JGR Biogeosciences

METHOD

10.1029/2023JG007702

Key Points:

- We estimated soil organic carbon (SOC) stocks at the continental scale by coupling multivariate geographic clustering with machine learning
- Independent validation showed our approach ($R^2 = 0.51$) outperformed Harmonized World Soil Database ($R^2 = 0.23$) and SoilGrids 2.0 ($R^2 = 0.39$) for the topsoil
- Low representativeness and high uncertainty regions can inform future sampling efforts to further improve SOC estimation

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

D. Sihi,
debjani.sihi@emory.edu

Citation:

Wang, Z., Kumar, J., Weintraub-Leff, S. R., Todd-Brown, K., Mishra, U., & Sihi, D. (2024). Upscaling soil organic carbon measurements at the continental scale using multivariate clustering analysis and machine learning. *Journal of Geophysical Research: Biogeosciences*, *129*, e2023JG007702. https://doi.org/10.1029/2023JG007702

Received 10 JUL 2023
Accepted 15 JAN 2024

Author Contributions:

Conceptualization: Jitendra Kumar, Samantha R. Weintraub-Leff, Debjani Sihi
Data curation: Zhuonan Wang, Samantha R. Weintraub-Leff, Katherine Todd-Brown, Umakant Mishra
Formal analysis: Zhuonan Wang
Funding acquisition: Jitendra Kumar, Samantha R. Weintraub-Leff, Debjani Sihi
Investigation: Zhuonan Wang
Methodology: Jitendra Kumar, Katherine Todd-Brown
Project Administration: Debjani Sihi
Supervision: Debjani Sihi

Upscaling Soil Organic Carbon Measurements at the Continental Scale Using Multivariate Clustering Analysis and Machine Learning

Zhuonan Wang[1], Jitendra Kumar[2], Samantha R. Weintraub-Leff[3], Katherine Todd-Brown[4], Umakant Mishra[5], and Debjani Sihi[1]

[1]Department of Environmental Sciences, Emory University, Atlanta, GA, USA, [2]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA, [3]National Ecological Observatory Network, Battelle, Boulder, CO, USA, [4]Department of Environmental Engineering Science, University of Florida, Gainesville, FL, USA, [5]Computational Biology & Biophysics, Sandia National Laboratories, Livermore, CA, USA

Abstract Estimates of soil organic carbon (SOC) stocks are essential for many environmental applications. However, significant inconsistencies exist in SOC stock estimates for the U.S. across current SOC maps. We propose a framework that combines unsupervised multivariate geographic clustering (MGC) and supervised Random Forests regression, improving SOC maps by capturing heterogeneous relationships with SOC drivers. We first used MGC to divide the U.S. into 20 SOC regions based on the similarity of covariates (soil biogeochemical, bioclimatic, biological, and physiographic variables). Subsequently, separate Random Forests models were trained for each SOC region, utilizing environmental covariates and SOC observations. Our estimated SOC stocks for the U.S. (52.6 ± 3.2 Pg for 0–30 cm and 108.3 ± 8.2 Pg for 0–100 cm depth) were within the range estimated by existing products like Harmonized World Soil Database, HWSD (46.7 Pg for 0–30 cm and 90.7 Pg for 0–100 cm depth) and SoilGrids 2.0 (45.7 Pg for 0–30 cm and 133.0 Pg for 0–100 cm depth). However, independent validation with soil profile data from the National Ecological Observatory Network showed that our approach ($R^2 = 0.51$) outperformed the estimates obtained from Harmonized World Soil Database ($R^2 = 0.23$) and SoilGrids 2.0 ($R^2 = 0.39$) for the topsoil (0–30 cm). Uncertainty analysis (e.g., low representativeness and high coefficients of variation) identified regions requiring more measurements, such as Alaska and the deserts of the U.S. Southwest. Our approach effectively captures the heterogeneous relationships between widely available predictors and the current SOC baseline across regions, offering reliable SOC estimates at 1 km resolution for benchmarking Earth system models.

Plain Language Summary Soils represent the largest terrestrial carbon (C) pool. To understand how soil C will change under a changing climate, we first need to have an accurate estimate of how much soil organic carbon (SOC) is present. However, SOC maps for the U.S. are highly variable. In this study, we developed a new framework for estimating SOC stocks across the entire U.S. using data from site-level measurements. We first divided the U.S. into 20 regions based on environmental conditions and then created machine-learning models for each region to make an accurate, continuous map. Our method was able to capture different relationships between environmental variables and SOC across regions and improved the overall estimates for the amount of SOC found in U.S. soils. While in all regions, climate was an important variable for predicting SOC, soil properties, plant inputs, and elevation played key roles in some regions. As part of this effort, we identified areas with high uncertainty, which could be target regions for additional measurements in the future (e.g., Alaska and the deserts of the Southwest U.S.). Our method provides new insights for the soil mapping community and yields robust SOC estimates that can inform the terrestrial C cycle in models.

1. Introduction

Soils represent the largest terrestrial carbon (C) pool, containing approximately twice as much C as the atmosphere and vegetation combined (Canadell et al., 2021; Lal, 2004b). Three-fifths of the global soil C is soil organic carbon (SOC)—the main component of soil organic matter with essential effects on promoting soil health, the functioning of terrestrial ecosystems, and the global C cycle (FAO, 2018; Lal, 2004b). However, in the face of climate change, land use change, and other environmental pressures, it remains unclear if soils will continue to be a C sink or instead become a C source (Canadell et al., 2021; Friedlingstein et al., 2022; Gautam et al., 2022). A

Visualization: Zhuonan Wang
Writing – original draft: Zhuonan Wang
Writing – review & editing: Jitendra Kumar, Samantha R. Weintraub-Leff, Katherine Todd-Brown, Umakant Mishra, Debjani Sihi

small relative change in either direction would significantly affect atmospheric carbon dioxide (CO_2) concentrations, resulting in a strong feedback effect on future climate (Cox et al., 2000).

Resolving this question is complicated partially due to uncertainties in the size of current SOC pools (Schrumpf et al., 2011; Todd-Brown et al., 2013). Significant discrepancies exist in estimates of the amount and spatial distribution of SOC stocks (Köchy et al., 2015; Scharlemann et al., 2014). For instance, there is disagreement in the estimation of SOC stocks across commonly used gridded databases, such as SoilGrids, the Harmonized World Soil Database (HWSD), and the Northern Circumpolar Soil Carbon Database (NCSCD), both regionally and globally (Tifafi et al., 2018). Also, there are notable mismatches in the spatial patterns of SOC stocks between the Unified North American Soil Map and HWSD version 1.21 across North America (Liu et al., 2013). Having SOC maps and gridded products that are as accurate as possible is critical because they are relevant to many applications, from setting land management and carbon policy to Earth system model benchmarking.

SOC stocks vary spatially due to factors such as climate, soil type, and land use. Mapping SOC stocks can help identify areas with high C sequestration potential (Rumpel et al., 2020; Smith et al., 2020; Vågen & Winowiecki, 2013) or regions more susceptible to climate change impacts (Ahmed et al., 2017). Significant efforts have been made to collect and upscale soil profile data for mapping SOC stocks at regional or global scale (Amundson, 2001; Batjes, 1996; Chaney et al., 2019; FAO & ITPS, 2020; Guevara et al., 2020; Hengl et al., 2014, 2017; Mishra et al., 2022; Ramcharan et al., 2018; Scharlemann et al., 2014; Stockmann et al., 2015; Tarnocai et al., 2009). Soil C mapping methods have been constantly evolving, leveraging refinement and innovation from various fields to enhance map accuracy. Initially, soil C maps relied on time-consuming, labor-intensive field soil surveys (Brevik et al., 2016). With the development of computer systems and geographic information systems, conventional upscaling methods (class- and geo-matching approaches) were used to derive soil C maps (Batjes, 2000; FAO, 2018; Lettens et al., 2004). However, these maps relied on expert-informed mapping units and did not consider spatially explicit uncertainty assessments.

Subsequently, algorithmically generated mapping units rapidly improved soil C mapping due to advancements in remote-sensing technologies, geospatial data sets, and machine learning (McBratney et al., 2003; Minasny & McBratney, 2016; Mishra et al., 2010; Scull et al., 2003). Broadly, machine learning algorithms define soil mapping units via statistical relationships between remote-sensing data on environmental factors and observed soil C at georeferenced sample locations (aka Digital Soil Mapping [DSM]). Random Forests and variations on this algorithm have shown promise in estimating SOC stocks (Li et al., 2022; Padarian et al., 2020; Zhang et al., 2023). For example, the widely used global soil properties data sets, SoilGrids250m and SoilGrids 2.0, were generated using Random Forests and Quantile Regression Forests (QRF) (Hengl et al., 2017; Poggio et al., 2021). QRF was also used to predict the spatial distribution of SOC stocks (0–30 cm depth) across Mexico and the conterminous United States (CONUS) at 250 m resolution (Guevara et al., 2020).

Despite these advances, the way Random Forests approaches have been applied to mapping SOC to date has some limitations. Researchers often implement a single model across a large areal extent when fitting quantitative relationships between SOC and covariates. A DSM based on a single model may not capture the spatially heterogeneous environmental factors influencing SOC stocks at global or continental scales. Multiple studies working at broad scales have highlighted that dominant environmental controllers of SOC stocks vary spatially (Gonçalves et al., 2021; Mishra et al., 2021; Rasmussen et al., 2018; Vitharana et al., 2019). For instance, country-specific predictors for SOC and their respective weights varied across Latin America and no universal predictive algorithm was established among these countries (Guevara et al., 2018).

Recent studies demonstrated that segmenting a region of interest into mapping units with similar environmental conditions, then predicting SOC within each unit using unique models, is a valid approach to capturing the spatially diverse environmental factors that control SOC stocks (Chen et al., 2019; Song et al., 2020). Multivariate geographic clustering (MGC), a statistical algorithm, can contribute to existing DSM approaches by revealing otherwise hard-to-capture patterns in SOC data (Hargrove & Hoffman, 2004). Hargrove and Hoffman (1999) first developed MGC to define clustered regions based on their representativeness (i.e., similarity) across multiple variables. Initially, MGC was applied to delineate and visualize ecoregions in CONUS and evaluate the representativeness of the AmeriFlux network (Hargrove et al., 2003; Hargrove & Hoffman, 2004). Since then, the ecological and environmental sciences communities have successfully used MGC analysis for other applications. For example, implementing MGC on nine climate variables partitioned the U.S. into 20 National Ecological Observatory Network (NEON) ecoclimatic domains that represent distinct regions of ecosystem dynamics (Keller et al., 2008; Schimel et al., 2007). Utilizing MGC analysis, a representativeness-based sampling network was designed in Alaska, which optimized sampling strategies, and offered a framework for up-scaling measurements (Hoffman et al., 2013). Similarly, MGC analysis was used to evaluate the representativeness of FLUXNET observations and develop a representativeness-based upscaling approach for CO_2 fluxes from eddy covariance measurements (Kumar et al., 2016). Lastly, MGC analysis was well suited for identifying and assigning decomposition functional types based on global climatic, edaphic, gross primary production, and topographic characteristics to estimate heterotrophic respiration at large scales (Bond-Lamberty et al., 2016).

The Earth and environmental sciences community are calling for more precise information on SOC stocks to promote the understanding of spatio-temporal dynamics of SOC and sustainable soil management practices (Amelung et al., 2020; Billings et al., 2021; Malhotra et al., 2019; Todd-Brown et al., 2022). In this study, we build on DSM approaches and introduce a novel method that combines MGC analysis and Random Forests regression (aka representativeness-based Random Forests) for spatial estimation of SOC stocks across the United States. We first divided the United States into 20 clusters, defined as SOC regions, at 30 arc second (∼1 km) spatial resolution. Then, we upscaled point-based SOC measurements (0–30 cm and 0–100 cm depths) to the continental scale using separate Random Forests in each SOC region. Our specific objectives were to (a) investigate the critical environmental predictors in different SOC regions; (b) estimate and map SOC stocks in the United States at a scale suitable for comparison with Earth System Model outputs (e.g., 30 arc seconds, ∼1 km resolution); (c) compare our SOC estimates with existing estimates and validate our maps; (d) identify the regions with low representativeness and high uncertainty to inform future monitoring efforts.

2. Materials and Methods

In this study, we followed a DSM workflow as illustrated in Figure 1. We first used the MGC approach to partition the United States into SOC regions based on principal components analysis (PCA) of 36 environmental covariates (Table 1, Figures S1 and S2 in Supporting Information S1). We then used Random Forests regressions to map SOC stocks within each of these SOC regions.

2.1. Environmental Covariates

SOC is controlled by multiple independent variables, including climate, parent material, topography, organisms, and time (Jenny, 1994; McBratney et al., 2003). A set of environmental variables that span these factors were collected to conduct MGC and Random Forests analyses, including bioclimatic, soil biogeochemical, biological, and physiographic variables (Table 1). We acknowledge that not all variables that directly influence SOC are included in our predictor set (mineralogy, soil metals, microbes, etc.), but we focus on broad-scale controls with widely available gridded data sets.

2.1.1. Bioclimatic Variables

We selected 19 bioclimatic variables (Table 1) with a spatial resolution of 30 arc seconds from WorldClim Version 2 to represent bioclimatic conditions across the United States (Fick & Hijmans, 2017). The bioclimatic variables are ecologically meaningful, representing annual trends (e.g., mean annual temperature, Annual Precipitation [AP]), seasonality (e.g., annual range in temperature and precipitation), and extreme or limiting environmental factors (e.g., temperature of the coldest and warmest months, and precipitation of the wet and dry quarters). Bioclimatic variables are derived from monthly temperature and rainfall values, and we used the average for the years 1970–2000 (Fick & Hijmans, 2017).

2.1.2. Physiographic Variables

For physiographic variables (or topographic characteristics), we used elevation data from North America Elevation 1-Kilometer Resolution GRID (U.S. Geological Survey, 2023). This North America elevation data was derived from the Global 30 Arc-Second Elevation data set (Earth Resources Observation and Science Center, 2017). We further derived slope (unit: degrees) and aspect (unitless) data at 30 arc seconds resolution from elevation data using ArcGIS software (version 10.8.1).

2.1.3. Soil Biogeochemical Properties

The gridded National Soil Survey Geographic Database (gNATSGO), composed by the United States Department of Agriculture, Natural Resources Conservation Service (NRCS), and Soil and Plant Science Division (USDA-NRCS-SPSD), provides complete coverage of the best available soil information for the entire United

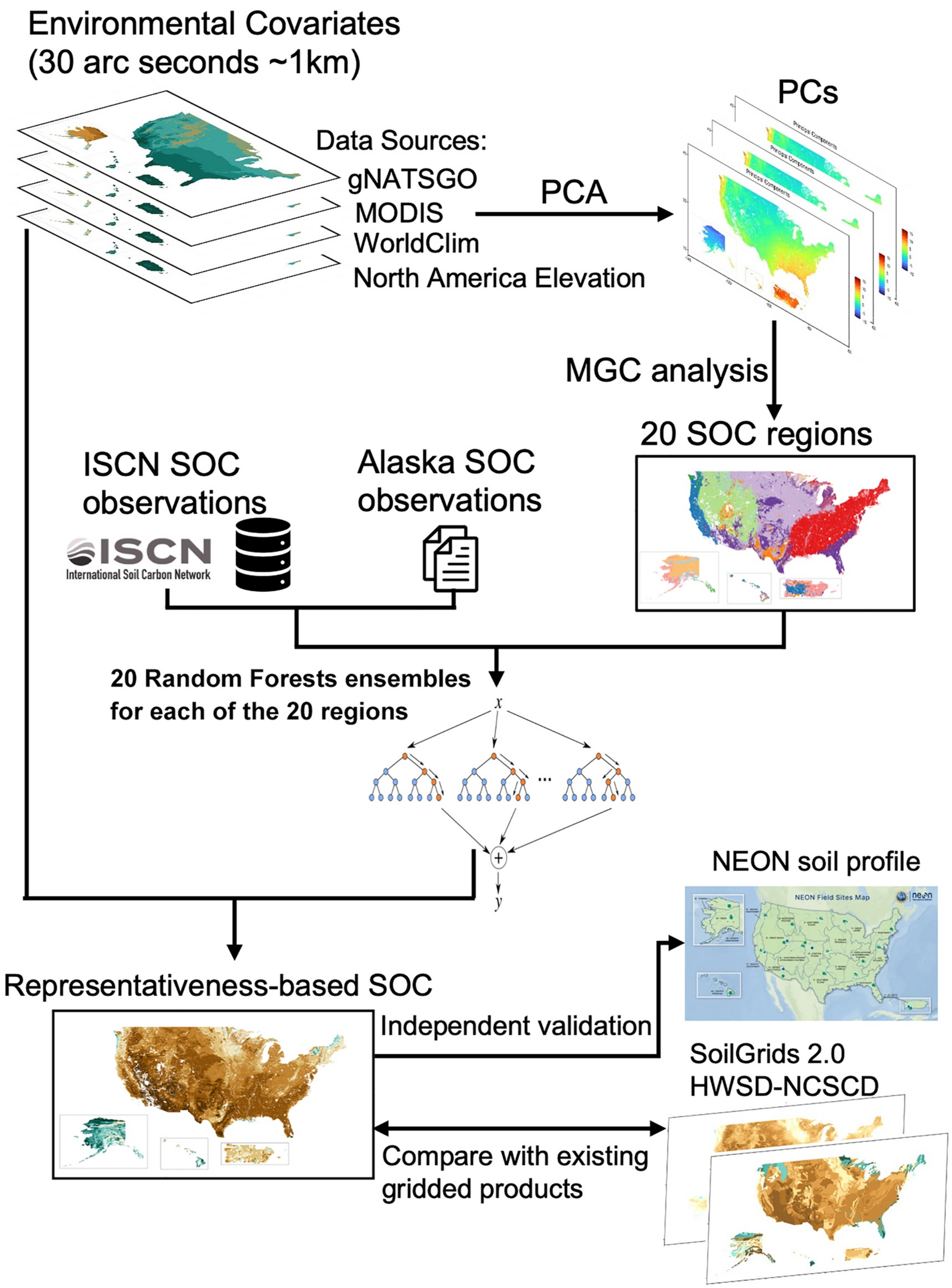

Figure 1. Workflow of generating representativeness-based soil organic carbon (SOC) stocks in the United States using multivariate geographic clustering (MGC) analysis and machine learning (Random Forests) models under the Digital Soil Mapping (DSM) framework. gNATSGO, the gridded National Soil Survey Geographic Database; MODIS, Moderate Resolution Imaging Spectroradiometer; PCA, principal component analysis; PCs, principal components; HWSD-NCSCD, Harmonized World Soil Database (HWSD v1.2) with Alaska replaced by Northern Circumpolar Soil Carbon Database (NCSCD).

Table 1
Environmental Covariates Used for Soil Organic Carbon Region Delineation and to Train Random Forests Models

Environmental covariates	Brief description	Data source
Bioclimatic variables		
Mean Annual Temperature (MAT)		WorldClim; Fick and Hijmans (2017)
Mean Diurnal Range (MDR)	Mean of monthly (max temperature − min temperature)	
Isothermality	(MDR/TAR) × 100	
Temperature Seasonality (TS)	Standard deviation × 100	
Max Temperature of Warmest Month (MaxTWM)		
Min Temperature of Coldest Month (MinTCM)		
Temperature Annual Range (TAR)	MaxTWM- MinTCM	
Mean Temperature of Wettest Quarter (MeanTWetQ)		
Mean Temperature of Driest Quarter (MeanTDQ)		
Mean Temperature of Warmest Quarter (MeanTWQ)		
Mean Temperature of Coldest Quarter (MeanTCQ)		
Annual Precipitation (AP)		
Precipitation of Wettest Month (PWM)		
Precipitation of Driest Month (PDM)		
Precipitation Seasonality (PS)	Coefficient of variation	
Precipitation of Wettest Quarter (PWetQ)		
Precipitation of Driest Quarter (PDQ)		
Precipitation of Warmest Quarter (PWQ)		
Precipitation of Coldest Quarter (PCQ)		
Physiographic variables		
Elevation	North America Elevation 1-Kilometer Resolution GRID	US Geological Survey
Aspect	Downslope direction of the maximum rate of change in value from each cell to its neighbors	Derived from the elevation
Slope	Steepest downhill descent from the cell	
Soil biogeochemical properties		

Table 1
Continued

Environmental covariates	Brief description	Data source
Calcium Carbonate ($CaCO_3$) equivalent	% of carbonates in the fraction of the soil <2 mm	Derived from gNATSGO (Soil Survey Staff, 2020)
Cation Exchange Capacity (CEC)	Total amount of extractable cations that can be held by soil (millieq. per 100 g of soil)	
Erosion Factor K	Susceptibility of a soil to sheet/rill erosion	
pH		
Sand content	Expressed as % of the soil material <2 mm	
Silt content	Expressed as % of the soil material <2 mm	
Clay content	Expressed as % of the soil material <2 mm	
Depth to Water Table (DTW)	Water table is a saturated zone in soil	
Ponding Frequency (PF)	Number of times that ponding occurs over a given period. Frequency is expressed as none, rare, occasional, and frequent	
Water Content	Amount of soil water retained at a tension of 1/3 bar, expressed as a volumetric % of the whole soil	
Available Water Storage (AWS)	Total volume of water (in cm) that should be available to plants at field capacity	
Biological variables		
Net Primary Productivity (NPP)		MODIS (Zhao et al., 2005)
Normalized Difference Vegetation Index (NDVI)		MODIS MOD13A2 V6.1 (Didan, 2021)
Leaf area index (LAI)		MODIS MOD15A2H Version 6.1 (Myneni et al., 2021)

Note. All data sets were resampled to a 30 arc seconds (~1 km) raster grid. gNATSGO, the gridded National Soil Survey Geographic Database; MODIS, Moderate Resolution Imaging Spectroradiometer.

States and Island Territories (Soil Survey Staff, 2020). The gNATSGO combines data from the Soil Survey Geographic Database (SSURGO), State Soil Geographic Database, and Raster Soil Survey Databases into a single seamless ESRI file geodatabase. Using a custom set of ArcTools, the "Soil Data Development Toolbox" in ArcGIS (10.8.1), we created a data set of spatial soil biogeochemical properties for the United States at 30 arc seconds from the gNATSGO (30 m resolution), including clay, sand, and silt content, soil erosion factor (K), pH, cation exchange capacity (CEC), water content, available water storage (AWS), calcium carbonate ($CaCO_3$) content, depth of water table, and ponding frequency. All soil properties were aggregated at 0–30 cm and 0–100 cm layers.

2.1.4. Biological Variables

We used Net Primary Productivity (NPP), leaf area index (LAI), and Normalized Difference Vegetation Index (NDVI) from the Moderate Resolution Imaging Spectroradiometer (MODIS) to represent biological characteristics. NPP data was produced by the Numerical Terradynamic Simulation Group, University of Montana (Zhao et al., 2005). This improved MODIS NPP Project (MOD17) is a post-reprocessed MODIS NPP data set where the contaminated MODIS Fraction of Photosynthetically Active Radiation and LAI inputs to the MOD17 algorithm have been cleaned (Zhao et al., 2005). The mean annual NPP data set covers the years 2000–2015. We collected the average LAI and NDVI data spanning 2000 to 2015 using Google Earth Engine (Gorelick et al., 2017) and the "geemap" python package (Wu, 2020). LAI data was obtained from the MODIS MOD15A2H Version 6.1 LAI product (Myneni et al., 2021); NDVI data were obtained from the MODIS MOD13Q1 V6.1 NDVI product (Didan, 2021).

2.2. SOC Observational Data

2.2.1. International Soil Carbon Network (ISCN)

SOC observations (27,976 for 0–30 cm depth and 21,807 for 0–100 cm depth) across the United States were obtained from the International Soil Carbon Network (ISCN) version 3 Database (ISCN; Nave et al., 2022). SOC observations were collected over several decades, ranging from the 1910s to the 2010s. We considered all soil profiles containing continuous soil layers at 0–30 cm and 0–100 cm intervals of soil depth in the United States (see Text S1 in Supporting Information S1 for more information). ISCN is one of the largest, most wide-ranging, and most diverse repositories of measured soil data (Harden et al., 2018; Malhotra et al., 2019). The majority of the ISCN data is from the NRCS and National Cooperative Soil Survey. It is worth noting that ISCN data is not completely independent of soil biogeochemical properties from gNATSGO, which we used to produce the soil properties described above. While the data sets are not completely independent, they are not entirely overlapping either. Since our response variable is SOC while the predictors are physical and chemical properties, this primary data source overlap has undue influence on our study.

2.2.2. An Independent Alaska Soil Profile Data

We collected georeferenced Alaska soil profile observations (113 for 0–100 cm depth) from a published database (Michaelson et al., 2013). This database includes data collected by the University of Alaska Fairbanks Northern Latitudes Soils Program from 1991 through 2011 in Alaska (Vitharana et al., 2017).

2.2.3. National Ecological Observatory Network (NEON)

NEON is a continental-scale observation facility designed to collect long-term open-access ecological data to improve our understanding of ecosystems across the United States. It was advantageously designed as a network of sites with a suite of common measurements including SOC (Weintraub et al., 2019), as opposed to data compilation of independent efforts. NEON statistically partitioned the continental U.S., Hawaii, and Puerto Rico into 20 ecoclimatic domains to ensure sampling would occur across the full range of U.S. ecological and climatic diversity. Here, we used soil profile data from the distributed initial characterization data set (Soil physical and chemical properties, distributed initial characterization, DP1.10047.001), which was collected from a number of 1 m deep pits distributed throughout each terrestrial site for a one-time characterization of site-wide soil properties (National Ecological Observatory Network, 2023).

2.2.4. Calculation of SOC Stocks

The SOC stocks for 0–30 cm and 0–100 cm depths were calculated using the following equation:

$$\mathrm{SOC} = \%\mathrm{C} \times \mathrm{BD} \times D \times (1 - \mathrm{CF}) \tag{1}$$

where SOC is the SOC stock (kg m^{-2}), %C is the concentration (g 100 g^{-1}) of C in the sample, BD is the soil bulk density (kg m^{-3}), D is the soil layer thickness (m), and CF is the volumetric fraction of coarse fragments. For soil

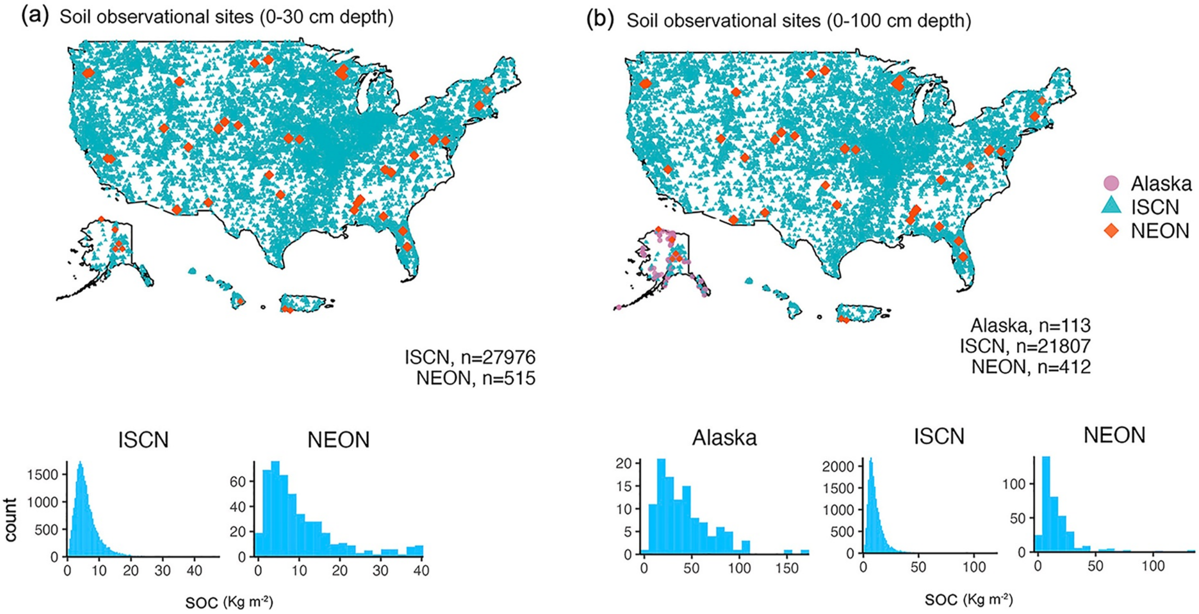

Figure 2. Locations of soil organic carbon (SOC) observation sites and distribution of SOC values in the United States from three data sets (Alaska data set, ISCN, and NEON). Point-based SOC measurements were obtained for two depths: (a) 0–30 cm and (b) 0–100 cm. ISCN, International Soil Carbon Network; NEON, National Ecological Observatory Network.

profiles missing bulk density (20,261 for 0–30 cm depth and 16,111 for 0–100 cm depth), we used a pedotransfer function to estimate it (Drew, 1973; Guevara et al., 2020; Yigini et al., 2018).

$$\mathrm{BD} = \frac{1}{(0.6268 + 0.0361 \times \mathrm{OM})} \tag{2}$$

The OM (organic matter) content was estimated as OM = SOC concentration × 1.724. This pedotransfer function was used based on findings from Guevara et al. (2020). They conducted an extensive analysis of the residual variance of six conventional pedotransfer functions for estimating bulk density in relation to SOC stocks. Their findings suggested that the equation proposed by Drew (1973, Equation 2) exhibited the strongest correlation with SOC prediction across Mexico and CONUS. Hence, we incorporated this pedotransfer function into our analysis and tested it using the NEON data, yielding an R^2 value of 0.39, a reasonable prediction accuracy for pedotransfer functions (Abdelbaki, 2018; Tranter et al., 2007). All SOC observation values were log-transformed while developing Random Forests models to reduce the right-skewed distribution of SOC values (see histograms of SOC in Figure 2). By implementing the log-transformation, we aimed to reduce the relative impact of these high values, thereby achieving a more acceptable level of error in prediction, while simultaneously enhancing the accuracy for the majority of the data set.

2.3. Principal Component Analysis (PCA)

Some statistical methods, such as *k*-means clustering based MGC, are sensitive to multicollinearity in their input variables. We performed principal component analysis (PCA) on the 36 environmental covariates to address that issue. PCA is a multivariate analysis technique for dimensionality reduction. It increases interpretability but minimizes information loss by extracting the most important information from non-independent variables to compress the size of the data set (Abdi & Williams, 2010; Jolliffe & Cadima, 2016). The PCA applied an orthogonal transformation that converted our set of possibly correlated variables into a set of values of linearized and uncorrelated variables called principal components (Figures S1 and S2 in Supporting Information S1) that were most appropriate for use in the clustering algorithm.

2.4. Multivariate Geographic Clustering (MGC) and Representativeness Analysis

We implemented the MGC (Hargrove & Hoffman, 1999, 2004) approach to delineate SOC regions across the United States. MGC uses an iterative *k*-means clustering algorithm, which starts with a set of initial centroids and calculates the Euclidean distance of each point to every centroid, classifying it to the closest centroid creating *k* clusters. Once all points are classified into *k* clusters, new centroids of the clusters are calculated and iterations are continued till the convergence criteria of less than 0.05% of cells changing their cluster assignment since previous iteration is reached. We applied MGC to 12 principal components (explaining 89% of variance), derived from 36 original environmental covariates (Table 1). MGC identifies clusters containing spatial grid cells based solely on their environmental characteristics without explicitly using their spatial coordinates. Thus, it does not force or require any spatial contiguity in the distribution of clusters, rather it lets them evolve based on the underlying environmental covariates.

The *k*-means clustering algorithm requires the desired number of clusters (*k*) to be prescribed. Traditionally, the optimal number of clusters (*k*) is determined using the elbow method by examining the mean distance of points from the centroid within each cluster, which is a measure of cluster compactness. However, in our study, the goal of MGC was not just to identify an optimal number of clusters based on environmental variables, but to also segment the available SOC observations to develop models for estimating SOC stocks. Thus, we conducted clustering at varying numbers of *k* and these initial tests helped us identify an appropriate *k* value, most suitable for developing machine learning models for estimating SOC stocks. At $k = 10$ (Figure S3 in Supporting Information S1), a large portion of the eastern US was grouped into one SOC region, masking known heterogeneity in important soil forming factors. On the other hand, at $k = 30$ (Figure S3 in Supporting Information S1) there were six SOC regions that had less than 100 data points for both 0–30 cm and 0–100 cm layers, limiting our ability to create accurate SOC prediction models. We adopted $k = 20$ (Figure S3 in Supporting Information S1) in our study as it limited only three SOC regions that had less than 100 data points for both 0–30 cm and 0–100 cm layers, allowing sufficient SOC observations within most SOC regions for this modeling application while still allowing the data to be grouped along key axes of environmental variation. Thus, our approach struck a balance between the number of observations within each region and the number of SOC regions. NEON also segmented the U.S. into 20 ecoclimatic domains, but our 20 SOC regions are distinct from those as we used multiple sources of input data for MGC while NEON only used climate variables.

Since the SOC observations available within each SOC region were not taken with the SOC regions in mind, they represent a random sampling of the environmental conditions across the SOC region and thus provide heterogenous representation for grid cells within and across the study regions. Representativeness provided by the available SOC observations, or lack thereof, is likely to have important implications for accuracy and uncertainty of models trained using them and is important to quantify and understand. We therefore quantified how well each grid cell is represented by the available SOC observations using a point-based representativeness approach (Hoffman et al., 2013). At each site with SOC observations, the environmental covariables (the principal components here) were extracted from the gridded data sets and the Euclidean distance to every other pixel was computed in the environmental variables data space. We selected the inverse of the closest Euclidean distance between the pixel and the observation sites among all observation sites, in the environmental data space defined by 12 principal components, as the representativeness of that pixel. Higher representativeness values indicate better representation of grid cells by observation sites' environmental conditions while lower values represent poor representation.

2.5. Random Forests and Model Performance

Random Forests is a commonly used machine learning algorithm that combines the output of multiple decision trees to yield a single result. It is suited for handling non-linear relationships between response variables and predictors without requiring predefined functional forms or a normal sample distribution (Breiman, 2001). Unlike MGC, Random Forests algorithms have been demonstrated to be robust and less sensitive to the presence of multicollinearity in the predictor variables. Thus, we used the original environmental covariates ($n = 36$) instead of principal components to generate data-driven SOC estimation models, allowing for better interpretability of predictor importance scores. Separate Random Forests models were trained to predict SOC in each SOC region, similar to Chen et al. (2019) and Song et al. (2020). SOC observations from ISCN ($n = 27{,}976$ and 21,807 for 0–30 cm and 0–100 cm depths, respectively) and Alaska soil profile data ($n = 113$ for 0–100 cm depth) were randomly split into 80% training and 20% testing sets. To quantify the uncertainties caused by random sampling, we repeated the random splitting 20 times, generating 20 groups of training and testing sets for each SOC region. We then trained 20 Random Forests using the training sets with all predictors and evaluated model performance using the testing sets. In our study, the ntree (number of trees) was set to 1,000. The hyperparameter

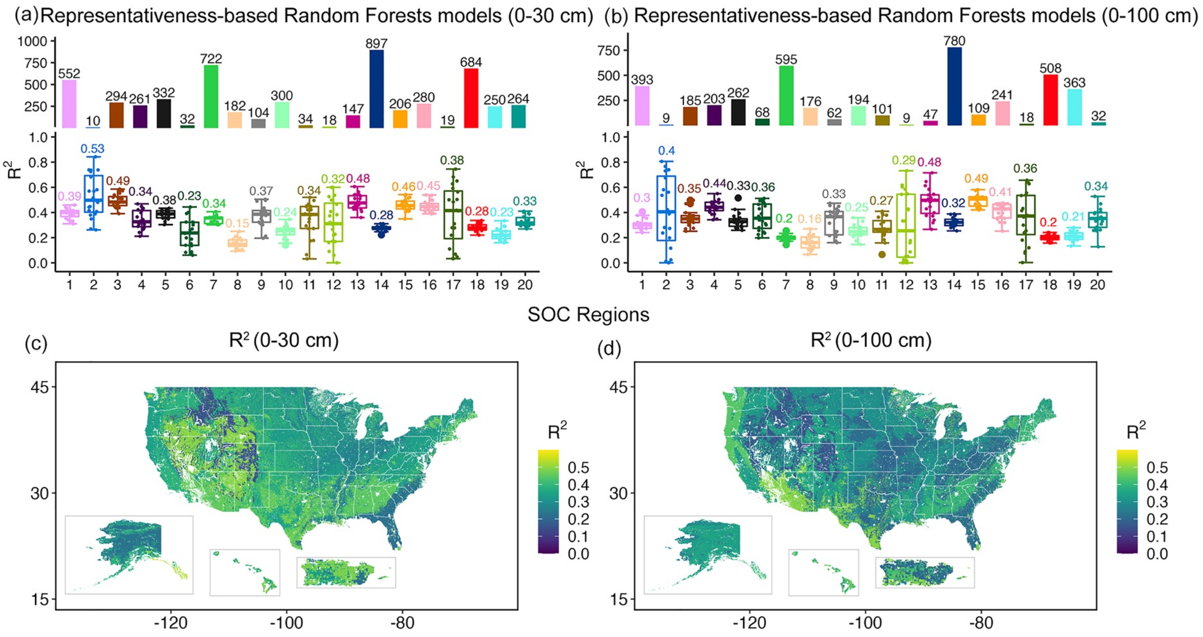

Figure 3. Model performances were evaluated using testing data in 20 soil organic carbon (SOC) regions. Box plots represent R^2 values for 20 representativeness-based Random Forests models for (a) 0–30 cm and (b) 0–100 cm depths. Maps represent spatial distribution of mean R^2 of all SOC regions for (c) 0–30 cm and (d) 0–100 cm depths. The bar plot and numbers above each bar (a, b) indicate the number of observations in testing sets. The numbers above each box are the mean value of the 20 R^2 from 20 Random Forests runs. The box and whiskers plot display the minimum, first, median, third, and maximum quartile values.

mtry (the number of variables considered at each split in prediction trees, mtry = 3, 5, 7, or 9) was optimized using out-of-bag (OOB) samples with the caret package (Kuhn, 2008) for the R software (version 4.1.2; R Core Team, 2022).

Model performance was evaluated based on the coefficient of determination (R^2), the mean absolute error (MAE), and the root mean squared error (RMSE) in testing sets. To compare our new SOC maps with existing gridded products, we downloaded SoilGrids 2.0 (Poggio et al., 2021), HWSD v1.2 (FAO et al., 2012), and NCSCD (Hugelius et al., 2013) and used NCSCD to replace HWSD v1.2 values in Alaska to correct the high latitudes SOC stocks bias (Georgiou et al., 2021), naming it HWSD-NCSCD. Then, we extracted the SOC values from the pixels that matched the locations (latitude and longitude) of the testing data from ISCN. We accordingly compared testing data to SoilGrids 2.0 and HWSD-NCSCD values, generating R^2, MAE, and RMSE that could be compared to our model metrics. Additionally, the NEON soil profile data were used for independent validation, where we compared measured values to predictions from our method as well as those from SoilGrids 2.0 and HWSD-NCSCD at the same locations. The uncertainties of our SOC estimates were quantified using the coefficients of variation (CV) of the predictions from 20 Random Forests in each SOC region. Lastly, we compared our overall estimates of SOC stocks with SoilGrids 2.0 and HWSD-NCSCD.

3. Results

3.1. Model Performance

The mean R^2 for our representativeness-based Random Forests from testing sets (Figure 3) in 20 SOC regions (Figure S3 and Table S1 in Supporting Information S1) varied between 0.15–0.53 and 0.16–0.49 for 0–30 cm and 0–100 cm depths, respectively. There was a large difference in the number of ground truth observations in our testing sets among the 20 SOC regions, ranging between 10 and 897 for 0–30 cm depth and 9 to 780 for 0–100 cm depth, respectively. Random Forests runs in the SOC regions with a small number of observations tended to have a more extensive range of R^2 than SOC regions with a large number of observations. For reference, we trained a Random Forests model using a single SOC region (e.g., no clusters), and the resulting R^2 was 0.47 and 0.41 for 0–30 cm and 0–100 cm depths, respectively (Figures S4c and S4f in Supporting Information S1). The single SOC region-based Random Forests represents the commonly used method but tends to mask the variable model performances in different regions.

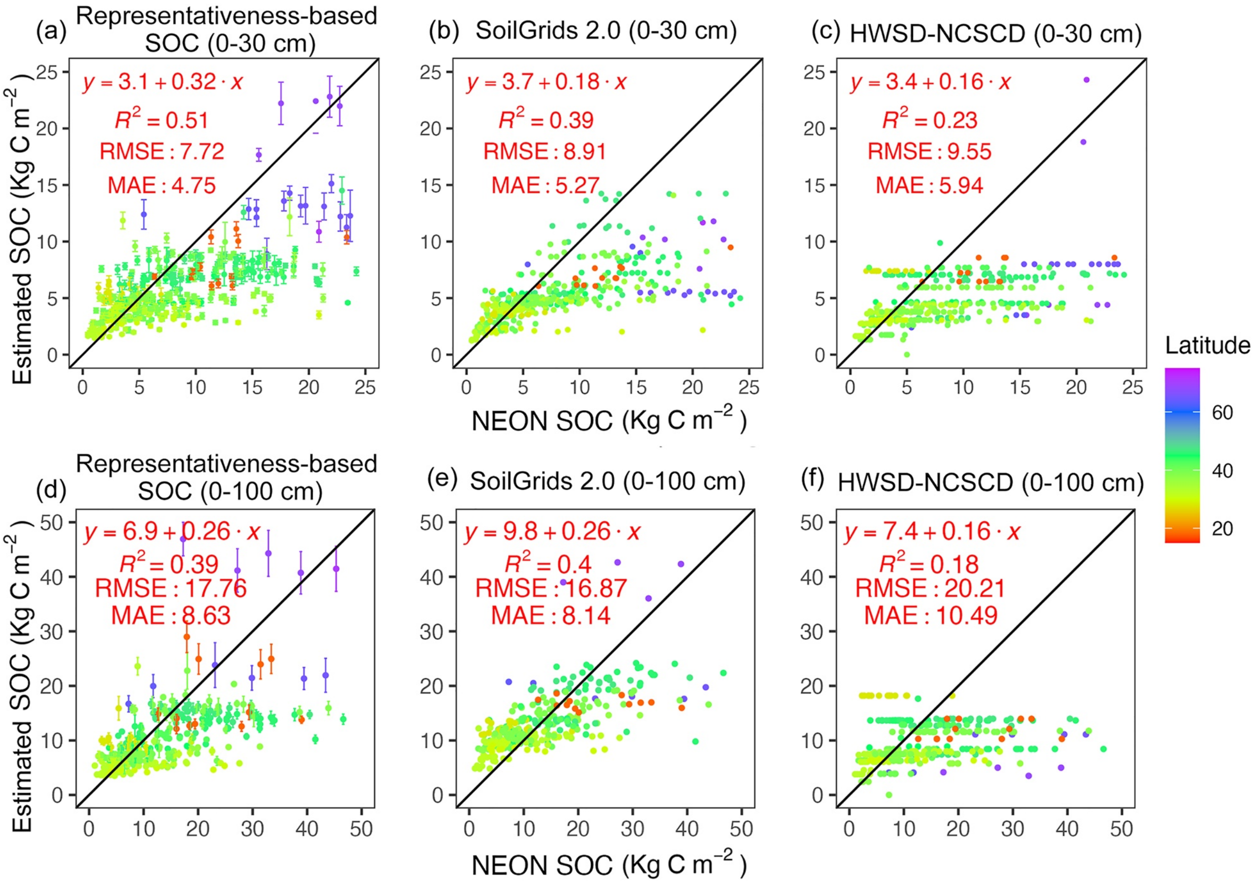

Figure 4. Independent validation using National Ecological Observatory Network (NEON) soil profile soil organic carbon (SOC) and comparison with HWSD-NCSCD and SoilGrids 2.0 SOC data sets. Scatter plots represent predicted versus observed SOC stocks at NEON sites for (a, d) representativeness-based SOC; (b, e) SoilGrids 2.0; and (c, f) HWSD-NCSCD. Top panels (a, b, c) represent 0–30 cm depth and bottom panels (d, e, f) represent 0–100 cm depth. Black lines represent 1:1 lines. Error bars in (a) and (d) indicate the standard deviation of 20 Random Forests ensembles in representativeness-based SOC estimates. The Units of root mean squared error and mean absolute error are kg C m^{-2}. HWSD-NCSCD, Harmonized World Soil Database (HWSD v1.2) with Alaska replaced by Northern Circumpolar Soil Carbon Database (NCSCD).

Compared to SoilGrids 2.0 and HWSD-NCSCD, our method showed better performance in reproducing observations in most SOC regions. While our approach always resulted in higher R^2 values than HWSD-NCSCD, 16 and 13 out of 20 SOC regions also resulted in higher R^2 values than SoilGrids 2.0 for 0–30 cm and 0–100 cm depths, respectively (Figure 3, Figure S4 in Supporting Information S1). Compared to SoilGrids 2.0 and HWSD-NCSCD, our representativeness-based estimates had lower RMSE in 14 SOC regions and 8 SOC regions for 0–30 cm and 0–100 cm depths, respectively (Figure S5 in Supporting Information S1). For MAE, our estimates showed lower values than SoilGrids 2.0 and HWSD-NCSCD in 12 SOC regions and 14 SOC regions for 0–30 cm and 0–100 cm depths, respectively (Figure S6 in Supporting Information S1).

We used NEON soil profile SOC stocks (0–30 cm and 0–100 cm depths) to independently validate our representativeness-based SOC estimates (Figure 4). For 0–30 cm depth, representativeness-based SOC (R^2 = 0.51; RMSE = 7.72 kg C m^{-2}; MAE = 4.75 kg C m^{-2}) showed a higher R^2, lower RMSE and MAE than SoilGrids 2.0 (R^2 = 0.39; RMSE = 8.91 kg C m^{-2}; MAE = 5.27 kg C m^{-2}) and HWSD-NCSCD (R^2 = 0.23; RMSE = 9.55 kg C m^{-2}; MAE = 5.94 kg C m^{-2}). For 0–100 cm depth, representativeness-based SOC estimates (R^2 = 0.39; RMSE = 17.76 kg C m^{-2}; MAE = 8.63 kg C m^{-2}) were nearly the same as SoilGrids 2.0 (R^2 = 0.4; RMSE = 16.87 kg C m^{-2}; MAE = 8.14 kg C m^{-2}) but better than HSWD-NCSCD (R^2 = 0.18; RMSE = 20.21 kg C m^{-2}; MAE = 10.49 kg C m^{-2}). The representativeness-based SOC stocks captured larger SOC values in surface soil (0–30 cm) at high latitudes than SoilGrids 2.0 and HWSD-NCSCD, in which high latitudes SOC stocks were underestimated (Figure 4). Overall, our combined MGC analysis and Random Forests regression approach outperformed HWSD and SoilGrids 2.0.

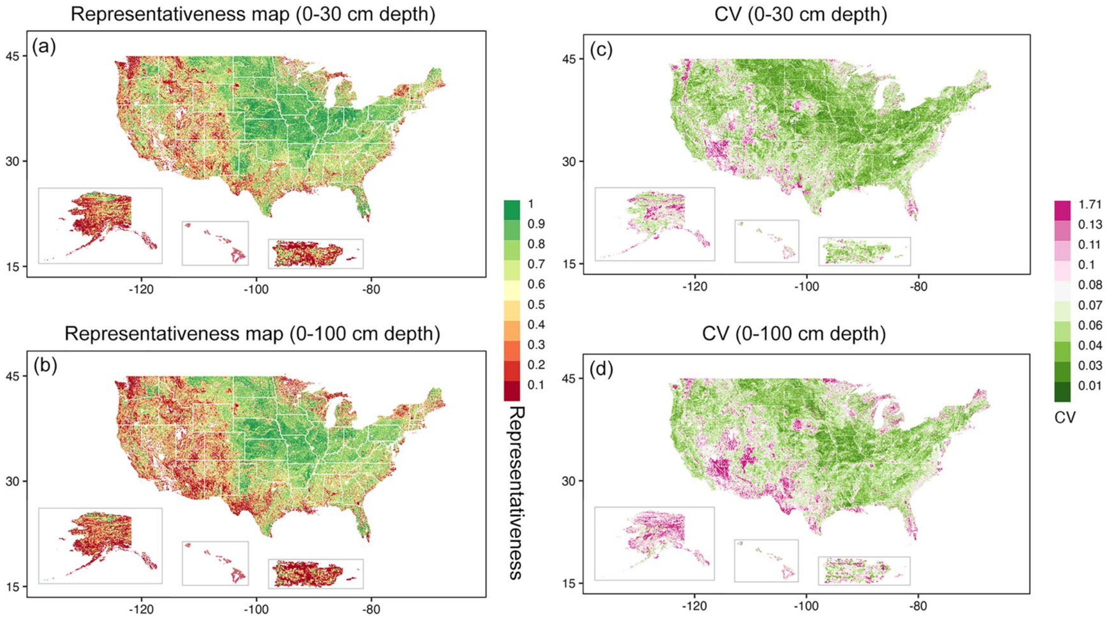

Figure 5. Representativeness maps and coefficient of variations (CV) identify areas that need more ground truth observations. Left panels (a, b) represent representativeness maps comparing each pixel to observation pixels based on the closest Euclidean distance in environmental covariates data space at 0–30 cm and 0–100 cm depths, respectively. Right panels (c, d) represent spatial distributions of CV of representativeness-based soil organic carbon for 0–30 cm and 0–100 cm depths, respectively. Top panels (a, c) represent 0–30 cm depth and bottom panels (b, d) represent 0–100 cm depth.

3.2. Representativeness and Coefficient of Variation

The representativeness of the United States is low in the Southeast coastal area, the Great Lakes basin, the Southwest deserts, the Rocky Mountains, the Pacific Northwest, and most parts of Alaska, Hawaii, and Puerto Rico (Figures 5a and 5b). All coastlines are likely under-represented. Overall, 0–30 cm depth was more representative than 0–100 cm depth, indicated by higher representativeness values (or fewer red pixels) in representativeness maps in the former (Figures 5a and 5b).

To evaluate uncertainty in our products, we analyzed the CV of the representativeness-based SOC estimates. The higher CV indicated more uncertainties introduced by the random sampling and splitting of data. We found larger CVs located in Alaska, the Great Lakes basin, Nebraska Sand Hills, the far northeast US, Southwest deserts, the Rocky Mountains, and southeast Florida (Figures 5c and 5d). These areas had fewer observations, and the random sampling resulted in wider variability of Random Forests models. More areas with high CV values (or more pink pixels) were found for 0–100 cm than for 0–30 cm in our SOC maps, especially in Alaska, Southwest desert areas, and the Great Lakes basin.

3.3. Spatial Variability in Dominant Environmental Factors Across SOC Regions

The importance of different environmental predictors varied across SOC regions (Figure 6, Figure S7 in Supporting Information S1). For 0–30 cm depth, all SOC region's top five important covariates included more than one bioclimatic variable (Figure 6, Table 1), and in many regions, they were the majority of the top predictors. In SOC regions 3, 6, 8, 9, and 11, biological covariates were more important in predicting SOC than in other regions. This mainly covers drylands such as Nevada, interior Alaska, Idaho, Utah, Colorado, and Texas (Figure S3a in Supporting Information S1). Soil biogeochemical properties were most important for predicting SOC in regions 2 and 19 (e.g., some areas in southeast Alaska Pacific coastal and southeast and southern US coastal plain) and emerged as important predictors in regions 4, 9, 17, and 18 as well (e.g., some areas in Pacific Northwest, Utah, south Alaska, Corn Belt plains). Elevation was an important predictor in SOC regions 7, 13, 14, 18, and 20 (e.g., some areas in the Appalachians, Mississippi River floodplain, Corn Belt plains, and Nebraska Sand Hills) and co-occurred with temperature-related bioclimatic covariates.

For 0–100 cm depth, bioclimatic covariates were relatively important in all regions except SOC region 4 (Pacific Northwest) (Figures S3 and S7 in Supporting Information S1). Similar to 0–30 cm depth, biological covariates

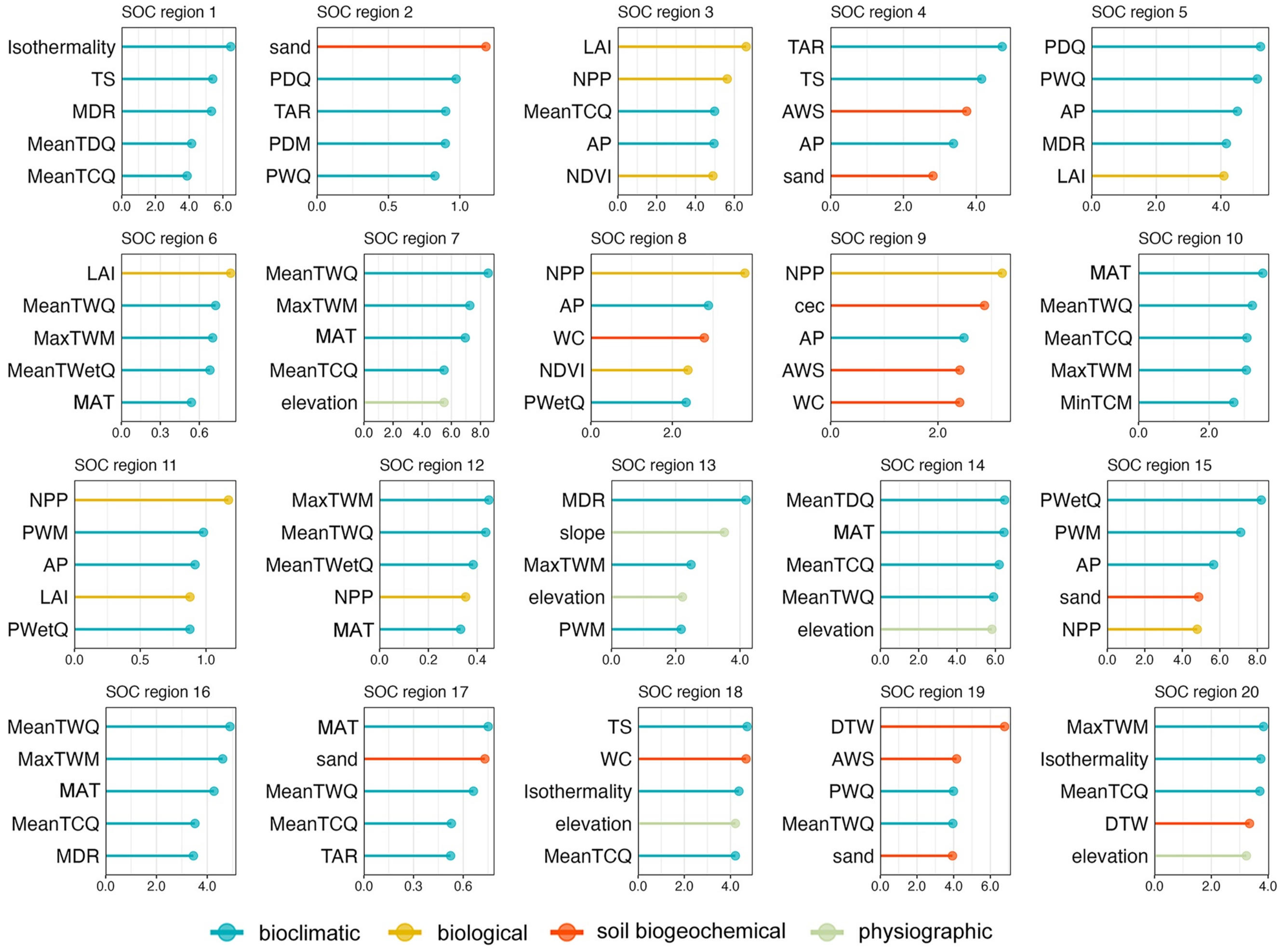

Figure 6. The relative importance of predictor variables in 20 soil organic carbon (SOC) regions for the 0–30 cm depth. More information and spatial distribution of SOC regions can be found in Supporting Information S1 (Figure S3, Table S1, and Text S2 in Supporting Information S1). Only the top five most important variables are shown here. MAT, Mean Annual Temperature; MDR, Mean Diurnal Range; TS, Temperature Seasonality; MaxTWM, Max Temperature of Warmest Month; MinTCM, Min Temperature of Coldest Month; TAR, Temperature Annual Range; MeanTWetQ, Mean Temperature of Wettest Quarter; MeanTDQ, Mean Temperature of Driest Quarter; MeanTWQ, Mean Temperature of Warmest Quarter; MeanTCQ, Mean Temperature of Coldest Quarter; AP, Annual Precipitation; PWM, Precipitation of Wettest Month; PDM, Precipitation of Driest Month; PS, Precipitation Seasonality; PWetQ, Precipitation of Wettest Quarter; PDQ, Precipitation of Driest Quarter; PWQ, Precipitation of Warmest Quarter; PCQ, Precipitation of Coldest Quarter; AWS, Available Water Storage; WC, Water Content; DTW, Depth to Water Table; PF, Ponding Frequency. See the relative importance of prediction variables in 20 SOC regions for the 0–100 cm depth in Figure S7 in Supporting Information S1.

were found to be most important in the dryland and Alaska regions (SOC regions 15, 17, and 20, areas in Southwest deserts regions, Southeastern plains, southern Alaska, and interior Alaska, Figures S3 and S7 in Supporting Information S1). Soil biogeochemical properties were more broadly important at the deeper depth, appearing in the top 5 variables in 11 out of the 20 SOC regions (Figure S7 in Supporting Information S1), as opposed to nine regions for surface soil (Figure 6). Like 0–30 cm, elevation was an important predictor in the Appalachians and Mississippi River floodplain (SOC regions 7 and 14, Figures S3 and S7 in Supporting Information S1) as well as in areas of the Rocky Mountains and interior Alaska (SOC regions 3 and 9, Figures S3 and S7 in Supporting Information S1).

3.4. Estimates of SOC Stocks

The spatial pattern of mean SOC stocks at 30 arc seconds resolution is shown in Figure 7. In the United States, well-known spatial patterns of SOC stocks were reproduced. SOC stock generally decreased from high-latitude to low-latitude areas. The Great Lakes basin, northeast CONUS, and Alaska tended to have high SOC stocks.

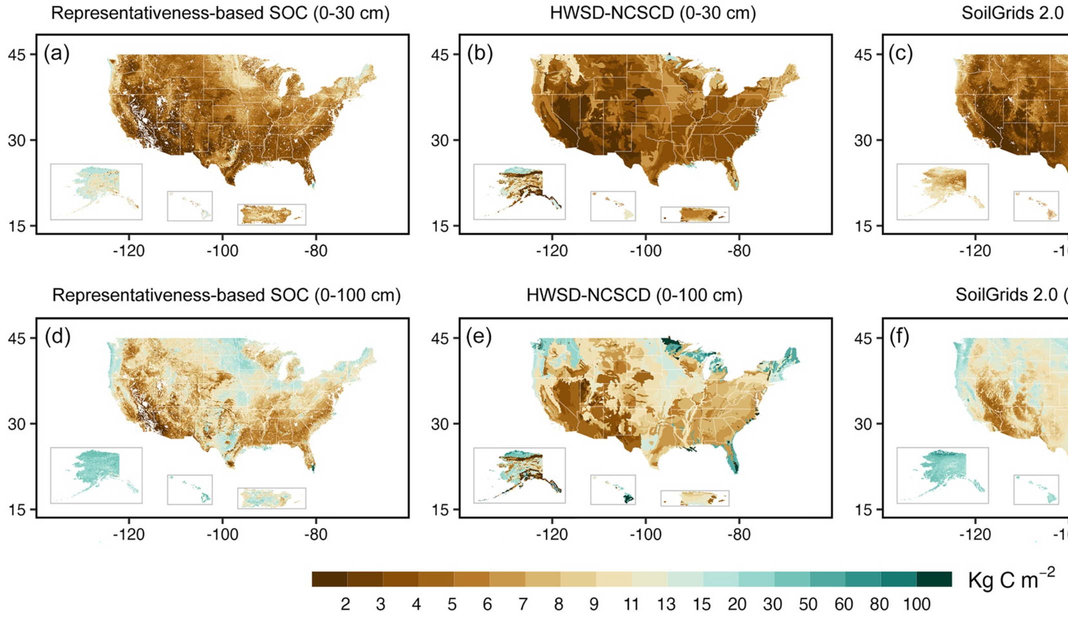

Figure 7. Spatial distribution of soil organic carbon (SOC) stocks for (a–c) 0–30 cm and (d–f) 0–100 cm depths in the United States. The top panels represent SOC stocks for 0–30 cm depth for (a) representativeness-based SOC, (b) HWSD-NCSCD, and (c) SoilGrids 2.0. The bottom panels represent SOC stocks for 0–100 cm depth for (d) representativeness-based SOC, (e) HWSD-NCSCD, and (f) SoilGrids 2.0. HWSD-NCSCD, Harmonized World Soil Database (HWSD v1.2) with Alaska replaced by Northern Circumpolar Soil Carbon Database (NCSCD).

In contrast, the Southwest deserts and the Rocky Mountains tended to show low SOC stocks. For the 0–30 cm layer, we estimated 52.6 ± 3.2 (Mean ± SD) Pg C in the United States, which was 12.7% and 15.1% larger than HWSD-NCSCD and SoilGrids 2.0, respectively. Our estimation for the 0–100 cm layer was 108.3 ± 8.2 (Mean ± SD) Pg C, which was 19.4% larger than HWSD-NCSCD and 18.6% smaller than SoilGrids 2.0 (Table 2). In Alaska, representativeness-based SOC stock for the 0–30 cm depth was larger than HWSD-NCSCD and SoilGrids 2.0 by 30.8% and 46.6%, respectively. In contrast, in CONUS, representativeness-based SOC for the 0–30 cm depth was only larger than HWSD-NCSCD and SoilGrids 2.0 by 5.9% and 4.3%, respectively. For Alaska and CONUS, representativeness-based SOC for 0–100 cm depth was larger than HWSD-NCSCD by 71.7% and 2% but lower than SoilGrids 2.0 by 11.8% and 21.9%, respectively. SOC stocks had more discrepancies at high latitudes and deeper layers across different data sets (Figure 7 and Table 2).

4. Discussion

Our DSM framework, coupling MGC with Random Forests to estimate SOC, shows promise in both generating accurate SOC maps at the continental scale and identifying areas that need more observational data. Additionally,

Table 2
Estimated Soil Organic Carbon Stocks (Mean ± SD) in the United States, Alaska, and Continental US for 0–30 cm and 0–100 cm Depths

Depth (cm)	Data set	Total stock (Pg)	Alaska stock	Continental US stock
0–30	Representativeness-based SOC	52.6 ± 3.2	17.0 ± 1.4	35.4 ± 1.8
0–30	HWSD-NCSCD	46.7	13.0	33.4
0–30	SoilGrids 2.0	45.7	11.6	33.9
0–100	Representativeness-based SOC	108.3 ± 8.2	38.6 ± 3.7	69.3 ± 4.4
0–100	HWSD-NCSCD	90.7	22.5	67.8
0–100	SoilGrids 2.0	133.0	43.8	88.7

Note. HWSD-NCSCD, Harmonized World Soil Database (HWSD v1.2) with Alaska replaced by Northern Circumpolar Soil Carbon Database (NCSCD); Pg, Peta-gram.

we revealed new insights into how different environmental factors influence SOC stocks in different regions. Finally, we generated gridded SOC stock estimates (and associated uncertainties) for the United States with 30 arc second resolutions that are more accurate, especially for surface soils, compared to other commonly used gridded products according to independent validations.

4.1. Low Representativeness and High Uncertainty Areas for Further Survey

We bootstrapped uncertainties in the SOC estimates using 20 repeated random samplings data sets (80% training and 20% testing sets). We found a general pattern of converging uncertainty based on the spatial density of observations. In SOC regions defined by covariates with fewer observations (generally, $n < 500$; Figure S8 in Supporting Information S1), the 20 repeated random samplings resulted in a larger range of R^2 than regions with more observations (Figure 3 and Figure S8 in Supporting Information S1), indicating that model performances in SOC stock estimates are more unstable and uncertain in these SOC regions. Instead of simply merging those SOC regions into other regions, we chose to keep the 20 SOC regions to highlight these regions with unique properties and high uncertainty, emphasizing the pressing need for additional observations in these areas. These uncertainties from bootstrapping indicate the need to structure sample collection to be more representative. In theory, if there are enough well-distributed observations covering the entire "population" of SOC, the random separation into training and testing sets should result in only minor differences in prediction results. Spatially, the CV maps highlight higher uncertainty areas (Figure 5), which tend to be where observations are sparse (Figure 2), and representativeness values are low (Figure 5). These areas of low representativeness and high uncertainties occurred in Alaska, the Great Lakes basin, Nebraska Sand Hills, far northeast US, Southwest Deserts, the Rocky Mountains, and southeast Florida (Figure S3 in Supporting Information S1). In these areas, coordinated and systematic field sampling efforts are needed to fill data gaps and increase our confidence in estimates of SOC stocks. Additionally, the SOC map for 0–100 cm depth has more areas with higher uncertainty and less representativeness than the SOC map for 0–30 cm (Figure 5). This can be partially attributed to the scarcity of SOC observations at greater depths since obtaining samples from these deeper layers is a resource-intensive and challenging endeavor (Billings et al., 2021; Jandl et al., 2014). Meanwhile, estimated SOC at 0–30 cm depth by our approach and HSWD-NCSCD showed better performance (larger R^2 values) compared to estimated SOC at 0–100 cm depth (Figure 4). This aligns with the observation that DSM generally performs better in topsoil than subsoil for SOC concentrations and SOC stocks, as reported by Chen et al. (2022).

We also quantified the representativeness of each grid cell in the data space (Figure 5). The areas with larger CV values were generally matched with the underrepresented areas, indicating that representativeness could be a helpful method for identifying areas needing more ground-based SOC measurements. Increasing SOC measurements, specifically in underrepresented grid cells, could better center the SOC regions in data space, leading to improved representation within each SOC region. Several studies have used the MGC point-based representativeness method to locate measurement sites to better monitor large-scale ecosystem dynamics (Hoffman et al., 2013). This kind of representativeness analysis could guide future field campaigns as far as where to do more monitoring and, in turn, help modelers identify areas where more observations are needed to better constrain Earth system models.

4.2. The Impacts of Environmental Factors Across Different SOC Regions

Multiple environmental factors control SOC, including climate, plant productivity, edaphic properties, and topographic variables (McBratney et al., 2003). Moreover, understanding the spatial distribution of what variables drive SOC is crucial for predicting how ecosystems will respond to changing climate conditions (Doetterl et al., 2015; Gautam et al., 2022; Gonçalves et al., 2021; Mishra et al., 2022). Thus, it is critical to understand the key predictive drivers across different regions. Our variable importance analysis indicates that bioclimatic covariates were the most common top five predictors in all SOC regions for both 0–30 cm and 0–100 cm depths (Figure 6 and Figure S7 in Supporting Information S1). However, other types of predictors were as or more important in specific regions.

Biological predictors tended to be more important in the southwest drylands (SOC regions 3, 8, 9, and 11, for 0–30 cm depth and regions 15 for 0–100 cm depth; Figure 6), which have arid or semi-arid climates (Lauenroth & Bradford, 2009). Our finding is in line with results from a CONUS-scale study, which revealed NPP was the most important predictor for SOC stocks in the driest ecosystems (Gonçalves et al., 2021). In drylands, SOC accrual is more limited by inputs (i.e., plant production, NPP) since decomposition rate is relatively slow (Lal, 2004a; Schulze & Freibauer, 2005). Despite exhibiting temperature and precipitation patterns contrary to those found in drylands, biological covariates (e.g., NPP, NDVI) also played a significant role in predicting SOC in Alaska (e.g., SOC regions 17 and 20) at 0–100 cm depth. This could be explained by limited carbon input in high-latitude regions due to short and cold growing seasons (Bjorkman et al., 2020; Hobbie et al., 2000). Although not often considered in soil studies across Alaska (Bliss & Maursetter, 2010; Johnson et al., 2011; Mishra & Riley, 2012), including more explicit vegetation-related variables, such as NPP, NDVI, could be helpful in estimating SOC stocks.

Soil biogeochemical covariates were important predictors in some regions. For example, depth to water table was a key predictor for 0–30 cm SOC stock in SOC region 19 (mainly found in the southeast and southern coastal plain) and for 0–100 cm SOC stocks in SOC regions 1, 2, and 5 (mainly found in the Northwest US). In these areas with high AP, depth to water table likely plays a critical role in SOC dynamics (Ise et al., 2008). For instance, anoxic conditions resulting from a rising water table could lower SOC decomposition rates, leading to high SOC stocks (Fenton et al., 2005; Ise et al., 2008). Soil variables such as sand content, CEC, and AWS were important predictors in surface soils (0–30 cm) for SOC regions 2, 4, 9, and 17, which were distributed among the Rocky Mountains and mountainous areas of the south and southeast Alaska. Mountain soils are likely to be shallow, thin, and coarse-textured, so variation in these properties may have more influence on SOC dynamics there (Egli & Poulenard, 2016).

Other soil biogeochemical covariates (e.g., $CaCO_3$, CEC, pH) emerged as important predictors for 0–100 cm depth more so than for 0–30 cm depth (Figure 6 and Figure S7 in Supporting Information S1). For 0–100 cm depth, 11 out of 20 SOC regions have at least one edaphic variable ranked in the top five important variables. Similar to what we observed here, another continental-scale study across sub-Saharan Africa reported that climate variables played a larger role in the topsoil, and geochemical predictors had a larger influence on the subsoil (von Fromm et al., 2021). Other global scale studies also supported that climate variables dominated SOC in surface soil, while edaphic properties were more important controllers in deeper layers (Jobbágy & Jackson, 2000; Luo et al., 2021). However, a recent study from the United States demonstrated the generally consistent relative importance of geochemical and climate predictors of SOC across the soil depth (Yu et al., 2021). Thus, the relative importance of soil biogeochemical and climate predictors of SOC in top versus sub-layers might change depending on the scale of analysis or the specific region.

Regions related to the Appalachians and the Mississippi River had physiographic variables (e.g., elevation) as influential predictors. In mountains, elevation has an effect on vegetation productivity and decomposition rates due to decreasing temperature along the elevation gradient, which is likely to influence organic carbon inputs to and losses from soils (Garten et al., 1999; Shedayi et al., 2016; Sheikh et al., 2009). In the Mississippi River floodplain, flooding has a significant influence on SOC by changing hydrology and sediment transportation processes (De Jager et al., 2012; Elsey-Quirk et al., 2019; Grubaugh & Anderson, 1989). Regions with higher elevations are less impacted by flooding, and lower-elevation land experiences more flooding. Even though it is generally a low-relief area, this could explain the critical role of elevation in predicting SOC in areas along the Mississippi River.

4.3. Comparison of Representativeness-Based SOC Stocks With Other Existing Gridded Products

We estimated larger SOC stocks for the 0–30 cm layer in the United States (52.6 ± 3.2 Pg) compared to HWSD-NCSCD (46.7 Pg) and SoilGrids 2.0 (45.7 Pg) at 30 arc-seconds (1 km) resolution. This is primarily due to relatively high estimates of SOC stocks in Alaska by our approach compared to other products (17.0 ± 1.4 Pg for representativeness-based SOC; 13.04 Pg for HWSD-NCSCD; and 11.6 Pg for SoilGrids 2.0). In Alaska, we also observed a large range of SOC estimates across the three data sets for 0–100 cm depth (Table 2). Discrepancies among SOC stock data sets at high latitudes were also detected by other studies (Lin et al., 2022; Tifafi et al., 2018). Resolving this uncertainty in Alaska and high-latitude boreal regions is paramount due to their disproportionately large SOC stocks (Ping et al., 2008) and faster warming (IPCC, 2022; Schuur et al., 2008) compared to other regions. Variations in estimated SOC stocks across Alaska most likely arise from differences in upscaling and statistical modeling approaches between the data sets we compared. In our DSM framework, instead of using a single Random Forests model globally (e.g., SoilGrids), we used a clustering method, building Random Forests based on individual SOC regions, to better capture the spatial heterogeneity between key predictors and SOC. In Alaska, our MGC analysis identified over 10 distinct pixel-specific SOC regions. These regions were sporadically distributed across geographic areas, with some being small in size (Figure S3 in Supporting Information S1). We argue that these detailed data-based, pixel-specific SOC regions better represent varying relationships between SOC and predictors, making our method a more robust way to upscale SOC stock estimates. Still, more observational data are needed to confirm this, as some of the SOC regions in Alaska were very data-poor.

Our estimated SOC stocks and spatial distributions for the 0–30 cm layer in the United States (52.6 ± 3.2 Pg) were close to Global Soil Organic Carbon Map v1.5 (GSOC) (FAO & ITPS, 2020) (52.8 Pg 0–30 cm depth, Figure S9 in Supporting Information S1). GSOC is the first global SOC map ever produced through a consultative and participatory process involving member countries, generating global SOC maps at 30 arc-seconds (1 km) resolution.

We also compared our results with country-specific and continental efforts that used both conventional and DSM across the US at various scales. At 30 m spatial resolution, Probabilistic Remapping of SSURGO (POLARIS) (Chaney et al., 2019), predicted 59.2 Pg and 105.0 Pg SOC for 0–30 cm and 0–100 cm depth, respectively across CONUS. The spatial patterns observed in the POLARIS SOC align closely with our estimations; however, notably higher values were evident in regions with high SOC levels at both depths, including the Pacific Northwest, the Northern Lake States, and the Northeast (Figure S9 in Supporting Information S1). Calculation from soil property and class maps of CONUS at 100 m spatial resolution (Ramcharan et al., 2018) indicated 106.2 Pg and 186.7 Pg for 0–30 cm and 0–100 cm depth, respectively. SOC estimation derived from Ramcharan et al. (2018) revealed the highest SOC value among all data sets across the continental US (Figure S9 in Supporting Information S1). Another study implemented geographically weighted regression (GWR) at 800 m resolution and estimated SOC stock at 0–100 cm depth across CONUS as 75.2 Pg (Gonçalves et al., 2021), which was also higher than our estimate. However, GWR cannot simulate the nonlinear relationships between environmental covariates and SOC, which may explain why their estimate is high. On the other hand, the SSURGO with data gaps filled at 1:250,000-scale (~125 m resolution) Digital General Soil Map (Bliss et al., 2014) estimated smaller SOC stocks in CONUS for 0–30 cm (29.3 Pg) and 0–100 cm (57 Pg) depths compared to our estimates. This might be attributed to the fact that the SSURGO data set is structured based on soil map units. However, assuming SOC stocks are homogenous within map units underrepresents the spatial variability of SOC (Adams & Wilde, 1976; Thomas et al., 1989). Our 0–30 cm estimate in CONUS (35.4 ± 1.8 Pg, Table 2) was also larger than an estimate across Mexico and CONUS at 250 m resolution (28.9 Pg) from Guevara et al. (2020), despite exhibiting similar spatial patterns (Figure S9 in Supporting Information S1). While both studies used similar training data (ISCN observations) and machine learning models (RF and QRF), they operated at different scales and Guevara et al. (2020) did not utilize MGC as we have here.

While our calculated SOC stocks at the CONUS scale are generally comparable to other estimates, they have the advantage of revealing heterogeneous spatial patterns of the drivers of SOC stocks, as well as highlighting regions of high uncertainty/low representativeness. Nevertheless, it is interesting that such large discrepancies exist in SOC estimates across the US among different upscaling methods and at different scales. Future studies should more deeply explore the root cause of this, probing both scale uncertainties and method uncertainties, in order to harmonize and reconcile community efforts toward the prediction of SOC.

4.4. Limitations and Perspectives

We acknowledge that there were certain limitations in this study. First, the selection of the 30 arc second (~1 km) resolution was primarily based on its status as the most commonly employed finest resolution in ESMs for modeling SOC. It is also a common resolution for large-scale SOC mapping efforts. For example, HWSD and GSOCmap also adopted the 1 km resolution. However, we fully recognize that the estimation of SOC might be influenced by variations in pixel sizes, potentially introducing uncertainties and impacting the magnitudes and trends of the estimated SOC stocks. As demonstrated in prior studies (Adhikari et al., 2020), the key predictors of SOC stocks display variability across spatial scales throughout the CONUS. Understanding the scaling behavior of SOC stocks and their environmental controls is important. Future studies should explore uncertainties, including scaling uncertainties, in estimated SOC stocks.

Second, we aimed to evaluate the current "baseline" SOC stock and its spatial distribution across the US. The scope of the new map encompasses contemporary and general conditions in the US. However, the soil observational database (ISCN) we used consists of soil profiles collected over several decades, ranging from the 1910s to the 2010s. Although total SOC stocks are expected to be fairly stable over time, such a diverse time range of data is likely to introduce additional uncertainties, considering potential variations in soil conditions over the years. Thus, using legacy data may limit the full potential of our approach to accurately represent baseline SOC stocks (Guevara et al., 2020). In future research, further exploration of time series and time-dependent estimates should be considered.

Third, this study utilized a single pedotransfer function for estimating bulk density in cases where BD data for certain soil observations were missing. Precise bulk density data are crucial for obtaining realistic SOC estimates and minimizing uncertainties. Recently, several studies have adopted machine learning-based models, achieving R^2 values exceeding 0.5 for more accurate bulk density estimations (Jalabert et al., 2010; Zihao et al., 2022). Nonetheless, the scarcity of accurate bulk density data remains a significant challenge in current country-to-global SOC assessments. The soil community should increase efforts to provide bulk density data along with estimates of SOC concentrations (Billings et al., 2021).

5. Conclusions

This study presents a novel framework integrating MGC with machine learning to produce robust estimates of SOC stocks on a continental scale across the United States. Our approach effectively captured the diverse relationships between environmental covariates and SOC stocks across different SOC regions. We identified regions with low representativeness and high uncertainty, where additional measurements are required to produce more accurate SOC maps. We generated gridded SOC stock estimates for the United States at a 30 arc-second (~1 km) spatial resolution. These estimates, with quantified uncertainties, indicated SOC stocks of 52.6 ± 3.2 Pg and 108.3 ± 8.2 Pg for 0–30 cm and 0–100 cm depths, respectively. Our work has provided valuable insights into the complex relationships between SOC and predictor variables in different regions. Furthermore, our approach can contribute to model benchmarking activities by improving continental-scale SOC estimates and informing Earth system models to accurately represent terrestrial C pools and processes.

Acknowledgments

This research was sponsored by the National Science Foundation, Macrosystem Biology and NEON-enabled Science program (Award # DEB-2106137 and DEB-2106138). This research used resources from the Compute and Data Environment for Science (CADES) at Oak Ridge National Laboratory. ORNL is managed by UT-Battelle, LLC, under contract DEAC05-00OR22725 with the US Department of Energy. The National Ecological Observatory Network is a program sponsored by the National Science Foundation and operated under cooperative agreement by Battelle. This material is based in part upon work supported by the National Science Foundation through the NEON Program. This research was partially supported by the Reducing Uncertainties in Biogeochemical Interactions through Synthesis and Computation (RUBISCO) Science Focus Area, which is sponsored by the Regional and Global Model Analysis (RGMA) activity of the Earth & Environmental Systems Modeling (EESM) Program in the Earth and Environmental Systems Sciences Division (EESSD) of the Office of Biological and Environmental Research (BER) in the US Department of Energy Office of Science. This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The publisher acknowledges the US government license to provide public access under the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

Statistical Software: Data processing and statistical analysis were performed in R software (R Core Team, 2022) version 4.2.2 using packages “randomForest” version 4.7.1.1 (Liaw & Wiener, 2002), “caret” version 6.0-92 (Kuhn, 2008), “terra” version 1.6-5 (Hijmans, 2022), “tidyverse” version 1.3.1 (Wickham et al., 2019), “data.table” version 1.14.3 (Barrett et al., 2022).

Data and Code Availability: The data and code used for creating SOC regions and the SOC estimates in the study are available from Zenodo (Wang et al., 2024). Code used for the MGC and Representativeness Analysis is also available from Zenodo (Kumar, 2023).

References

Abdelbaki, A. M. (2018). Evaluation of pedotransfer functions for predicting soil bulk density for U.S. soils. *Ain Shams Engineering Journal*, *9*(4), 1611–1619. https://doi.org/10.1016/j.asej.2016.12.002

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/wics.101

Adams, J. A., & Wilde, R. H. (1976). Variability within a soil mapping unit mapped at the soil type level in the Wanganui district. *New Zealand Journal of Agricultural Research*, *19*(4), 435–442. https://doi.org/10.1080/00288233.1976.10420972

Adhikari, K., Mishra, U., Owens, P. R., Libohova, Z., Wills, S. A., Riley, W. J., et al. (2020). Importance and strength of environmental controllers of soil organic carbon changes with scale. *Geoderma*, *375*, 114472. https://doi.org/10.1016/j.geoderma.2020.114472

Ahmed, Z. U., Woodbury, P. B., Sanderman, J., Hawke, B., Jauss, V., Solomon, D., & Lehmann, J. (2017). Assessing soil carbon vulnerability in the Western USA by geospatial modeling of pyrogenic and particulate carbon stocks. *Journal of Geophysical Research: Biogeosciences*, *122*(2), 354–369. https://doi.org/10.1002/2016jg003488

Amelung, W., Bossio, D., de Vries, W., Kögel-Knabner, I., Lehmann, J., Amundson, R., et al. (2020). Towards a global-scale soil climate mitigation strategy. *Nature Communications*, *11*(1), 5427. https://doi.org/10.1038/s41467-020-18887-7

Amundson, R. (2001). The carbon budget in soils. *Annual Review of Earth and Planetary Sciences*, *29*(1), 535–562. https://doi.org/10.1146/annurev.earth.29.1.535

Barrett, T., Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., & Hocking, T. (2022). data.table: Extension of 'data.frame'. R package version 1.14.3. Retrieved from https://Rdatatable.gitlab.io/data.table

Batjes, N. H. (1996). Total carbon and nitrogen in the soils of the world. *European Journal of Soil Science*, *47*(2), 151–163. https://doi.org/10.1111/j.1365-2389.1996.tb01386.x

Batjes, N. H. (2000). Effects of mapped variation in soil conditions on estimates of soil carbon and nitrogen stocks for South America. *Geoderma*, *97*(1), 135–144. https://doi.org/10.1016/s0016-7061(00)00031-8

Billings, S. A., Lajtha, K., Malhotra, A., Berhe, A. A., de Graaff, M.-A., Earl, S., et al. (2021). Soil organic carbon is not just for soil scientists: Measurement recommendations for diverse practitioners. *Ecological Applications*, *31*(3), e02290. https://doi.org/10.1002/eap.2290

Bjorkman, A. D., García Criado, M., Myers-Smith, I. H., Ravolainen, V., Jónsdóttir, I. S., Westergaard, K. B., et al. (2020). Status and trends in Arctic vegetation: Evidence from experimental warming and long-term monitoring. *Ambio*, *49*(3), 678–692. https://doi.org/10.1007/s13280-019-01161-6

Bliss, N. B., & Maursetter, J. (2010). Soil organic carbon stocks in Alaska estimated with spatial and pedon data. *Soil Science Society of America Journal*, *74*(2), 565–579. https://doi.org/10.2136/sssaj2008.0404

Bliss, N. B., Waltman, S., West, L. T., Neale, A., & Mehaffey, M. (2014). Distribution of soil organic carbon in the conterminous United States. *Soil Carbon*, 85–93. https://doi.org/10.1007/978-3-319-04084-4_9

Bond-Lamberty, B., Epron, D., Harden, J., Harmon, M. E., Hoffman, F., Kumar, J., et al. (2016). Estimating heterotrophic respiration at large scales: Challenges, approaches, and next steps. *Ecosphere*, *7*(6), e01380. https://doi.org/10.1002/ecs2.1380

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brevik, E. C., Calzolari, C., Miller, B. A., Pereira, P., Kabala, C., Baumgarten, A., & Jordán, A. (2016). Soil mapping, classification, and pedologic modeling: History and future directions. *Geoderma*, *264*, 256–274. https://doi.org/10.1016/j.geoderma.2015.05.017

Canadell, J. G., Monteiro, P. M. S., Costa, M. H., Cotrim da Cunha, L., Cox, P. M., Eliseev, A. V., et al. (2021). Global carbon and other biogeochemical cycles and feedbacks. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, et al. (Eds.), *Climate change 2021: The physical science basis. Contribution of Working Group I to the sixth assessment report of the Intergovernmental Panel on Climate Change* (pp. 673–816). Cambridge University Press.

Chaney, N. W., Minasny, B., Herman, J. D., Nauman, T. W., Brungard, C. W., Morgan, C. L. S., et al. (2019). POLARIS soil properties: 30-m probabilistic maps of soil properties over the contiguous United States. *Water Resources Research*, *55*(4), 2916–2938. https://doi.org/10.1029/2018wr022797

Chen, S., Arrouays, D., Angers, D. A., Chenu, C., Barré, P., Martin, M. P., et al. (2019). National estimation of soil organic carbon storage potential for arable soils: A data-driven approach coupled with carbon-landscape zones. *Science of the Total Environment*, *666*, 355–367. https://doi.org/10.1016/j.scitotenv.2019.02.249

Chen, S., Arrouays, D., Leatitia Mulder, V., Poggio, L., Minasny, B., Roudier, P., et al. (2022). Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. *Geoderma*, *409*, 115567. https://doi.org/10.1016/j.geoderma.2021.115567

Cox, P. M., Betts, R. A., Jones, C. D., Spall, S. A., & Totterdell, I. J. (2000). Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature*, *408*(6809), 184–187. https://doi.org/10.1038/35041539

De Jager, N. R., Thomsen, M., & Yin, Y. (2012). Threshold effects of flood duration on the vegetation and soils of the Upper Mississippi River floodplain, USA. *Forest Ecology and Management*, *270*, 135–146. https://doi.org/10.1016/j.foreco.2012.01.023

Didan, K. (2021). MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid V061 [Dataset]. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD13A2.061

Doetterl, S., Stevens, A., Six, J., Merckx, R., Van Oost, K., Casanova Pinto, M., et al. (2015). Soil carbon storage controlled by interactions between geochemistry and climate. *Nature Geoscience*, *8*(10), 780–783. https://doi.org/10.1038/ngeo2516

Drew, L. A. (1973). Bulk density estimation based on organic matter content of some Minnesota soils.

Earth Resources Observation and Science (EROS) Center. (2017). Global 30 arc-second elevation (GTOPO30) [Dataset]. U.S. Geological Survey. https://doi.org/10.5066/F7DF6PQS

Egli, M., & Poulenard, J. (2016). Soils of mountainous landscapes. In *International encyclopedia of geography* (pp. 1–10).

Elsey-Quirk, T., Graham, S. A., Mendelssohn, I. A., Snedden, G., Day, J. W., Twilley, R. R., et al. (2019). Mississippi river sediment diversions and coastal wetland sustainability: Synthesis of responses to freshwater, sediment, and nutrient inputs. *Estuarine, Coastal and Shelf Science*, *221*, 170–183. https://doi.org/10.1016/j.ecss.2019.03.002

FAO, & ITPS. (2020). *Global soil organic carbon Map V1.5: Technical report*. FAO. https://doi.org/10.4060/ca7597en

FAO, IIASA, ISRIC, ISSCAS, & JRC. (2012). *Harmonized World Soil Database - HWSD (version 1.2)*. International Institute for Applied Systems Analysis (IIASA).

FAO. (2018). Soil organic carbon mapping cookbook.

Fenton, N., Lecomte, N., Légaré, S., & Bergeron, Y. (2005). Paludification in black spruce (*Picea mariana*) forests of eastern Canada: Potential factors and management implications. *Forest Ecology and Management*, *213*(1), 151–159. https://doi.org/10.1016/j.foreco.2005.03.017

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Gregor, L., Hauck, J., et al. (2022). Global carbon budget 2022. *Earth System Science Data*, *14*(11), 4811–4900. https://doi.org/10.5194/essd-14-4811-2022

Garten, C. T., Post, W. M., Hanson, P. J., & Cooper, L. W. (1999). Forest soil carbon inventories and dynamics along an elevation gradient in the southern Appalachian Mountains. *Biogeochemistry*, *45*(2), 115–145. https://doi.org/10.1007/BF01106778

Gautam, S., Mishra, U., Scown, C. D., Wills, S. A., Adhikari, K., & Drewniak, B. A. (2022). Continental United States may lose 1.8 petagrams of soil organic carbon under climate change by 2100. *Global Ecology and Biogeography*, *31*(6), 1147–1160. https://doi.org/10.1111/geb.13489

Georgiou, K., Malhotra, A., Wieder, W. R., Ennis, J. H., Hartman, M. D., Sulman, B. N., et al. (2021). Divergent controls of soil organic carbon between observations and process-based models. *Biogeochemistry*, *156*(1), 5–17. https://doi.org/10.1007/s10533-021-00819-2

Gonçalves, D. R. P., Mishra, U., Wills, S., & Gautam, S. (2021). Regional environmental controllers influence continental scale soil carbon stocks and future carbon dynamics. *Scientific Reports*, *11*(1), 6474. https://doi.org/10.1038/s41598-021-85992-y

Gorelick, N., Matt, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Grubaugh, J. W., & Anderson, R. V. (1989). Upper Mississippi River: Seasonal and floodplain forest influences on organic matter transport. *Hydrobiologia*, *174*(3), 235–244. https://doi.org/10.1007/BF00008163

Guevara, M., Arroyo, C., Brunsell, N., Cruz, C. O., Domke, G., Equihua, J., et al. (2020). Soil organic carbon across Mexico and the conterminous United States (1991–2010). *Global Biogeochemical Cycles*, *34*(3), e2019GB006219. https://doi.org/10.1029/2019GB006219

Guevara, M., Olmedo, G. F., Stell, E., Yigini, Y., Aguilar Duarte, Y., Arellano Hernández, C., et al. (2018). No silver bullet for digital soil mapping: Country-specific soil organic carbon estimates across Latin America. *Soil*, *4*(3), 173–193. https://doi.org/10.5194/soil-4-173-2018

Harden, J. W., Hugelius, G., Ahlström, A., Blankinship, J. C., Bond-Lamberty, B., Lawrence, C. R., et al. (2018). Networking our science to characterize the state, vulnerabilities, and management opportunities of soil organic matter. *Global Change Biology*, *24*(2), e705–e718. https://doi.org/10.1111/gcb.13896

Hargrove, W. W., & Hoffman, F. M. (1999). Using multivariate clustering to characterize ecoregion borders. *Computing in Science & Engineering*, *1*(4), 18–25. https://doi.org/10.1109/5992.774837

Hargrove, W. W., & Hoffman, F. M. (2004). Potential of multivariate quantitative methods for delineation and visualization of ecoregions. *Environmental Management*, *34*(1), S39–S60. https://doi.org/10.1007/s00267-003-1084-0

Hargrove, W. W., Hoffman, F. M., & Law, B. E. (2003). New analysis reveals representativeness of the AmeriFlux network. *Eos, Transactions American Geophysical Union*, *84*(48), 529–535. https://doi.org/10.1029/2003eo480001

Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., et al. (2014). SoilGrids1km — Global soil information based on automated mapping. *PLoS One*, *9*(8), e105992. https://doi.org/10.1371/journal.pone.0105992

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., et al. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One*, *12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Hijmans, R. (2022). terra: Spatial data analysis. R package version 1.6-5. Retrieved from https://CRAN.R-project.org/package=terra

Hobbie, S. E., Schimel, J. P., Trumbore, S. E., & Randerson, J. R. (2000). Controls over carbon storage and turnover in high-latitude soils. *Global Change Biology*, *6*(S1), 196–210. https://doi.org/10.1046/j.1365-2486.2000.06021.x

Hoffman, F. M., Kumar, J., Mills, R. T., & Hargrove, W. W. (2013). Representativeness-based sampling network design for the State of Alaska. *Landscape Ecology*, *28*(8), 1567–1586. https://doi.org/10.1007/s10980-013-9902-0

Hugelius, G., Tarnocai, C., Broll, G., Canadell, J. G., Kuhry, P., & Swanson, D. K. (2013). The Northern Circumpolar Soil Carbon Database: Spatially distributed datasets of soil coverage and soil carbon storage in the northern permafrost regions. *Earth System Science Data*, *5*(1), 3–13. https://doi.org/10.5194/essd-5-3-2013

IPCC. (2022). *Climate change 2022: Impacts, adaptation, and vulnerability. Contribution of Working Group II to the sixth assessment report of the Intergovernmental Panel on Climate Change* (p. 3056). Cambridge University Press.

Ise, T., Dunn, A. L., Wofsy, S. C., & Moorcroft, P. R. (2008). High sensitivity of peat decomposition to climate change through water-table feedback. *Nature Geoscience*, *1*(11), 763–766. https://doi.org/10.1038/ngeo331

Jalabert, S. S. M., Martin, M. P., Renaud, J.-P., Boulonne, L., Jolivet, C., Montanarella, L., & Arrouays, D. (2010). Estimating forest soil bulk density using boosted regression modelling. *Soil Use and Management*, *26*(4), 516–528. https://doi.org/10.1111/j.1475-2743.2010.00305.x

Jandl, R., Rodeghiero, M., Martinez, C., Cotrufo, M. F., Bampa, F., van Wesemael, B., et al. (2014). Current status, uncertainty and future needs in soil organic carbon monitoring. *Science of the Total Environment*, *468–469*, 376–383. https://doi.org/10.1016/j.scitotenv.2013.08.026

Jenny, H. (1994). *Factors of soil formation: A system of quantitative pedology*. Courier Corporation.

Jobbágy, E. G., & Jackson, R. B. (2000). The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecological Applications*, *10*(2), 423–436. https://doi.org/10.1890/1051-0761(2000)010[0423:tvdoso]2.0.co;2

Johnson, K. D., Harden, J., McGuire, A. D., Bliss, N. B., Bockheim, J. G., Clark, M., et al. (2011). Soil carbon distribution in Alaska in relation to soil-forming factors. *Geoderma*, *167–168*, 71–84. https://doi.org/10.1016/j.geoderma.2011.10.006

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Keller, M., Schimel, D. S., Hargrove, W. W., & Hoffman, F. M. (2008). A continental strategy for the National Ecological Observatory Network. *Frontiers in Ecology and the Environment*, *6*(5), 282–284. https://doi.org/10.1890/1540-9295(2008)6[282:acsftn]2.0.co;2

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, Articles*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Kumar, J. (2023). Multivariate quantitative representativeness and constituency analysis of ecological observation networks (v1.0.1). *Zenodo*. https://doi.org/10.5281/zenodo.8048530

Kumar, J., Hoffman, F. M., Hargrove, W. W., & Collier, N. (2016). Understanding the representativeness of FLUXNET for upscaling carbon flux from eddy covariance measurements. *Earth System Science Data Discussions*, *2016*, 1–25. Retrieved from https://essd.copernicus.org/preprints/essd-2016-36/

Köchy, M., Hiederer, R., & Freibauer, A. (2015). Global distribution of soil organic carbon – Part 1: Masses and frequency distributions of SOC stocks for the tropics, permafrost regions, wetlands, and the world. *Soil*, *1*(1), 351–365. https://doi.org/10.5194/soil-1-351-2015

Lal, R. (2004a). Carbon sequestration in dryland ecosystems. *Environmental Management*, *33*(4), 528–544. https://doi.org/10.1007/s00267-003-9110-9

Lal, R. (2004b). Soil carbon sequestration impacts on global climate change and food security. *Science*, *304*(5677), 1623–1627. https://doi.org/10.1126/science.1097396

Lauenroth, W. K., & Bradford, J. B. (2009). Ecohydrology of dry regions of the United States: Precipitation pulses and intraseasonal drought. *Ecohydrology*, *2*(2), 173–181. https://doi.org/10.1002/eco.53

Lettens, S., Van Orshoven, J., van Wesemael, B., & Muys, B. (2004). Soil organic and inorganic carbon contents of landscape units in Belgium derived using data from 1950 to 1970. *Soil Use and Management*, *20*(1), 40–47. https://doi.org/10.1111/j.1475-2743.2004.tb00335.x

Li, H., Wu, Y., Liu, S., Xiao, J., Zhao, W., Chen, J., et al. (2022). Decipher soil organic carbon dynamics and driving forces across China using machine learning. *Global Change Biology*, *28*(10), 3394–3410. https://doi.org/10.1111/gcb.16154

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22. Retrieved from https://CRAN.R-project.org/doc/Rnews/

Lin, Z., Dai, Y., Mishra, U., Wang, G., Shangguan, W., Zhang, W., & Qin, Z. (2022). On the magnitude and uncertainties of global and regional soil organic carbon: A comparative analysis using multiple estimates. *Earth System Science Data Discussions*, *2022*, 1–24. Retrieved from https://essd.copernicus.org/preprints/essd-2022-232/

Liu, S., Wei, Y., Post, W. M., Cook, R. B., Schaefer, K., & Thornton, M. M. (2013). The Unified North American Soil Map and its implication on the soil organic carbon stock in North America. *Biogeosciences*, *10*(5), 2915–2930. https://doi.org/10.5194/bg-10-2915-2013

Luo, Z., Viscarra-Rossel, R. A., & Qian, T. (2021). Similar importance of edaphic and climatic factors for controlling soil organic carbon stocks of the world. *Biogeosciences*, *18*(6), 2063–2073. https://doi.org/10.5194/bg-18-2063-2021

Malhotra, A., Todd-Brown, K., Nave, L. E., Batjes, N. H., Holmquist, J. R., Hoyt, A. M., et al. (2019). The landscape of soil carbon data: Emerging questions, synergies and databases. *Progress in Physical Geography: Earth and Environment*, *43*(5), 707–719. https://doi.org/10.1177/0309133319873309

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*(1), 3–52. https://doi.org/10.1016/s0016-7061(03)00223-4

Michaelson, G. J., Ping, C.-L., & Clark, M. (2013). Soil pedon carbon and nitrogen data for Alaska: An analysis and update. *Open Journal of Soil Science*, *03*(02), 11–142. https://doi.org/10.4236/ojss.2013.32015

Minasny, B., & McBratney, A. B. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma*, *264*, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017

Mishra, U., & Riley, W. J. (2012). Alaskan soil carbon stocks: Spatial variability and dependence on environmental factors. *Biogeosciences*, *9*(9), 3637–3645. https://doi.org/10.5194/bg-9-3637-2012

Mishra, U., Hugelius, G., Shelef, E., Yang, Y., Strauss, J., Lupachev, A., et al. (2021). Spatial heterogeneity and environmental predictors of permafrost region soil organic carbon stocks. *Science Advances*, *7*(9), eaaz5236. https://doi.org/10.1126/sciadv.aaz5236

Mishra, U., Lal, R., Liu, D., & Van Meirvenne, M. (2010). Predicting the spatial variation of the soil organic carbon pool at a regional scale. *Soil Science Society of America Journal*, *74*(3), 906–914. https://doi.org/10.2136/sssaj2009.0158

Mishra, U., Yeo, K., Adhikari, K., Riley, W. J., Hoffman, F. M., Hudson, C., & Gautam, S. (2022). Empirical relationships between environmental factors and soil organic carbon produce comparable prediction accuracy to machine learning. *Soil Science Society of America Journal*, *86*(6), 1611–1624. https://doi.org/10.1002/saj2.20453

Myneni, R., Knyazikhin, Y., & Park, T. (2021). MODIS/Terra leaf area index/FPAR 8-day L4 global 500m SIN grid V061. Distributed by NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD15A2H.061

National Ecological Observatory Network. (2023). Soil physical and chemical properties, distributed initial characterization (DP1.10047.001). Retrieved from https://data.neonscience.org/data-products/DP1.10047.001

Nave, L., Johnson, K., van Ingen, C., Agarwal, D., Humphrey, M., & Beekwilder, N. (2022). *International Soil Carbon Network version 3 database (ISCN3) ver 1*. Environmental Data Initiative.

Padarian, J., McBratney, A. B., & Minasny, B. (2020). Game theory interpretation of digital soil mapping convolutional neural networks. *Soil*, *6*(2), 389–397. https://doi.org/10.5194/soil-6-389-2020

Ping, C.-L., Michaelson, G. J., Jorgenson, M. T., Kimble, J. M., Epstein, H., Romanovsky, V. E., & Walker, D. A. (2008). High stocks of soil organic carbon in the North American Arctic region. *Nature Geoscience*, *1*(9), 615–619. https://doi.org/10.1038/ngeo284

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *Soil*, *7*(1), 217–240. https://doi.org/10.5194/soil-7-217-2021

R Core Team. (2022). R: A language and environment for statistical computing.

Ramcharan, A., Hengl, T., Nauman, T., Brungard, C., Waltman, S., Wills, S., & Thompson, J. (2018). Soil property and class maps of the conterminous United States at 100-meter spatial resolution. *Soil Science Society of America Journal*, *82*(1), 186–201. https://doi.org/10.2136/sssaj2017.04.0122

Rasmussen, C., Heckman, K., Wieder, W. R., Keiluweit, M., Lawrence, C. R., Berhe, A. A., et al. (2018). Beyond clay: Towards an improved set of variables for predicting soil organic matter content. *Biogeochemistry*, *137*(3), 297–306. https://doi.org/10.1007/s10533-018-0424-3

Rumpel, C., Amiraslani, F., Chenu, C., Garcia Cardenas, M., Kaonga, M., Koutika, L.-S., et al. (2020). The 4p1000 initiative: Opportunities, limitations and challenges for implementing soil organic carbon sequestration as a sustainable development strategy. *Ambio*, *49*(1), 350–360. https://doi.org/10.1007/s13280-019-01165-2

Scharlemann, J. P. W., Tanner, E. V. J., Hiederer, R., & Kapos, V. (2014). Global soil carbon: Understanding and managing the largest terrestrial carbon pool. *Carbon Management*, *5*(1), 81–91. https://doi.org/10.4155/cmt.13.77

Schimel, D., Hargrove, W., Hoffman, F., & MacMahon, J. (2007). NEON: A hierarchically designed national ecological network. *Frontiers in Ecology and the Environment*, *5*(2), 59. https://doi.org/10.1890/1540-9295(2007)5[59:nahdne]2.0.co;2

Schrumpf, M., Schulze, E. D., Kaiser, K., & Schumacher, J. (2011). How accurately can soil organic carbon stocks and stock changes be quantified by soil inventories? *Biogeosciences*, *8*(5), 1193–1212. https://doi.org/10.5194/bg-8-1193-2011

Schulze, E. D., & Freibauer, A. (2005). Carbon unlocked from soils. *Nature*, *437*(7056), 205–206. https://doi.org/10.1038/437205a

Schuur, E. A. G., Bockheim, J., Canadell, J. G., Euskirchen, E., Field, C. B., Goryachkin, S. V., et al. (2008). Vulnerability of permafrost carbon to climate change: Implications for the global carbon cycle. *BioScience*, *58*(8), 701–714. https://doi.org/10.1641/B580807

Scull, P., Franklin, J., Chadwick, O. A., & McArthur, D. (2003). Predictive soil mapping: A review. *Progress in Physical Geography: Earth and Environment*, *27*(2), 171–197. https://doi.org/10.1191/0309133303pp366ra

Shedayi, A. A., Xu, M., Naseer, I., & Khan, B. (2016). Altitudinal gradients of soil and vegetation carbon and nitrogen in a high altitude nature reserve of Karakoram ranges. *SpringerPlus*, *5*(1), 320. https://doi.org/10.1186/s40064-016-1935-9

Sheikh, M. A., Kumar, M., & Bussmann, R. W. (2009). Altitudinal variation in soil organic carbon stock in coniferous subtropical and broadleaf temperate forests in Garhwal Himalaya. *Carbon Balance and Management*, *4*(1), 6. https://doi.org/10.1186/1750-0680-4-6

Smith, P., Soussana, J.-F., Angers, D., Schipper, L., Chenu, C., Rasse, D. P., et al. (2020). How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal. *Global Change Biology*, *26*(1), 219–241. https://doi.org/10.1111/gcb.14815

Soil Survey Staff. (2020). Gridded National Soil Survey Geographic (gNATSGO) database for the conterminous United States. Retrieved from https://nrcs.app.box.com/v/soils

Song, X.-D., Wu, H.-Y., Ju, B., Liu, F., Yang, F., Li, D.-C., et al. (2020). Pedoclimatic zone-based three-dimensional soil organic carbon mapping in China. *Geoderma*, *363*, 114145. https://doi.org/10.1016/j.geoderma.2019.114145

Stockmann, U., Padarian, J., McBratney, A., Minasny, B., de Brogniez, D., Montanarella, L., et al. (2015). Global soil organic carbon assessment. *Global Food Security*, *6*, 9–16. https://doi.org/10.1016/j.gfs.2015.07.001

Tarnocai, C., Canadell, J. G., Schuur, E. A. G., Kuhry, P., Mazhitova, G., & Zimov, S. (2009). Soil organic carbon pools in the northern circumpolar permafrost region. *Global Biogeochemical Cycles*, *23*(2), GB2023. https://doi.org/10.1029/2008GB003327

Thomas, P. J., Baker, J. C., & Simpson, T. W. (1989). Variability of the Cecil Map Unit in Appomattox County, Virginia. *Soil Science Society of America Journal*, *53*(5), 1470–1474. https://doi.org/10.2136/sssaj1989.03615995005300050028x

Tifafi, M., Guenet, B., & Hatté, C. (2018). Large differences in global and regional total soil carbon stock estimates based on SoilGrids, HWSD, and NCSCD: Intercomparison and evaluation based on field data from USA, England, Wales, and France. *Global Biogeochemical Cycles*, *32*(1), 42–56. https://doi.org/10.1002/2017gb005678

Todd-Brown, K. E. O., Abramoff, R. Z., Beem-Miller, J., Blair, H. K., Earl, S., Frederick, K. J., et al. (2022). Reviews and syntheses: The promise of big diverse soil data, moving current practices towards future potential. *Biogeosciences*, *19*(14), 3505–3522. https://doi.org/10.5194/bg-19-3505-2022

Todd-Brown, K. E. O., Randerson, J. T., Post, W. M., Hoffman, F. M., Tarnocai, C., Schuur, E. A. G., & Allison, S. D. (2013). Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations. *Biogeosciences*, *10*(3), 1717–1736. https://doi.org/10.5194/bg-10-1717-2013

Tranter, G., Minasny, B., Mcbratney, A. B., Murphy, B., Mckenzie, N. J., Grundy, M., & Brough, D. (2007). Building and testing conceptual and empirical models for predicting soil bulk density. *Soil Use and Management*, *23*(4), 437–443. https://doi.org/10.1111/j.1475-2743.2007.00092.x

U.S. Geological Survey. (2023). North America elevation 1-kilometer resolution grid. *ScienceBase*. Retrieved from https://www.sciencebase.gov/catalog/item/4fb5495ee4b04cb937751d6d

Vitharana, U. W. A., Mishra, U., & Mapa, R. B. (2019). National soil organic carbon estimates can improve global estimates. *Geoderma*, *337*, 55–64. https://doi.org/10.1016/j.geoderma.2018.09.005

Vitharana, U. W. A., Mishra, U., Jastrow, J. D., Matamala, R., & Fan, Z. (2017). Observational needs for estimating Alaskan soil carbon stocks under current and future climate. *Journal of Geophysical Research: Biogeosciences*, *122*(2), 415–429. https://doi.org/10.1002/2016jg003421

von Fromm, S. F., Hoyt, A. M., Lange, M., Acquah, G. E., Aynekulu, E., Berhe, A. A., et al. (2021). Continental-scale controls on soil organic carbon across sub-Saharan Africa. *Soil*, *7*(1), 305–332. https://doi.org/10.5194/soil-7-305-2021

Vågen, T.-G., & Winowiecki, L. A. (2013). Mapping of soil organic carbon stocks for spatially explicit assessments of climate change mitigation potential. *Environmental Research Letters*, *8*(1), 015011. https://doi.org/10.1088/1748-9326/8/1/015011

Wang, Z., kumar, J., Weintraub-Leff, S. R., Todd-Brown, K., Mishra, U., & Sihi, D. (2024). Soil organic carbon stocks distribution over continental United States [Dataset]. In *Journal of Geophysical Research: Biogeosciences* (Vol. 129). Zenodo. https://doi.org/10.5281/zenodo.10602250

Weintraub, S. R., Flores, A. N., Wieder, W. R., Sihi, D., Cagnarini, C., Gonçalves, D. R. P., et al. (2019). Leveraging environmental research and observation networks to advance soil carbon science. *Journal of Geophysical Research: Biogeosciences*, *124*(5), 1047–1055. https://doi.org/10.1029/2018jg004956

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wu, Q. (2020). geemap: A Python package for interactive mapping with Google Earth Engine. *Journal of Open Source Software*, *5*(51), 2305. https://doi.org/10.21105/joss.02305

Yigini, Y., Olmedo, G., Reiter, S., Baritz, R., Viatkin, K., & Vargas, R. (2018). Soil organic carbon mapping: Cookbook.

Yu, W., Weintraub, S. R., & Hall, S. J. (2021). Climatic and geochemical controls on soil carbon at the continental scale: Interactions and thresholds. *Global Biogeochemical Cycles*, *35*(3), e2020GB006781. https://doi.org/10.1029/2020GB006781

Zhang, X., Chen, S., Xue, J., Wang, N., Xiao, Y., Chen, Q., et al. (2023). Improving model parsimony and accuracy by modified greedy feature selection in digital soil mapping. *Geoderma*, *432*, 116383. https://doi.org/10.1016/j.geoderma.2023.116383

Zhao, M., Heinsch, F. A., Nemani, R. R., & Running, S. W. (2005). Improvements of the MODIS terrestrial gross and net primary production global data set. *Remote sensing of Environment*, *95*(2), 164–176. https://doi.org/10.1016/j.rse.2004.12.011

Zihao, H., Shaofei, J., & Ku, W. (2022). Application of machine learning methods for estimation soil bulk density. In *2022 2nd Asia-Pacific conference on communications technology and computer science (ACCTCS)* (pp. 194–198).