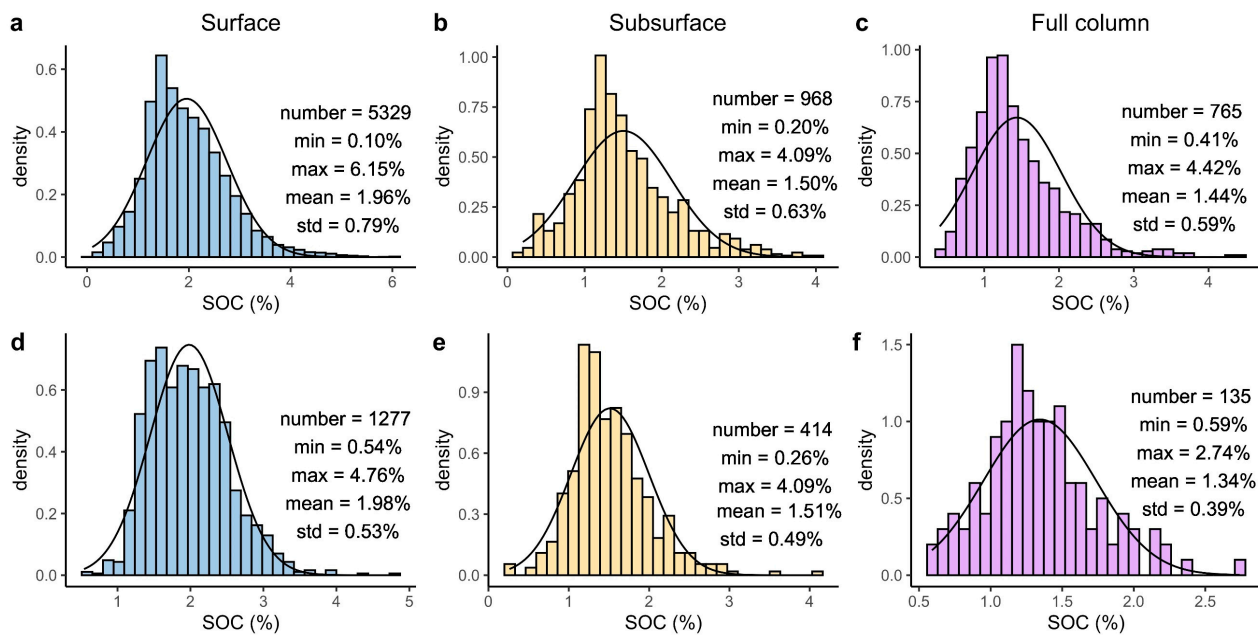


Figure 4. The density distribution of SOC measurements at the surface (0–5 cm, (a,d)), subsurface (5–30 cm, (b,e)), and full column levels (0–30, (c,f)). The bin area represents the relative frequency of the corresponding value range. The first row (a–c) refers to the unfiltered soil samples (with corresponding environmental covariates containing both missing and non-missing values) and the second row (d–f) represents filtered soil samples with the corresponding environmental covariates having complete values (no missing values). Soil samples from the second row were used in the model evaluation and selection. min: minimum, max: maximum, std: standard deviation.

5.2. Model Performance and Comparisons

Table 2 shows the performance, indicated by MAE, ME, RMSE, r^2 , and MEC, of the different types of models (A1–A3) used to characterize SOC percentage values at 0–30 cm. The intercept (b_0) and slope (b_1) of the fitting line between predictions and observations are also included in Table 2. The cross-validated ME values (close to 0.00) were not a useful indicator for model performance across all models. This is reasonable since machine learning and statistical models often produce predictions centered around the mean of measured SOC values. In A1, the regression models based on machine learning, i.e., ANN, RF, and XGB, built at the surface level exhibited lower RMSE and higher r^2 and MEC values than those at the subsurface level. Validations in these predictions (integrated into full column using Equation (5)) against the full column (0–30 cm) observations showed that the XGB model had the best performance with an r^2 of 0.39, an RMSE of 0.31%, an MAE of 0.25%, and an MEC of 0.31 (Table 2 and Figure 5a) and produced a slope value of 0.85 and an intercept value of 0.12. In A2, GAMs were used to quantify 0–30 cm SOC measurements and produced r^2 values between 0.22 and 0.34 and RMSE values between 0.32 and 0.34% based on the separate full column test dataset (Table 2). However, the MEC values were all negative, ranging between -0.15 and -1.91 , suggesting that the SOC predictions and sampled measurements were not aligned well along the 1:1 line. The differences in the model performance among A2 were attributed to the different ways SOC measurements from depth intervals were converted to point observations (Figure 3). While the cross-validation results suggested that D2 (center + upper and lower boundary point depth values) was the best approach (with an r^2 value of 0.68, an RMSE value of 0.32%, an MAE of 0.24%, and an MEC of 0.68), the testing results showed that D1 (only the center point depth value was used) was the best approach (with an r^2 value of 0.34, an RMSE value of 0.32%, an MAE of 0.56%, and an MEC of -1.91), as shown in Table 2 and Figure 5b. With the D1 strategy, the slope and intercept values were 0.49 and 0.76, respectively. In A3, when depth was used as a feature variable along with other environmental covariates, the machine learning models built with D3 (full column depth assignment in Figure 3) generally exhibited better

performance (*RMSE* values) than their counterparts using D1 or D2 in both the cross-validation and testing phases. For example, as shown in Table 2, A3-RF-D3 produced an r^2 value of 0.70, an *RMSE* value of 0.28%, an *MAE* of 0.21%, and an *MEC* of 0.70 in the cross-validation phase, better than A3-RF-D2 (with an r^2 value of 0.64, an *RMSE* value of 0.30%, an *MAE* of 0.22%, and an *MEC* of 0.64) and A3-RF-D1 (with an r^2 value of 0.65, an *RMSE* value of 0.33%, an *MAE* of 0.24%, and an *MEC* of 0.65). Both the cross-validation and testing results suggested that the best modeling approach among A1–A3 was A3-XGB-D3. In the testing phase, the A3-XGB-D3 model produced an *RMSE* value of 0.29%, an *ME* of 0.06%, an *MAE* of 0.25%, an r^2 value of 0.48, an *MEC* of 0.36 (Table 2 and Figure 4c), and a less biased slope value (0.99) compared with the other slope values (Table 2 column b_1). Figure 6 shows the cross-validation relationship between predictions and observations from fields with a number of soil samples larger than five as produced by the best modeling approach (A3-XGB-D3). When predictions were aggregated at the field level, the best modeling approach yielded an r^2 of 0.64, an *RMSE* value of 0.32%, an *MEC* of 0.48, an *ME* of −0.20%, an *MAE* of 0.28%, and a slope value of 0.86.

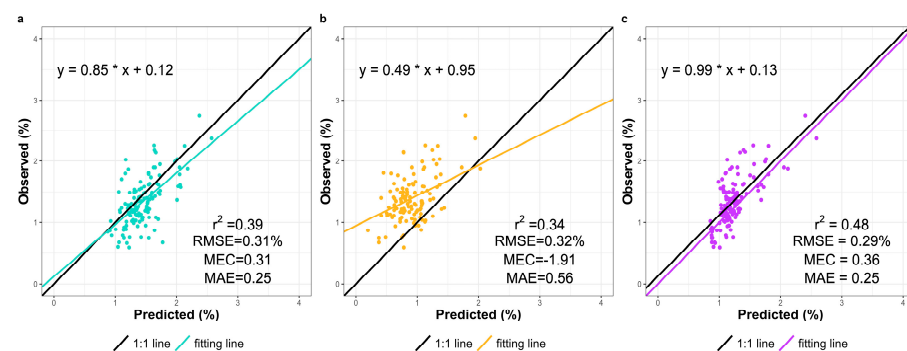


Figure 5. Comparisons between full column observations (0–30 cm) from the test dataset and predictions modeled by the best regression model in each approach (A1–A3), as shown in Table 2. (a) Predictions made using A1-XGB (XGB built at two depth intervals), (b) predictions made using A2-D1 (GAM with the center depth value as a feature variable), and (c) predictions made using A3-XGB-D3 (XGB with the column depth value as a feature variable).

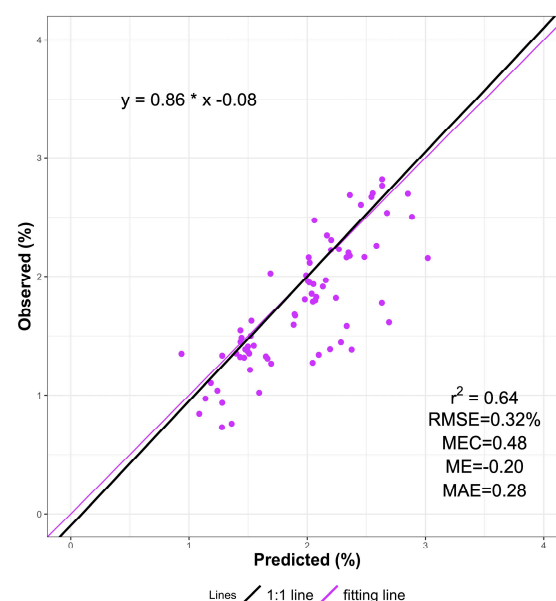


Figure 6. Comparisons between predictions and observations for fields with a number of soil samples larger than 5 using the leave-one-field-out cross-validation. These cross-validations were made using the XGB regression model with depth as a feature variable (i.e., A3-XGB-D3). The number of fields shown in this figure is 73, and each point in the figure is a mean value of SOC values within the corresponding field.

5.3. The Impact of RaCA Samples on Prediction Performance

As shown in Table 2, the best modeling approach to estimate SOC content at 0–30 cm is A3-XGB-D3. This model was then trained by using both RaCA and soil samples collected between 2020 and 2022. Figure 7 shows the impact of RaCA samples on prediction performance based on this model (A3-XGB-D3). The model metrics, r^2 , RMSE, ME, and MEC were derived based on the validation of model predictions against the separate full column dataset at 0–30 cm. The weight 0 represents the scenario in which the XGB model was trained using only the soil samples collected between 2020 and 2022 (as shown in Table 2) and then model predictions were validated against the separate full column test dataset. As the weight varied from 0.05 to 1, there was a general trend of deteriorating model performance. This decline was apparent through a reduction in r^2 and MEC, along with an elevation in RMSE and the emergence of negative ME values. Despite this overall pattern, there were instances of non-monotonic variations. The best model performance with an r^2 of 0.5, an RMSE of 0.28%, an MEC of 0.48, and an ME of 0.00 was achieved when the weight on the RaCA soil samples was set at 0.05. A weight value of 0.10–0.25 on RaCA soil samples yielded a model with similar prediction performance to that built only on soil samples collected between 2020 and 2022. When the weight value was set at 0.30 or higher, the model exhibited even worse prediction performance (lower r^2 and MEC, higher RMSE, and increasingly negative ME) than the model trained only on the soil samples collected between 2020 and 2022. When the weight value was set at 0.30 or higher, a discernable downward trend in MEC and ME values was observed, indicating that MEC and ME exhibited greater sensitivity to changes in weight values compared with r^2 and RMSE.

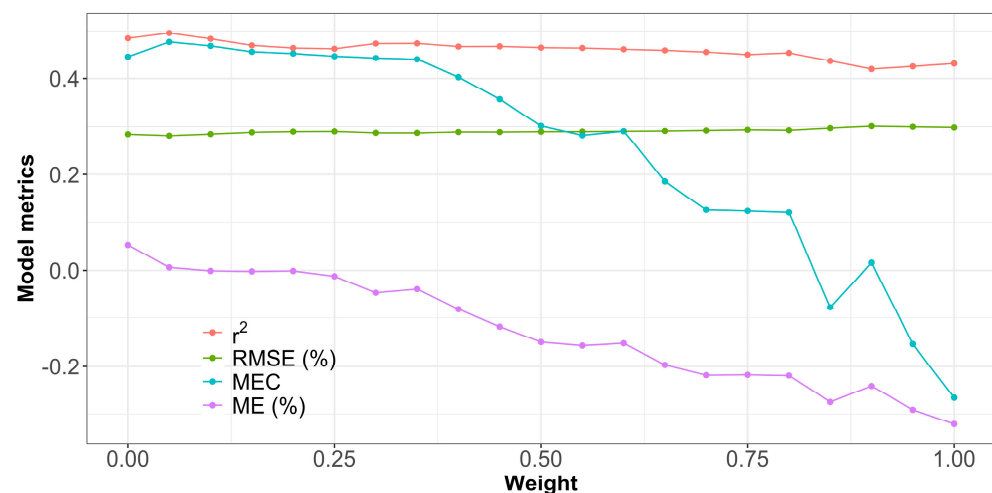


Figure 7. The performance of A3-XGB-D3 (i.e., the XGB model using depth and other environmental covariates as independent variables) in characterizing 0–30 cm SOC for different weights set on the RaCA samples. The model metrics r^2 , RMSE, ME, and MEC were derived using a separate full column test dataset. The weight 0 means that only soil samples collected between 2020 and 2022 were used for model training.

6. Discussion

6.1. Model Selection and Depth Strategies

The characterization of SOC content to a certain depth in the spatial domain at both the pixel (or sampling site) and field levels is a prerequisite for estimating SOC stocks. This information is critical to carbon offset programs incentivizing farmers to implement sustainable management practices to reduce SOC losses and increase SOC stocks [5] in agricultural fields. At present, four types of DSM-based methods (Table 1) in three depth strategies, i.e., the pseudo-3D mapping, parametric soil depth functions, 3D geostatistics, and full 3D depth-based models, can be potentially used to quantify SOC percentage at

0–30 cm. Prediction performances from these types of models may be diversified when specific machine learning techniques and the ways to convert measurements from depth intervals to point observations, as shown in Figure 3, are considered. Thus, there is a necessity to understand what method(s) performs best in quantifying SOC percentage at 0–30 cm and what depth strategies should be used for modeling, especially when dealing with spatially non-random sampled datasets.

In this study, three different types of methods in two depth strategies, i.e., the pseudo-3D mapping (A1), 3D GAMs with depth as a variable (A2), and 3D machine learning with depth as a variable (A3), were examined and compared as soil samples from the surface (0–5 cm) and subsurface (5–30 cm) levels were available. The results shown in Table 2 and Figure 5 suggested that the pseudo-3D mapping exhibited a similar, if not better, prediction performance to the 3D GAMs. Since the residual kriging regression was not used in GAMs because of the lack of soil profile observations in this study, depth values were treated as an independent variable in these models, similar to those in A3. As such, the heterogeneous prediction performance among models in A2 and A3 was attributed to the following two factors: regression models and the conversion of depth-interval-based measurements into point observations. The selection of machine learning models had an impact on prediction performance. In general, the RF- and XGB-based models performed better than the NN-based models. This finding is in agreement with previous studies suggesting that tree-based algorithms have now been dominating the establishment of relationships between soil attributes and environmental covariates [17,23,63].

The prediction performance of different machine learning models was also impacted by the conversion of depth-interval measurements into point observations. Three ways to convert depth-interval measurements into point observations (Figure 3) were used including (1) the center depth value, (2) center + lower/upper boundary depth values, and (3) the full column depth value (e.g., set depth for soil samples at the 0–30 cm level as 30 cm). The first two required the model to deal with surface and subsurface soil samples directly, whereas the last one demanded the integration of surface and subsurface soil samples into the full column observations (as shown in Equation (5)) before the model training phase. In (2), observations were replicated two times; however, such a replication did not lead to improved prediction performance (as D2-associated methods were not among the best in each category in either the cross-validation or testing phase (Table 2)). When 3D GAM was used, the best way to convert depth-interval measurements into point observations was to assign depth values as the center depth value of soil horizons. This finding was consistent with [32], in which 2.5D and 3D models for estimating 3D SOC stocks were compared, and the 3D GAM using D1 performed the best with an r^2 of 0.39. Compared with [32], this study also examined the full column depth assignment strategy in both A2 and A3.

The best modeling approach among A1–A3 was A3-XGB-D3 (an r^2 of 0.48, an RMSE of 0.29%, an ME of 0.06%, and an MEC of 0.36 in the testing phase and an r^2 of 0.72, an RMSE of 0.23%, an ME of 0.00%, and an MEC of 0.72% in the cross-validation). This prediction performance at 10 m spatial resolution was similar to those reported in previous studies at the continental and global scales [23,44,45] at various spatial resolutions (30–100 m). The use of depth as a feature variable in the machine learning models reduced further efforts to build models at more horizons. It used both the surface (0–5 cm) and full column soil observations as input data to build the relationship between SOC content and environmental covariates and thus could make predictions at any depth. However, this study only evaluated the model prediction performance based on SOC measurements at 0–5 cm (cross-validation) or 0–30 cm (testing phase). Thus, future work can evaluate whether the model can reconstruct SOC profiles at much finer resolutions (e.g., at every 5 cm interval) should data become available and soil profile variation be of interest. It is also worth further exploring the prediction performance of the 3D XGB-based depth model in characterizing SOC beyond the top 30 cm (e.g., 60 cm, 100 cm, and 200 cm).

Despite the analysis of modeling performance, it is essential to note that the current study does not extend to quantify the uncertainty that may arise from various sources

associated with inherent model inaccuracies and the input data such as data noise resulting from the resampling of a variable from coarse to fine spatial resolution and measurement errors. Thus, further research should quantify the uncertainty associated with model predictions, aiming to enhance the fidelity of SOC predictions. In addition, the selection of an optimal depth strategy for SOC mapping may vary by the specific datasets available and the ultimate objectives of the soil mapping efforts. Recognizing this variability, our study aims to serve as a foundational resource for researchers and practitioners navigating the complexities of depth strategy selection in DSM.

6.2. Variable Importance

To further explain the performance of the best modeling approach (i.e., A3-XGB-D3), feature importance was provided using the importance score calculated by the XGB model and further summed into and averaged within the following six categories: (1) location, i.e., latitude and longitude, (2) long-term physical climate proxies, (3) short-term physical climate and weather data, (4) topographic and edaphic information, (5) remotely sensed data consisting of Sentinel-1 synthetic aperture radar (SAR) and Sentinel-2 surface reflectance and its derived spectral indices, and (6) depth, as shown in Figure 8. The panel (a) in Figure 8 only shows the top 20 most important variables used in A3-XGB-D3. These top 20 variables included MODIS LSTs, depth, SMAP soil moisture, NCEP weather, and Sentinel remotely sensed datasets. The most important variable for predicting SOC percentage measurements at 0–30 cm was the MODIS daytime LST winter mean (with an importance value of 0.1238), followed by depth (0.0816) and other variables. Among the six categories of variables, Sentinel remotely sensed data were the most important variables with an aggregated importance value of 0.4648 (mean value of 0.0020), followed by short-term physical climate and weather variables with an importance value of 0.4070 (mean value of 0.0024), depth with an importance value of 0.0819, topographic and edaphic information with an importance value of 0.0408 (mean value of 0.0026), long-term physical climate proxies with an importance value of 0.0031 (mean value of 0.0010), and location variables with an importance value of 0.0024 (mean value of 0.0012). These results suggested that the inclusion of remotely sensed and short-term physical climate and weather variables, together contributing to more than 85% importance, were highly needed in characterizing 3D SOC percentage measurements. However, it is important to consider the potential impact of multicollinearity among these variables, which, while generally less influential in models like XGBoost, could still affect the interpretation of feature importance. Thus, caution should be taken when interpreting feature importance results, suggesting a deeper analysis might be necessary to fully understand the contributions of each variable amidst interdependencies.

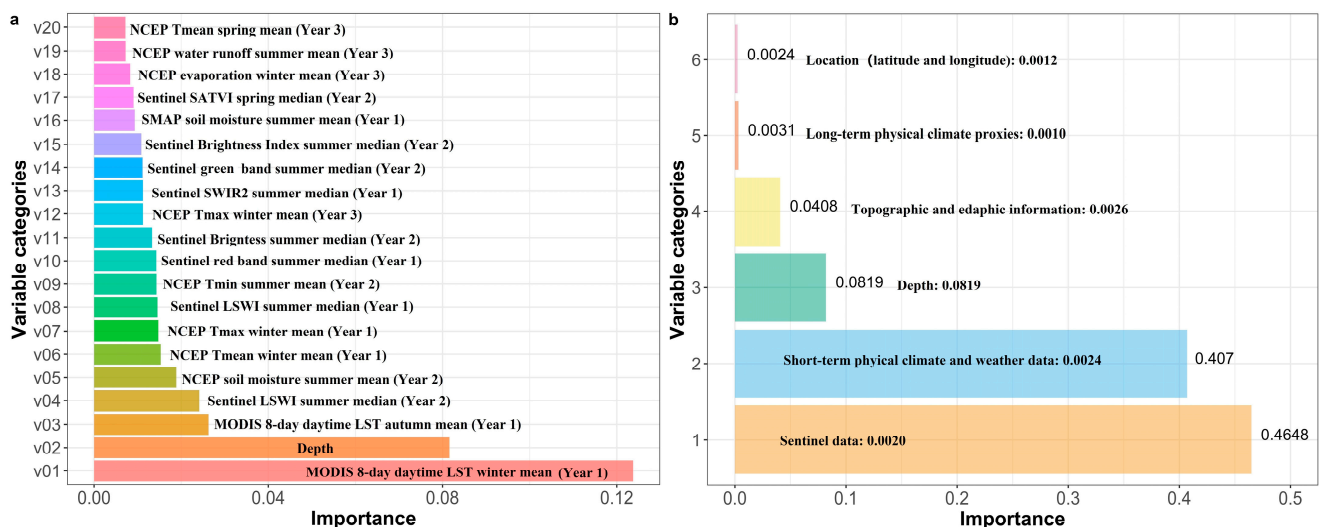


Figure 8. Importance of environmental variables used in A3-XGB-D3. (a) The top 20 most important variables and their importance values, and (b) importance of variables grouped into six categories

including Sentinel data, short-term physical climate and weather data, depth, topographic and edaphic information, long-term physical climate proxies, and location. Numbers after the bar plot in Panel *b* show the sum of variable importance of all variables in each category, and numbers after text labels in Panel *b* show the mean importance among variables in each category. Year 1 refers to the immediate past (or first) year prior to the soil sampling date, Year 2 is the immediate second year prior to the soil sampling date, and Year 3 represents the immediate third year prior to the soil sampling date. For example, if the soil samples were collected in September 2022, the immediate first year would be 1 September 2022–31 August 2022, the immediate second year would be 1 September 2021–31 August 2021, and the immediate third year would be 1 September 2020–31 August 2020.

6.3. Legacy Soil Samples to Improve Model Performance

Typically, legacy soil samples are used in two ways. First, these soil samples can be used as a baseline survey for studies aimed at quantifying changes in SOC over time [64,65]. This baseline information can facilitate understanding the soil's potential to sequester atmospheric carbon. Second, legacy soil samples are more commonly used to build predictive models because of the cost of collecting new field samples across broad scales [66]. However, there are many challenges associated with the inclusion of legacy soil samples in training predictive models. For example, these soil samples may be collected among different projects with various sampling strategies, different laboratory methods are used to obtain SOC percentage measurements, and inconsistencies exist in the sampling support such as data collected from different horizons or specific depth intervals [67]. Thus, efforts have been devoted to the harmonization of various soil samples for digital soil mapping efforts at regional and global scales [27,68]. Similarly, in the current study, legacy soil samples from RaCA were screened (soil samples at 0–5 and 0–30 cm) and selected (970 soil samples were selected) for model building. The use of RaCA soil samples greatly expanded the spatial coverage of the existing soil samples collected between 2020 and 2022. As shown in Figure 4a,d, SOC percentage values from the RaCA soil samples had a wider range (0.1–6.15%) than those (0.54–4.76%) from the soil samples collected between 2020 and 2022.

Theoretically, this wider range or spatial coverage of SOC percentage measurements, if used properly, should improve model performance since it helps with sample representativeness in both the model calibration (sample representativeness can also potentially help build a representative relationship between SOC and environmental covariates) and testing phases [69]. However, this improvement in model performance was complicated by missing values in the feature variables, i.e., remotely sensed and/or other variables, corresponding to the RaCA soil samples. In this study, missing values occurred mainly because the RaCA soil samples were collected in 2010–2011 when the corresponding Sentinel remotely sensed variables were not available yet. Landsat images can be an alternative to avoid missing values but need to be further examined for their synergy with Sentinel data in predicting SOC over time. As such, the predictive model used to leverage this wider range of SOC values or the broader spatial coverage that the RaCA samples provided needed to have the ability to deal with missing data in feature variables. Our results suggested that such an improvement in model prediction performance could be achieved with the XGB model using a down-weighting value on the RaCA soil samples. Although the enhancement in model prediction performance was only marginal, as depicted in Figure 7 (increases in model r^2 , RMSE, and MEC by 0.02, 0.01%, and 0.03 when set weight value at 0.05), the utilization of RaCA soil samples demonstrated the feasibility of incorporating historical soil survey data into model development. This suggests that soil samples from previous surveys, like RaCA, hold potential for cost-effective integration into carbon accounting frameworks for agriculture. This promising outcome hints at the possibility of achieving further cost reduction in carbon accounting processes within the agricultural sector.

In the XGB regression, instances can be provided with weight values to differentiate their importance in model training. The weight travels with the corresponding instance throughout the entire training phase to directly impact the split points. Compared with

GAM and other machine learning models, the XGB model can deal with missing values. Specifically, XGB uses the sparsity-aware split-finding algorithm [56] to deal with missing values in the training data. The algorithm not only returns the best split at each step but also the default direction for splitting should missing values occur. This step requires the calculation of the loss function twice to consider possible directions that missing values for a feature can take. It is also worth mentioning that there are methods such as mean or median imputation and k-nearest neighbors that can be used to deal with missing values before the modeling training. Nevertheless, these methods can induce further issues such as the distribution of imputed variables can become highly distorted or imputation values vary in space following an assumed pattern [70]. These issues can greatly affect model training and, in the worst scenario, completely deteriorate model performance.

7. Conclusions

Increases in organic carbon in soils have been considered one of the most important nature-based solutions to climate change mitigation and an important agenda in climate-smart agriculture. Accurate quantification of SOC content and stocks over agricultural lands is critical to the effective implementation of improved agricultural land management practices, such as cover crop planting and reductions in tillage, to increase SOC storage within agricultural lands. Although there are different types of approaches available to estimate SOC at depth, there is a lack of understanding regarding what depth strategy should be used to map SOC at a fine spatial resolution (10 m) over agricultural lands in the CONUS. In this study, three different types of approaches in two depth strategies were evaluated and compared for quantifying SOC percentage values for the top 30 cm agricultural soils, involving GAM and the ANN, RF, and XGB machine learning algorithms. Both the machine learning algorithms and the ways to convert depth-interval measurements into point observations greatly affected prediction performance. The best predictive model identified in this study was XGBoost regression using the full column depth assignment strategy, exhibiting an r^2 of 0.48, an RMSE of 0.29%, an ME of 0.06%, an MAE of 0.25%, and an MEC of 0.36 at the pixel level and yielded an r^2 of 0.64, an RMSE of 0.32%, an ME of −0.20%, an MAE of 0.28%, and an MEC of 0.48 at the field level. The analysis of variable importance in our model revealed that remotely sensed and short-term climate/weather variables were more important than variables such as long-term climate, location, and depth in characterizing SOC percentage at 0–30 cm. Further analysis suggested that legacy soil samples such as those from RaCA could be used in model building to further improve prediction performance. Therefore, this encouraging result suggests the potential for additional cost reductions in carbon accounting within the agricultural sector. Overall, this study showed that the full column depth strategy should be used in machine learning models to characterize 0–30 cm SOC content.

Author Contributions: Conceptualization, P.F., J.M. and J.R.K.; data curation, V.G., M.K.-Y., K.L. and L.A.W.; formal analysis, P.F., J.M. and J.R.K.; funding acquisition, D.W.S. and J.R.K.; investigation, P.F., C.C. and L.A.W.; methodology, P.F., J.M. and J.R.K.; project administration, D.W.S. and J.R.K.; resources, L.G. and D.W.S.; software, V.G., M.K.-Y., J.M. and K.L.; supervision, J.R.K.; validation, K.M.D. and L.G.; visualization, P.F.; writing—original draft, P.F.; writing—review and editing, C.C., J.M., D.W.S. and J.R.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by a grant from the Agricultural Products Utilization Commission Grant Program of the North Dakota Department of Agriculture to Perennial Climate Incorporated, by an Advanced Industries Early-Stage Capital and Retention award from the Colorado Office of Economic Development and International Trade to Perennial Climate Incorporated, and by Perennial Climate Incorporated in Boulder, Colorado.

Data Availability Statement: The original contributions presented in this study are included in the article material. Further inquiries can be directed to the corresponding author.

Acknowledgments: We would like to acknowledge Alexandre M. Wadoux for providing comments and suggestions on an earlier version of this manuscript.

Conflicts of Interest: PF and JRK are employed by Perennial Climate Inc. in addition to their respective university affiliations. All remaining authors are employed by Perennial Climate Inc., and all authors own stock in the company.

Appendix A

Table A1. Sampling dates for soil measurements (year-month-day).

Sampling Date	Sampling Date	Sampling Date	Sampling Date	Sampling Date
2010-11-03	2011-02-08	2010-10-14	2011-04-06	2011-06-22
2010-10-27	2010-11-01	2011-08-17	2011-04-01	2011-04-12
2021-04-06	2022-04-15	2010-11-02	2020-11-04	2010-11-18
2020-05-01	2011-02-17	2011-06-06	2011-03-31	2011-05-03
2011-06-14	2010-11-19	2021-04-01	2011-05-16	2021-03-30
2011-03-25	2010-11-09	2011-02-22	2011-04-07	2011-06-01
2010-12-02	2011-05-19	2011-07-01	2010-09-21	2010-11-05
2010-12-07	2011-04-25	2011-06-30	2011-04-18	2011-03-08
2010-09-30	2011-03-23	2010-10-29	2011-03-22	2011-03-14
2022-05-26	2011-08-09	2011-05-02	2011-05-04	2011-04-14
2011-01-12	2011-06-08	2022-04-13	2011-04-28	2011-02-02
2011-05-10	2011-03-07	2020-11-05	2011-03-11	2010-11-17
2010-10-12	2011-01-04	2011-06-29	2011-05-17	2011-04-13
2022-07-20	2010-10-22	2020-11-12	2021-04-23	2011-08-03
2021-10-13	2010-10-18	2010-09-03	2011-03-30	2010-12-14
2010-10-06	2010-10-21	2010-12-06	2011-09-21	2011-06-15
2011-06-07	2010-10-05	2010-11-22	2011-06-02	2010-09-28
2011-05-06	2010-11-04	2021-04-19	2010-10-28	2020-11-17
2020-05-12	2020-04-15	2020-11-03	2021-11-08	2022-06-09
2020-05-07	2021-03-31	2011-05-24	2011-04-11	2020-11-25
2022-04-25	2021-04-08	2010-10-15	2020-11-20	2020-11-26
2022-06-10	2022-04-22	2020-11-18	2022-04-18	2021-04-02
2020-11-06	2022-04-08	2021-04-21	2020-11-23	2011-06-20
2020-11-07	2021-10-16	2011-04-20	2020-12-02	2021-10-23
2011-04-04	2011-02-15	2011-04-15	2010-11-23	2011-04-27
2011-07-08	2011-01-06	2010-11-10	2010-12-13	2010-10-26
2011-04-26	2011-03-28	2011-05-23	2011-03-01	2011-09-08
2010-11-16	2011-05-25	2010-12-08	2011-02-16	2011-03-24
2011-05-26	2011-04-21	2011-09-02	2011-05-31	2011-03-29
2011-08-10	2011-03-21	2011-06-09	2011-07-12	2010-09-29
2011-05-11	2010-12-10	2011-05-09	2010-12-16	2010-10-25
2011-06-28	2010-09-22	2011-03-18	2010-10-20	2011-02-10
2011-08-01	2011-08-08	2010-11-12	2011-09-27	2011-04-22
2010-10-07	2010-12-15	2011-07-06	2011-02-18	2010-09-27
2011-06-03	2011-07-22	2010-12-09	2011-01-24	2011-07-27
2011-03-02	2010-10-19	2011-02-03	2010-11-08	2010-11-30
2011-09-19	2010-11-15	2011-06-23	2010-11-29	2010-12-03
2011-02-23	2010-10-13	2011-11-18	2011-11-03	2010-12-22

Table A1. *Cont.*

Sampling Date	Sampling Date	Sampling Date	Sampling Date	Sampling Date
2011-01-26	2011-05-20	2011-03-10	2011-09-12	2011-07-29
2011-01-20	2011-01-05	2011-03-15	2011-05-05	2011-05-18
2011-02-11	2011-01-27	2011-07-18	2010-09-16	2011-03-09
2011-09-20	2011-03-17	2011-07-28	2011-06-13	2010-10-03
2011-07-13	2011-06-17	2011-07-20	2011-06-10	2011-07-26
2010-12-21	2011-02-04	2011-02-24	2011-02-07	2011-01-07
2011-12-07	2010-12-01	2011-10-03	2011-05-12	2011-04-19
2021-04-07	2020-04-18	2010-12-28	2011-09-23	2010-12-27
2011-07-07	2011-09-26	2011-03-03	2010-08-25	2011-10-14
2011-05-29	2020-04-27	2010-08-30	2011-04-05	2011-06-21
2011-09-29	2011-03-04	2011-07-11	2011-01-28	2011-01-11
2011-06-16	2022-05-24	2010-09-14	2011-03-16	2011-08-29
2011-06-24	2011-05-22	2010-07-21	2011-10-11	2011-06-12
2021-04-05	2021-04-22	2011-09-22	2022-06-01	2021-11-18
2011-08-22	2010-10-08	2011-04-29	2011-02-28	2011-01-13
2022-04-09	2020-11-19	2020-05-03	2011-08-30	2011-01-03
2011-05-13	2011-08-02	2011-12-05	2010-10-01	2011-06-11
2011-01-19	2011-02-14	2011-08-18	2011-11-01	2011-07-14
2010-11-11	2010-09-15	2011-02-09	2010-12-29	2010-10-09
2020-04-30	2022-04-24	2020-11-02	2011-10-06	2022-05-25
2021-04-09	2020-05-16	2022-07-19	2020-11-15	2020-12-01
2011-08-04	2011-07-15	2011-09-01	2011-01-25	2011-01-31
2010-09-23	2010-12-20	2010-10-04	2011-04-02	2022-04-11
2011-07-25	2011-07-21	2011-07-19	2011-04-16	2010-09-20
2011-08-24	2011-01-21	2021-05-05	2020-10-30	2020-11-27
2011-02-01	2011-08-05	2021-11-20	2021-11-19	2021-11-01
2022-04-04	2020-05-11	2022-04-10	2020-12-09	2020-04-22
2021-11-04	2020-11-30	2020-05-21	2021-10-22	2021-04-10
2020-04-20	2020-04-23	2020-04-24		

Table A2. A list of environmental and remotely sensed data used in this study.

Data Type	Variable Name	Spatial Resolution
Long-term physical climate proxies	BIO1 (mean annual temperature)	~1 km
	BIO6 (precipitation of wettest quarter)	
	BIO17 (precipitation of the driest quarter)	
Short-term physical climate and weather data	Soil moisture	0.2°
	Mean air temperature	
	Potential evapotranspiration	
	Transpiration	
	Maximum air temperature	
	Minimum air temperature,	
	Shortwave radiation net flux	
	Sensible heat flux	
	Precipitation	
Topographic and edaphic information	Water runoff	10 m
	3DEP evaluation	
	SoilGrids gridded clay content	250 m
	SoilGrids gridded sand content	
	SoilGrids gridded silt content	

Table A2. Cont.

Data Type	Variable Name	Spatial Resolution
Remote sensing data	Sentinel-1 SAR imagery (VH and VV)	20 m
	Sentinel-2 Blue (B2, 458–533 nm)	10 m
	Sentinel-2 Green (B3, 543–578 nm)	10 m
	Sentinel-2 Red (B4, 650–680 nm)	10 m
	Sentinel-2 Near-Infrared (NIR, B8, 785–900 nm)	10 m
	Sentinel-2 Shortwave Infrared 1 (SWIR-1, B11, 1565–1655 nm)	20 m
	Sentinel-2 Shortwave Infrared 2 (SWIR-2, B12, 2100–2280 nm) bands	20 m
	Sentinel-2 spectral indices listed in Appendix A Table A3	10–20 m
	MODIS 8-day LST composite product (MOD11A2)	1 km
	SMAP L3 daily product	9 km

Table A3. The list of spectral indices, equations, and sources.

Spectral Indices	Equation	Source
Normalized difference vegetation index (NDVI)	$\frac{NIR - Red}{NIR + Red}$	[71]
Soil adjusted vegetation index (SAVI)	$1.5 * \frac{NIR - Red}{NIR + Red + 0.5}$	[72]
Soil adjusted total vegetation index (SATVI)	$1.5 * \frac{SWIR1 - Red}{SWIR1 + Red + 0.5} - \frac{SWIR2}{2}$	[73]
Bare Soil Index (BSI)	$\frac{(Red + SWIR1) - (NIR + Blue)}{(Red + SWIR1) + (NIR + Blue)}$	[74]
Normalized burn ratio (NBR2)	$\frac{NIR - SWIR2}{NIR + SWIR2}$	[75]
Normalized difference tillage index (NDTI)	$\frac{SWIR1 - SWIR2}{SWIR1 + SWIR2}$	[76]
Brightness index (BI)	$\sqrt{\frac{Red^2 + Green^2}{2}}$	[77]
Land surface water index (LSWI)	$\frac{NIR - SWIR1}{NIR + SWIR1}$	[78]
Tasseled cap brightness	$0.3510 * Blue + 0.3813 * Green + 0.3437 * Red + 0.7196 * NIR + 0.2396 * SWIR1 + 0.1949 * SWIR2$	[79]
Tasseled cap greenness	$-0.3599 * Blue - 0.3813 * Green - 0.3437 * Red + 0.7196 * NIR + 0.2396 * SWIR1 + 0.2856 * SWIR2$	
Tasseled cap wetness	$0.2578 * Blue + 0.2305 * Green + 0.0833 * Red + 0.1071 * NIR - 0.7611 * SWIR1 - 0.5308 * SWIR2$	

Note: For all the remotely sensed variables, the seasonal median values were computed and used as feature variables. Here, the seasonal median refers to the median value of each variable in a season, i.e., spring, summer, autumn, and winter. In this study, three years of remotely sensed data prior to the soil sampling date were used, and seasonal median composite images were generated in each of the three years.

Appendix B

To mitigate potential overfitting during model training, the recursive feature elimination (RFE) technique was applied to the top-performance model, specifically, A3-XGB-D3, as shown in Table 1. RFE was specifically applied to A3-XGB-D3, as it exhibited the smallest difference between training and testing r^2 . The steps for implementing RFE are listed below.

- Step 1: The model is trained on the initial set of features, and the importance of each feature is determined.
- Step 2: The least important feature(s) are removed from the current set of features.
- Step 3: The model is then re-trained on the reduced set of features, and the process is repeated.

This recursive elimination of features continued until either a specified number of features was reached, or further removal resulted in a decline in model performance, as indicated by a chosen metric such as RMSE or r^2 . The results of this analysis are in Table A4.

Table A4. The model performance for XGB with the use of recursive feature elimination.

Type	Cross-Validation		Testing (0–30 cm)	
	RMSE (%)	r^2	RMSE (%)	r^2
A3-XGB-D3	0.30	0.63	0.31	0.43

Comparison of this model to the version originally reported in the manuscript (Table 2) is instructive. The original model exhibited training and testing r^2 of 0.72 and 0.48, respectively, a difference of 0.24. The updated model, based on RFE, exhibited a training and testing r^2 of 0.63 and 0.43, respectively, a difference of 0.2. RFE was able to reduce the difference between training and testing performance (0.2 vs. 0.24) but did so at an absolute cost of testing performance (0.43 vs. 0.48). This strongly suggests that the original model was not overfitted and that the difference between training and testing performance is due to inherent differences between these two data sources.

References

1. Billings, S.A.; Lajtha, K.; Malhotra, A.; Berhe, A.A.; Graaff, M.-A.; Earl, S.; Fraterrigo, J.; Georgiou, K.; Grandy, S.; Hobbie, S.E.; et al. Soil Organic Carbon Is Not Just for Soil Scientists: Measurement Recommendations for Diverse Practitioners. *Ecol. Appl.* **2021**, *31*, e02290. [\[CrossRef\]](#)
2. Lal, R. Carbon Sequestration. *Phil. Trans. R. Soc. B* **2008**, *363*, 815–830. [\[CrossRef\]](#)
3. Lorenz, K.; Lal, R.; Ehlers, K. Soil Organic Carbon Stock as an Indicator for Monitoring Land and Soil Degradation in Relation to U Nited N Ations' S Ustainable D Evelopment G Oals. *Land Degrad Dev* **2019**, *30*, 824–838. [\[CrossRef\]](#)
4. Amelung, W.; Bossio, D.; de Vries, W.; Kögel-Knabner, I.; Lehmann, J.; Amundson, R.; Bol, R.; Collins, C.; Lal, R.; Leifeld, J.; et al. Towards a Global-Scale Soil Climate Mitigation Strategy. *Nat Commun* **2020**, *11*, 5427. [\[CrossRef\]](#)
5. Minasny, B.; Malone, B.P.; McBratney, A.B.; Angers, D.A.; Arrouays, D.; Chambers, A.; Chaplot, V.; Chen, Z.-S.; Cheng, K.; Das, B.S.; et al. Soil Carbon 4 per Mille. *Geoderma* **2017**, *292*, 59–86. [\[CrossRef\]](#)
6. Rumpel, C.; Amiraslani, F.; Koutika, L.-S.; Smith, P.; Whitehead, D.; Wollenberg, E. Put More Carbon in Soils to Meet Paris Climate Pledges. *Nature* **2018**, *564*, 32–34. [\[CrossRef\]](#)
7. Ramifehiarivo, N.; Chevallier, T.; Defrance, D.; Brossard, M.; Chotte, J.-L. Framing the Future of the Koronivia Joint Work on Agriculture from Science-Based Evidence. A Review. *Agron. Sustain. Dev.* **2022**, *42*, 102. [\[CrossRef\]](#)
8. Rietra, R.; Lesschen, J.; Porre, R. *Recarbonizing Global Soils: A Technical Manual of Recommended Management Practices: Volume 3-Cropland, Grassland, Integrated Systems and Farming Approaches-Practices Overview*; FAO: Rome, Italy, 2021.
9. Eggleston, H.S.; Buendia, L.; Miwa, K.; Ngara, T.; Tanabe, K. 2006 IPCC Guidelines for National Greenhouse Gas Inventories; IPCC: Bannockburn, IL, USA, 2006.
10. Smith, P.; Soussana, J.; Angers, D.; Schipper, L.; Chenu, C.; Rasse, D.P.; Batjes, N.H.; Egmond, F.; McNeill, S.; Kuhnert, M.; et al. How to Measure, Report and Verify Soil Carbon Change to Realize the Potential of Soil Carbon Sequestration for Atmospheric Greenhouse Gas Removal. *Glob. Change Biol.* **2020**, *26*, 219–241. [\[CrossRef\]](#)
11. Minasny, B.; McBratney, A.B.; Malone, B.P.; Wheeler, I. Digital Mapping of Soil Carbon. In *Advances in Agronomy*; Elsevier: Amsterdam, The Netherlands, 2013; Volume 118, pp. 1–47, ISBN 978-0-12-405942-9.
12. McBratney, A.B.; Mendonça Santos, M.L.; Minasny, B. On Digital Soil Mapping. *Geoderma* **2003**, *117*, 3–52. [\[CrossRef\]](#)
13. Florinsky, I.V. The Dokuchaev Hypothesis as a Basis for Predictive Digital Soil Mapping (on the 125th Anniversary of Its Publication). *Eurasian Soil Sc.* **2012**, *45*, 445–451. [\[CrossRef\]](#)
14. Lacoste, M.; Minasny, B.; McBratney, A.; Michot, D.; Viaud, V.; Walter, C. High Resolution 3D Mapping of Soil Organic Carbon in a Heterogeneous Agricultural Landscape. *Geoderma* **2014**, *213*, 296–311. [\[CrossRef\]](#)
15. Grimm, R.; Behrens, T.; Märker, M.; Elsenbeer, H. Soil Organic Carbon Concentrations and Stocks on Barro Colorado Island—Digital Soil Mapping Using Random Forests Analysis. *Geoderma* **2008**, *146*, 102–113. [\[CrossRef\]](#)
16. Aitkenhead, M.J.; Coull, M.C. Mapping Soil Carbon Stocks across Scotland Using a Neural Network Model. *Geoderma* **2016**, *262*, 187–198. [\[CrossRef\]](#)
17. Hengl, T.; Mendes de Jesus, J.; Heuvelink, G.B.M.; Ruiperez Gonzalez, M.; Kilibarda, M.; Blagotić, A.; Shangguan, W.; Wright, M.N.; Geng, X.; Bauer-Marschallinger, B.; et al. SoilGrids250m: Global Gridded Soil Information Based on Machine Learning. *PLoS ONE* **2017**, *12*, e0169748. [\[CrossRef\]](#)
18. Padarian, J.; Minasny, B.; McBratney, A.B. Using Deep Learning for Digital Soil Mapping. *SOIL* **2019**, *5*, 79–89. [\[CrossRef\]](#)
19. Arrouays, D.; Grundy, M.G.; Hartemink, A.E.; Hempel, J.W.; Heuvelink, G.B.M.; Hong, S.Y.; Lagacherie, P.; Lelyk, G.; McBratney, A.B.; McKenzie, N.J.; et al. GlobalSoilMap. In *Advances in Agronomy*; Elsevier: Amsterdam, The Netherlands, 2014; Volume 125, pp. 93–134, ISBN 978-0-12-800137-0.
20. Minasny, B.; McBratney, A.B.; Mendonça-Santos, M.L.; Odeh, I.O.A.; Guyon, B. Prediction and Digital Mapping of Soil Carbon Storage in the Lower Namoi Valley. *Soil Res.* **2006**, *44*, 233. [\[CrossRef\]](#)
21. Veronesi, F.; Corstanje, R.; Mayr, T. Landscape Scale Estimation of Soil Carbon Stock Using 3D Modelling. *Sci. Total Environ.* **2014**, *487*, 578–586. [\[CrossRef\]](#)
22. Poggio, L.; Gimona, A. National Scale 3D Modelling of Soil Organic Carbon Stocks with Uncertainty Propagation—An Example from Scotland. *Geoderma* **2014**, *232–234*, 284–299. [\[CrossRef\]](#)
23. Sothe, C.; Gonsamo, A.; Arabian, J.; Snider, J. Large Scale Mapping of Soil Organic Carbon Concentration with 3D Machine Learning and Satellite Observations. *Geoderma* **2022**, *405*, 115402. [\[CrossRef\]](#)

24. Liu, F.; Rossiter, D.G.; Song, X.-D.; Zhang, G.-L.; Yang, R.-M.; Zhao, Y.-G.; Li, D.-C.; Ju, B. A Similarity-Based Method for Three-Dimensional Prediction of Soil Organic Matter Concentration. *Geoderma* **2016**, *263*, 254–263. [\[CrossRef\]](#)
25. Bishop, T.F.A.; McBratney, A.B.; Laslett, G.M. Modelling Soil Attribute Depth Functions with Equal-Area Quadratic Smoothing Splines. *Geoderma* **1999**, *91*, 27–45. [\[CrossRef\]](#)
26. Rentschler, T.; Werban, U.; Ahner, M.; Behrens, T.; Gries, P.; Scholten, T.; Teuber, S.; Schmidt, K. 3D Mapping of Soil Organic Carbon Content and Soil Moisture with Multiple Geophysical Sensors and Machine Learning. *Vadose Zone J.* **2020**, *19*, e20062. [\[CrossRef\]](#)
27. Hengl, T.; de Jesus, J.M.; MacMillan, R.A.; Batjes, N.H.; Heuvelink, G.B.M.; Ribeiro, E.; Samuel-Rosa, A.; Kempen, B.; Leenaars, J.G.B.; Walsh, M.G.; et al. SoilGrids1km—Global Soil Information Based on Automated Mapping. *PLoS ONE* **2014**, *9*, e105992. [\[CrossRef\]](#)
28. Orton, T.G.; Pringle, M.J.; Bishop, T.F.A.; Menzies, N.W.; Dang, Y.P. Increment-Averaged Kriging for 3-D Modelling and Mapping Soil Properties: Combining Machine Learning and Geostatistical Methods. *Geoderma* **2020**, *361*, 114094. [\[CrossRef\]](#)
29. Orton, T.G.; Pringle, M.J.; Bishop, T.F.A. A One-Step Approach for Modelling and Mapping Soil Properties Based on Profile Data Sampled over Varying Depth Intervals. *Geoderma* **2016**, *262*, 174–186. [\[CrossRef\]](#)
30. Wadoux, A.M.J.-C.; Minasny, B.; McBratney, A.B. Machine Learning for Digital Soil Mapping: Applications, Challenges and Suggested Solutions. *Earth-Sci. Rev.* **2020**, *210*, 103359. [\[CrossRef\]](#)
31. Aitkenhead, M.; Coull, M. Mapping Soil Profile Depth, Bulk Density and Carbon Stock in Scotland Using Remote Sensing and Spatial Covariates. *Eur. J. Soil Sci.* **2020**, *71*, 553–567. [\[CrossRef\]](#)
32. Ma, Y.; Minasny, B.; McBratney, A.; Poggio, L.; Fajardo, M. Predicting Soil Properties in 3D: Should Depth Be a Covariate? *Geoderma* **2021**, *383*, 114794. [\[CrossRef\]](#)
33. Nauman, T.W.; Duniway, M.C. Relative Prediction Intervals Reveal Larger Uncertainty in 3D Approaches to Predictive Digital Soil Mapping of Soil Properties with Legacy Data. *Geoderma* **2019**, *347*, 170–184. [\[CrossRef\]](#)
34. Minasny, B.; Stockmann, U.; Hartemink, A.E.; McBratney, A.B. Measuring and Modelling Soil Depth Functions. In *Digital Soil Morphometrics*; Hartemink, A.E., Minasny, B., Eds.; Progress in Soil Science; Springer International Publishing: Cham, Switzerland, 2016; pp. 225–240, ISBN 978-3-319-28294-7.
35. Aldana Jague, E.; Sommer, M.; Saby, N.P.A.; Cornelis, J.-T.; Van Wesemael, B.; Van Oost, K. High Resolution Characterization of the Soil Organic Carbon Depth Profile in a Soil Landscape Affected by Erosion. *Soil Tillage Res.* **2016**, *156*, 185–193. [\[CrossRef\]](#)
36. Mishra, U.; Lal, R.; Slater, B.; Calhoun, F.; Liu, D.; Van Meirvenne, M. Predicting Soil Organic Carbon Stock Using Profile Depth Distribution Functions and Ordinary Kriging. *Soil Sci. Soc. Am. J.* **2009**, *73*, 614–621. [\[CrossRef\]](#)
37. Malone, B.P.; Odgers, N.P.; Stockmann, U.; Minasny, B.; McBratney, A.B. Digital Mapping of Soil Classes and Continuous Soil Properties. In *Pedometrics*; McBratney, A.B., Minasny, B., Stockmann, U., Eds.; Progress in Soil Science; Springer International Publishing: Cham, Switzerland, 2018; pp. 373–413, ISBN 978-3-319-63437-1.
38. Chen, C.; Hu, K.; Li, H.; Yun, A.; Li, B. Three-Dimensional Mapping of Soil Organic Carbon by Combining Kriging Method with Profile Depth Function. *PLoS ONE* **2015**, *10*, e0129038. [\[CrossRef\]](#)
39. Bradford, M.A.; Carey, C.J.; Atwood, L.; Bossio, D.; Fenichel, E.P.; Gennet, S.; Fargione, J.; Fisher, J.R.B.; Fuller, E.; Kane, D.A.; et al. Soil Carbon Science for Policy and Practice. *Nat. Sustain.* **2019**, *2*, 1070–1072. [\[CrossRef\]](#)
40. Gillenwater, M.; Broekhoff, D.; Trexler, M.; Hyman, J.; Fowler, R. Policing the Voluntary Carbon Market. *Nat. Clim. Change* **2007**, *1*, 85–87. [\[CrossRef\]](#)
41. Boryan, C.; Yang, Z.; Mueller, R.; Craig, M. Monitoring US Agriculture: The US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program. *Geocarto Int.* **2011**, *26*, 341–358. [\[CrossRef\]](#)
42. Rabenhorst, M.C. Determination of Organic and Carbonate Carbon in Calcareous Soils Using Dry Combustion. *Soil Sci. Soc. Am. J.* **1988**, *52*, 965–968. [\[CrossRef\]](#)
43. Wills, S.; Loecke, T.; Sequeira, C.; Teachman, G.; Grunwald, S.; West, L.T. Overview of the U.S. Rapid Carbon Assessment Project: Sampling Design, Initial Summary and Uncertainty Estimates. In *Soil Carbon*; Hartemink, A.E., McSweeney, K., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 95–104, ISBN 978-3-319-04083-7.
44. Ramcharan, A.; Hengl, T.; Nauman, T.; Brungard, C.; Waltman, S.; Wills, S.; Thompson, J. Soil Property and Class Maps of the Conterminous United States at 100-Meter Spatial Resolution. *Soil Sci. Soc. Am. J.* **2018**, *82*, 186–201. [\[CrossRef\]](#)
45. Sanderman, J.; Hengl, T.; Fiske, G.; Solvik, K.; Adame, M.F.; Benson, L.; Bukoski, J.J.; Carnell, P.; Cifuentes-Jara, M.; Donato, D.; et al. A Global Map of Mangrove Forest Soil Carbon at 30 m Spatial Resolution. *Environ. Res. Lett.* **2018**, *13*, 055002. [\[CrossRef\]](#)
46. Fick, S.E.; Hijmans, R.J. WorldClim 2: New 1-km Spatial Resolution Climate Surfaces for Global Land Areas. *Int. J. Clim.* **2017**, *37*, 4302–4315. [\[CrossRef\]](#)
47. Saha, S.; Moorthi, S.; Pan, H.-L.; Wu, X.; Wang, J.; Nadiga, S.; Tripp, P.; Kistler, R.; Woollen, J.; Behringer, D.; et al. The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.* **2010**, *91*, 1015–1058. [\[CrossRef\]](#)
48. Saha, S.; Moorthi, S.; Wu, X.; Wang, J.; Nadiga, S.; Tripp, P.; Behringer, D.; Hou, Y.-T.; Chuang, H.; Iredell, M.; et al. The NCEP Climate Forecast System Version 2. *J. Clim.* **2014**, *27*, 2185–2208. [\[CrossRef\]](#)
49. Stoker, J.; Miller, B. The Accuracy and Consistency of 3D Elevation Program Data: A Systematic Analysis. *Remote Sens.* **2022**, *14*, 940. [\[CrossRef\]](#)

50. Filipponi, F. Sentinel-1 GRD Preprocessing Workflow. *Proceedings* **2019**, *18*, 11. [\[CrossRef\]](#)
51. Main-Knorn, M.; Pflug, B.; Louis, J.; Debaecker, V.; Müller-Wilm, U.; Gascon, F. Sen2Cor for Sentinel-2. In *Proceedings of the Image and Signal Processing for Remote Sensing XXIII*; Bruzzone, L., Bovolo, F., Benediktsson, J.A., Eds.; SPIE: Warsaw, Poland, 2017; p. 3.
52. Skakun, S.; Wevers, J.; Brockmann, C.; Doxani, G.; Aleksandrov, M.; Batič, M.; Frantz, D.; Gascon, F.; Gómez-Chova, L.; Hagolle, O.; et al. Cloud Mask Intercomparison eXercise (CMIX): An Evaluation of Cloud Masking Algorithms for Landsat 8 and Sentinel-2. *Remote Sens. Environ.* **2022**, *274*, 112990. [\[CrossRef\]](#)
53. Wan, Z. New Refinements and Validation of the Collection-6 MODIS Land-Surface Temperature/Emissivity Product. *Remote Sens. Environ.* **2014**, *140*, 36–45. [\[CrossRef\]](#)
54. Ahirwal, J.; Nath, A.; Brahma, B.; Deb, S.; Sahoo, U.K.; Nath, A.J. Patterns and Driving Factors of Biomass Carbon and Soil Organic Carbon Stock in the Indian Himalayan Region. *Sci. Total Environ.* **2021**, *770*, 145292. [\[CrossRef\]](#)
55. Were, K.; Bui, D.T.; Dick, Ø.B.; Singh, B.R. A Comparative Assessment of Support Vector Regression, Artificial Neural Networks, and Random Forests for Predicting and Mapping Soil Organic Carbon Stocks across an Afromontane Landscape. *Ecol. Indic.* **2015**, *52*, 394–403. [\[CrossRef\]](#)
56. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: San Francisco, CA, USA, 2016; pp. 785–794.
57. Klein, A.; Falkner, S.; Bartels, S.; Hennig, P.; Hutter, F. Fast Bayesian Optimization of Machine Learning Hyperparameters on Large Datasets. In Proceedings of the Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017; PMLR: London, UK, 2017; pp. 528–536.
58. Breiman, L. Random Forest. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
59. Jain, A.K.; Mao, J.; Mohiuddin, K.M. Artificial Neural Networks: A Tutorial. *Computer* **1996**, *29*, 31–44. [\[CrossRef\]](#)
60. Kimes, D.S.; Nelson, R.F.; Manry, M.T.; Fung, A.K. Review Article: Attributes of Neural Networks for Extracting Continuous Vegetation Variables from Optical and Radar Measurements. *Int. J. Remote Sens.* **1998**, *19*, 2639–2663. [\[CrossRef\]](#)
61. Hastie, T.; Tibshirani, R. Generalized Additive Models. *Stat. Sci.* **1986**, *1*, 297–318. [\[CrossRef\]](#)
62. Augustin, N.H.; Musio, M.; von Wilpert, K.; Kublin, E.; Wood, S.N.; Schumacher, M. Modeling Spatiotemporal Forest Health Monitoring Data. *J. Am. Stat. Assoc.* **2009**, *104*, 899–911. [\[CrossRef\]](#)
63. Hounkpatin, K.O.L.; Stendahl, J.; Lundblad, M.; Karlton, E. Predicting the Spatial Distribution of Soil Organic Carbon Stock in Swedish Forests Using a Group of Covariates and Site-Specific Data. *SOIL* **2021**, *7*, 377–398. [\[CrossRef\]](#)
64. Karunaratne, S.B.; Bishop, T.F.A.; Odeh, I.O.A.; Baldock, J.A.; Marchant, B.P. Estimating Change in Soil Organic Carbon Using Legacy Data as the Baseline: Issues, Approaches and Lessons to Learn. *Soil Res.* **2014**, *52*, 349. [\[CrossRef\]](#)
65. Mukumbuta, I.; Chabala, L.M.; Sichinga, S.; Lark, R.M. Accessing and Assessing Legacy Soil Information, an Example from Two Provinces of Zambia. *Geoderma* **2022**, *420*, 115874. [\[CrossRef\]](#)
66. Odeh, I.O.; Leenaars, J.; Hartemink, A.; Amapu, I. The Challenges of Collating Legacy Data for Digital Mapping of Nigerian Soils. *Digit. Soil Assess. Beyond* **2012**, *453–458*.
67. Schillaci, C.; Acutis, M.; Vesely, F.; Saia, S. A Simple Pipeline for the Assessment of Legacy Soil Datasets: An Example and Test with Soil Organic Carbon from a Highly Variable Area. *Catena* **2019**, *175*, 110–122. [\[CrossRef\]](#)
68. Zádorová, T.; Skála, J.; Žižala, D.; Vaněk, A.; Penížek, V. Harmonization of a Large-Scale National Soil Database with the World Reference Base for Soil Resources 2014. *Geoderma* **2021**, *384*, 114819. [\[CrossRef\]](#)
69. Borovicka, T.; Jirina Jr, M.; Kordik, P.; Jirina, M. Selecting Representative Data Sets. *Adv. Data Min. Knowl. Discov. Appl.* **2012**, *12*, 43–70.
70. Lodder, P. To Impute or Not Impute: That's the Question. In *Advising on Research Methods: Selected Topics 2013*; Johannes van Kessel Publishing: Huizen, The Netherlands, 2013; pp. 1–7.
71. Tucker, C.J. Red and Photographic Infrared Linear Combinations for Monitoring Vegetation. *Remote Sens. Environ.* **1979**, *8*, 127–150. [\[CrossRef\]](#)
72. Huete, A.R. A Soil-Adjusted Vegetation Index (SAVI). *Remote Sens. Environ.* **1988**, *25*, 295–309. [\[CrossRef\]](#)
73. Marsett, R.C.; Qi, J.; Heilman, P.; Biedenbender, S.H.; Carolyn Watson, M.; Amer, S.; Weltz, M.; Goodrich, D.; Marsett, R. Remote Sensing for Grassland Management in the Arid Southwest. *Rangel. Ecol. Manag.* **2006**, *59*, 530–540. [\[CrossRef\]](#)
74. Diek, S.; Fornallaz, F.; Schaepman, M.E. Rogier De Jong Barest Pixel Composite for Agricultural Areas Using Landsat Time Series. *Remote Sens.* **2017**, *9*, 1245. [\[CrossRef\]](#)
75. Demattê, J.A.M.; Fongaro, C.T.; Rizzo, R.; Safanelli, J.L. Geospatial Soil Sensing System (GEOS3): A Powerful Data Mining Procedure to Retrieve Soil Spectral Reflectance from Satellite Images. *Remote Sens. Environ.* **2018**, *212*, 161–175. [\[CrossRef\]](#)
76. Van Deventer, A.P.; Ward, A.D.; Gowda, P.H.; Lyon, J.G. Using Thematic Mapper Data to Identify Contrasting Soil Plains and Tillage Practices. *Photogramm. Eng. Remote Sens.* **1997**, *63*, 87–93.
77. Marques, M.J.; Alvarez, A.; Carral, P.; Sastre, B.; Bienes, R. The Use of Remote Sensing to Detect the Consequences of Erosion in Gypsiferous Soils. *Int. Soil Water Conserv. Res.* **2020**, *8*, 383–392. [\[CrossRef\]](#)

-
78. Chandrasekar, K.; Sesha Sai, M.V.R.; Roy, P.S.; Dwevedi, R.S. Land Surface Water Index (LSWI) Response to Rainfall and NDVI Using the MODIS Vegetation Index Product. *Int. J. Remote Sens.* **2010**, *31*, 3987–4005. [[CrossRef](#)]
 79. Shi, T.; Xu, H. Derivation of Tasseled Cap Transformation Coefficients for Sentinel-2 MSI At-Sensor Reflectance Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 4038–4048. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.