Geoderma 375 (2020) 114472

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Importance and strength of environmental controllers of soil organic carbon changes with scale

K. Adhikari[a,*], U. Mishra[b], P.R. Owens[c], Z. Libohova[d], S.A. Wills[d], W.J. Riley[e], F.M. Hoffman[f], D.R. Smith[a]

[a] USDA-ARS, Grassland, Soil and Water Research Laboratory, Temple, TX 76502, USA
[b] Argonne National Laboratory, Environmental Sciences Division, Argonne, IL 60439, USA
[c] USDA-ARS, Dale Bumpers Small Farms Research Center, Booneville, AR 72927, USA
[d] USDA-NRCS, National Soil Survey Center, Lincoln, NE 68508, USA
[e] Lawrence Berkeley National Laboratory, Earth Sciences Division, Berkeley, CA 94720, USA
[f] Oak Ridge National Laboratory, Computational Sciences & Engineering Division and Climate Change Science Institute, Oak Ridge, TN 37831, USA

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Soil organic carbon
Digital soil mapping
Scaling
Earth system models

ABSTRACT

Spatial heterogeneity in environmental factors on the land surface moderates exchanges of water, energy, and greenhouse gases between the land and the atmosphere. However, appropriately representing this heterogeneity in earth system models remains a critical scientific challenge. We used a large dataset of environmental factors (n = 31) representing soil-forming factors, field observations of soil organic carbon (SOC) (n = 6213), and a machine-learning algorithm (Cubist) to analyze the scaling behavior of SOC across the conterminous United States. We found that various environmental factors are significant predictors of SOC stocks at different spatial scales. Out of the 31 environmental factors we investigated, only 13 were significant predictors of SOC stocks at spatial scales ranging from 100 m to 50 km. Overall, topographic variables had higher influence at finer scales, whereas climatic variables were more important at coarser scales. The model performance worsened with increasing scale or the spatial resolution of prediction (R^2 = 0.38–0.65). The strength of environmental controls (median regression coefficient) on SOC weakened with scale, and we represented them using mathematical functions (R^2 = 0.38–0.98). Both the mean and variance of SOC stocks decreased linearly with increasing the scale in soils of the conterminous United States. Fitted linear functions accounted for 81% and 82% of the variability in the mean and variance of SOC, respectively. We also found linear relationships among mean and high-order moments of SOC (R^2 = 0.51–0.97). Improved understanding of the scaling behavior of SOC stocks and their environmental controllers can improve earth system model benchmarking and may eventually improve representation of the spatial heterogeneity of land surface biogeochemistry.

1. Introduction

Observation-based estimates of global soil organic carbon (SOC) stocks show large spatial heterogeneity (Batjes, 2016; FAO and ITPS, 2018; Hengl et al., 2014). This heterogeneity in SOC is primarily controlled by soil-forming factors: climate, topography, organisms, parent material, and time (Jenny, 1941; McBratney et al., 2003). Very often, it is also conditioned by soil use and management (Follett, 2001; Paustian et al., 1997). As a result, these environmental factors have been widely used to predict soil properties including SOC at a variety of spatial scales (Adhikari and Hartemink, 2015; Adhikari et al., 2014; Adhikari et al., 2013; Minasny et al., 2013; Mishra et al., 2017).

Despite their key roles in determining the spatial heterogeneity of SOC and regulating the rate of SOC decomposition, many soil-forming factors and pedogenic processes are not adequately represented in current land surface models. As a result, current land surface models poorly represent the baseline SOC spatial heterogeneity (Carvalhais et al., 2014; Todd-Brown et al., 2013) and show large uncertainties in predicting future carbon climate feedbacks (Friedlingstein et al., 2014). Burke et al. (2012) reported that the quantity, spatial distribution, and decomposability of SOC stocks accounted for half of the overall uncertainty in predicting future carbon climate feedbacks and associated climate changes. Therefore, to reduce the uncertainty in future carbon climate feedback projections, it is critical to appropriately represent environmental controllers and the spatial heterogeneity of SOC in land surface models.

* Corresponding author.
E-mail address: Kabindra.Adhikari@usda.gov (K. Adhikari).

https://doi.org/10.1016/j.geoderma.2020.114472
Received 26 February 2020; Received in revised form 15 May 2020; Accepted 21 May 2020
Available online 23 June 2020

One way to improve the spatial heterogeneity of SOC stocks in land surface models could be to represent the environmental controls on SOC stocks consistent with field observations. To achieve this, scaling functions could be developed to examine the relationship between the strength of environmental controllers and SOC stocks as they scale. We could then use these scaling functions to infer the appropriate environmental controllers of SOC across scales (Mishra and Riley, 2015).

Although several digital soil mapping (DSM) studies have used different environmental controllers to predict the spatial variation of SOC stocks (Adhikari et al., 2014; Adhikari et al., 2019; Hengl et al., 2014; Ramcharan et al., 2018; Viscarra Rossel et al., 2019), examination of the quantitative scaling relationship between SOC and its environmental controllers is sparse in the literature. In spatial prediction of soil properties, mathematical or statistical relationships are typically developed using a limited number of soil observations and environmental predictors without paying much attention to scaling behavior; the derived relationship is then applied using environmental predictors across the study area to produce spatially explicit estimates of soil properties (Lagacherie et al., 2007; McBratney et al., 2003). Several geospatial approaches have been used to predict the spatial heterogeneity of SOC, depending upon the available data density and environmental data on soil-forming factors (Adhikari et al., 2014; Minasny et al., 2013; Mishra et al., 2010).

Recently, Wiesmeier et al. (2019) reported on the variables that control SOC at different spatial scales. They reported that there are eight main drivers of SOC stocks operating at multiple spatial scales (soil particles or pedons to global scales): climate, topography, vegetation, microorganisms, soil physicochemistry, parent material, soil texture, and land use and management. Among these drivers, climate was highly influential at sub-regional to global scales, whereas topography was more closely related to SOC at local to sub-regional scales. Vegetation and climate showed comparable control on SOC stocks, except in areas smaller than 10,000 m^2. Parent material influenced SOC stocks at sub-regional to sub-conterminous scales, whereas land use and management was mostly related to scales other than global and micro-scale. Guo et al. (2019) showed that the spatial distribution of SOC scales because the controllers are scale dependent. Miller et al. (2015) also reported that SOC controllers are scale dependent, and that this influences model performance-models exhibit better performance when they include controllers from multiple scales rather than those from a single scale.

Many studies have also evaluated the scaling behavior and environmental controllers of soil moisture (e.g., Biswas and Si, 2011; Blöschl and Sivapalan, 1995; Famiglietti et al., 2008; Kachanoski and de Jong, 1988; Western et al., 2002). Biswas and Si (2011) characterized the spatial variation of soil water storage and its controllers across scales using empirical mode decomposition method. Sun et al. (2019) applied a similar approach to terrain attributes and mapped soil properties with increased model accuracy. Additional studies modeled the spatial variability of soil moisture patterns from relatively fine to coarse scales and attempted to characterize the spatial structure based on system properties and climate forcing (Crow et al., 2012; Pau et al., 2014; Riley and Shen, 2014; Shen et al., 2016; Western et al., 2002). For some systems, the spatial variance of soil moisture follows a power-law decay as a function of spatial area (Manfreda et al., 2007); in other systems, there are clear scale breaks in this relationship (e.g., Das and Mohanty, 2008; Joshi and Mohanty, 2010; Pau et al., 2014). Gebremichael et al. (2009) reported that, in a watershed located in the U.S. Southern Great Plains, soil moisture showed scale invariance, and that if the scaling parameters could be estimated from large-scale soil moisture fields, it might be feasible to change spatial soil moisture representations between scales. Consistent spatial scaling behavior was also recently reported for the spatial structure of high-latitude lake distributions (Muster et al., 2019). Despite the progress made in modeling these scaling properties, to our knowledge, no study has examined the statistical structure of SOC scaling behavior in temperate soils at large spatial scales.

The land surface interacts with the atmosphere at multiple spatial scales (Anderegg et al., 2019; Zhou et al., 2016). As a result, land surface spatial heterogeneity affects land–atmosphere exchanges of energy, moisture, and greenhouse gases (Clark et al., 2011; Riley and Shen, 2014). The current generation of earth system models (ESMs) typically operate at spatial scales larger than 50 km and use a nested sub-grid hierarchy approach to represent land surface heterogeneity (Lawrence et al., 2012). Attempts are being made to increase the spatial scale of these models to more accurately represent localized features of Earth systems that affect energy, water, and greenhouse gas fluxes (Pan et al., 2016; Pau et al., 2014).

Our goal is to model changes in SOC heterogeneity that result from changing the scales of environmental factors. Therefore, we used field observations of SOC, high-resolution environmental information, and a machine-learning approach to predict SOC at different spatial scales (from S = 100 m to 50 km). Throughout this paper, we refer to the "scale" (S) as either the area across which SOC properties are assumed to be homogeneous or the square root of the pixel area satisfying that criteria; note that the terms "scale" and "resolution" are often interchangeable in this context. We use the term "scaling" to mean the transfer of information about environmental controls (aggregation of environmental factors) and the statistical properties of SOC stocks from one scale to another. For the first time, in this study, we used a large set of the most recent SOC field observations available across the conterminous United States and a wide range of spatially explicit soil-forming factors to characterize the scaling behavior of SOC, and we developed scaling functions that could be used to improve land-surface representation of SOC in models. The specific objectives of our study were to (1) identify the dominant controllers of SOC at various scales, (2) quantify the change in the strength of environmental controllers of SOC at different spatial scales, and (3) develop scaling functions that relate the statistical properties of SOC to scale.

2. Materials and methods

2.1. SOC observations

SOC measurements in this study came from the rapid carbon assessment project initiated by the Natural Resources Conservation Service's Soil Science Division of the U.S. Department of Agriculture (USDA) (Soil Survey Staff and Loecke, 2016). The main goal of the rapid carbon assessment project was to produce a robust estimate of SOC stocks across the conterminous United States based on state-of-the-art soil sampling and modeling. More than 6200 sites across the conterminous United States (Fig. 1) were established according to a multi-level stratified random sampling scheme-a hierarchical sampling design that included region, land-use and land-cover classes, and groups of similar soil types (nested within regions).

The first level of sampling strata corresponds to the USDA's major land resource area (MLRA) grouped into regions (17 MLRAs across the conterminous United States). Each region was further divided into 8 to 20 soil groups by clustering of numerical scores calculated for soil taxonomy, particle size, soil depth, soil temperature regime, and drainage class (Wills et al., 2013). The soil groups were then combined with land use and land cover classes of national land-use and land-cover data, and sampling sites were randomly assigned to each combination.

At each site, five pedons were sampled: one at the plot center and one 30 m away in each cardinal direction. However, this study considered only the sample from the central pedon. Soil samples were collected from each soil horizon and analyzed for SOC concentration according to the Soil Survey Laboratory Methods Manual (Burt, 2004) and bulk density (volumetric method). SOC stock for a fixed soil depth (0–30 cm, in this case) was then calculated after correcting for non-soil materials (rock or coarse fragments) (Eq. (1)). Although bulk density is a key soil property in SOC stock calculations and should be a routine measurement, not all samples had bulk density measurements. The

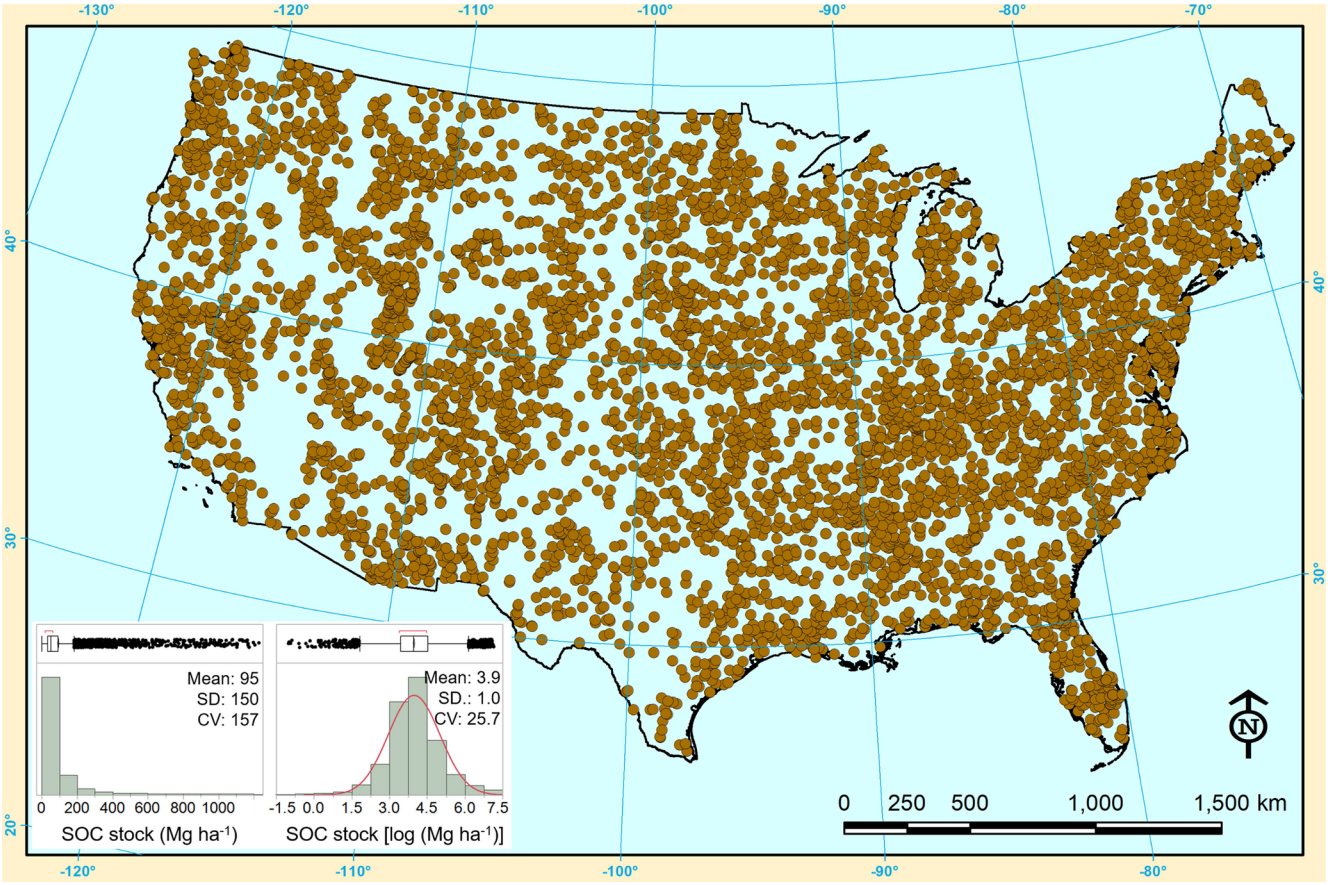

Fig. 1. Geographical distribution of soil sampling sites ($n = 6213$) across the conterminous United States. Inset: Histogram and box plots of SOC measurements (left) and after the log-transformation (right).

missing values for bulk density were estimated using pedotransfer functions in which the prediction error ranged from 0.10 to 0.15 g cm^{-3}. The pedotransfer functions were derived through a random forest model using 20,045 soil horizon data from 2680 pedons, and nine variables were used as predictors: four from the sampled horizon (horizon designation, textural class, depth [at the middle of the horizon], and thickness), and five from a neighbor horizon (bulk density, horizon designation, textural class, depth, and thickness) (Sequeira et al., 2014). Additional details about the rapid carbon assessment methodology can be found in Soil Survey Staff and Loecke (2016). The SOC stock for sampled depth was calculated using the following equation:

$$SOC_{stk} = \left[(SOC \times BD \times D) \times (1 - \frac{CF}{100}) \right] \quad (1)$$

where SOC_{stk} is the SOC stock (Mg ha^{-1}), SOC is the SOC content (g 100 g^{-1}), BD is the soil bulk density (Mg m^{-3}), D is the given soil layer thickness (cm), and CF is the volumetric fraction of the coarse fragments.

2.2. SOC predictors at multiple scales

A wide range of environmental variables were collected and evaluated as SOC predictors. These variables fall into four main categories and represent the state factors of soil formation as proposed by Jenny (1941): climate (*cl*), land use and land cover (*lulc*), soil or soil-related variables (*s*), and topographic (*r*) variables (Table 1). The *cl* variables included a 30-yr (1981 to 2010) annual average of temperature (minimum, mean, maximum, dew point), precipitation (rainfall, rainfall during the wettest and driest quarter in a year), and potential evapotranspiration. The *lulc* variables included the national land-use and land-cover database of the conterminous United States, potential vegetation cover, remote sensing imagery (spectral bands and vegetation index), net primary production, and ecological regions-areas representative of specific flora and fauna whose presence is conditioned and mediated by regional topography and prevailing climate. The *s*-related variables included soil types, geology, drainage condition, hydrological unit, and soil temperature regime. Similarly, *r* variables corresponded to terrain attributes (e.g., slope aspect, wetness index) derived from the national digital elevation model (DEM) of 30-m spatial resolution obtained from the U.S. Geological Survey. The original DEM was resampled to a 100-m raster and was hydrologically corrected by removing unnecessary pits and sinks before deriving terrain attributes.

To generate topographic variables at multiple scales, the 100-m raster DEM was subsequently resampled to the eight different spatial scales (grid sizes) considered in this study (i.e., grid sizes of 250 m, 500 m, 1 km, 2.5 km, 5 km, 10 km, 25 km, and 50 km), and the new DEM thus generated at each scale was used to derive terrain attributes for that scale ($n = 9$). Similarly, the multiscale *cl*-, *lulc*-, and *s*-related variables were compiled by resampling the original raster layer to the nine different spatial resolutions. Resampling of *cl*-, *t*-, and *lulc*-related variables (e.g., spectral band, vegetation index, and net primary production) was based on bilinear interpolation (distance-weighted average value of the surrounding 4 pixels), whereas that of *s*- and other *lulc*-related variables (e.g., national land use and land cover, potential vegetation cover, and ecological regions) was based on the nearest neighbor. Altogether, there were 31 environmental predictors at each of the nine scales, and all the raster layers and point SOC observations were remapped to a common projection to extract predictor values at the sampling locations. Table 1 provides information about the

Table 1
Environmental variables used as predictors of SOC across the conterminous United States.

Environmental variable and its abbreviation	Brief description	Data source	Resolution
Climate variable (*cl*)			
Precipitation (PPT)	30-yr (1981 to 2010) annual average precipitation	http://www.prism.oregonstate.edu/normals	800 m
Precipitation of the driest season (PDRY)	30-yr (1971–2000) annual average precipitation of the driest month	http://worldclim.org/bioclim	1 km
Potential evapotranspiration (PET)	30-yr (1971–2000) potential evapotranspiration	Trabucco, Antonio; Zomer, Robert (2019): Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v2. Figshare, https://doi.org/10.6084/m9.figshare.7504448.v3	1 km
Precipitation of the wettest season (PWET)	30-yr (1971–2000) annual average precipitation of the wettest month	http://worldclim.org/bioclim	1 km
Dew point temperature (TD)	30-yr (1981–2010) annual average dew point temperature	http://www.prism.oregonstate.edu/normals	800 m
Minimum temperature (TMIN)	30-yr (1981–2010) annual average minimum temperature	http://www.prism.oregonstate.edu/normals	800 m
Mean temperature (TMEAN)	30-yr (1981–2010) annual average temperature	http://www.prism.oregonstate.edu/normals	800 *m*
Maximum temperature (TMAX)	30-yr (1981–2010) annual average maximum temperature	http://www.prism.oregonstate.edu/normals	800 m
Land use and land cover variable (*lulc*)			
Ecological region (ECOL3)	Ecological zone map at level 3 legend	Derived from gSSURGO.	100 m
Net primary production (NETPP)	Annual terrestrial primary production (kg C m^{-2}) for 2018	Derived from Landsat	30 m
Landsat Band 3 (RED)	Landsat Band 3 for 2014	http://earthenginepartners.appspot.com/science-2013-global-forest/download_v1.6.html	30 m
Landsat Band 5 (SW1)	Landsat Band 5 for 2014	http://earthenginepartners.appspot.com/science-2013-global-forest/download_v1.6.html	30 m
Landsat Band 7 (SW2)	Landsat Band 7 for 2014	http://earthenginepartners.appspot.com/science-2013-global-forest/download_v1.6.html	30 m
National land cover database (NLCD)	Land cover of the United States for 2011		30 m
Potential vegetation (PVEG)	U.S. Potential natural vegetation	Original Kuchler Types, v2.0	5 km
Normalized difference vegetation index (NDVI)	Calculated as (NIR – RED)/(NIR + RED), where, NIR is near-infrared band (Landsat Band 4)	http://earthenginepartners.appspot.com/science-2013-global-forest/download_v1.6.html	30 m
Topographic variable (*r*)			
Elevation (DEM)	Land surface elevation	Derived from the national digital elevation dataset (NDEM) from U.S. Geological Survey	100 m
Slope aspect (ASPECT)	Direction of the steepest slope from the north	Derived from the DEM	100 m
Slope length factor (LSFACTOR)	Slope length factor calculated as in the USLE (universal soil-loss equation)	Derived from the DEM	100 m
Multi-resolution valley bottom flatness index (MRVBF)	Potential depositional areas	Derived from the DEM	100 m
Melton ruggedness number (MRN)	Melton ruggedness number	Derived from the DEM	100 m
Mid-slope position (MSPOS)	Covers the warmer zones of slopes	Derived from the DEM	100 m
Wetness index (SAGAWI)	Topographic wetness index with modified catchment area	Derived from the DEM	100 m
Slope height (SLOPEHT)	Height of the local slope	Derived from the DEM	100 *m*
Slope gradient (SLOPE)	Local slope gradient in percent	Derived from the DEM	100 m
Valley depth (VALDEP)	Calculates the extent of valley depth	Derived from the DEM	100 m
Soil variable (*s*)			
Drainage class (DRNG)	Natural soil drainage class	Derived from gSSURGO	100 m
Surface geology (GEOSUR)	Surficial geology class	Derived from gSSURGO	1 km
Hydrological group (HYDRO)	Hydrologic soil group class	Derived from gSSURGO	100 m
Soil order class (SOIL)	Taxonomy soil order class	Derived from gSSURGO	100 m
Soil temperature regime (SOILTR)	Soil temperature regime class	Derived from gSSURGO	100 m

environmental variables used in this study and their data sources.

2.3. Selection of SOC predictors, model development, and accuracy assessment

From the pool of 31 SOC predictors, only the significant predictors (p-value < 0.1) were selected to build a prediction model at each scale. We applied a stepwise regression technique that uses forward selection and backward elimination to select significant variables in the model. Once the significant variables were identified, we used data-mining algorithms from the Cubist tool (Quinlan, 1993) to develop SOC prediction models at each scale. For this purpose, we extracted the values of environmental predictors at SOC observation locations, and the entire dataset was randomly divided into training (75%) and test (25%) datasets to calibrate and validate prediction models, respectively. The Cubist model built a set of hierarchical regression rules such that each rule was defined by certain environmental conditions based on the prevailing soil-landscape relationships. Once each condition was met, a rule-specific multiple linear regression (MLR) function was used to predict SOC stocks (Adhikari and Hartemink, 2015; Minasny and McBratney, 2008; Ugbaje and Reuter, 2013).

The performance of the prediction model at each scale was evaluated using the test dataset. To test model accuracy across scales, we used statistical indices of the coefficient of determination (R^2), root-mean square error (RMSE), mean error or bias in prediction (ME), and relative error (RE). The RE was calculated following Eq. (2) (Ugbaje and

Reuter, 2013). RE is the ratio of the average error (AE) value—which is equivalent to the mean absolute error—to the AE that would result from always predicting the observation mean (AE_m). All these indices were calculated for both the training and test datasets.

$$RE = \frac{AE}{AE_m} = \frac{\frac{1}{n}\sum_{i=1}^{n}|obs_i - pred_i|}{\frac{1}{n}\sum_{i=1}^{n}|obs_i - \overline{pred_i}|} \quad (2)$$

where obs_i, $pred_i$, and $\overline{pred_i}$ are observed, predicted, and mean predicted SOC values, respectively, and n is the number of observations.

2.4. Key SOC predictors and their scaling properties

To identify key SOC predictors, the relative importance (RI) of the variables in predicting SOC stocks at each spatial scale was quantified as its relative usage in the prediction model. As the prediction model developed a set of conditions and associated MLR functions, the RI was calculated for both. Similarly, to quantify the strength of a variable in SOC prediction, we calculated its median regression coefficient (β) associated with each MLR function in the set of hierarchical regression rules for each scale. We then normalized it by multiplying it with the average raster value of the corresponding predictor. The β values for each predictor were plotted against prediction scales, and a mathematical function was fitted to model its scaling behavior across scales. To further analyze the scaling properties of SOC, and to characterize SOC variability across scales, the mean and variance of predicted SOC stocks were plotted against the spatial scales and against each other. Similarly, the mean of predicted SOC stocks was plotted against standard deviation, coefficient of variation, skewness, and with the standard error of mean. In addition, we grouped the range of predicted SOC means into bins equivalent to an SOC stock of 1 [log(Mg ha^{-1})]. To quantify the magnitude of spatial heterogeneity at different ranges of SOC stocks, the mean SOC from each bin was plotted against its standard deviation, coefficient of variation, skewness, and the standard error of the mean.

3. Results

3.1. SOC field observations

SOC in the study area was extremely variable (CV = 157%) and log-normally distributed. SOC measurements ranged from 0.1 to over 1200 Mg ha^{-1}, with mean and standard deviations of 95 and 150 Mg ha^{-1}, respectively (Fig. 1, inset). The distribution was highly and positively skewed (skewness coefficient: 4.3; median: 51.6 Mg ha^{-1}). After the data were log-transformed, the skewness coefficient and mean dropped to −0.04 and 3.9 Mg ha^{-1}, respectively, with a standard deviation of 1.0 and a CV of 25.7%. For modeling, SOC data were randomly split into training and test datasets; Table 2 shows the general statistics and distributions of SOC in the datasets. Most of the statistical parameter values for these datasets are comparable to each other, which validates the data split (Table 2).

3.2. Significance of environmental predictors of SOC stocks changes with scale

Among the 31 environmental variables used as SOC predictors, only 13 were significant ($p < 0.1$) at all spatial scales (Fig. 2), which shows that a clear scale dependency on the SOC distribution. For example, temperature-related variables, such as maximum temperature (TMAX), minimum temperature (TMIN), and dew point temperature (TD), were not significant at 100 m but were significant at spatial scales greater than 1 km. Almost all topographic variables were significant at scales ≤250 m except for melton ruggedness number (MRN) and slope height (SLOPEHT). The effect of SLOPEHT was significant at scales greater than 1 km. On the other hand, precipitation-related variables, such as precipitation (PPT), potential evapotranspiration (PET), precipitation of the wettest season (PWET), and precipitation of the driest season (PDRY), were significant at all scales, indicating a significant control of moisture on SOC distributions across scales. Similarly, variables like drainage class (DRNG), surface geology (GEOSUR), hydrological group (HYDRO), soil order class (SOIL), and soil temperature regime (SOILTR) were significant at all scales, suggesting that soil and its drainage played significant roles in the spatial distribution of SOC across the conterminous United States. Land-use- and land-cover-related variables, like ecological regions (ECOL3), landsat band 3 (RED), and landsat band 7 (SW2), were significant at all scales. Normalized difference vegetation index (NDVI) was found to be significant between the 100-m and 1-km scales, and it was not significant at scales greater than 1 km. National land cover database (NLCD) showed an intermittent contribution, playing a significant role at 100 m, 5 km, and 50 km, but not at other scales. Potential vegetation (PVEG), on the other hand, was significant at all scales except at 10 km. In general, most topographic variables were significant at smaller scales (< 5 km), whereas climate variables, particularly those related to temperature, were significant at larger scales (> 1 km). Soil and land-use- and land-cover-related variables were important at all scales.

Table 2
Summary statistics of measured SOC data across the study area before and after the data split into training and tests sets for model building and validation.

Parameter	SOC stock (Mg ha^{-1})		
	All data	Training data	Test data
Number of observations	6213	4660	1553
Mean	94.9	94.7	95.7
Std. Dev.	149.9	148.0	152.1
Std. Error Mean	1.8	2.1	3.8
Skewness	4.3	4.3	4.3
Kurtosis	21.8	21.8	21.7
Coeff. of variation	156.9	156.2	158.9
Minimum	0.1	0.1	0.4
Maximum	1268.2	1268.2	1257.7
Median	51.6	51.5	52.2
Interquartile range	60.7	60.8	60.7

3.3. Importance and strength of environmental predictors of SOC changes with scale

The importance of environmental predictors quantified as relative importance in the prediction model is shown in Fig. 3. Because the prediction model was based on regression rules, Fig. 3A represents the RI of variables in rule-setting conditions, and Fig. 3B shows the RI in MLR functions within rule-setting conditions. In principle, the condition rules divided the study area into different sub-units or strata where the SOC distribution could be predicted using the specific MLR function associated with the rule. Results showed that ECOL3 was one of the main prediction variables, which was used in setting condition rules at all scales, followed by RED and GEOSUR.

For the rule-setting conditions, ECOL3 RI ranged from 87% at 500 m to 100% at the 50 km scale. SOIL and DRNG were also important in rule setting; however, DRNG was important at scales below 2.5 km with a maximum RI of 63% at 500 m. The prediction variable SOIL was important at almost all scales with its maximum RI (53%) at 2.5 km and its minimum (9%) at 25 km. PET was found to be important at scales beyond 250 m; a maximum contribution of 56% was reported at 10 km and a minimum (20%) at 250 m. PPT showed a similar RI trend; its importance began at 1 km with the lowest RI (16%) and reached the highest RI (56%) at 10 km. The model also identified DEM as a main prediction variable in rule-setting conditions at all scales; however, its contribution was lower than ECOL3, RED, and GEOSUR, with an RI ranging from 17 to 26%. Temperature-related environmental predictors were also used in rule-setting condition, but their contributions were

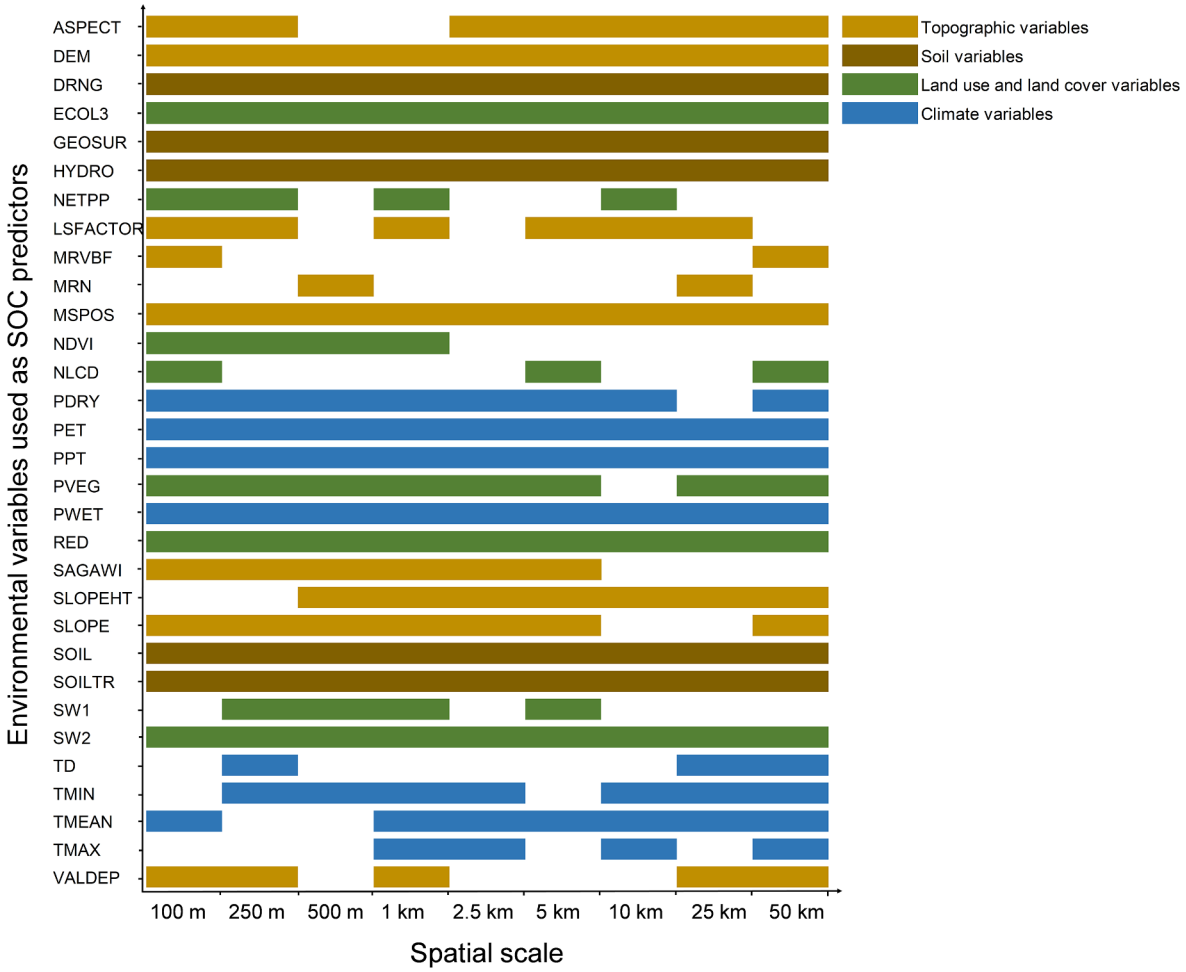

Fig. 2. Scales over which environmental variables are significant predictors (p-value $<$ 0.1) of SOC distribution across the conterminous United States. ASPECT: slope aspect; DEM: digital elevation model; DRNG: natural soil drainage class; ECOL3: ecological region at level 3 legend; GEOSUR: surface geology; HYDRO: hydrological group; NETPP: net primary production; LSFACTOR: slope-length factor; MRVBF: multi-resolution valley bottom flatness index; MRN: melton ruggedness number; MSPOS: mid-slope position; NDVI: normalized difference vegetation index; NLCD: national land cover data; PDRY: total precipitation of the driest season; PET: potential evapotranspiration; PPT: annual precipitation; PVEG: potential vegetation; PWET: total precipitation of the wettest season; RED: red band; SAGAWI: wetness index; SLOPE: slope gradient; SLOPEHT: slope height; SOIL: soil order; SOILTR: soil temperature regime;; SW1: landsat band 5; SW2: landsat band 7; TD: dew point temperature; TMIN: minimum temperature; TMEAN: mean temperature; TMAX: maximum temperature; VALDEP: valley depth.

minimal and started at scales greater than 2.5 km.

For the MLR prediction function, topographic predictors such as DEM and mid-slope position (MSPOS) were important at all scales with RI ranging from 23% (50 km) to 51% (10 km) for DEM, and 25% (50 km) to 76% (1 km) for MSPOS (Fig. 3B). Other topographic predictors such as wetness index (SAGAWI) were important at scales $\leq$5 km, and multi-resolution valley bottom flatness index (MRVBF) and MRN were important at scales $\leq$500 m. Slope aspect (ASPECT) was found to be important at all scales except for 1 km, with a maximum RI at 100 m (RI 37%) and a minimum (3%) at 2.5 km and 10 km. The contribution of temperature-related variables started from scales greater than 1 km, reaching a maximum RI of 100% at 50 km for TMAX, TMIN, and mean temperature (TMEAN). TD was only important at scales of 25 and 50 km. Precipitation-related variables, such as PPT, PET, PDRY, and PWET, were important in predicting SOC at all scales, with RI ranging from 52 to 100%; higher RI values were reported for scales greater than 1 km. We also noticed that there were no temperature-related predictors in the MLR function at the 100-m scale; their significant contributions started at scales greater than 1 km, except for TMIN, which was also important at 250 and 500 m. Of the land-use- and land-cover-related predictors, NDVI and net primary production (NETPP) were important up to 1 km, SW2 and RED at all scales, and landsat band 5 (SW1) between the 250-m and 5-km scales. Maximum RI values for NDVI (100%), NETPP (61%), and SW1 (75%) were found at 1 km, and the minimum was at 250 m.

Fig. 3C–F show the RI of the variables at the 100-m and 50-km scales. A total of 11 environmental predictors were used to set rule conditions at 100 m. Five were related to land-use and land-cover types, three were related to soil properties, two were related to topographic attributes, and one was related to climate (Fig. 3C). The RI of land-use- and land-cover-related variables was much higher than that of the other environmental variables, and climate-related variables had the lowest RI of all. On the other hand, out of 17 variables used in the MLR prediction function at the same scale, eight were related to topographic attributes. These had a higher RI than climatic variables other than PWET, which had an RI of 98% (Fig. 3D). Unlike at the 100-m scale, the climate variables showed a higher influence (~30% RI) even in setting prediction rule conditions at this scale. The RI of topographic variables ranged from 23% for valley depth (VALDEP) to 75% for MRVBF, whereas land-use- and land-cover-type variables (four variables) ranged from 43% (NETPP) to 97% (RED). Similarly, at the 50-km scale (Fig. 3E,F), only seven variables were used to set rule conditions, and 17 were used in the MLR prediction function. Out of these, almost 50% were climate variables; two were land-use and land-cover types and the rest were topographic variables. Among all variables, climate variables had a maximum RI reaching up to 100%: six out of seven variables had

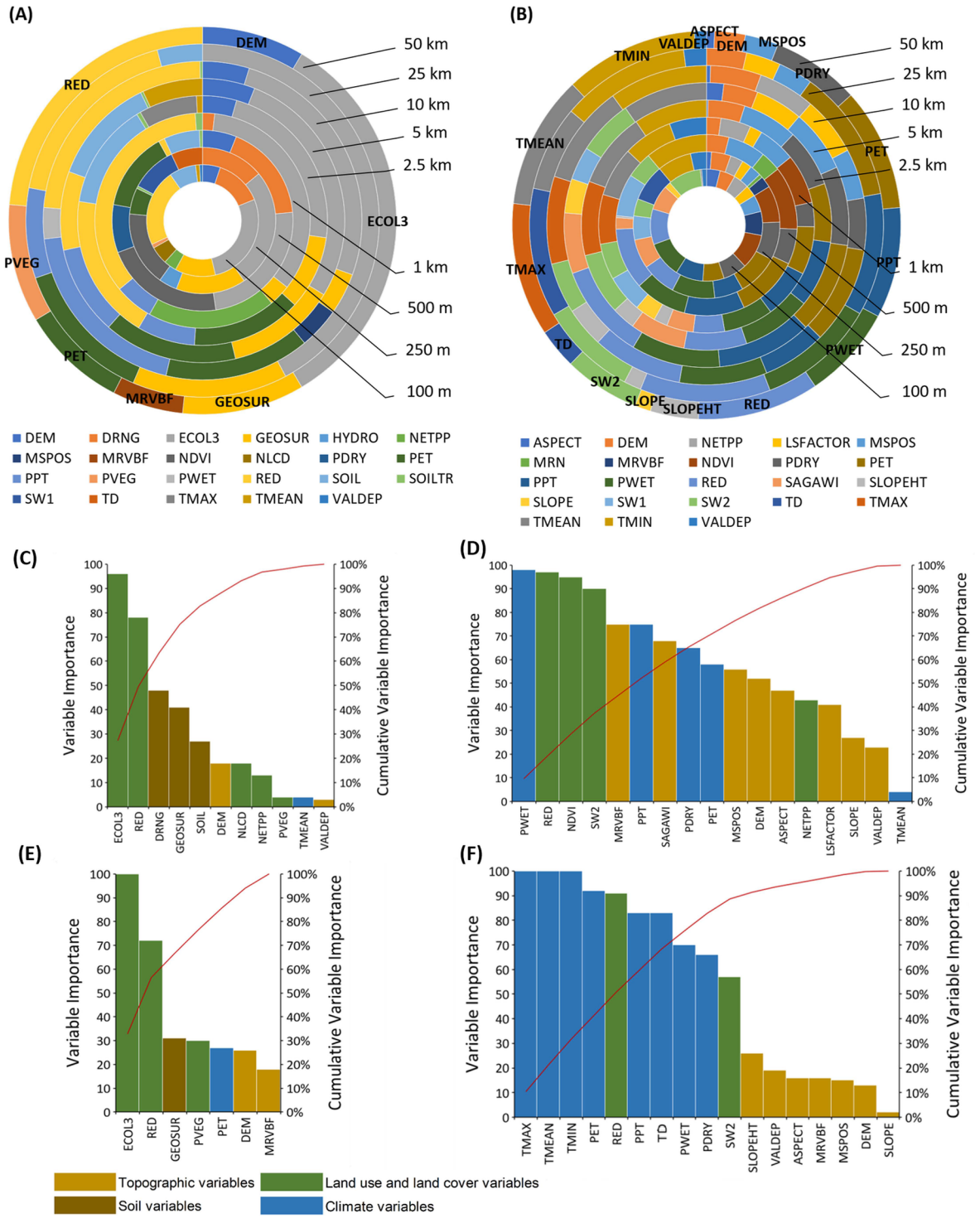

Fig. 3. Relative importance (RI) of SOC predictors across scales. (A) RI for rule-setting conditions; (B) RI in the MLR prediction function; (C) and (D) RI for rule-setting conditions, and MLR function at 100 m, respectively; and (E) and (F) RI for rule-setting conditions, and MLR function at 50 km, respectively. ASPECT: slope aspect; DEM: digital elevation model; DRNG: natural soil drainage class; ECOL3: ecological region at level 3 legend; GEOSUR: surface geology; HYDRO: hydrological group; NETPP: net primary production; LSFACTOR: slope-length factor; MRVBF: multi-resolution valley bottom flatness index; MRN: melton ruggedness number; MSPOS: mid-slope position; NDVI: normalized difference vegetation index; NLCD: national land cover data; PDRY: total precipitation of the driest season; PET: potential evapotranspiration; PPT: annual precipitation; PVEG: potential vegetation; PWET: total precipitation of the wettest season; RED: red band; SAGAWI: wetness index; SLOPE: slope gradient; SLOPEHT: slope height; SOIL: soil order; SOILTR: soil temperature regime; SW1: landsat band 5; SW2: landsat band 7; TD: dew point temperature; TMIN: minimum temperature; TMEAN: mean temperature; TMAX: maximum temperature; VALDEP: valley depth.

RI ≥ 97%. Topographic variables had a minimum RI as low as 17%: six out of seven variables had RI ≤ 54%. Overall, a higher influence of topographic variables was found at finer scales, whereas climatic variables were more important at larger scales.

The regression coefficient (β) of SOC predictors in the MLR function quantified the strength of predictors in SOC distributions across scales, and the median β of the selected predictors (i.e., three topographic, one land-use and land-cover, and four climate variables) are shown in

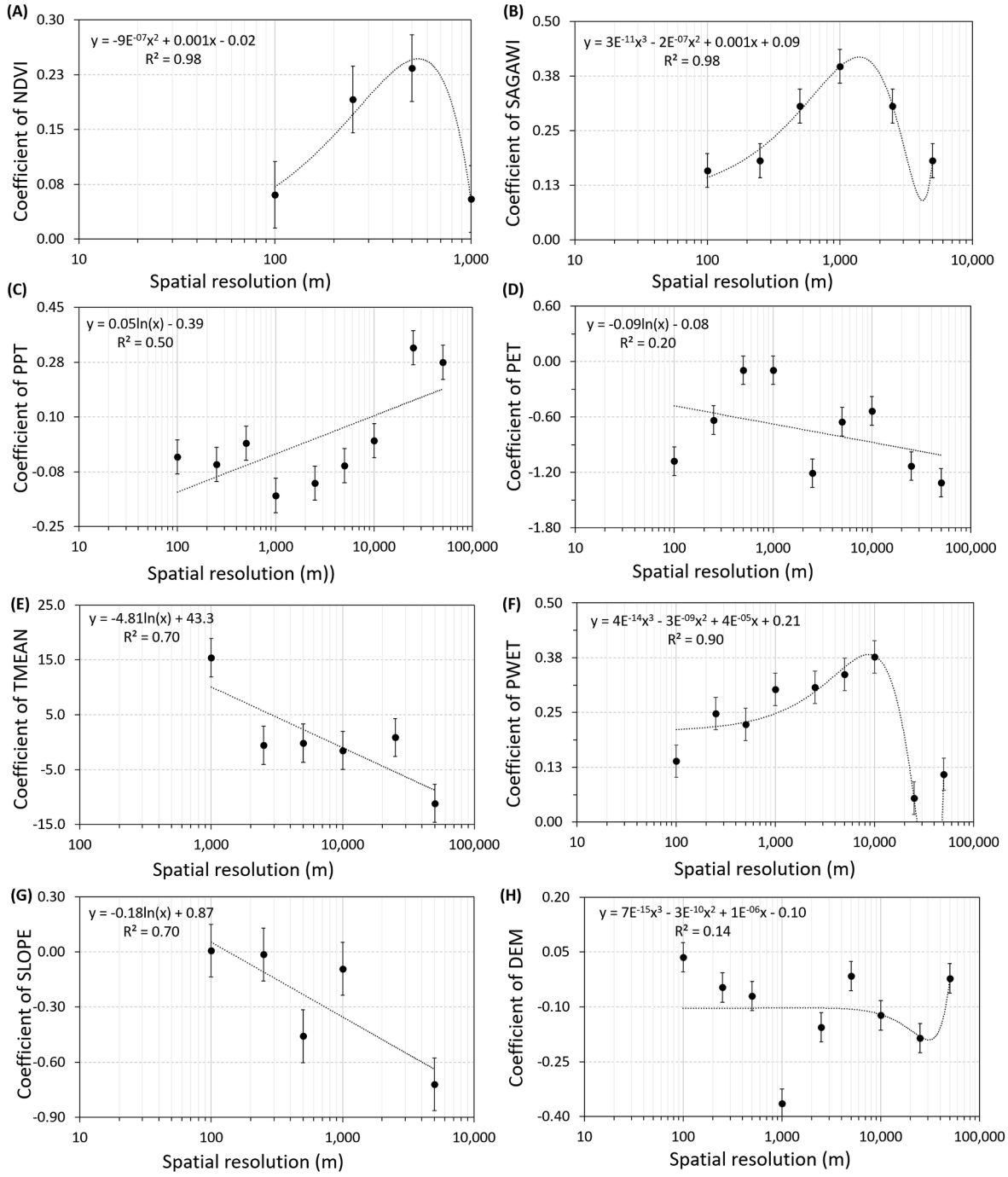

Fig. 4. Control of environmental factors on SOC stocks as a function of spatial scale. Each dot is a median regression coefficient multiplied by the average value of the environmental predictor across the conterminous United States. Error bars represent standard error. NDVI: normalized difference vegetation index; SAGAWI: wetness index; PPT: annual precipitation; PET: potential evapotranspiration; TMEAN: annual mean temperature; PWET: total precipitation of the wettest season; SLOPE: slope gradient; DEM: elevation.

Fig. 4. As mentioned above, NDVI was important at spatial scales between 100 m and 1 km, and its median β values ranged from 0.07 at 100 m to 0.24 at 500 m. Similarly, SAGAWI was important at scales ≤5 km, and its strength increased linearly; it reached its peak at 1 km and decreased thereafter to a minimum of 0.18 at 5 km. The lowest control of SAGAWI in the SOC distribution was at 100 m. Similarly, the influence of slope gradient (SLOPE) was negative, occurred at scales less than 5 km, and increased (more negative) with scale. DEM exerted an influence at all scales (e.g., higher influence at 1 and 2.5 km; lower influence at 5 and 50 km), and the relationship was negative except at 100 m. where the relationship was positive. DEM's strength was highest at 1 km ($\beta = -0.36$). It decreased thereafter to 50 km, which suggests 1 km as a cutoff scale from which to observe DEM's influence.

For the climatic variables, PPT and PWET both had positive coefficients at all scales. However, a small negative coefficient was recorded for PPT at the 1- and 2.5-km scales. Overall, the coefficients increased with increasing scale, reaching a maximum value at 25 km for PPT and at 10 km for PWET. On the other hand, PET had a negative coefficient with SOC; its influence was negative at scales greater than 1 km. The influence of TMEAN was almost always negative, and the influence increased with increasing scale. These results further reinforced that the SOC distributions at finer scales were mostly controlled by topographic

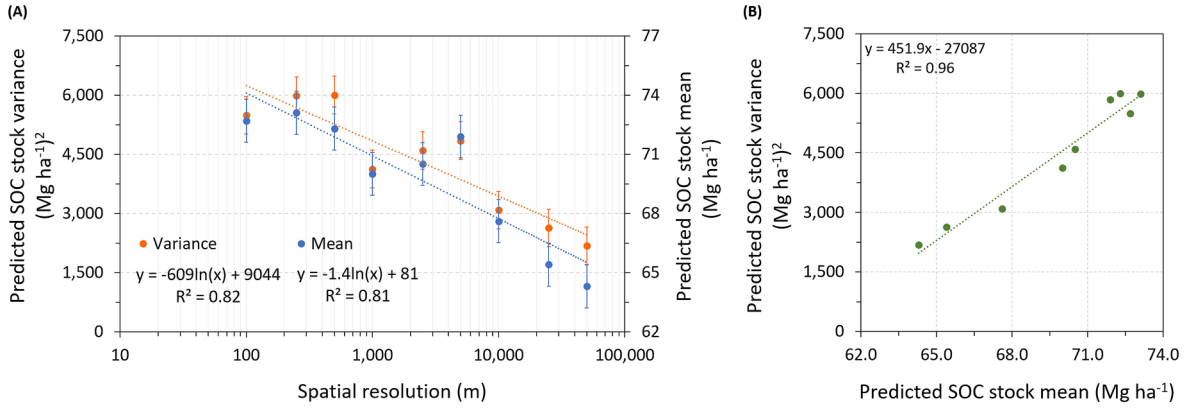

Fig. 5. (A) Predicted SOC mean and variance across multiple scales. (B) Relationship between mean and variance. Error bars in panel A represent standard error.

variables, whereas climatic variables had greater influence at coarser scales. This generally agrees with recent results from Wiesmeier et al. (2019).

3.4. Scaling impacts on the spatial heterogeneity of SOC

Results showed that the mean and variance of predicted SOC stocks decreased with scale, and this relationship could be modeled with a linear function ($R^2 > 0.80$) (Fig. 5A). The highest predicted SOC mean and variance were at 250 m, and the lowest at 50 km. It was fairly constant between 100 and 500 m, then decreased. The predicted mean and variance had a strong positive linear relationship ($R^2 = 0.96$) (Fig. 5B). Similarly, the relationships among mean predicted SOC and its standard deviation, CV, skewness, and standard error of mean at each spatial scale showed that these relationships could be described with linear functions (Fig. 6) that increased positively. The R^2 ranged between 0.51 and 0.97, with the highest value for the standard deviation and the lowest for the skewness coefficient. Those for CV and standard error were 0.96 and 0.94, respectively. The kurtosis coefficient was weakly and linearly related with the mean predicted SOC stocks ($R^2 = 0.15$) (not shown in Fig. 6).

We examined relationships among the statistical properties of the SOC distributions and the mean stock to identify spatial scaling properties. To compare SOC variability across scales, the range of log-transformed mean SOC stock between 0 and 7.5 (log [Mg ha^{-1}]) was divided into a bin size of 1 (log [Mg ha^{-1}]), and the standard deviation, CV, and skewness were derived and plotted against the mean for each bin (Fig. 7). Fig. 7A and 7B compare bin-averaged SOC standard deviation and CV across scales.

A wider range of standard deviation and CV was observed at SOC less than 1.5 (log [Mg ha^{-1}]); a convex and upward trend in standard deviation and a decreasing trend in CV were observed with increasing predicted SOC mean. For the same mean range, the standard deviation increased with scale, except at 100 m and 10 and 25 km. It reached a peak at an SOC between 3.0 and 4.5 and decreased thereafter. At all scales, a maximum CV was found at SOC < 1.5 and a minimum at SOC > 6.0 (log [Mg ha^{-1}]).

The skewness coefficient, on the other hand, had a mixed response to the SOC mean at different scales (Fig. 7C). The range of skewness was smaller for SOC between 3.0 and 4.5, and remained greater beyond that mean SOC range. A similar trend was observed for the standard error of mean (Fig. 7D), which was at its minimum for SOC between 3.0 and 4.5 (log [Mg ha^{-1}]). The scaling impact of predictors was more obvious at a specific SOC range (i.e., < 3.5 and > 4.5). Outside that range, the influence was less pronounced. Moreover, we observed similar patterns in all the fitted curves for all scales and all statistics, indicating a general pattern across scales.

3.5. Model validation and prediction accuracy

Prediction model accuracy was assessed using common validation indices such as R^2, ME, RMSE, and RE. The results for both the training and the test datasets are listed in Table 3. R^2 values ranged from 0.38 to 0.65, and decreased with the scale of prediction for both datasets-the highest value at 100 m and the lowest at 50 km. The ME fluctuated around zero, and at a spatial scale of 5 km, the prediction was negatively biased. The RMSE values ranged between 0.41 and 0.54 and changed little across the predicted scales. A minimum RMSE was observed at 100 m for both datasets. With regard to RE, 100 m had the lowest and 50 km had the highest value. Overall, the model was more accurate in predicting SOC at 100 m and less accurate at 50 km, indicating that prediction accuracy decreased with increasing spatial scale.

4. Discussion

Most SOC resides in the soil surface, and thus it can be rapidly altered by anthropogenic and climatic factors. Therefore, the spatial heterogeneity of SOC impacts the magnitude of greenhouse gas fluxes from the land surface. We predicted SOC distributions across the conterminous United States using recent SOC measurement data, a suite of environmental variables, and a data-mining technique proved promising in several previous studies (e.g., Adhikari et al., 2014, 2019; Bonfatti et al., 2016; Dorji et al., 2014; Lacoste et al., 2014). However, the knowledge found using such techniques must be treated with caution, especially in soil attribute predictions where the selection of pedologically relevant variables, and map interpretation is crucial (Wadoux et al., 2020). We document different environmental factors that control SOC at different scales. The strength of the control of different environmental factors on SOC stocks weakened as scale increased; as a result, the model performance (R^2) decreased as scale increased. The relationship between variance and mean values of SOC stocks and scale can be modeled using simple linear functions. We also observed a linear relationship between the mean and the variance of the predicted SOC stocks, and nonlinear but consistent relationships with its higher-order moments.

In this study, we used a variety of environmental factor datasets from various sources (Table 1). These datasets had different spatial resolutions but were resampled to specific spatial resolutions for modeling. The mismatch in original spatial detail among the environmental covariates could have influenced the prediction performance and model outputs. However, this issue seems inevitable at regional-/continental-scale studies that use a large number of secondary datasets. We believe future DSM activities should prioritize research in harmonizing multi-scale-multisource data together with increasing the usefulness of legacy data for a seamless product.

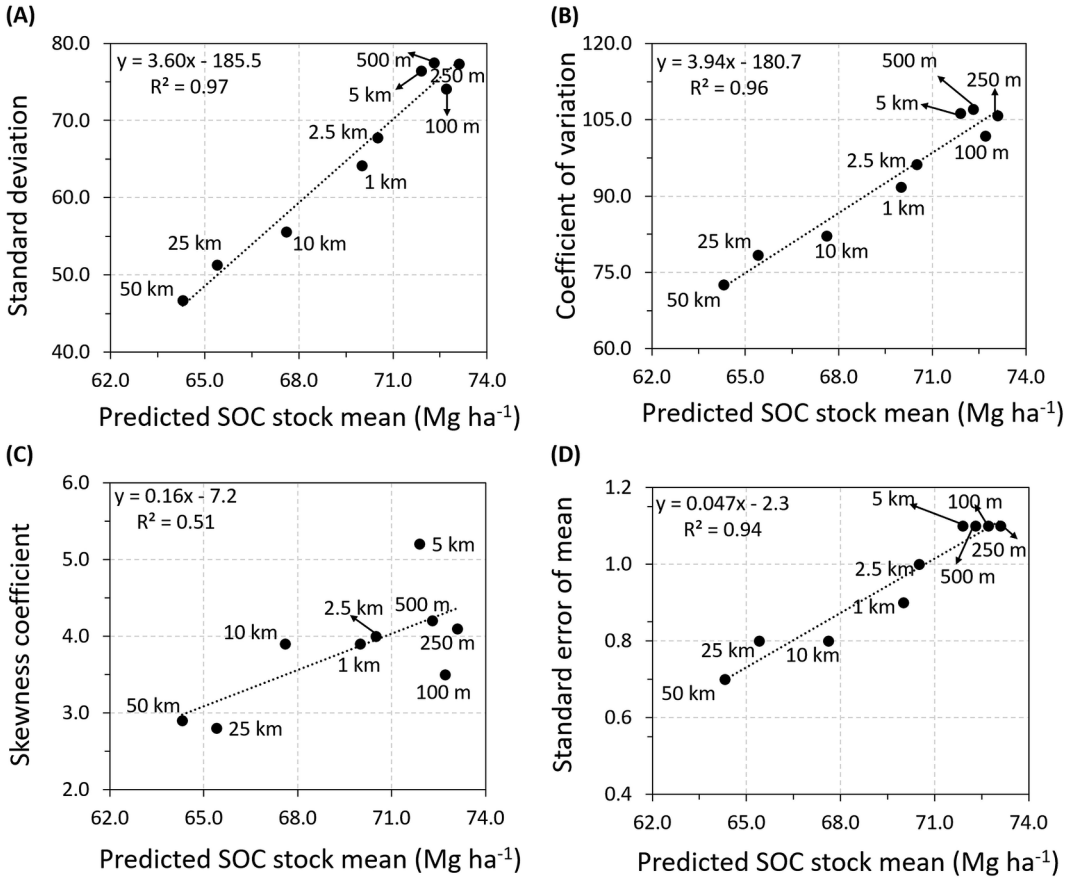

Fig. 6. Relationship between predicted SOC and (A) standard deviation, (B) coefficient of variation, (C) skewness coefficient, and (D) standard error of mean at corresponding spatial scales.

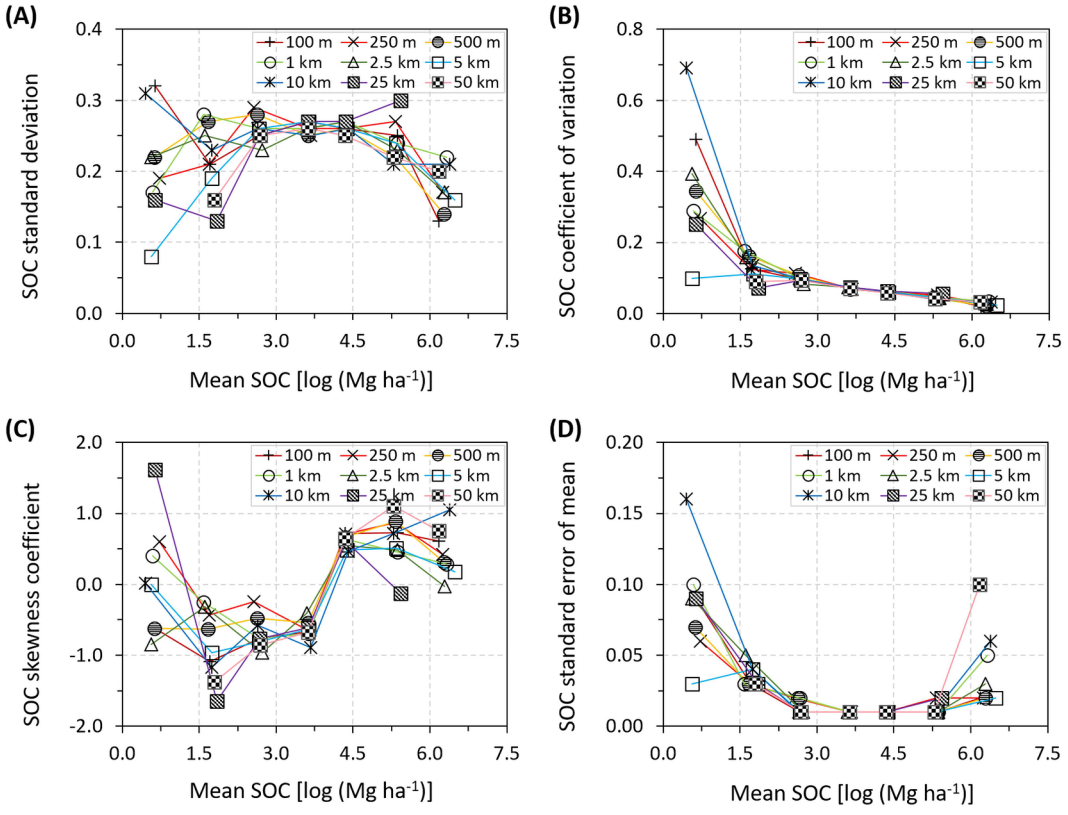

Fig. 7. Predicted SOC mean (log [Mg ha^{-1}]) plotted against (A) standard deviation, (B) coefficient of variation, (C) skewness, and (D) standard error of mean derived for the binned data of 1 log (Mg ha^{-1}) across spatial scales.

Table 3
Model validation indices derived for training and test datasets at multiple spatial scales.

Validation index		100 m	250 m	500 m	1 km	2.5 km	5 km	10 km	25 km	50 km
R^2	Training	0.65	0.62	0.60	0.56	0.55	0.53	0.52	0.49	0.48
	Test	0.53	0.51	0.49	0.48	0.47	0.47	0.41	0.40	0.38
ME	Training	0.006	−0.007	0.004	0.008	0.004	−0.003	0.01	0.014	0.018
	Test	0.035	0.02	0.025	−0.001	0.01	−0.01	−0.006	0.035	0.01
RMSE	Training	0.41	0.43	0.45	0.46	0.46	0.45	0.45	0.45	0.44
	Test	0.51	0.52	0.53	0.53	0.52	0.52	0.51	0.54	0.54
RE	Training	0.59	0.62	0.63	0.66	0.66	0.67	0.68	0.70	0.71
	Test	0.67	0.69	0.69	0.71	0.73	0.71	0.75	0.76	0.76

R^2: coefficient of determination; ME: mean error; RMSE: root mean squared error; RE: relative error.

4.1. Decrease of the strength of environmental controls on SOC stocks

We observed that the strength of environmental controls on SOC stocks decreased with increased spatial scale. Among the environmental predictors, DEM and SLOPE had maximum strengths (median β) at 100 m and minimum strengths at 5 and 50 km. SLOPE and SAGAWI had important controls only between 100 m and 5 km; the latter had its maximum strength at 1 km and decreased thereafter. Similarly, the strength of NDVI was observed only at scales between 100 m and 1 km, reaching its maximum value at 500 m and a minimum at 1 km. Beyond this point, no strength was observed.

Among the climatic variables, PET and TMEAN had the largest strength at 1 km and the smallest at 50 km. However, an opposite trend was observed for PPT, which had the highest strength at 25 km and lowest at 1 km. Overall, the largest strength or control on the SOC stocks distribution was governed by temperature (TMEAN, TMIN, and TMAX), followed by land use and land cover (NDVI) and topography (SLOPE and MSPOS). Mishra and Riley (2015) reported a similar decreasing trend for the strength of environmental controls of SOC stocks of arctic/boreal soils. They found elevation, temperature, potential evapotranspiration, and land cover to be significant environmental predictors of SOC at all scales. In addition to these four environmental predictors, we also found that drainage, soil type, precipitation, ecological zone, and surficial geology were significant predictors of SOC at all investigated spatial scales in the conterminous United States. These environmental variables represent the major soil-forming factors (Jenny, 1941), or "scorpan" factors (McBratney et al., 2003), driving the spatial heterogeneity of SOC. The systematic reviews of Minasny et al. (2013), Wiesmeier et al. (2019), and Lamichhane et al. (2019) listed these variables as key SOC predictors that control SOC spatial variability. Vasques et al. (2012) showed inconsistent controls of environmental factors in predicting SOC stocks as the scale of environmental factors increased from 30 to 1920 m. Our findings are partially consistent with these results, and we report mathematical functions that represent the scaling behaviors of several additional environmental factors on SOC distributions.

We found that changing the scale of predictors greatly affected prediction accuracy; that is, the accuracy decreased with increasing spatial scale. Guo et al. (2019) also reported that Cubist model performance decreased while predicting SOC using 22 terrain attributes at 71 different scales ranging from 12.8 to 2304 m. They also observed that scale influences variable importance, which is consistent with our results. There is a wealth of literature on the relationship between scale and soil property predictions (e.g., Florinsky and Kuryakova, 2000; Kuo et al., 1999; McBratney, 1998; Thompson et al., 2001; Zhang and Montgomery, 1994), and which suggested that the coarsening of scale progressively diminishes the information contained, and thereby affects soil-predictor correlations and prediction accuracies (Behrens et al., 2010; Maynard and Johnson, 2014). This idea was also verified in this study.

4.2. Scaling impacts on spatial heterogeneity of SOC

We observed a linear decrease in variance and mean of SOC stocks with spatial scale. In contrast with our findings, Mishra and Riley (2015) reported a nonlinear (exponentially decreasing) relationship between the variance of SOC and spatial scale. They reported that in arctic and boreal systems, the spatial heterogeneity of SOC stocks decreased exponentially up to the 500-m spatial scale and then remained constant. There is a lack of SOC scaling studies in different systems, but the scaling behavior of soil moisture is well documented (Crow et al., 2012; Famiglietti et al., 2008; Li and Rodell, 2013). Some studies report that the variance of soil moisture follows a power-law relationship with scale (Manfreda et al., 2007; Rodriguez-Iturbe et al., 1995), while others show complex scaling behaviors (Joshi and Mohanty, 2010; Pau et al., 2014; Riley and Shen, 2014). These soil moisture scaling studies suggest that further work is needed to understand the influence of environmental factors (topography, vegetation, soil properties, and rainfall) on soil moisture to enhance the mechanistic understanding of scaling properties; we argue here that future studies should prioritize similar analyses to better understand the scaling behavior of SOC, other soil attributes, and environmental controllers.

4.3. Scaling impacts on statistical properties of SOC

One way to represent the spatial heterogeneity of soil properties in ESMs could be to relate their statistical properties to the mean state. For example, several soil moisture studies examined the relationships among soil moisture mean and higher-order statistics, such as skewness and kurtosis (Famiglietti et al., 2008; Li and Rodell, 2013; Riley and Shen, 2014; Ryu and Famiglietti, 2005). Results from these studies suggest that mean soil moisture is often related to its skewness and kurtosis. However, they have different functional forms that depend on various ecosystem properties and scales. Mishra and Riley (2015) reported moderate but statistically significant linear relationships among the mean and higher-order moments of SOC (i.e., variance, skewness, and kurtosis). Consistent with these results, our study also found that mean SOC had linear relationships with variance, skewness, and CV and its standard deviation across scales ($R^2 > 0.96$). However, the linear relationship was moderate with skewness and weak with kurtosis. These results suggest that with the known average value of SOC stocks in an area at a given spatial scale, the statistical distribution of SOC stocks could be predicted using linear functions.

4.4. Implications for ESMs and DSM

Current land models use a nested sub grid hierarchy approach to represent land-surface heterogeneity (Koven et al., 2013; Lawrence et al., 2012; Tang et al., 2013). In this approach, the model grid cells are divided into non-spatially explicit land units, such as natural

vegetation, lakes, urban, glaciers, and crops. We demonstrated that this type of representation cannot characterize the environmental controls and scaling properties of SOC; therefore, rectifying this problem would require substantial restructuring of the model's sub-grid hierarchy. One potential application of the relationships we developed in this study could be to apply them with coarse-resolution ESM results to generate fine-scale spatial heterogeneity parameters of SOC that are more representative of the natural landscape.

Currently, most ESM land models have a spatial scale of 50 or 100 km. In the next 5–10 years, ESM land models will likely function within the spatial resolutions we identify in our study as being representative of the observed SOC landscape heterogeneity (i.e., between 100 m and 50 km). As model resolution becomes finer in the next generation of ESMs, datasets such as the one we describe in this study will be critical for model benchmarking. We note that many environmental factors that we found significant at various scales are not represented in current land models. However, representing these factors in future land model developments could improve the prediction and understanding of SOC dynamics.

Digital mapping of soil carbon utilizes a wide range of SOC predictors; however, there are no fixed rules for selecting an appropriate scale at which SOC should be mapped. Our study tested nine different scales ranging from 100 m to 50 km for SOC predictions, and we believe our results could inform the pedometrics and DSM community about scale-dependent environmental variable selection and appropriate scale considerations for spatial predictions. The scaling knowledge relationships we developed here could benefit the modeling community by identifying scale-dependent SOC predictors for research and management applications. Moreover, our study highlighted some knowledge gaps regarding scaling issues for development and verification of research relevant to the soil mapping and modeling communities.

5. Conclusions

Understanding the causes and consequences of spatial heterogeneity in ecosystem function is challenging. We modeled the spatial relationships of observed SOC stocks and their prediction variables at nine different spatial scales, ranging from 100 m to 50 km, by using machine learning-based regression rules and quantified the scaling behavior of SOC stocks across the conterminous United States. Key SOC predictors at each prediction scale were identified, and their strengths of influence were quantified. Based on the results, the following conclusions can be drawn:

- SOC distribution in the study area was highly variable (CV = 157%) with a mean and standard deviation of 95 and 150 Mg ha^{-1}, respectively.
- Out of the 31 predictors used, only 13 were significant at all scales. Almost all topographic variables were significant at scales < 250 m; however, precipitation-, land-use- and land-cover-, and soil-related variables were significant at all scales.
- Topographic variables had higher importance at finer scales, whereas climatic variables were more important at coarser scales. Specifically, at a spatial scale of 100 m, close to 50% of the variables in the MLR prediction function were related to topography, and almost 50% were related to climate at 50 km. Similarly, at setting condition rules, climate variables had the least impact at 100 m (RI: 4%) and the most impact at 50 km (RI: ~30%).
- The strengths of SOC predictors depend upon spatial scale or grid size. For example, the largest influence of NDVI, which was significant at scales ≤1 km, was at 500 m. Similarly, SAGAWI was a key predictor at scales up to 5 km, showing a maximum influence at 1 km. The strengths of PPT and PET increased with scale, with PPT showing positive and PET showing negative influences on SOC distribution.
- Predicted SOC mean and variance decreased linearly with scale (R^2 > 0.80), and both mean, and variance had a strong, positive linear relationship (R^2 = 0.96). Mean SOC also had strong linear relationships with higher order moments.
- The scaling impact of predictors was more obvious at a specific SOC range (i.e., < 3.5 and > 4.5 log [Mg SOC ha^{-1}]); the influence was less pronounced between these SOC values.
- Prediction model performance decreased with spatial scale.

Acknowledgements

We thank the editor and three anonymous reviewers for their valuable suggestions to improve this manuscript. We thank Amanda Ramcharan for providing some environmental covariate data. Contributions from U. Mishra were supported through a grant from U.S. Department of Energy under Argonne National Laboratory contract DE-AC02-06CH11357. W.J. Riley was supported by the U.S. Department of Energy under contract DE-AC02-05CH11231 with Lawrence Berkeley National Laboratory as part of the Regional & Global Modeling Analysis program of the RUBISCO SFA. Mention of tradenames or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2020.114472.

References

Adhikari, K., Hartemink, A.E., 2015. Digital mapping of topsoil carbon content and changes in the Driftless Area of Wisconsin, USA. Soil Sci. Soc. Am. J. 79 (1), 155–164.
Adhikari, K., Hartemink, A.E., Minasny, B., Bou Kheir, R., Greve, M.B., Greve, M.H., 2014. Digital mapping of soil organic carbon contents and stocks in Denmark. PLoS One 9 (8), e105519.
Adhikari, K., Kheir, R.B., Greve, M.B., Bøcher, P.K., Malone, B.P., Minasny, B., McBratney, A.B., Greve, M.H., 2013. High-resolution 3-D mapping of soil texture in Denmark. Soil Sci. Soc. Am. J. 77, 860–876.
Adhikari, K., Owens, P.R., Libohova, Z., Miller, D.M., Wills, S.A., Nemecek, J., 2019. Assessing soil organic carbon stock of Wisconsin, USA and its fate under future land use and climate change. Sci. Total Environ. 667, 833–845.
Anderegg, W.R.L., Trugman, A.T., Bowling, D.R., Salvucci, G., Tuttle, S.E., 2019. Plant functional traits and climate influence drought intensification and land–atmosphere feedbacks. Proc. Natl. Acad. Sci. 116 (28), 14071–14076.
Batjes, N.H., 2016. Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. Geoderma 269, 61–68.
Behrens, T., Zhu, A.X., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155 (3–4), 175–185.
Biswas, A., Si, B.C., 2011. Revealing the controls of soil water storage at different scales in a hummocky landscape. Soil Sci. Soc. Am. J. 75 (4), 1295–1306.
Blöschl, G., Sivapalan, M., 1995. Scale issues in hydrological modelling: a review. Hydrol. Process. 9 (3–4), 251–290.
Bonfatti, B.R., Hartemink, A.E., Giasson, E., Tornquist, C.G., Adhikari, K., 2016. Digital mapping of soil carbon in a viticultural region of Southern Brazil. Geoderma 261, 204–221.
Burke, E.J., Hartley, I.P., Jones, C.D., 2012. Uncertainties in the global temperature change caused by carbon release from permafrost thawing. CRYOSPHERE 6 (5), 1063–1076.
Burt, R., 2004. Soil Survey Laboratory Methods Manual. Soil Survey Investigations Report No. 42. Department of Agriculture, Natural Resources Conservation Service.
Carvalhais, N., Forkel, M., Khomik, M., Bellarby, J., Jung, M., Migliavacca, M., Mu, M., Saatchi, S., Santoro, M., Thurner, M., Weber, U., Ahrens, B., Beer, C., Cescatti, A., Randerson, J.T., Reichstein, M., 2014. Global covariation of carbon turnover times with climate in terrestrial ecosystems. Nature 514 (7521), 213–217.
Clark, J.S., Bell, D.M., Hersh, M.H., Kwit, M.C., Moran, E., Salk, C., Stine, A., Valle, D., Zhu, K., 2011. Individual-scale variation, species-scale differences: inference needed to understand diversity. Ecol. Lett. 14 (12), 1273–1287.
Crow, W.T., Berg, A.A., Cosh, M.H., Loew, A., Mohanty, B.P., Panciera, R., de Rosnay, P., Ryu, D., Walker, J.P., 2012. Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products. Rev. Geophys. 50 (2).
Das, N.N., Mohanty, B.P., 2008. Temporal dynamics of PSR-based soil moisture across spatial scales in an agricultural landscape during SMEX02: a wavelet approach. Remote Sens. Environ. 112 (2), 522–534.
Dorji, T., Odeh, I.O.A., Field, D.J., Baillie, I.C., 2014. Digital soil mapping of soil organic carbon stocks under different land use and land cover types in montane ecosystems,

Eastern Himalayas. For. Ecol. Manage. 318, 91–102.
Famiglietti, J.S., Ryu, D., Berg, A.A., Rodell, M., Jackson, T.J., 2008. Field observations of soil moisture variability across scales. Water Resour. Res. 44 (1).
FAO and ITPS, 2018. Global Soil Organic Carbon Map (GSOCmap), FAO, Rome. Pp 162.
Florinsky, I.V., Kuryakova, G.A., 2000. Determination of grid size for digital terrain modelling in landscape investigations—exemplified by soil moisture distribution at a micro-scale. Int. J. Geogr. Inf. Sci. 14 (8), 815–832.
Follett, R.F., 2001. Soil management concepts and carbon sequestration in cropland soils. Soil Tillage Res. 61 (1), 77–92.
Friedlingstein, P., Andrew, R.M., Rogelj, J., Peters, G.P., Canadell, J.G., Knutti, R., Luderer, G., Raupach, M.R., Schaeffer, M., van Vuuren, D.P., Le Quéré, C., 2014. Persistent growth of CO2 emissions and implications for reaching climate targets. Nat. Geosci. 7 (10), 709–715.
Gebremichael, M., Rigon, R., Bertoldi, G., Over, T., 2009. On the scaling characteristics of observed and simulated spatial soil moisture fields. Nonlinear Processes Geophys. 16 (1), 141.
Guo, Z., Adhikari, K., Chellasamy, M., Greve, M.B., Owens, P.R., Greve, M.H., 2019. Selection of terrain attributes and its scale dependency on soil organic carbon prediction. Geoderma 340, 303–312.
Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km–global soil information based on automated mapping. PLoS One 9 (8) e105992-e105992.
Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, New York.
Joshi, C., Mohanty, B.P., 2010. Physical controls of near-surface soil moisture across varying spatial scales in an agricultural landscape during SMEX02. Water Resour. Res. 46 (12).
Kachanoski, R.G., de Jong, E., 1988. Scale dependence and the temporal persistence of spatial patterns of soil water storage. Water Resour. Res. 24 (1), 85–91.
Koven, C.D., Riley, W.J., Subin, Z.M., Tang, J.Y., Torn, M.S., Collins, W.D., Bonan, G.B., Lawrence, D.M., Swenson, S.C., 2013. The effect of vertically resolved soil biogeochemistry and alternate soil C and N models on C dynamics of CLM4. Biogeosciences 10 (11), 7109–7131.
Kuo, W.-L., Steenhuis, T.S., McCulloch, C.E., Mohler, C.L., Weinstein, D.A., DeGloria, S.D., Swaney, D.P., 1999. Effect of grid size on runoff and soil moisture for a variable-source-area hydrology model. Water Resour. Res. 35 (11), 3419–3428.
Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., Walter, C., 2014. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma 213, 296–311.
Lagacherie, P., McBratney, A., Voltz, M., 2007. Digital Soil Mapping: An Introductory Perspective. Elsevier, Amsterdam, The Netherlands.
Lamichhane, S., Kumar, L., Wilson, B., 2019. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: a review. Geoderma 352, 395–413.
Lawrence, P.J., Feddema, J.J., Bonan, G.B., Meehl, G.A., O'Neill, B.C., Oleson, K.W., Levis, S., Lawrence, D.M., Kluzek, E., Lindsay, K., Thornton, P.E., 2012. Simulating the biogeochemical and biogeophysical impacts of transient land cover change and wood harvest in the community climate system model (CCSM4) from 1850 to 2100. J. Clim. 25 (9), 3071–3095.
Li, B., Rodell, M., 2013. Spatial variability and its scale dependency of observed and modeled soil moisture over different climate regions. Hydrol. Earth Syst. Sci. 17 (3), 1177–1188.
Manfreda, S., McCabe, M.F., Fiorentino, M., Rodríguez-Iturbe, I., Wood, E.F., 2007. Scaling characteristics of spatial patterns of soil moisture from distributed modelling. Adv. Water Resour. 30 (10), 2145–2150.
Maynard, J.J., Johnson, M.G., 2014. Scale-dependency of LiDAR derived terrain attributes in quantitative soil-landscape modeling: effects of grid resolution vs. neighborhood extent. Geoderma 230–231, 29–40.
McBratney, A.B., 1998. Some considerations on methods for spatially aggregating and disaggregating soil information. Nutr. Cycl. Agroecosyst. 50 (1), 51–62.
McBratney, A.B., Santos, M.L.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52.
Miller, B.A., Koszinski, S., Wehrhan, M., Sommer, M., 2015. Impact of multi-scale predictor selection for modeling soil properties. Geoderma 239–240, 97–106.
Minasny, B., McBratney, A.B., 2008. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. Chem. Intelligent Lab. Syst. 94 (1), 72–79.
Minasny, B., McBratney, A.B., Malone, B.P., Wheeler, I., 2013. Chapter One - Digital Mapping of Soil Carbon. In: Sparks, D.L. (Ed.), Advances in Agronomy. Academic Press, pp. 1–47.
Mishra, U., Drewniak, B., Jastrow, J.D., Matamala, R.M., Vitharana, U.W.A., 2017. Spatial representation of organic carbon and active-layer thickness of high latitude soils in CMIP5 earth system models. Geoderma 300, 55–63.
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 Sci. Soc. Am. J. 74 (3), 906–914.
Mishra, U., Riley, W.J., 2015. Scaling impacts on environmental controls and spatial heterogeneity of soil organic carbon stocks. Biogeosciences 12 (13), 3993–4004.
Muster, S., Riley, W.J., Roth, K., Langer, M., Cresto Aleina, F., Koven, C.D., Lange, S., Bartsch, A., Grosse, G., Wilson, C.J., Jones, B.M., Boike, J., 2019. Size distributions of arctic waterbodies reveal consistent relations in their statistical moments in space and time. Front. Earth Sci. 7 (5).
Pan, M., Cai, X., Chaney, N.W., Entekhabi, D., Wood, E.F., 2016. An initial assessment of SMAP soil moisture retrievals using high-resolution model simulations and in situ observations. Geophys. Res. Lett. 43 (18), 9662–9668.
Pau, G.S.H., Bisht, G., Riley, W.J., 2014. A reduced-order modeling approach to represent subgrid-scale hydrological dynamics for land-surface simulations: application in a polygonal tundra landscape. Geosci. Model Dev. 7 (5), 2091–2105.
Paustian, K., Collins, H., Paul, E.A., 1997. Management controls on soil carbon. In: Paul, E.A., Paustian, K., Elliott, E.T., Cole, C.V. (Eds.), Soil Organic Matter in Temperate Agroecosystems: Long-term Experiments in North America. CRC Press, Boca Raton, Florida, pp. 15–49.
Quinlan, J.R., 1993. C4. 5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, USA.
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 Sci. Soc. Am. J. 82 (1), 186–201.
Riley, W.J., Shen, C., 2014. Characterizing coarse-resolution watershed soil moisture heterogeneity using fine-scale simulations. Hydrol. Earth Syst. Sci. 18 (7), 2463–2483.
Rodriguez-Iturbe, I., Vogel, G.K., Rigon, R., Entekhabi, D., Castelli, F., Rinaldo, A., 1995. On the spatial organization of soil moisture fields. Geophys. Res. Lett. 22 (20), 2757–2760.
Ryu, D., Famiglietti, J.S., 2005. Characterization of footprint-scale surface soil moisture variability using Gaussian and beta distribution functions during the Southern Great Plains 1997 (SGP97) hydrology experiment. Water Resour. Res. 41 (12).
Sequeira, C.H., Wills, S.A., Seybold, C.A., West, L.T., 2014. Predicting soil bulk density for incomplete databases. Geoderma 213, 64–73.
Shen, C., Riley, W.J., Smithgall, K.R., Melack, J.M., Fang, K., 2016. The fan of influence of streams and channel feedbacks to simulated land surface water and carbon dynamics. Water Resour. Res. 52 (2), 880–902.
Soil Survey Staff and T. Loecke, 2016. Rapid Carbon Assessment: Methodology, Sampling and Summary, U.S. Department of Agriculture, Natural Resources Conservation Service.
Sun, X.-L., Wang, Y., Wang, H.-L., Zhang, C., Wang, Z.-L., 2019. Digital soil mapping based on empirical mode decomposition components of environmental covariates. Eur. J. Soil Sci. 70 (6), 1109–1127.
Tang, J.Y., Riley, W.J., Koven, C.D., Subin, Z.M., 2013. CLM4-BeTR, a generic biogeochemical transport and reaction module for CLM4: model development, evaluation, and application. Geosci. Model Dev. 6 (1), 127–140.
Thompson, J.A., Bell, J.C., Butler, C.A., 2001. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. Geoderma 100 (1), 67–89.
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.
Ugbaje, S.U., Reuter, H.I., 2013. Functional digital soil mapping for the prediction of available water capacity in Nigeria using legacy data. Vadose Zone J. 12 (4).
Vasques, G.M., Grunwald, S., Myers, D.B., 2012. Influence of the spatial extent and resolution of input data on soil carbon models in Florida, USA. J. Geophys. Res. Biogeosci. 117 (G4).
Viscarra Rossel, R.A., Lee, J., Behrens, T., Luo, Z., Baldock, J., Richards, A., 2019. Continental-scale soil carbon composition and vulnerability modulated by regional environmental controls. Nat. Geosci. 12 (7), 547–552.
Wadoux, A.M.J.-C., Samuel-Rosa, A., Poggio, L., Mulder, V.L., 2020. A note on knowledge discovery and machine learning in digital soil mapping. Eur. J. Soil Sci. 71 (2), 133–136.
Western, A.W., Grayson, R.B., Blöschl, G., 2002. Scaling of soil moisture: a hydrologic perspective. Annu. Rev. Earth Planet. Sci. 30 (1), 149–180.
Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., von Lützow, M., Marin-Spiotta, E., van Wesemael, B., Rabot, E., Ließ, M., Garcia-Franco, N., Wollschläger, U., Vogel, H.-J., Kögel-Knabner, I., 2019. Soil organic carbon storage as a key function of soils - a review of drivers and indicators at various scales. Geoderma 333, 149–162.
Wills, S., Seybold, C., Chiaretti, J., Sequeira, C., West, L., 2013. Quantifying tacit knowledge about soil organic carbon stocks using soil taxa and official soil series descriptions. Soil Sci. Soc. Am. J. 77 (5), 1711–1723.
Zhang, W., Montgomery, D.R., 1994. Digital elevation model grid size, landscape representation, and hydrologic simulations. Water Resour. Res. 30 (4), 1019–1028.
Zhou, Y., Wu, D., Lau, W.K.-M., Tao, W.-K., 2016. Scale dependence of land-atmosphere interactions in wet and dry regions as simulated with NU-WRF over the Southwestern and South-Central United States. J. Hydrometeorol. 17 (8), 2121–2136.