PLOS One

RESEARCH ARTICLE

Digital soil mapping in support of voluntary carbon market programs in agricultural land

James R. Kellner[1,2,3*], **Christian Clanton**[1], **Kirk M. Demuth**[1], **Mitchell Donovan**[1], **Y. Katherina Feng**[1], **Mage Khim-Young**[1], **Julia Maddalena**[1], **Rose Rustowicz**[1], **David Schurman**[1]

1 Perennial Climate Inc., Boulder, Colorado, United States of America, **2** Institute at Brown for Environment and Society, Brown University, Providence, Rhode Island, United States of America, **3** Department of Ecology, Evolution and Organismal Biology, Brown University, Providence, Rhode Island, United States of America

* jim@perennial.earth

OPEN ACCESS

Citation: Kellner JR, Clanton C, Demuth KM, Donovan M, Feng YK, Khim-Young M, et al. (2025) Digital soil mapping in support of voluntary carbon market programs in agricultural land. PLoS One 20(9): e0327895. https://doi.org/10.1371/journal.pone.0327895

Editor: Mattias Gaglio, University of Ferrara, ITALY

Received: December 14, 2024

Accepted: June 23, 2025

Published: September 2, 2025

Data availability statement: Measurements of soil organic carbon from the United States Department of Agriculture Rapid Carbon Assessment Program are available at https://www.nrcs.usda.gov/resources/data-and-reports/rapid-carbon-assessment-raca. All other publicly available covariate data analyzed in

Abstract

Voluntary carbon market (VCM) programs in agriculture depend on accurate measurements of soil organic carbon (SOC) that can be deployed at scale efficiently, but barriers are preventing widespread adoption. To overcome these challenges, we developed a digital soil mapping (DSM) framework driven by machine-learning and numerous spatial covariates, including long-term climate proxies, short-term climate and weather-related variables, topographic and edaphic measurements, and remote sensing time-series summaries. We show that the model can predict SOC content in the top 30 cm of soil using 5,230 measurements of SOC in agricultural land within 47 states in the contiguous United States (CONUS). Model predictions closely matched independent measured values. The intercept and slope of the cross-validated relationship at the agricultural field level were −0.179 and 1.095. The coefficient of determination was $R^2 = 0.811$, and the RMSE was 0.041. In contrast, comparison of independent field measurements to four publicly available SOC data products using 165 fields that contained 3,285 in-situ soil samples showed poor ability of existing public SOC maps to reproduce measured values, underscoring the importance of quantification technologies developed specifically for agricultural land and with recent soil measurements. Three prior SOC data products underestimated SOC content at small values and overestimated it at large ones, while one underestimated SOC content at all values examined. Analysis of feature importance showed that time series summaries from Sentinel-2 are the strongest predictors, followed by temperature variables and features related to surface hydrology. These findings underscore the value of geographically representative training and validation data for quantifying SOC content in agricultural land and demonstrate that feature engineering can increase the sensitivity of SOC quantification to optical remote sensing summaries. Data-driven algorithms can generate accurate estimates of field-level SOC content in agricultural land in CONUS that overcome barriers to scale in the VCM.

this study is accessible through repositories referenced in citations that appear in the bibliography.

Funding: This study was financially supported by the Agricultural Products Utilization Commission Grant Program of the North Dakota Department of Agriculture through grants APUC 21-333A and APUC 23-436A to JRK, Cloud Agronomics Incorporated, and Perennial Climate Incorporated; by the Advanced Industries Accelerator Program of the Colorado Office of Economic Development and International Trade through grant CTGG1 2022-3049 to Cloud Agronomics Incorporated; and by Perennial Climate Incorporated. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors would like to declare the following patent applications associated with this research: Perennial Climate Incorporated has filed two provisional patent applications that are currently under consideration, one of which is based on this research. Perennial Climate Incorporated is a private company that provides digital soil mapping as a service. The company markets digital soil mapping products related to this research. There are no additional patents, products in development or marketed products associated with this research to declare. This does not alter our adherence to PLOS ONE policies on sharing data and materials. JRK completed this work as an employee of Perennial Climate Inc. in compliance with the Brown University Policy on Outside Professional Activities for Faculty.

1. Introduction

Soil organic carbon (SOC) plays an important role in the voluntary carbon market (VCM) due to its ability to mitigate and offset greenhouse gas (GHG) emissions through implementation of regenerative agricultural practices [1]. Activities such as cover cropping, reduced tillage and changes to crop rotation increase SOC storage and remove GHG from the atmosphere [2]. The VCM incentivizes these practices by paying for GHG reductions and removals, which can be used to offset emissions elsewhere or sold on a secondary market.

A key challenge to the development of a robust VCM in agricultural soils is quantification. Current methods are uncertain and expensive to scale. Direct measurement, such as in-situ soil sampling, is locally accurate and can be extrapolated to larger regions [e.g., 3], but may require large sample sizes to overcome spatial variation [4–6]. An alternative to direct measurement is biogeochemical simulation. Simulation models forecast SOC sequestration using physical, chemical, and biological principles, but they are difficult to deploy at scale and in regions without a long history of academic research, including small-holder farms and crop types other than globally significant row crops [7–9].

Digital soil mapping (DSM) is able to overcome these measurement challenges. DSM refers to the practice of using empirical statistical and machine learning models to predict SOC content by associating in-situ soil samples with spatial covariates, including climate proxies, weather, soils, topography, and remote sensing data [10–14]. This approach is similar to methods used to estimate aboveground forest carbon in VCM programs [e.g., 15–18]. Numerous studies have shown that environmental and remote sensing variables can map SOC content in soils [e.g., 11,14].

Two criteria necessary for the success of the VCM in agricultural soils are the accuracy of area-based summaries and sensitivity of predictions to variables related to on-the-ground farm practice. VCM programs credit SOC sequestration within areas that can be hundreds of hectares or larger, requiring measurement techniques to be validated at the unit of agricultural land management, which may be the individual field, ranch, or collections of management units that include hundreds to thousands of individual parcels. Sensitivity to management practices is necessary to ensure that changes in SOC are attributable to changes in land management – a requirement of VCS carbon programs [19–21].

Here we develop the Advanced Terrestrial machine-Learning Analysis System for Soil Organic Carbon (ATLAS-SOC). ATLAS-SOC is a data-driven DSM framework to quantify SOC content in the top 30 cm of soil in non-tree row crops in the contiguous United States (CONUS) for use in VCM programs. We design time-series features that increase the sensitivity of remote sensing variables to SOC prediction using high-resolution satellite data from Sentinel-2 and other environmental covariates [12]. Model performance is evaluated with a geographically dependent cross-validation at both sample and field levels and benchmarked against existing publicly available SOC data products. Our analysis shows that publicly available SOC data products perform poorly when validated against recent in-situ soil samples, whereas

ATLAS-SOC produces accurate estimates of field-level SOC content, underscoring the potential of data-driven SOC quantification in support of robust VCM programs in agricultural soils.

2. Materials and methods

2.1. In-situ soil samples

Data-driven calculation of SOC depends on the relationship between independently measured variables and SOC content from in-situ soil samples. SOC content is the amount of SOC in a given soil sample expressed as a percentage of oven-dry mass. The approach developed here allows in-situ soil samples collected at different locations, depths below the surface, and times to be used for model calibration and validation. It does this by using locally measured covariates generated at times close to sample collection, and by treating depth as a continuous covariate feature [13,14,22–24]. This means that sample data is always associated with covariate features that coincide in space and time, even though samples themselves have been collected at different times [14,25,26].

Our approach uses 5,230 in-situ soil samples collected within 410 agricultural fields (Fig 1; Table 1). Most of these samples (4,260) were collected within actively cultivated, conventional row-crop agriculture in the states of Arkansas, Colorado, Illinois, Iowa, Kansas, Minnesota, Nebraska, New Mexico, Oklahoma, South Dakota, Texas, and Wisconsin during the years 2020 and 2021. The remaining 970 samples were collected in 2010 and 2011 under the USDA Rapid Carbon Assessment (RaCA) program [3; S1 Table]. At the time of collection, samples represented corn (2,249), soybeans (1,342), sorghum (400), and alfalfa (253), with the remaining 1,239 samples among less common conditions with < 100 samples each (e.g., grassland and pasture, cotton, spring wheat, and rice; Table 2). Crop types were identified using the United States Department of Agriculture (USDA) Cropland Data Layer (CDL) associated with the calendar year of sample collection. Measurements in 2020 and 2021 were acquired at randomly selected locations in fields that could be accessed before the growing season or after harvest, subject to the constraint that no sample location was within 5 m of a field boundary. The data set consists of 4,168 measurements in the 0–5 cm depth range, and 1,062 measurements of 0–30 cm. All samples were shipped to an analytical laboratory where SOC was determined as a percentage by mass using the method of dry combustion [27]. Most samples represent Mollisol soils (3,478), followed by Alfisols (913), Entisols (318), Aridosols (227) and Inceptisols (123). The

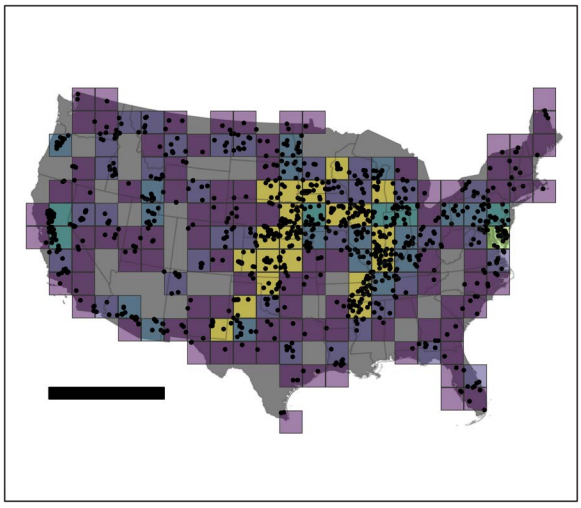

Fig 1. The spatial distribution of 5,230 in-situ soil samples used to develop a data-driven model of soil organic carbon as a percentage by mass in agricultural soil. Colors are proportional to the density of points within each grid cell. Scale bar is 1,000 km. State boundaries are reprinted from the Global Administrative Areas Database (GADM) under a CC BY license with permission from GADM, original copyright 2025.

https://doi.org/10.1371/journal.pone.0327895.g001

Table 1. The number of in-situ soil samples in 8 cropland classes in CONUS. Crop classification is from the USDA CDL during the calendar year of sample collection.

	Alfalfa	Corn	Cotton	Grass/Pasture	Sorghum	Soybeans	Winter Wheat	Other	Total
Alabama	0	0	1	0	0	1	0	1	**3**
Arizona	4	2	11	0	0	0	0	9	**26**
Arkansas	0	131	26	1	0	37	0	35	**230**
California	4	1	2	2	0	0	2	49	**60**
Colorado	64	28	0	2	5	0	10	51	**160**
Connecticut	0	0	0	0	0	0	0	1	**1**
Delaware	0	4	0	0	0	5	0	1	**10**
Florida	0	0	0	3	0	0	0	10	**13**
Georgia	0	2	5	0	0	0	0	1	**8**
Idaho	6	4	0	6	0	0	4	10	**30**
Illinois	0	138	0	3	0	106	0	5	**252**
Indiana	0	20	0	2	0	14	1	3	**40**
Iowa	0	108	0	2	0	214	0	1	**325**
Kansas	0	224	0	3	319	145	172	6	**869**
Kentucky	0	10	0	11	0	7	0	4	**32**
Louisiana	0	0	0	1	0	2	0	5	**8**
Maine	0	0	0	0	0	0	0	5	**5**
Maryland	0	5	0	2	0	3	0	8	**18**
Massachusetts	0	0	0	0	0	0	0	1	**1**
Michigan	1	12	0	1	0	4	2	6	**26**
Minnesota	3	221	0	1	0	253	0	7	**485**
Mississippi	0	0	3	1	0	11	0	2	**17**
Missouri	0	10	2	2	0	11	0	2	**27**
Montana	4	0	0	4	0	0	1	19	**28**
Nebraska	0	811	0	7	0	371	3	4	**1196**
Nevada	11	0	0	5	0	0	0	9	**25**
New Hampshire	0	1	0	0	0	0	0	0	**1**
New Jersey	0	2	0	0	0	1	0	2	**5**
New Mexico	51	30	2	3	0	0	2	90	**178**
New York	1	5	0	0	0	0	1	7	**14**
North Carolina	0	1	1	2	0	3	0	2	**9**
North Dakota	1	3	0	7	0	7	0	7	**25**
Ohio	0	6	0	2	0	7	0	1	**16**
Oklahoma	1	102	0	0	7	2	30	26	**168**
Oregon	6	0	0	2	0	0	1	13	**22**
Pennsylvania	3	9	0	0	0	4	0	11	**27**
Rhode Island	0	0	0	0	0	0	0	0	**0**
South Carolina	0	1	1	0	0	0	0	3	**5**
South Dakota	1	133	0	3	0	7	1	8	**153**
Tennessee	0	3	3	1	0	4	0	2	**13**
Texas	1	35	23	4	69	0	42	11	**185**
Utah	14	3	0	0	0	0	0	4	**21**
Vermont	1	3	0	0	0	0	0	2	**6**
Virginia	0	4	4	1	0	2	0	1	**12**
Washington	0	0	0	2	0	0	3	8	**13**

(Continued)

Table 1. (Continued)

	Alfalfa	Corn	Cotton	Grass/Pasture	Sorghum	Soybeans	Winter Wheat	Other	Total
West Virginia	0	2	0	1	0	1	0	0	**4**
Wisconsin	73	174	0	6	0	120	1	72	**446**
Wyoming	3	1	0	3	0	0	0	5	**12**
Total	**253**	**2,249**	**84**	**96**	**400**	**1,342**	**276**	**530**	**5,230**

https://doi.org/10.1371/journal.pone.0327895.t001

Table 2. The number of in-situ soil samples in 8 cropland classes and 10 soil orders in CONUS. Crop classification is from the USDA CDL during the calendar year of sample collection. Soil orders are from the USDA-NCSS soil survey data (SSURGO).

	Alfisols	Andisols	Aridisols	Entisols	Histosols	Inceptisols	Mollisols	Spodosols	Ultisols	Vertisols	Total
Alfalfa	75	0	68	77	0	4	25	3	1	0	**253**
Corn	453	0	34	79	1	30	1,628	2	21	1	**2,249**
Cotton	23	0	5	11	0	7	17	0	12	9	**84**
Grass/Pasture	29	0	7	9	0	7	35	2	5	2	**96**
Sorghum	10	0	0	4	0	0	386	0	0	0	**400**
Soybeans	175	0	0	27	4	28	1,072	0	14	22	**1,342**
Winter Wheat	21	0	3	14	0	7	230	1	0	0	**276**
Other	127	1	110	97	2	40	85	8	23	37	**530**
Total	**913**	**1**	**227**	**318**	**7**	**123**	**3,478**	**16**	**76**	**71**	**5,230**

https://doi.org/10.1371/journal.pone.0327895.t002

remaining samples are among soil orders with < 100 samples each (Table 2). Soil orders were identified using the USDA National Resource Conservation Service (NRCS) Soil Survey Geographic Database (SSURGO).

2.2. Developing the covariate data set

Predictors include features related to climate, topography, and biophysically meaningful variables derived from optical remote sensing that are correlated with soil properties and formation [12,14,28]. Here we identified 90 features using data products related to climate and weather, soil properties and topography, land use, vegetation, management, and other characteristics (S2 Table).

Previous work demonstrates that combinations of long-term climate variables, topography and optical remote sensing are predictive of organic carbon content in soils [12,14,28]. Because our analysis addresses agricultural land in particular and was designed to evaluate whether SOC content in agricultural soils can be predicted using DSM methods, we focused on the selection of covariate features known to be associated with SOC content through biological, geological, or farm-practice relationships.

Following McBratney et al. [12], we describe covariate data using the concept of resolution. Resolution is equivalent to pixel size, where native resolution refers to the source data product before reprojection or sampling, and fundamental resolution is the scale at which the feature varies in the real world. For example, the mean annual temperature from WorldClim may have a fundamental resolution of tens of kilometers, even though it is gridded at 30 arcseconds (about 1 km). Although we resampled all covariate features to 10 m resolution using a cubic spline or nearest-neighbor procedure, resampling does not change fundamental resolution. Differences in fundamental resolution among covariate features are characteristics of the environment that are consistent with the state-factor framework extended by McBratney [12] in the context of DSM. Other data-driven approaches to carbon quantification also use features at varying fundamental and

native resolution [e.g., 13,24]. Hierarchical variation in fundamental resolution among features is a well-defined characteristic of data-driven analysis of environmental properties [e.g., 29,30].

We extracted values from covariate data sets at the location of each sample point. The covariate data set includes upper and lower points of soil sample measurement as quantitative predictors [22]. This allows the algorithm to use measurements collected at different depths in the soil profile to produce standardized output over the 0–30 cm depth range. For the features described below, we assign all features to SOC measurements acquired in 2020 and 2021. Data collected under the USDA RaCA program in 2010 and 2011 precede the availability of Sentinel-1 synthetic aperture radar (SAR), Sentinel-2 bottom of atmosphere surface reflectance, and Soil Moisture Active Passive (SMAP) data products. For samples acquired prior to 2020, these covariates are treated as missing values.

Missing values occur when no observation is possible or due to spatial or temporal gaps in the data record. This can occur when cloud cover results in data loss [e.g., 31]. For example, consider the following simplified feature set x = [1, 3, NA, 5]. This feature set represents four covariate features that could be associated with a single response. The first, second and fourth features have numerical values, but the third feature is missing, indicated by NA. One solution to missingness is to eliminate records that contain at least one missing feature value, called complete-case analysis. This is not ideal, because many records contain at least one missing value, resulting in large data losses. An alternative approach to dealing with missing data imputes missing observations using a predictive model and then analyzes the combined set of complete cases and imputed values using the same methods that are applied in a complete-case analysis [32]. Here we use a third approach developed in the context of gradient-boosted regression trees [33]. The approach identifies the default direction of the split in the decision tree for cases where the required feature is missing using non-missing records for that feature, and then selects the default direction [33].

2.2.1. Long-term physical-climate proxies. Physical-climate proxies include the WorldClim bioclimatic variables [34], which contain summaries of mean annual temperature and precipitation, temperature and precipitation seasonality and extremes at 30 arcsecond resolution. These variables represent conditions in the 1970–2000 time interval and therefore define typical climate characteristics of a given location, not the weather on any specific day. Exploratory data analysis indicated that some WorldClim variables introduced spatial mapping artifacts into model predictions. We therefore included three WorldClim bioclimatic variables as covariates: BIO1 (mean annual temperature), BIO16 (precipitation of the wettest quarter), and BIO17 (precipitation of the driest quarter).

2.2.2. Short-term physical climate and weather proxies. We generated summaries of gridded daily records from the National Centers for Environmental Prediction Climate Forecast System version 1 [NCEP CFS; 35,36]. The NCEP CFS data product is a 6 hourly summary of weather-related variables at 0.3-degree resolution (about 30 km). The variables are precipitation, soil moisture, water runoff, minimum temperature, maximum temperature, mean temperature, sensible heat net flux, downward shortwave radiation net flux, potential evapotranspiration, and transpiration. We aggregated the 6-hour data product within days using the arithmetic mean for soil moisture, mean temperature, potential evapotranspiration, transpiration, minimum temperature, maximum temperature, sensible heat net flux and downward shortwave radiation net flux, and the sum for precipitation and water runoff. Six-month and three-year summaries were generated for each variable using the arithmetic mean for the corresponding time interval prior to the target date. The target date is the date of collection of an in-situ soil sample or the date for which a prediction is desired. The NCEP data record is not available prior to April 2011. We treat all NCEP summaries as missing observations for records associated with sample dates that precede the NCEP data record (604 or 605 records, depending on the NCEP variable).

Although the NCEP summaries provide gridded temperature data at 0.3-degree resolution, variation at finer resolution influences biogeochemical processes and SOC. We therefore supplemented NCEP summaries with the 1 km MODIS land surface temperature and emissivity data product [MOD11A2 V006; 37]. MOD11A2 is an 8-day composite with 1 km native resolution. For each 1 km pixel in the MOD11A2 V006 data set, we computed the arithmetic mean in the six-month interval prior to the target date.

2.2.3. Topographic and edaphic variables. We derived surface elevation from the United States Geological Survey (USGS) 3D Elevation Program (3DEP) 10 m digital elevation model. We derived six-month and three-year soil-moisture summaries from the SMAP L3 daily 9 km soil moisture data product [38]. Summaries were the arithmetic mean within six-month and three-year windows prior to the target date. The covariate data set contained estimates of soil properties from the SoilGrids version 2.0 global 250 m data product [13]. We used gridded clay, sand and silt content as covariates.

2.2.4. Synthetic aperture radar. We used Sentinel-1 measurements derived from Level-1 Ground Range Detected (GRD) data products [39]. Sentinel-1 is a C-band radar that acquires images in all weather conditions at a variety of spatial resolutions. Here we used images acquired with dual polarizations (VH and VV) in the Interferometric Wideswath (IW) mode from ascending orbits. For each 20 m pixel in the Sentinel-1A data set, we computed the median within a six-month window prior to the target date.

2.2.5. Sentinel-2 remote sensing. We used remote sensing summaries from Sentinel-2 bottom of atmosphere surface reflectance data products [40]. For a given location there is a set of candidate Sentinel-2 observations within a time interval that could be used for training and prediction. We refer to the problem of selecting and summarizing this set of observations as image reduction, and we refer to the derived summaries as features. Engineering an image reducer to generate features requires selecting three criteria: the time interval over which the reducer will be applied, filtering that determines whether a candidate observation is included in the summary and choosing the reducing function. For example, Hansen et al. [41] applied cloud shadow and cloud cover filtering criteria to time series Landsat observations over the 2000–2012 time interval to generate a global map of forest cover change. Dvorakova et al. [42] filtered time series Sentinel-2 surface reflectance using the normalized burn ratio (NBR2) and normalized difference vegetation index (NDVI) to generate a bare soil composite. For data-driven prediction, where a model from training data is applied in the real world, the reducer must produce summaries that are spatially and temporally consistent with minimal spatial mapping artifacts. Spatial mapping artifacts can occur when there are spatially structured errors in covariate features that propagate through model predictions.

We used individual bands and derived indices from Sentinel-2 level 2 surface reflectance data products to derive features from optical remote sensing using two image reducers. These two image reducers allow the ATLAS-SOC algorithm to use covariates that represent different intervals of time. The first of these, which we call the simple reducer, computes the arithmetic median over all candidate observations for every individual location within the six-month window prior to the target date. The simple reducer thus considers the relationship between image features and SOC immediately prior to the SOC measurement. For example, if an in-situ soil sample was collected on April 9, 2021, the simple reducer would use October 9, 2020 – April 9, 2021. If a prediction is desired using a calendar date of October 16, 2021, the simple reducer would use April 16, 2021 – October 16, 2021. The second version, which we call the time-series summary reducer, computes the arithmetic median within sequential, three-month time windows over the previous 24 months. The time-series summary reducer generates statistical summaries of the land surface that are related to agricultural practices over the previous 24 months, including planting and harvest time, cover cropping, and tillage status.

We selected indices that track green and senescent vegetation cover, tillage status, and the presence of water on the land surface. Below we define each index. Band definitions for Sentinel-2 are blue (B2; 460–530 nm), green (B3; 540–0.580 nm), red (B4; 650–680 nm), near-infrared (B8; 780–890 nm), SWIR1 (B11; 1,570–1,660 nm), and SWIR2 (B12; 2,110–2,290 nm).

NDVI is correlated with cover of green vegetation and therefore distinguishes vegetated from bare-soil conditions [43]. NDVI is computed using the simple reducer and time-series summary reducer according to,

$$NDVI = \frac{NIR - red}{NIR + red} \tag{1}$$

The soil adjusted vegetation index (SAVI) is a modification of NDVI that corrects the index in the presence of minimal vegetation cover [44] SAVI is positively associated with vegetation cover. We computed SAVI using the simple reducer according to,

$$SAVI = 1.5 \times \frac{NIR - red}{NIR + red + 0.5} \tag{2}$$

Marsett et al. [45] developed the soil adjusted total vegetation index (SATVI). The index is positively associated with green and dry vegetation. SATVI was calculated using the simple reducer according to,

$$SATVI = 1.5 \times \frac{SWIR1 - red}{SWIR1 + red + 0.5} - \frac{SWIR2}{2} \tag{3}$$

We used the BSI to identify pixels with minimal vegetation cover and exposed soil. BSI has been used to develop bare-soil composites of agricultural land using blue, red, and shortwave-infrared reflectance [42,46]. BSI is negatively associated with vegetation cover. We computed BSI using the simple reducer and time-series summary reducer according to,

$$BSI = \frac{(red + SWIR1) - (NIR + blue)}{(red + SWIR1) + (NIR + blue)} \tag{4}$$

Dvorakova et al. [42] used the normalized burn ratio (NBR2) index to generate bare soil composites for prediction of organic carbon content using Sentinel-2. Larger values of NBR2 are typically associated with moist soils or crop residues [42]. The NBR2 index was calculated using the simple reducer according to,

$$NBR2 = \frac{NIR - SWIR2}{NIR + SWIR2} \tag{5}$$

The normalized difference tillage index (NDTI) takes advantage of changes in shortwave infrared reflectance in water absorption regions to detect surface roughness associated with tillage status [47]. Larger values of the index are more likely to be associated with conventional tillage status. We computed NDTI using the simple reducer according to,

$$NDTI = \frac{SWIR1 - SWIR2}{SWIR1 + SWIR2} \tag{6}$$

The brightness index (BI) is associated with total brightness. Under bare soil conditions, brighter soils have less organic matter content than darker soils do [47]. The BI was calculated using the simple reducer,

$$BI = \sqrt{\frac{red^2 + green^2}{2}} \tag{7}$$

The land surface water index (LSWI) is positively correlated with the total content of liquid water in vegetation and soil [48]. It was calculated using the simple reducer according to,

$$LSWI = \frac{NIR - SWIR1}{NIR + SWIR1} \tag{8}$$

We computed tasseled cap brightness, greenness and wetness using the Sentinel-2 coefficients [49]. Tasseled cap brightness was calculated using the simple reducer according to,

$$brightness = 0.3510 \times blue + 0.3813 \times green + 0.3437 \times red + 0.7196 \times NIR + 0.2396 \times SWIR1 + 0.1949 \times SWIR2 \quad (9)$$

Tasseled cap greenness was calculated using the simple reducer according to,

$$greenness = -0.3599 \times blue - 0.3813 \times green - 0.3437 \times red + 0.7196 \times NIR + 0.2396 \times SWIR1 + 0.2856 \times SWIR2 \quad (10)$$

Tasseled cap wetness was calculated using the simple reducer according to,

$$wetness = 0.2578 \times blue + 0.2305 \times green + 0.0883 \times red + 0.1071 \times NIR - 0.7611 \times SWIR1 - 0.5308 \times SWIR2 \quad (11)$$

2.3. Machine learning algorithm

The algorithm is a weighted gradient-boosted regression tree [XGBoost; 33]. XGBoost was selected because tree-based regression consistently performs well on a wide-range of prediction tasks using tabular data, including SOC mapping [11,14,50]. Off-the-shelf implementations are efficient to train, able to handle missing values, and produce interpretable feature-importance metrics. The XGBoost algorithm works by assembling an ensemble of weak learners corrected iteratively to reduce bias and variance. Every weak learner is a single regression tree developed using recursive binary splitting and a user-defined objective function. Here the objective function was the mean absolute error,

$$MAE = \frac{1}{n}\sum_{i=1}^{n} abs\,(y_i - \hat{y}_i) \quad (12)$$

where n is the number of training samples in the algorithm, y_i is measured SOC as a percentage by mass for sample i, and $\hat{y}_i$ is the predicted value of SOC as a percentage by mass for sample i.

The model was fitted to all 5,230 in-situ soil samples and their associated covariates. The 4,260 data records collected in CONUS row-crop agriculture were weighted equally with a value of 1. For the remaining 970 measurements collected under the USDA RaCA program, we explored weights of 0–1 in increments of 0.05 and examined cross-validated model performance under each weighting scenario. The purpose of down weighting RaCA samples was to allow previously collected data to contribute broad spatial coverage while maintaining the importance of recently collected data in row-crop agriculture and during the period of Sentinel-2. We evaluated the ability of the model to predict SOC associated with individual in-situ soil samples and aggregated field means. Because the model was trained using physical samples, the physical-sample performance is simply the geographically-dependent cross-validated regression described below. To evaluate the performance of the model at the field level, we computed the mean SOC content using in-situ soil samples in fields that contained ≥ 5 samples (n = 165 fields and 3,285 in-situ soil samples) and the associated mean of predicted values at the same sample locations in each field.

2.3.1. Cross-validation. When models are trained using a nonrandom sample of the prediction domain, cross-validation should use geographically dependent partitions [16,51–54]. The idea is to understand how the algorithm will perform when transferred to a geographic subset of the prediction domain that is different from training data. We used individual units of land management (agricultural fields) as cross-validation folds. Most samples in our data collected under the USDA RaCA program are spatially isolated (mean nearest-neighbor distance = 30.6 km, range = 0.3–198.1 km). In contrast, new measurements collected in 2020 and 2021 were acquired on individual fields with multiple in-situ soil samples (mean = 10.4 samples per field, range = 3–133).

We evaluated model performance using the intercept and slope of the cross-validated linear regression, the root mean squared error (RMSE), and the coefficient of determination (R^2). The RMSE was computed using,

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

where y_i indicates a measurement of SOC using physical soil sampling, and $\hat{y}_i$ is the predicted value derived from machine-learning. The coefficient of determination was computed according to,

$$R^2 = \frac{SS_{total} - SS_{residual}}{SS_{total}} \tag{14}$$

where SS_{total} and $SS_{residual}$ are the total and residual sums of squares from the cross-validated regression, respectively. Because SOC as a percentage by mass is bounded at 0 and 100, we assumed a truncated Gaussian error term when computing the cross-validated regression. We quantified whether model performance varied with depth of the soil measurement using indicator-variables regression [55,56]. The regression model was,

$$y_i = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2 + \in_i \tag{15}$$

where X_1 is the predicted value derived from machine-learning in units of percentage-by-mass and X_2 is binary indicator variable that dictates whether y_i was measured over the 0–5 ($X_2 = 0$) or 0–30 ($X_2 = 1$) cm depth range. The parameter $\in_i$ is a Gaussian error truncated over the range 0–100. When the value of $X_2 = 0$, the equation collapses to the ordinary linear regression on samples over the 0–5 cm depth range. When $X_2 = 1$, b_2 and b_3 test the null hypothesis that the cross-validated intercept and slope are equivalent between 0–5 and 0–30 cm depth ranges, respectively. The parameters b_2 and b_3 are the change in the intercept and slope for the case where $X_2 = 1$.

2.3.2. Hyperparameter tuning. XGBoost supports numerous hyperparameters that can be defined prior to fitting the model. These hyperparameters can be assigned default values or optimized in a process called hyperparameter tuning. Here we tuned five hyperparameters: the number of trees in the ensemble, the learning rate, the proportion of columns used to identify the split at each node, the proportion of samples used by each tree, and the maximum depth of each tree. We used Bayesian optimization to identify hyperparameter values [57]. We selected values that resulted in the smallest sample-level mean absolute error using the cross-validation described above.

2.4. Evaluation of publicly available SOC data products

Our objective is to produce a DSM algorithm that can predict SOC as a percentage by mass in North American agricultural land. However, we also need to determine whether existing data products accurately estimate field-level SOC as a percentage by mass. Although existing publicly available maps of SOC that include CONUS agricultural land were not designed for agricultural land in particular, they might be useful to VCM programs for establishing baselines in carbon removal projects, or for initialization of biogeochemical simulations of SOC fluxes. However, the accuracy of these data products in agricultural land in particular has not been assessed. These data products are generally developed using legacy soil sample data. Validating them against recent soil samples provides a useful assessment of their real-world applicability in current agricultural conditions. We compared four publicly available SOC data products to independent measurements of SOC as a percentage by mass.

2.4.1. SoilGrids version 2.0. SoilGrids version 2.0 is a global 250 m data product that contains estimates of SOC as a percentage by mass at multiple soil depths [13]. The SoilGrids 2.0 data product was developed using a Quantile

Random Forest algorithm trained on about 240,000 observations worldwide and > 400 environmental covariates. Covariates included climate and temperature, geology and landcover, vegetation indices and raw measurements from the Landsat and MODIS sensors. The SoilGrids 2.0 data set contains predictions of SOC and other properties within six depth intervals. We used predictions from 0–5, 5–15 and 15–30 cm beneath the soil surface to derive an estimate of SOC within the top 30 cm of soil. We did this by computing a weighted average, where the weight applied to each depth interval was proportional to the range of that interval over 0–30 cm. For example, the weights applied to 0–5, 5–15 and 15–30 cm were 1/6, 1/3, and 1/2.

2.4.2. Soil property maps at 100 m resolution. Ramcharan et al. [58] generated 100 m resolution soil property and class maps using an approach similar to Poggio et al. [13]. The approach differs by integrating multiple soil data sets within a common analytical framework and by predicting SOC at specific depths beneath the soil surface, rather than over depth ranges. Soil data sets include the National Cooperative Soil Survey Characterization Database, the National Soil Information System, and measurements from the USDA RaCA program. We used linear interpolation among SOC predictions at 0 cm (surface), 5 cm, 15 cm and 30 cm to produce an SOC profile with 1 cm increments over the 0–30 cm depth range. We then computed the unweighted arithmetic average over all 30 increments to produce a single estimate for the top 30 cm of soil in every 100 m pixel.

2.4.3. POLARIS. Chaney et al. [59] remapped the Soil Survey Geographic Database (SSURGO) at 30 m resolution using machine learning and environmental covariates (POLARIS). The POLARIS data product does not contain estimates of SOC as a percentage by mass. We converted POLARIS estimates of organic matter within three depth ranges to SOC as a percentage by mass using the van Bemmelen factor of 0.58. Some recent work has challenged whether a single value can be used to convert organic matter into SOC [e.g., 60,61]. We considered values other than 0.58 to determine whether alignment between POLARIS data product and in-situ soil samples was sensitive to the conversion factor used (S1 Appendix). We used predictions from 0–5, 5–15 and 15–30 cm beneath the soil surface to derive an estimate of SOC within the top 30 cm of soil using the same weighted average procedure applied to the SoilGrids version 2.0 data product described in 2.4.1.

2.4.4. Harmonized world soil database version 2.0. The Harmonized World Soil Database version 2.0 (HWSD2) is a 2023 compilation of national soil surveys and legacy maps distributed as a global 30-arc-second raster [about 1 km resolution, 62]. Each pixel (called a soil mapping unit in HWSD2) is linked to attribute tables that report soil properties for seven depth layers. Because a given mapping unit can contain > 1 soil unit, each of which has distinct properties, we first calculated area-weighted SOC for each mapping unit. The area weights for each mapping unit were based on the SHARE percentage, which is the area of each soil mapping unit contained within each soil unit. Negative values in the SOC variable were ignored in subsequent analysis. To estimate mean SOC as a percentage by mass in the top 30 cm, we combined the 0–20 cm layer (D1) with half of the 20–40 cm layer (D2) using weights of 2/3 and 1/3, respectively.

2.4.4. Extracting the field mean from publicly available SOC data products. We calculated the field mean SOC over the 0–30 cm depth range using each publicly available SOC data product within each of 165 fields with ≥ 5 in-situ soil samples. We did this by re-projecting field polygons from the World Geodetic System 1984 coordinate reference system (EPSG:4326) to the native projection associated with each data product. For each of the four data sets, we extracted pixels that intersected the field geometry. Because some pixels were on boundary edges, we calculated the proportion, w_i, of each pixel, x_i, that was contained within field *j,* and used these proportions to compute the weighted mean SOC_j as a percentage by mass according to,

$$SOC_j = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{16}$$

3. Results

SOC can be predicted with a high degree of precision and accuracy in North American agricultural land using machine learning. Analysis of the indicator-variables regression showed that model performance was independent of the measured depth range. At the sample level, parameter estimates were $b_0 = –0.199$ ($P = 0.959$), $b_1 = 1.109$ ($P = 0.028$), $b_2 = –0.068$, ($P = 0.978$), and $b_3 = 0.010$ ($P = 0.994$). We therefore dropped the indicator variable and fit the linear relationship between measured and predicted values. The intercept and slope of the cross-validated regression without indicator variables were $b_0 = –0.249$ ($P < 0.001$) $b_1 = 1.130$ ($P < 0.001$), respectively. The coefficient of determination was $R^2 = 0.487$ and RMSE was 0.316 in units of percentage by mass (Fig 2).

Aggregation to the field level demonstrated that mean predictions are accurate (Fig 3). We evaluated field-level predictions by computing the arithmetic mean in fields with ≥ 5 in-situ soil samples (n = 165 fields, 3,285 in-situ soil samples). The observed value (response variable) was the mean from sampled soil cores in the field, and the predicted value (independent variable) was the mean from pixel-level predictions at the same locations in each field. Predicted field means were always from an instance of the geographically dependent cross validation that did not contain any in-situ soil samples from the observed field. The intercept and slope of the cross-validated relationship were $b_0 = –0.179$ ($P = 0.033$) $b_1 = 1.095$ ($P < 0.001$). The coefficient of determination was $R^2 = 0.811$, and the RMSE was 0.041 in units of percentage by mass.

Comparison of field measurements of SOC to four publicly available SOC data products showed poor ability of publicly available data products to reproduce measured values (Table 3). Ranked by RMSE, the best-performing publicly available data product was the SoilGrids version 2.0 dataset, followed by HWSD2, the 100 m resolution soil property and class map developed by Ramcharan et al. [58] and POLARIS. All of these data products exhibited substantial bias in comparison to measured values, and none of them out-performed ATLAS-SOC (Table 3; Fig 3). The SoilGrids version 2.0 data product, HWSD2, and the 100 m soil property and class map underestimated field measurements at values < 2.17%, 1.79%, and 1.61%, respectively, and overestimated field measurements at other values. The POLARIS dataset underestimated

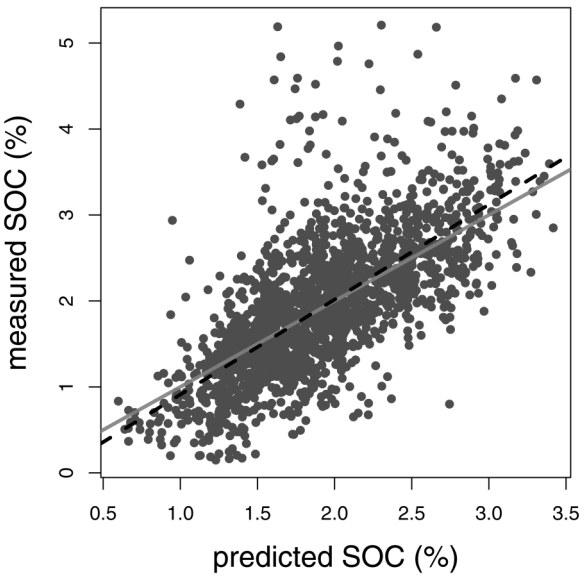

Fig 2. Cross-validated relationship between measured in-situ SOC as a percentage by mass and predicted values from a data-driven model in agricultural soils. Grey line is the 1:1 relationship, and the dashed line is the best-fit linear regression. This plot shows a random sample of 2,000 points for clarity, but the full data set contains 5,230 observations.

https://doi.org/10.1371/journal.pone.0327895.g002

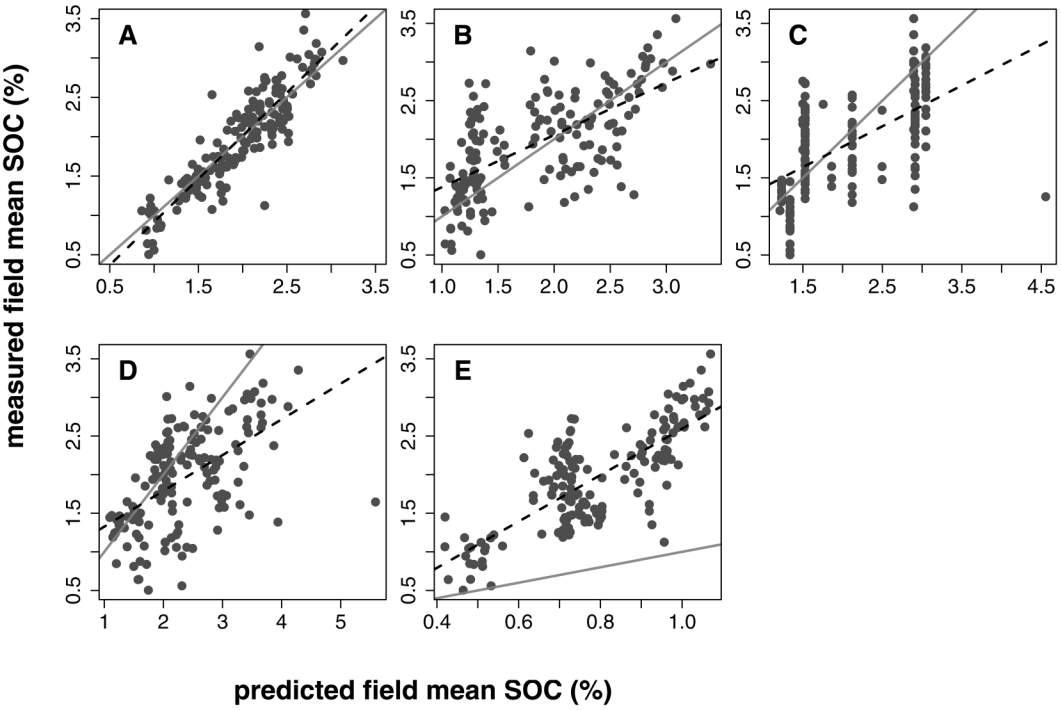

Fig 3. Relationships between field mean SOC and predicted values. Points are 165 fields with ≥ 5 samples (total number of samples = 3,285). (A) ATLAS-SOC (this study). (B) The 100 m soil properties and class map from Ramcharan et al. (2018). (C) the Harmonized World Soil Database version 2.0. (D) SoilGrids version 2.0. (E) POLARIS. Grey line is the one-to-one relationship, and the dashed line is the best-fit linear regression.

https://doi.org/10.1371/journal.pone.0327895.g003

Table 3. Validation of mean SOC as a percentage by mass in 165 agricultural fields measured using 3,285 in-situ soil samples. b_0 **and** b_1 **are the intercept and slope of the linear regression relating mean SOC from in-situ soil samples (response) to the predicted value from four publicly available data products and ATLAS-SOC.** R^2 **is the coefficient of determination and RMSE is the root mean squared error.**

Data product	b_0	b_1	R^2	RMSE
ATLAS-SOC (this study)	−0.179	1.095	0.811	0.041
SoilGrids version 2.0	0.705	0.675	0.412	0.148
Harmonized World Soil Database version 2	0.845	0.529	0.346	0.201
Ramcharan et al. (2018)	0.863	0.463	0.317	0.306
POLARIS	−0.409	3.004	0.585	0.831

https://doi.org/10.1371/journal.pone.0327895.t003

the field mean SOC as a percentage by mass in all 165 fields, although the degree of correspondence improved when a different conversion factor was used (Fig 3; S1 Appendix).

We computed feature importance scores for every covariate in the training set. The top 20 features accounted for 68.6% of the overall importance in the data and included 7 variables from optical remote sensing, 8 weather-related proxies, 2 topographic and edaphic variables, and 2 climate variables. The most important feature type was Sentinel-2 optical remote sensing time series summaries, which collectively accounted for 28.2% of the importance in the data, followed by weather-related temperature (26.0%) and surface hydrology (11.6%; Table 4). Among feature classes, weather-related variables contributed 39.3% of the importance in the data, most of which was due to temperature (26.0%), followed by meteorology (9.5%) and precipitation (3.8%). Optical remote sensing summaries and their derivatives collectively contributed 36.3% of the importance in the data. Time series optical remote sensing summaries and their derivatives (28.2%)

Table 4. Relative importance for 90 features used to predict organic carbon content in CONUS agricultural soils. Importance is from a gradient-boosted regression fitted to 5,230 measurements of SOC linked to covariate features and summed within feature types. For example, there are 32 time-series features from Sentinel-2 optical remote sensing whose aggregate importance is 28.2%. Totals are marginal sums within feature classes.

Feature class	Feature type	N	Importance (%)
optical remote sensing	time series summary	32	28.2
	simple summary	18	8.1
Total			**36.3**
weather-related	meteorology	8	9.5
	precipitation	2	3.8
	temperature	8	26.0
Total			**39.3**
topographic and edaphic variables	surface elevation	1	0.3
	soil (clay, silt, sand)	3	2.5
	surface hydrology	12	11.6
	depth of measurement	1	4.2
Total			**18.6**
long-term physical climate proxies	temperature	1	2.3
	precipitation	2	3.1
Total			**5.4**
Sentinel-1 SAR	synthetic aperture radar (SAR)	2	0.5

https://doi.org/10.1371/journal.pone.0327895.t004

were about 3.5 times more important than summaries generated using the simple reducer (8.1%). Topographic and edaphic features were responsible for 18.6% of overall variable importance. Features related to surface hydrology were the most important topographic and edaphic features (11.6%), followed by the depth of the soil measurement (4.2%), silt, sand, and clay content (2.5%) and surface elevation (0.3%). Long-term physical climate proxies accounted for 5.4% of the importance in the data. Two measurements from Sentinel-1 SAR were relatively unimportant, contributing 0.5% to the overall importance in the data set (Table 4).

4. Discussion

4.1. Soil organic carbon measurement in support of the voluntary carbon market

VCM programs in agriculture require accurate measurement techniques that can be deployed efficiently at scale. Our analysis shows that data-driven, DSM methods can accurately predict mean SOC content in CONUS agricultural land. This analysis is based on 5,230 in-situ soil samples collected in actively cultivated, conventional row-crop agriculture, natural prairies and grasslands within 47 US states. Using a gradient-boosted regression tree and geographically dependent cross-validation, we developed a model driven by combinations of long-term climate proxies, topography and optical remote sensing variables that is both precise and accurate, overcoming challenges to widespread quantification that is necessary in support of VCM programs.

Models to predict SOC are trained using measurements from individual soil cores matched to remote sensing pixels, but VCM programs require estimates of SOC within regions. These regions could represent units of individual land management (agricultural fields) or aggregations of fields that total hundreds to thousands of hectares. Methodologies for carbon-credit generation penalize uncertainty in estimates of net GHG reductions at the scale of aggregation. Demonstrating that predictions are precise and accurate at the scale of aggregation is necessary to underpin rigorous carbon offsets in agricultural soil.

Our analysis, using a sample of 165 fields with ≥ 5 samples per field (n = 3,285 samples in total), shows that estimates of the field mean are precise and accurate when validated against fields that were excluded from model calibration (Table 3). However, comparison of field measurements of SOC to four publicly available SOC data products demonstrated poor ability of other products to reproduce measured values. Validation statistics for these data products were worse than that reported by the original authors in cases where summaries were available [13,58,59]. This is because these data products were developed to perform well throughout large regions in general, not agricultural land in particular. These products were also developed using legacy soil sample data, and it is likely that agricultural and soil conditions have changed. The data archive used to train ATLAS-SOC for this analysis includes 5,230 in-situ soil samples that were collected in 47 US states and exclusively in agricultural conditions, most of which were collected recently (S1 Table). The limited ability of publicly available data products to generalize to agricultural land in our study is an example of Simpson's paradox, a problem in statistics where a relationship changes, or even reverses, in the presence of new variables [63]. In the context of the VCM, Simpson's paradox underscores the importance of model validation within the domain of application. It should not be assumed, for example, that a model with generally acceptable performance in North American soils will perform well in any given subset of North America that was poorly represented by calibration data.

4.2. Data-driven quantification of SOC

Our analysis underscores the value of broadly distributed training data. Inclusion of geographically distributed in-situ soil samples collected under the USDA RaCA program decreased bias in the model in comparison to a version where broadly distributed data were absent (S3 Table). When the model was trained exclusively on actively cultivated, conventional row-crop agriculture in the states of Arkansas, Colorado, Illinois, Iowa, Kansas, Minnesota, Nebraska, New Mexico, Oklahoma, South Dakota, Texas, and Wisconsin, it exhibited moderate bias, as indicated by the intercept and slope of cross-validated predictions ($b_0 = 0.179$, $b_1 = 0.895$; S3 Table). But when broadly distributed in-situ soil samples were added to the training set, bias reduced (S3 Table). Geographically distributed in-situ soil samples reduced bias in the model even though they accounted for < 20% of the training set, and even though optical remote sensing and some weather variables were treated as missing values in broadly distributed data. This finding suggests that the value of geographically distributed in-situ soil samples in our analysis was primarily their contribution to resolving the relationship between SOC content, long-term physical climate proxies, and topographic and edaphic variables. This indicates that when combined with contemporary in-situ soil data, legacy soil samples can play an important role in DSM efforts.

Our findings strengthen the argument that optical remote sensing can be an important and independent source of information about SOC content in soils [10,64]. Other data-driven approaches to SOC quantification indicate that combinations of physical climate and topography are the strongest predictors of SOC content in soils of North America and China [14,65,66]. Features from optical remote sensing have been less important. The ranking of optical remote sensing in aggregate in two previous studies placed them at rank 3 or 4 out of 5 feature sets [14,66]. However, in our analysis weather-related variables and optical remote sensing summaries contributed 39.3% and 36.3% of the overall importance in the data, and 8 of the top 20 individual features were derived from optical remote sensing summaries (Table 4).

Time series summaries from Sentinel-2 were the strongest optical remote sensing predictors of SOC content (Table 4). In particular, median NDVI within three-month windows during the previous two years represented 5 of the top 20 most important features. Overall, time series summaries represented 28.2% of the importance in the data, a number 3.5 times larger than static summaries that do not consider changes in land cover over time. The sensitivity of remote sensing summaries from specific times of year suggests that land use and land cover changes are responsible for the strength of these summaries in comparison to other descriptors. One hypothesis for the importance of optical remote sensing variables in our analysis in comparison to previous work is that we increased the number of remote sensing features in comparison to other variables and used time-series feature engineering to generate predictors that are likely to be correlated with SOC

content. Point-in-time optical remote sensing measurements may be less strongly correlated with SOC content than summaries that target specific land management actions.

The ranking of optical remote sensing variables in a DSM framework to quantify SOC content is important, because optical remote sensing variables are a source of information about land management, including cover cropping and reduced or no-tillage practices that are expected to impact SOC sequestration [67]. A data-driven framework that is sensitive to climate, weather, and topographic and edaphic conditions, but insensitive to variables related to land management will be able to recapitulate known physical gradients, but may not detect changes related to farm practice. More work is needed to determine whether the inclusion of practice data or biogeochemical variables in a covariate feature set – such as cover-cropping and tillage status or variables related to microbial activity – improves performance, whether data-driven calculations are sensitive to those variables, and whether direct inclusion of these variables is redundant with optical remote sensing summaries.

Interpretation of feature importance can be influenced by the number and type of variables within a covariate set and how the feature summary is generated. Our analysis contained 50 optical remote sensing features and 18 weather-related variables, 17 topographic and edaphic variables, 3 long-term physical climate proxies, and 2 measurements from Sentinel-1 SAR. This means that the algorithm is being presented with more opportunities to model the relationship using optical remote sensing than other variable types, and that overall variable importance within categories can be influenced by imbalance in the feature set. Imbalance in a covariate feature set can be desirable when an objective is to ensure sensitivity to particular kinds of features that are overrepresented.

4.3. Alternatives to data-driven quantification

The approach developed here was designed to overcome the challenges of efficient quantification and scale in the VCM within agricultural soils. It can be contrasted with two alternative methods, physical soil sampling and biogeochemical simulation. Physical soil sampling produces locally accurate measurements of SOC, but these measurements are costly to acquire and may demand unreasonably large sample sizes to generate precise area-based summaries required by the VCM [4,6].

Biogeochemical simulations stand on decades of research in biogeochemistry and soil science. They use local information about weather, climate, soil characteristics and farm management data to simulate interactions in soil using physics and chemistry. For example, the DeNitrification-DeComposition (DNDC) biogeochemical simulation uses climate, soil and cropping variables [68–70]. Some of these variables must be supplied by the grower (e.g., the fraction of crop residue left as stubble in the field after harvest) and other variables are difficult to measure at scale (e.g., the background NH_3 concentration in the air). Data-driven methods from the field of DSM can help to reduce barriers to large-area estimation by allowing machine learning to model relationships between input variables and SOC content and acquiring input variables from publicly available sources. Advances by data-driven methods in comparison to traditional analyses based on physical knowledge have been achieved in other domains, including weather forecasting [71], skin cancer detection [72] and protein folding [73].

5. Conclusions

Data-driven calculation of SOC in agricultural soil is achievable at scale using DSM methods that are sensitive to land use and land cover change. The ATLAS-SOC algorithm developed here can predict SOC as a percentage by mass with errors that are small when evaluated at the physical sample and field mean levels. This algorithm was trained using a data set that explicitly represented row-crop agricultural land in CONUS. Field-level validation of four publicly available data products in CONUS demonstrated poor ability of publicly available SOC data products to reproduce independently measured values from in-situ soil samples.

By engineering biophysically meaningful time series features using optical remote sensing, we were able to increase sensitivity to optical remote sensing features in comparison to previous studies. Additional work is needed to determine whether optical remote sensing summaries or other variables are sensitive to land-management changes that drive SOC sequestration and baseline criteria that underpin rigorous carbon offsets.

Supporting information

S1 Appendix. Assessment of alternative values of the conversion factor used to express POLARIS SOM in units of SOC.
(DOCX)

S1 Fig. MAE between field mean SOC and converted POLARIS SOM. Purple line is the van Bemmelen factor, blue line is a value of 1 (no conversion), and yellow line is the value that minimizes the MAE (1.47).
(PDF)

S2 Fig. Relationships between field mean SOC and predicted values. Points are 165 fields with ≥ 5 samples (total number of samples = 3,285). (A) ATLAS-SOC (this study). (B) The 100 m soil properties and class map from Ramcharan et al. (2018). (C) the Harmonized World Soil Database version 2.0. (D) SoilGrids version 2.0. (E) POLARIS (based on a conversion factor of 1.47). Grey line is the one-to-one relationship, and the dashed line is the best-fit linear regression.
(PDF)

S1 Table. The number of physical soil samples collected in combinations of month and calendar year. 970 samples collected in 2010 and 2011 were acquired under the USDA Rapid Carbon Assessment Program in a wide range of land cover types, including row-crop agriculture, natural prairies and rangeland. Samples collected during 2020 and 2021 are exclusively within actively cultivated, conventional row-crop agriculture in the states of Arkansas, Colorado, Illinois, Iowa, Kansas, Minnesota, Nebraska, New Mexico, Oklahoma, South Dakota, Texas, and Wisconsin.
(DOCX)

S2 Table. 90 covariate features used to predict SOC as a percentage by mass using a gradient-boosted regression tree. All features were resampled from the native resolution to 10 m using a cubic spline. Some features represent band combinations from Sentinel-2A that have different native resolutions. The number reported is the coarsest resolution among all inputs to the given feature.
(DOCX)

S3 Table. Performance of 20 candidate models used to predict SOC as a percentage by mass using physical soil samples. Weight is the per-sample weight applied to 970 previously collected and broadly distributed measurements of SOC. The remaining 4,260 measurements were weighted equally with a value of 1. The parameters $b_0$and b_1 are the intercept and slope from a geographically dependent cross-validation. Inclusion of broadly distributed data improved model performance (cf. the intercept and slope for weight 0.00 with all other weights). The selected model with a weight of 0.10 was retrained prior to prediction.
(DOCX)

Acknowledgments

We thank Skye Wills at the United States Department of Agriculture for providing access to USDA RaCA data and explaining its use. We thank numerous current and former employees of Perennial Climate Inc. for constructive feedback and work that made this research possible.

Author contributions

Conceptualization: James R. Kellner.

Data curation: James R. Kellner, Kirk M. Demuth, Mitchell Donovan.

Formal analysis: James R. Kellner, Y. Katherina Feng, Mage Khim-Young, Julia Maddalena, Rose Rustowicz.

Funding acquisition: James R. Kellner, David Schurman.

Investigation: James R. Kellner.

Methodology: James R. Kellner, Christian Clanton.

Project administration: James R. Kellner, Kirk M. Demuth, Mitchell Donovan, David Schurman.

Resources: James R. Kellner.

Software: James R. Kellner, Y. Katherina Feng, Mage Khim-Young, Julia Maddalena, Rose Rustowicz.

Supervision: James R. Kellner, Kirk M. Demuth, Mitchell Donovan, David Schurman.

Validation: James R. Kellner.

Visualization: James R. Kellner.

Writing – original draft: James R. Kellner.

Writing – review & editing: James R. Kellner.

References

1. Minasny B, Malone BP, McBratney AB, Angers DA, Arrouays D, Chambers A, et al. Soil carbon 4 per mille. Geoderma. 2017;292:59–86. https://doi.org/10.1016/j.geoderma.2017.01.002
2. Buma B, Gordon DR, Kleisner KM, Bartuska A, Bidlack A, DeFries R, et al. Expert review of the science underlying nature-based climate solutions. Nat Clim Chang. 2024;14(4):402–6. https://doi.org/10.1038/s41558-024-01960-0
3. Wills S, Loecke T, Sequeira C, Teachman G, Grunwald S, West LT. Overview of the U.S. Rapid Carbon Assessment Project: Sampling Design, Initial Summary and Uncertainty Estimates. Soil Carbon. Springer International Publishing. 2014. p. 95–104. https://doi.org/10.1007/978-3-319-04084-4_10
4. Bradford MA, Eash L, Polussa A, Jevon FV, Kuebbing SE, Hammac WA, et al. Testing the feasibility of quantifying change in agricultural soil carbon stocks through empirical sampling. Geoderma. 2023;440:116719. https://doi.org/10.1016/j.geoderma.2023.116719
5. Kravchenko AN, Robertson GP. Whole-Profile Soil Carbon Stocks: The Danger of Assuming Too Much from Analyses of Too Little. Soil Science Soc of Amer J. 2011;75(1):235–40. https://doi.org/10.2136/sssaj2010.0076
6. Stanley P, Spertus J, Chiartas J, Stark PB, Bowles T. Valid inferences about soil carbon in heterogeneous landscapes. Geoderma. 2023;430:116323. https://doi.org/10.1016/j.geoderma.2022.116323
7. Ogle SM, Breidt FJ, Easter M, Williams S, Killian K, Paustian K. Scale and uncertainty in modeled soil organic carbon stock changes for US croplands using a process-based model. Global Change Biology. 2010;16(2):810–22. https://doi.org/10.1111/j.1365-2486.2009.01951.x
8. Paustian K, Lehmann J, Ogle S, Reay D, Robertson GP, Smith P. Climate-smart soils. Nature. 2016;532(7597):49–57. https://doi.org/10.1038/nature17174 PMID: 27078564
9. Smith P, Soussana J-F, Angers D, Schipper L, Chenu C, Rasse DP, et al. How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal. Glob Chang Biol. 2020;26(1):219–41. https://doi.org/10.1111/gcb.14815 PMID: 31469216
10. Gomez C, Viscarra Rossel RA, McBratney AB. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. Geoderma. 2008;146(3–4):403–11. https://doi.org/10.1016/j.geoderma.2008.06.011
11. Hengl T, Mendes de Jesus J, Heuvelink GBM, Ruiperez Gonzalez M, Kilibarda M, Blagotić A, et al. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One. 2017;12(2):e0169748. https://doi.org/10.1371/journal.pone.0169748 PMID: 28207752
12. McBratney AB, Mendonça Santos ML, Minasny B. On digital soil mapping. Geoderma. 2003;117(1–2):3–52. https://doi.org/10.1016/s0016-7061(03)00223-4
13. Poggio L, de Sousa LM, Batjes NH, Heuvelink GBM, Kempen B, Ribeiro E, et al. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. SOIL. 2021;7(1):217–40. https://doi.org/10.5194/soil-7-217-2021
14. Sothe C, Gonsamo A, Arabian J, Snider J. Large scale mapping of soil organic carbon concentration with 3D machine learning and satellite observations. Geoderma. 2022;405:115402. https://doi.org/10.1016/j.geoderma.2021.115402
15. Baccini A, Goetz SJ, Walker WS, Laporte NT, Sun M, Sulla-Menashe D, et al. Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. Nature Clim Change. 2012;2(3):182–5. https://doi.org/10.1038/nclimate1354
16. Duncanson L, Kellner JR, Armston J, Dubayah R, Minor DM, Hancock S, et al. Aboveground biomass density models for NASA's Global Ecosystem Dynamics Investigation (GEDI) lidar mission. Remote Sensing of Environment. 2022;270:112845. https://doi.org/10.1016/j.rse.2021.112845
17. Guizar-Coutiño A, Jones JPG, Balmford A, Carmenta R, Coomes DA. A global evaluation of the effectiveness of voluntary REDD+ projects at reducing deforestation and degradation in the moist tropics. Conserv Biol. 2022;36(6):e13970. https://doi.org/10.1111/cobi.13970 PMID: 35713105

18. Saatchi SS, Harris NL, Brown S, Lefsky M, Mitchard ETA, Salas W, et al. Benchmark map of forest carbon stocks in tropical regions across three continents. Proc Natl Acad Sci U S A. 2011;108(24):9899–904. https://doi.org/10.1073/pnas.1019576108 PMID: 21628575
19. Fearnside PM, Lashof DA, Moura-Costa P. Accounting for time in Mitigating Global Warming through land-use change and forestry. Mitigation and Adaptation Strategies for Global Change. 2000;5(3):239–70. https://doi.org/10.1023/a:1009625122628
20. Ruseva T, Marland E, Szymanski C, Hoyle J, Marland G, Kowalczyk T. Additionality and permanence standards in California's Forest Offset Protocol: A review of project and program level implications. J Environ Manage. 2017;198(Pt 1):277–88. https://doi.org/10.1016/j.jenvman.2017.04.082 PMID: 28477569
21. Thamo T, Pannell DJ. Challenges in developing effective policy for soil carbon sequestration: perspectives on additionality, leakage, and permanence. Climate Policy. 2015;16(8):973–92. https://doi.org/10.1080/14693062.2015.1075372
22. Fu P, Clanton C, Demuth KM, Goodman V, Griffith L, Khim-Young M, et al. Accurate Quantification of 0–30 cm Soil Organic Carbon in Croplands over the Continental United States Using Machine Learning. Remote Sensing. 2024;16(12):2217. https://doi.org/10.3390/rs16122217
23. Ma Y, Minasny B, McBratney A, Poggio L, Fajardo M. Predicting soil properties in 3D: Should depth be a covariate?. Geoderma. 2021;383:114794. https://doi.org/10.1016/j.geoderma.2020.114794
24. Ramcharan A, Hengl T, Beaudette D, Wills S. A Soil Bulk Density Pedotransfer Function Based on Machine Learning: A Case Study with the NCSS Soil Characterization Database. Soil Science Soc of Amer J. 2017;81(6):1279–87. https://doi.org/10.2136/sssaj2016.12.0421
25. Heuvelink GBM, Angelini ME, Poggio L, Bai Z, Batjes NH, van den Bosch R, et al. Machine learning in space and time for modelling soil organic carbon change. European J Soil Science. 2020;72(4):1607–23. https://doi.org/10.1111/ejss.12998
26. Ugbemuna Ugbaje S, Karunaratne S, Bishop T, Gregory L, Searle R, Coelli K, et al. Space-time mapping of soil organic carbon stock and its local drivers: Potential for use in carbon accounting. Geoderma. 2024;441:116771. https://doi.org/10.1016/j.geoderma.2023.116771
27. Rabenhorst MC. Determination of Organic and Carbonate Carbon in Calcareous Soils Using Dry Combustion. Soil Science Soc of Amer J. 1988;52(4):965–8. https://doi.org/10.2136/sssaj1988.03615995005200040012x
28. Hobley E, Wilson B, Wilkie A, Gray J, Koen T. Drivers of soil organic carbon storage and vertical distribution in Eastern Australia. Plant Soil. 2015;390(1–2):111–27. https://doi.org/10.1007/s11104-015-2380-1
29. Clark JS, Gelfand AE. Hierarchical modelling for the environmental sciences: statistical methods and applications. Oxford, New York: Oxford University Press. 2006.
30. Raudenbush S, Bryk AS. A Hierarchical Model for Studying School Effects. Sociology of Education. 1986;59(1):1. https://doi.org/10.2307/2112482
31. Whitcraft AK, Vermote EF, Becker-Reshef I, Justice CO. Cloud cover throughout the agricultural growing season: Impacts on passive optical earth observations. Remote Sensing of Environment. 2015;156:438–47. https://doi.org/10.1016/j.rse.2014.10.009
32. Nakagawa S, Freckleton RP. Missing inaction: the dangers of ignoring missing data. Trends in Ecology & Evolution. 2008;23(11):592–6. https://doi.org/10.1016/j.tree.2008.06.014
33. Chen T, Guestrin C. XGBoost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. 785–94. https://doi.org/10.1145/2939672.2939785
34. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol. 2005;25(15):1965–78. https://doi.org/10.1002/joc.1276
35. Saha S, Moorthi S, Wu X, Wang J, Nadiga S, Tripp P. NCEP Climate Forecast System Version 2 (CFSv2) 6-hourly Products. UCAR/NCAR - Research Data Archive. 2011. https://doi.org/10.5065/D61C1TXF
36. Saha S, Moorthi S, Pan H-L, Wu X, Wang J, Nadiga S, et al. The NCEP Climate Forecast System Reanalysis. Bull Amer Meteor Soc. 2010;91(8):1015–58. https://doi.org/10.1175/2010bams3001.1
37. Wan Z, Lewis SL, Hulley G. MOD11A2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V006. NASA EOSDIS Land Processes DAAC. 2015. https://doi.org/10.5067/MODIS/MOD11A2.006
38. Entekhabi D, Njoku EG, O'Neill PE, Kellogg KH, Crow WT, Edelstein WN, et al. The Soil Moisture Active Passive (SMAP) Mission. Proc IEEE. 2010;98(5):704–16. https://doi.org/10.1109/jproc.2010.2043918
39. Filipponi F. Sentinel-1 GRD Preprocessing Workflow. In: 3rd International Electronic Conference on Remote Sensing, 2019. 11. https://doi.org/10.3390/ecrs-3-06201
40. Main-Knorn M, Pflug B, Louis J, Debaecker V, Müller-Wilm U, Gascon F. Sen2Cor for Sentinel-2. In: Image and Signal Processing for Remote Sensing XXIII, 2017. https://doi.org/10.1117/12.2278218
41. Hansen MC, Potapov PV, Moore R, Hancher M, Turubanova SA, Tyukavina A, et al. High-resolution global maps of 21st-century forest cover change. Science. 2013;342(6160):850–3. https://doi.org/10.1126/science.1244693 PMID: 24233722
42. Dvorakova K, Heiden U, van Wesemael B. Sentinel-2 Exposed Soil Composite for Soil Organic Carbon Prediction. Remote Sensing. 2021;13(9):1791. https://doi.org/10.3390/rs13091791
43. Tucker CJ. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sensing of Environment. 1979;8(2):127–50. https://doi.org/10.1016/0034-4257(79)90013-0
44. Huete AR. A soil-adjusted vegetation index (SAVI). Remote Sensing of Environment. 1988;25(3):295–309. https://doi.org/10.1016/0034-4257(88)90106-x
45. Marsett RC, Qi J, Heilman P, Biedenbender SH, Carolyn Watson M, Amer S, et al. Remote Sensing for Grassland Management in the Arid Southwest. Rangeland Ecology & Management. 2006;59(5):530–40. https://doi.org/10.2111/05-201r.1

46. Diek S, Fornallaz F, Schaepman ME, De Jong R. Barest Pixel Composite for Agricultural Areas Using Landsat Time Series. Remote Sensing. 2017;9(12):1245. https://doi.org/10.3390/rs9121245
47. van Deventer AP, Ward AD, Gowda PH, Lyon JG. Using thematic mapper data to identify contrasting soil plains and tillage practices. Photogrammetric Engineering and Remote Sensing. 1997;63.
48. Chandrasekar K, Sesha Sai MVR, Roy PS, Dwevedi RS. Land Surface Water Index (LSWI) response to rainfall and NDVI using the MODIS Vegetation Index product. International Journal of Remote Sensing. 2010;31(15):3987–4005. https://doi.org/10.1080/01431160802575653
49. Shi T, Xu H. Derivation of Tasseled Cap Transformation Coefficients for Sentinel-2 MSI At-Sensor Reflectance Data. IEEE J Sel Top Appl Earth Observations Remote Sensing. 2019;12(10):4038–48. https://doi.org/10.1109/jstars.2019.2938388
50. Chen F, Feng P, Harrison MT, Wang B, Liu K, Zhang C, et al. Cropland carbon stocks driven by soil characteristics, rainfall and elevation. Sci Total Environ. 2023;862:160602. https://doi.org/10.1016/j.scitotenv.2022.160602 PMID: 36493831
51. Brenning A. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In: 2012 IEEE International Geoscience and Remote Sensing Symposium, 2012. 5372–5. https://doi.org/10.1109/igarss.2012.6352393
52. Kellner JR, Armston J, Duncanson L. Algorithm Theoretical Basis Document for GEDI Footprint Aboveground Biomass Density. Earth and Space Science. 2023;10(4). https://doi.org/10.1029/2022ea002516
53. Ploton P, Mortier F, Réjou-Méchain M, Barbier N, Picard N, Rossi V, et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat Commun. 2020;11(1):4540. https://doi.org/10.1038/s41467-020-18321-y PMID: 32917875
54. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40(8):913–29. https://doi.org/10.1111/ecog.02881
55. Cottingham KL, Lennon JT, Brown BL. Knowing when to draw the line: designing more informative ecological experiments. Frontiers in Ecology and the Environment. 2005;3(3):145–52. https://doi.org/10.1890/1540-9295(2005)003[0145:kwtdtl]2.0.co;2
56. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. Applied linear statistical models. 4th ed. Chicago: Irwin. 1996.
57. Klein A, Falkner S, Bartels S, Hennig P, Hutter F. Fast Bayesian Optimization of Machine Learning Hyperparameters on Large Datasets. 2017. http://arxiv.org/abs/1605.07079
58. Ramcharan A, Hengl T, Nauman T, Brungard C, Waltman S, Wills S, et al. Soil Property and Class Maps of the Conterminous United States at 100-Meter Spatial Resolution. Soil Science Soc of Amer J. 2018;82(1):186–201. https://doi.org/10.2136/sssaj2017.04.0122
59. Chaney NW, Wood EF, McBratney AB, Hempel JW, Nauman TW, Brungard CW, et al. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. Geoderma. 2016;274:54–67. https://doi.org/10.1016/j.geoderma.2016.03.025
60. Minasny B, McBratney AB, Wadoux AMJ-C, Akoeb EN, Sabrina T. Precocious 19th century soil carbon science. Geoderma Regional. 2020;22:e00306. https://doi.org/10.1016/j.geodrs.2020.e00306
61. Pribyl DW. A critical review of the conventional SOC to SOM conversion factor. Geoderma. 2010;156(3–4):75–83. https://doi.org/10.1016/j.geoderma.2010.02.003
62. Harmonized World Soil Database version 2.0. FAO; International Institute for Applied Systems Analysis (IIASA); 2023. https://doi.org/10.4060/cc3823en
63. Simpson EH. The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1951;13(2):238–41. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x
64. Castaldi F, Hueni A, Chabrillat S, Ward K, Buttafuoco G, Bomans B, et al. Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands. ISPRS Journal of Photogrammetry and Remote Sensing. 2019;147:267–82. https://doi.org/10.1016/j.isprsjprs.2018.11.026
65. He X, Yang L, Li A, Zhang L, Shen F, Cai Y, et al. Soil organic carbon prediction using phenological parameters and remote sensing variables generated from Sentinel-2 images. CATENA. 2021;205:105442. https://doi.org/10.1016/j.catena.2021.105442
66. Zhou T, Geng Y, Chen J, Liu M, Haase D, Lausch A. Mapping soil organic carbon content using multi-source remote sensing variables in the Heihe River Basin in China. Ecological Indicators. 2020;114:106288. https://doi.org/10.1016/j.ecolind.2020.106288
67. McCoy J. Regenerative agriculture's potential carbon storage in Nebraska soils. 2021. https://digitalcommons.unl.edu/envstudtheses/297
68. Giltrap DL, Li C, Saggar S. DNDC: A process-based model of greenhouse gas fluxes from agricultural soils. Agriculture, Ecosystems & Environment. 2010;136(3–4):292–300. https://doi.org/10.1016/j.agee.2009.06.014
69. Li C, Frolking S, Frolking TA. A model of nitrous oxide evolution from soil driven by rainfall events: 1. Model structure and sensitivity. J Geophys Res. 1992;97(D9):9759–76. https://doi.org/10.1029/92jd00509
70. Stange F, Butterbach-Bahl K, Papen H, Zechmeister-Boltenstern S, Li C, Aber J. A process-oriented model of N2O and NO emissions from forest soils: 2. Sensitivity analysis and validation. J Geophys Res. 2000;105(D4):4385–98. https://doi.org/10.1029/1999jd900948
71. Price I, Sanchez-Gonzalez A, Alet F, Andersson TR, El-Kadi A, Masters D, et al. Probabilistic weather forecasting with machine learning. Nature. 2025;637(8044):84–90. https://doi.org/10.1038/s41586-024-08252-9 PMID: 39633054
72. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8. https://doi.org/10.1038/nature21056 PMID: 28117445
73. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844