Soil Organic Carbon across Mexico and the conterminous United States (1991-2010)

Authors:

[1]Mario Guevara, [2]Carlos Arroyo, [3]Nathaniel Brunsell, [4]Carlos O. Cruz, [5]Grant Domke, [2]Julian Equihua, [6]Jorge Etchevers, [7]Daniel Hayes, [6]Tom Hengl, [4]Alejandro Ibelles, [5]Kris Johnson, [9]Ben de Jong, [10]Zamir Libohova, [1]Ricardo Llamas, [11]Lucas Nave, [4]Jose L. Ornelas, [7]Fernando Paz, [2]Rainer Ressl, [12]Anita Schwartz, [4]Arturo Victoria, [10]Skye Wills and [1]Rodrigo Vargas

Affiliations:

[1]Department of Plant and Soil Sciences, University of Delaware, Newark DE United States

[2]Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, México

[3]Department of Geography and Atmospheric Science, The University of Kansas, Lawrence, KS

[4]Instituto Nacional de Estadística y Geografía, Aguascalientes México

[5]United States Department of Agriculture, Forest Service

[6]EnvirometriX Ltd — Research, Innovation and Consultancy, Wageningen, NL

[7] Colegio de Postgraduados, Texcoco, México

[8]School of Forest Resources University of Maine, Maine United States

[9]El Colegio de la Frontera Sur, Campeche, México

[10]National Soil Resource Conservation Service, Lincoln, Nebraska

[11]The International Soil Carbon Network-University of Michigan Biological Station, Michigan

[12]Information Technologies, University of Delaware, Newark DE United States

Corresponding author:

Rodrigo Vargas, University of Delaware,

Telephone: 302-831-0605

Fax: 302-831-0605

Email: rvargas@udel.edu

Keywords: soil organic carbon, simulated annealing, spatial variability, uncertainty.

Key points:

Multi-source topsoil organic carbon prediction and prediction variance in Mexico and the conterminous United States.

Calculated stocks of 46-47 Pg of SOC (0-30cm depth, years 1991-2010) using a simulated annealing regression framework.

Predicted stocks >30% below recent global estimates that are largely based on legacy data.

Abstract

Soil Organic Carbon (SOC) information is fundamental for improving global carbon cycle modeling efforts, but discrepancies exist from country-to-global scales. We predicted the spatial distribution of SOC stocks (topsoil; 0-30 cm) and quantified modeling uncertainty across Mexico and the conterminous United States (CONUS). We used a multi-source SOC dataset (>10000 pedons, between 1991-2010) coupled with a simulated annealing regression framework that accounts for variable selection. Our model explained ~50% of SOC spatial variability (across 250m grids). We analyzed model variance, and the residual variance of six conventional pedotransfer functions for estimating bulk density (BD) to calculate SOC stocks. Two independent datasets confirmed that the SOC stock for both countries represents between 46 and 47 Pg with a total modeling variance of ±12 Pg. We report a residual variance of 10.4 ±5.1 Pg of SOC stocks against the six pedotransfer functions. When reducing training data to defined decades with relatively higher density of observations (1991-2000 and 2001-2010, respectively), model variance for predicted SOC stocks ranged between 41 and 55 Pg. We found nearly 42% of SOC across Mexico in forests and 24% in croplands; whereas 31% was found in forests and 28% in croplands across CONUS. Grasslands and shrublands stored 29 and 35% of SOC across Mexico and CONUS, respectively. We predicted SOC stocks >30% below recent global estimates that do not account for uncertainty and are based on legacy data. Our results provide Insights for interpretation of estimates based on SOC legacy data and benchmarks for improving regional-to-global monitoring efforts.

1 Introduction

Terrestrial ecosystems store >1500 Pg of soil organic carbon (SOC, approximate stock at 1m depth) worldwide, but accurate spatial representation of these stocks is needed for fully understanding the contribution of soils within the global carbon cycle (Crowther *et al.* 2017, FAO 2017). For global modeling and validating the SOC stored in terrestrial ecosystems, high-resolution gridded datasets such as the SoilGrids250m system (Hengl *et al.* 2017) are increasingly being used to describe spatial SOC patterns (Jackson *et al.* 2017; Harden *et al.* 2017) and trends (Naipal *et al*. 2018). Such datasets are also required to facilitate the formulation of reliable climate change adaptation guidelines and the establishment of regional to global carbon monitoring and information systems (Ciais *et al*. 2014, Stockmann *et al.* 2015, Vargas *et al.* 2017, Villarreal *et al.* 2018). Previous studies suggest that the greatest source of discrepancy across regional to global carbon cycling estimates is associated with the SOC pool (Jones *et al.* 2005, Jones and Fallon 2009, Murray-Tortarolo *et al.* 2016, Crowther *et al.* 2016, Tifati *et al.* 2017). Arguably, current scientific challenges associated with the discrepancy of the soil carbon pool are to quantify: 1) the size and distribution of local to regional SOC stocks at scales relevant to inform land management decisions (FAO, 2017, Banwart *et al.* 2017); 2) the amount of carbon losses from soils due to heterotrophic respiration (Bond-Lamberty *et al.* 2018), the amount of carbon removed from erosion (Naipal *et al*. 2018) or aquatic export (Tank *et al.* 2018), or changes in land use and land cover (Sanderman *et al*. 2017); and 3) the carbon emissions from impacts of past and future climate conditions (Walsh *et al.* 2017, Crowther *et al.* 2016, Delgado-Baquerizo *et al.* 2017). Solving these scientific challenges around carbon cycling requires a good understanding of different uncertainty sources around SOC datasets and SOC modeling efforts.

Major uncertainties in spatial SOC estimates that are extrapolated from points/pedons to continuous estimates across the land surface are related to several factors. These include measurement methods, data sources (SOC data and SOC environmental covariates) and their resolution and extent, the different periods of data collection, or using multiple modeling and evaluation strategies (Grunwald 2009, Stockmann *et al.*, 2013, Ogle *et al.*, 2010). Thus, there is a need for improving interoperability for compiling the best available information and describing SOC spatial variability across local to global scales (Vargas *et al.*, 2017).

Global modeling outputs for SOC represent the only estimates of SOC across large areas of the world without in situ ground observations. These global estimates rely on large datasets that combine multiple SOC data collection periods and methods for calculating SOC stocks. These inconsistencies represent a known but unquantified bias for calculating SOC stocks (Poeplau *et al.,* 2017). Thus, there is a need to test different modeling approaches across areas with high density of SOC observations to improve the accuracy, detail and reliability of global SOC estimates (Vitharana *et al.,* 2019).

The Harmonized World Soil Data Base (HWSD) or the harmonized soil property values for broadscale modeling (WISE30sec, Batjes, 2016), are probably the most commonly used datasets for spatially quantifying SOC stocks and its spatial variability patterns at the global scale (Köchy *et al.* 2015; O'Rourke *et al.* 2015). The HWSD provides the most complete global soil description from synthesizing many regional or national soil maps, but it uses a polygon-based approach that has intrinsic quality limitations such as coarse scale (e.g., >1km pixels), discrepancy between national datasets and broad categorical generalizations (Stoorvogel *et al.* 2016, Folberth *et al.* 2016). Regional to global efforts to improve the spatial representation of the global SOC pool also include those by the International Union of Soil Sciences (Arrouays *et al.* 2017), the International Soil Resource Information Centre (ISRIC, e.g., Hengl *et al.* 2014; Batjes *et al.* 2017; Hengl *et al.* 2017), the International Soil Carbon Network (ISCN, e.g., Nave and Johnson, 2012, Harden *et al.* 2017), the Land GIS project (https://landgis.opengeohub.org) and the GlobalSoilMap consortium (Arrouays *et al.* 2014; Sanchez *et al.* 2009). Another initiative is the recent call from the United Nations Food and Agricultural Organization (FAO) requesting the development of country specific frameworks for reporting continuous and spatially explicit SOC stocks and patterns (FAO, 2017). These efforts have contributed information for global estimates, along with methodologies useful for applying standardized protocols for harmonizing SOC measurements from multiple sources for SOC assessments. However, validating global SOC estimates and developing country or region specific (e.g., North America) SOC prediction frameworks are still needed for increasing knowledge by quantifying uncertainties while explaining the discrepancy of current SOC estimates (e.g., country specific to global scales).

Large discrepancies have been reported among global (Tifati *et al.* 2017) or country specific SOC estimates (Guevara *et al.* 2018). Consequently, reporting uncertainty and bias of SOC estimates will allow better parameterization of land surface models, improved local to regional monitoring baselines and informed policy and management decisions regarding SOC stocks (Viscarra Rossel *et al.* 2014). The current discrepancies around SOC estimates could be partially attributed to SOC sampling errors and bias in the SOC sampling locations, but this is information that may not be always available for improving SOC estimates. Other sources of errors and spatial artifacts are related with the use of different measurement methods (or analytical techniques) for quantifying SOC stocks, lack of information on bulk density or rock fragments, and different methods to apply pedotransfer functions may generate contrasting results. For predictive SOC mapping (McBratney, *et al.* 2003), the quality of SOC training data and the quality of SOC environmental covariates represent a potential source of uncertainty that will propagate to final predictions. Thus, increasing information about how and when SOC data is collected and selecting only the most effective SOC environmental covariates (i.e., from remote sensing, geomorphometry, climatology surfaces, thematic maps) will reduce the propagation of errors on further modeling efforts. Quantifying the errors from inputs and models that influence SOC predictions and identifying how they are spatially distributed will benefit planning for future SOC sampling strategies, by assuming that a larger sample is required across areas with higher discrepancies and modeling bias (FAO, 2017, Heuvelink, 2014). Optimizing soil sampling strategies is constantly required to validate/calibrate SOC predictions and reduce their uncertainties across unsampled areas.

North America is a region characterized by a long history of soil data collection that has produced unprecedented information of SOC. For example, SOC predictions and estimates across Mexico and the CONUS are available from a variety of methods and in different formats. These include soil type polygon maps, field observations and reflectance spectroscopy analysis, as well as global SOC variability gridded surfaces based on environmental correlation methods (e.g., Bliss *et al.* 2014; Hengl *et al.* 2014; Wijewardane *et al.* 2016, Hengl *et al.* 2017). Further examples include the use of linear geostatistics for the interpolation of SOC across Mexico (Cruz-Cárdenas *et al.* 2014), and SOC modeling efforts across the United States (Padarian *et al.* 2015). For increasing prediction accuracy of SOC models, flexible statistics (e.g., machine learning) have been proposed to better predict non-linear relationships between SOC observational data and their environmental predictors at global and continental scales (Hengl *et al.* 2017, Ramcharan et al. 2017). Thus, SOC environmental covariates (i.e., surrogates of climate, biota, topography and geology) and observational data can be coupled with machine learning algorithms to improve the representation of spatial variability and uncertainty in SOC stocks. Reducing uncertainties from different sources of information, increasing data-model agreement, and simplifying model complexity (by assessing variable importance and removing not informative SOC environmental covariates) are required to enable the fine-scale monitoring of SOC stocks across countries and regions of the world where no such information is otherwise available (Viscarra-Rossel *et al. 2014*, de Gruijter *et al.* 2016; Minasny *et al.* 2017).

In this study, we quantified the spatial variability and associated uncertainty of SOC stocks across different land use categories of CONUS and Mexico; two countries with rich information of SOC measurements. Previous analyses have shown large discrepancies in SOC stocks (0-30 cm, ranging from ~38 to ~92 Pg of SOC) derived from country-specific or global SOC estimates (Lajtha *et al.* 2018, Hengl *et al*, 2017, Paz-Pellat *et al.* 2016, Bliss *et al.* 2014, Wieder *et al.* 2014). Our main goal was to generate a spatial predictive model of SOC variability for the top 30 cm depth at 250 m spatial resolution across both countries with information collected between 1991 and 2010.

We asked the following interrelated questions: 1) Which are the best SOC environmental covariates (i.e., prediction factors increasing SOC modeling accuracy) across Mexico and CONUS? 2) How much variation in SOC can machine learning methods (e.g., tree-based, kernel based, probabilistic-based) explain across this region using repeated cross-validation? 3) What is the SOC variance associated with multiple calculation methods for estimating SOC stocks and what is the variance associated to multiple model predictions? and 4) What are the sensitivities of these predictions associated to different training datasets (i.e., decadal information from different collection periods; 1991-2000 and 2001-2010)? The value of considering different collection periods is to explore the sensitivity of model predictions associated to different training datasets and provide insights for better interpretation of decadal changes in SOC stocks. In summary, this study provides benchmark information about how SOC spatial distribution is constrained by the soil forming environment (i.e., climate, biota, topography and geology), and quantifies the variance (spatial and temporal) from using multiple SOC observational datasets.

2 Datasets and methods

We followed a digital soil mapping strategy (Figure 1) for the prediction of the spatial variability of SOC across both countries. Digital soil maps are generated using field and laboratory observational methods coupled with environmental data through quantitative relationships (McBratney *et al.* 2003, Minasny *et al.* 2008). We assumed that the spatial variability of SOC (represented by observational data) can be predicted across large geographical (unsampled) areas as a function of soil forming factors (climate, biota, topography and geology, Jenny, 1941). These factors (surrogates of the soil forming environment) are represented through three main sources of information: remote sensing sensors, gridded climatology products (e.g., precipitation and temperature) and digital terrain analysis (i.e., geomorphometry; see Pike *et al.* 2009 and Wilson 2012, McBratney, *et al.* 2003).

Figure 1 Flow diagram of the proposed methodology to predict the spatial variability of SOC stocks across Mexico and CONUS. Orange folders indicate the SOC data sources used for training and pink folders indicate SOC data sources used for validating. These sources were harmonized with the SoilGrids250m covariates. White folders indicate the main results of this methodology. Gray ovals indicate main methodological steps. CV: 5-fold cross validation, QRF: quantile regression forest, OK: Ordinary Kriging, BD: bulk density.

2.1 SOC observational data

Legacy SOC estimates across CONUS were obtained from the International Soil Carbon Network (ISCN latest version 2018, >18 000 pedons available, Harden *et al.* 2017). Data from Mexico was provided by the Instituto Nacional de Estadística y Geografía (INEGI, SERIES 1 & 2; n >65 000 pedons available, Krasilnikov *et al.* 2013). We used only the observations collected between 1991 and 2010 to minimize confounding factors (related to potential changes in the SOC pool; n = 10385, Figure 2). We considered all soil horizons containing upper and lower soil depth limit information. The combination of using soil depth continuous functions (Bishop *et al.* 1999; Malone *et al.* 2009) and deriving the weighted average (by depth) from the first sampled soil horizon at 0 cm depth to all soil horizons sampled within the first 30 cm of soil depth, allowed aggregating irregular soil horizons for calculating SOC stocks across both countries. The weights for calculating these stocks were selected defining the proportion of each horizon within this 0-30 cm interval of soil depth.

Most contributors (across CONUS) and INEGI (in Mexico) considered (or adapted) the United States Department of Agriculture Soil Taxonomy guidelines for interpreting soil surveys including SOC and other soil variables (Soil Survey Staff. 1999). For CONUS, the ISCN database provides a harmonized compilation from many contributors (e.g., Natural Resources Conservation Service, United States Geological Survey and site-specific research or academic groups; Harden *et al.* 2017). However, the largest contributor for this curated dataset is the United States Department of Agriculture Natural Resource Conservation Service, where the SOC concentration was mainly obtained by the Walkley-Black technique (Soil Survey Staff. 2014). All samples for Mexico were systematically collected and analyzed by INEGI (INEGI, 2014, Krasilnikov *et al.* 2013) and SOC concentration was also measured using the Walkley-Black technique (IUSS-WRB-FAO, 2014). Potential error propagation from the use of different methods to calculate SOC using information collected over long periods of time (before 1991) is beyond the scope of this study. We only considered the sensitivity of SOC models (i.e., model outputs) to variations in training data and inputs derived from different pedotransfer functions for estimating bulk density (section 2.6).

Figure 2 Distribution and descriptive statistics of available datasets. The point map shows the spatial sampling locations of available data for the period 1991-2010 (1991-2000 and 2001-2010) (a). The colored histograms are representing the statistical distribution of all datasets (i.e., combined CONUS and Mexico information) (b). The variogram (relation between distance and variance of observed values) and variogram parameters (nugget, sill and range) are representing the spatial structure captured with available data (c). *Ste* is a Stein model parameterization (and its associated Kappa value) for the covariance function between all pairs of points separated by distance units (range in meters) defining the spatial structure of SOC available datasets.

2.2 Calculation of SOC stocks

SOC stocks were derived by a linear combination of soil depth (0-30cm), coarse fragments (CF) data, SOC concentration (%), and soil bulk density (BD) following the method proposed by Nelson and Sommers (1982) as implemented by Hengl (2017). In general, CF across CONUS was measured in the field considering soil fragments >2 mm and direct gravimetric mass-methods. In Mexico, CF was also measured in field (considering soil fragments >2 mm), but expressed as percentage of gravel, stones and pebbles. For the CONUS dataset, the ISCN has calculated SOC stocks using both modeled (i.e., incomplete) and non-modeled (i.e., complete) information about the aforementioned variables (Harden *et al.* 2017). We only used information flagged as 'complete' by the ISCN, so no model or pedotransfer functions were used for estimating BD and consequently SOC stocks. In Mexico, BD was estimated in the field using soil type maps, soil texture, soil organic matter and soil structure following international soil mapping guidelines (FAO, 2006, p 51, Table 58). These guidelines are based in a rule-based approach originally described in the German soil-mapping guidelines (Ad-hoc-AG Boden, 2005), and have been applied to the collection and analysis of soils across Mexico (Siebe *et al*., 2006) and for the contribution of Mexico to the United Nations SOC map (FAO and ITPS, 2018).

2.3 The environmental covariate space

For spatially representing soil forming factors (Jenny, 1941) we used environmental covariates from the SoilGrids250m system (Hengl *et al.* 2017). This dataset represents over 150 variables of environmental gridded data including terrain derivatives from a digital elevation model (DEM), the enhanced vegetation index (EVI), climate (precipitation and land surface temperature) and other soil related gridded variables (Figure 3, Supplementary Table S1). This covariate space is representative for the analyzed period of time (1991-2010) and is described in detail by previous publications (Hengl, *et al.* 2017, Reuter and Hengl, 2012). We extracted this global information within the geographical limits of the NALCMS (North American Land Change Monitoring System, 77% of land use classification accuracy) at 250m spatial resolution (NRCan/CCRS-USGS-INEGI-CONABIO-CONAFOR, 2005).

2.4 Recursive feature elimination

We first performed a variable reduction strategy using of a recursive feature elimination technique (Kuhn *et al.* 2008) and multiple models were fitted repeatedly using all possible combinations of highly ranked predictors. Predictors were ranked using as indicator the cross-validated prediction error of a Random Forest tree ensemble. We selected the Random Forest as our overall accuracy indicator method because it showed the highest predictive capacity compared against different machine learning algorithms tested in our modeling selection strategy (Supplementary Figure S1). This method is based on bagging predictors and the combination of multiples regression trees derived from different random data subsets (Breiman, 2001). Each model grows with the number of trees for minimizing the prediction variance. Model parameters to define the number of predictors and subsets on each regression tree were automatically selected by the means of 10-fold cross validation (Figure 1). Cross validation is a re-sampling technique that we used for maximizing the accuracy of results while obtaining a robust and stable prediction error estimate used for further selecting the most informative predictors. In addition, Random Forests uses an out-of-bag cross-validation form for assessing the relevance of each predictor in the model. Thus, multiple lists of the "best" predictors are generated from each Random Forest model realization in the recursive feature elimination framework. This provides a probabilistic assessment to determine the best predictors to retain at the end of the algorithm (Kuhn *et al.* 2008). After a 5-times repeated 5-fold cross validation for the recursive feature elimination technique (to account for the model sensitivity to data variations and reduce overfitting), we selected the first 25 environmental covariates for SOC.

2.5 Simulated annealing

The 25 environmental covariates selected from the recursive feature elimination analysis were used on a simulated annealing regression framework for predicting SOC stocks (Kuhn and Johnson, 2013). Simulated annealing is a well-known optimization framework for soil sampling designs (Groenigen and Stein, 1998, Minasny and McBratney, 2006, Szatmári *et al.* 2018) and for validating digital soil maps (Biswas and Zhang, 2018). Simulated annealing is a framework from statistical mechanics and combinatorial optimization problems (Firstpatrick, *et al.* 1983) that here we apply for maximizing the feature selection and prediction accuracy of SOC relevant environmental covariates.

In a simulated annealing framework, used for prediction (i.e., regression), a global search is performed and random perturbations are induced to the dataset for identifying the variables that are more sensitive to data variations and that have higher prediction capacity for the target variable (i.e., SOC). We used the cross validated Random Forest error as indicator to analyze the effect of such perturbations. This process is constantly repeated, and many iterations are produced in a global learning search that should in theory result in better solutions (Kuhn *et al.* 2008). We used the Random Forest regression algorithm within the simulated annealing framework to improve the probability of detecting the main drivers of SOC spatial patterns. After a 5-times repeated 5-fold cross validation, the entire data set is used for generating a model in the last execution of the simulated annealing global search. This model is built on the predictor subset that is associated with the optimal number of iterations determined by the cross-validation resampling technique (Kuhn *et al.* 2008). We used the final model of the simulated annealing framework for making predictions across 250m grids reporting the first five ranked environmental covariates of each generated model. These environmental covariates were selected because they contributed the most to reducing the error in the global search of the simulated annealing iterative (i.e., tree ensemble learning) process.

Figure 3. Visualization of covariates across the political boundaries between California and Oregon in western CONUS. Land surface temperature (a); precipitation (b); precipitable water vapor (c); and a digital elevation model (d); see Supplementary Table S1 for detailed description and sources of these variables. Gray histograms represent variation across latitude or longitude.

2.6 Uncertainty analysis

We represented uncertainty of our modeling approach as the sensitivity of prediction models to multiple data inputs. We first explored the residual variance of our SOC training data against six SOC stocks calculated using six BD pedotransfer functions. Then, we analyzed the spatial structure of these residuals using Geostatistics, and computed a model residual error against fully independent SOC datasets. Finally, we computed the modeling SOC prediction variance and the full quantile response of residuals (from independent datasets and from the BD variance) to the highest ranked environmental covariates (Figure 1). We postulate that estimates of SOC fall within a range of errors, and it is therefore important to account for variation in model inputs and model outputs. Our main goal was to quantify the variability range around predicted SOC stocks using multiple uncertainty indicators.

2.6.1 Pedotransfer functions for bulk density variance

We predicted SOC stocks at the pedon locations available in the WoSIS system (Batjes *et al.* 2017). We calculated the residual variance of our predictions and independent SOC stocks. These independent stocks were calculated using the WoSIS SOC concentration data (%), and six conventional pedotransfer functions for estimating BD. This resulted in six different SOC stocks estimates from the following pedotransfer functions:

- Saini (1966): BD = 1.62 - 0.06 * OM , Jeffrey (1970): BD = 1.482 - 0.6786 * (log(OM)),
- Adams (1973): = 100 / (OM /0.244 + (100 - OM)/2.65),
- Drew (1973): BD = 1 / (0.6268 + 0.0361 * OM),
- Honeysett and Ratkowsky (1989): BD = 1/(0.564 + 0.0556 * OM),
- Grigal et al. (1989): BD = 0.669 + 0.941 * exp(1)^(-0.06 * OM).

The OM= organic matter content was estimated as OM=SOC concentration * 1.724. These functions applied to the WoSIS data were selected because they were developed for a variety of soil weathering environments using multiple analytical techniques for measuring SOC and because they were recently proposed for the development of the United Nations global SOC map (Yigini et al., 2018). In addition, the WoSIS data has been curated under different protocols for controlling data quality and global interoperability standards (Batjes et al. 2017) and thus, this residual variance will allow us to have an idea of possible dispersion of values around the SOC calculated stocks.

2.6.2 Independent datasets for model prediction

We calculated model residuals against two fully independent datasets across both countries (n=9239). Across CONUS we used 6179 SOC estimates (2010) from the Rapid Carbon Assessment Project (RaCA, Soil Survey Staff and Loecke, 2016; Wijewardane *et al.* 2017) and 3060 (2009-2011) SOC estimates from top soil samples extracted from the Mexican National Forest and Soils Inventory of the Mexican Forest Service (2009-2011, Supplementary Figure S2). These independent datasets have been collected using different sampling designs and using different SOC calculation methods from our initial training dataset (INEGI and ISCN). The residual analysis against these independent datasets provides an overall measure of the models' sensitivity to multiple SOC data sources.

2.6.3 Spatial autocorrelation of model residuals

We compared the spatial structure (i.e., spatial autocorrelation) of model residuals using linear geostatistics. The spatial structure accounts for the variance of values as a response of the geographical distance (e.g., meters) between SOC sampling points (Figure 2a). The spatial structure of a soil property can be quantified using variograms (a graphical method for modeling the relationship between distance between points and the variance of their values, Figure 2c) and the variogram parameters: nugget (uncorrelated variance), sill (spatially-autocorrelated variance) and range (distance to the maximum variance) as explained previously (Oliver and Webster, 2014). We used automated variogram fitting (Hiemstra, *et al.* 2008) for calculating the nugget:sill ratio. The nugget is an uncorrelated component of soil variation that cannot be explained by our data, it depends on the calculation methods, the sampling resolution, and the spatial variability of SOC. The sill is the distance between the nugget and the variance stabilization (y - axis) point while increasing distance (x – axis). The range is the distance of the variance stabilization point. As in previous studies (Cruz-Cárdenas *et al*. 2014), a nugget:sill of <0.25 was considered evidence of a strong spatial dependence, a relationship between 0.25 and 0.75 was considered a moderate spatial dependence, and a relationship > 0.75 was considered a weak spatial dependence (Cambardella *et al.* 1994). We then used these variogram parameters to generate error maps by the means of Ordinary Kriging (Oliver and Webster, 2014), as explained earlier (Hengl *et al.* 2004), accounting for the potential spatially autocorrelation of the model residuals.

2.6.4 Model residual limits

For analyzing the model-based uncertainty, we estimated the quantile conditional response of the aforementioned modeling residuals to the 'best' environmental covariates identified by our simulated annealing framework aiming to estimate model prediction limits. The main purpose of estimating model prediction limits is to identify the variance from the most probable predicted SOC stock for each pixel across the 250m grids. For this purpose, we used the quantile regression forest approach, which is a variant of Random Forests. This method is able to: a) maintain the value of all observations in each node for each tree, not just their mean (as is the case of Random Forests) and b) assesses the quantile conditional distribution at each predicted location (pixel). This method has the assumption that the full conditional estimated response is not different from the mean of the training dataset (Meinshausen 2006). This method allowed us to quantify the maximum possible range of SOC prediction limits (e.g., 95%) given available data and available environmental covariates.

All analyses were performed in R (R Core Team 2018) and were repeated using subsets of available SOC data for the period 1991 to 2000 (n=4877) and for the period 2001-2010 (n=5508) in order to identify possible sensitivities of model predictions associated with defined (i.e., decadal) variations in training datasets.

3 Results

3.1 Descriptive statistics

The harmonized SOC training dataset across both countries (n = 10 385, 1991-2010, Figure 2a) showed a right skewed distribution and most estimates were between 0 and 10 kg m^{-2} up to a maximum of 87.9 kg m^{-2}. While the Mexican dataset dominates the 0-10 kg m^{-2} range, the CONUS dataset has larger SOC values (>10 kg m^{-2}) (Figure 2b). We used a logarithmic transformation (i.e., log(1+x)) of the combined (Mexico-CONUS) dataset to reduce the skewed distribution for further analysis. The combined dataset shows a nugget:sill ratio of 0.55, suggesting a moderate spatial autocorrelation of its values, a sill of 0.9 and a nugget of 0.5 (units in log(kg m^{-2} + 1), Figure 2c).

3.2 Recursive feature elimination

The five times repeated 5-fold cross validation (applied to the recursive variable elimination framework) showed errors of 1.7 (R^2 = 0.30), 2.0 (R^2= 0.27) and 2.6 (R^2=0.34) kg m^{-2} for the models 1991-2010 (n=10385), 1991 to 2000 (n=4877) and 2001-2010 (n=5508), respectively. For the years 1991-2010, the highest ranked environmental covariates for predicting SOC were: the DEM and topographic terrain attributes (the wetness index, the valley bottom flatness index and the valley depth index) and a remotely sensed precipitable water vapor estimate. For the years 1991-2000, the highest ranked environmental covariates were: the land surface temperature, the DEM, the wetness index, the valley bottom flatness index and the standard deviation of the EVI (surrogate of vegetation seasonality). Finally, for the years 2001-2010, the highest ranked environmental covariates were: valley bottom flatness index, mean value of the EVI (surrogate of vegetation productivity), the valley depth index, the wetness index and the night-time land surface temperature. These results showed consistency on the highest ranked environmental covariates such as the DEM and other terrain derivatives, considering the three recursive feature elimination models and years.

The same technique applied to independent datasets showed similar results (i.e., when combined RaCA and the Mexican Forest Service datasets). This independent analysis (years 2010-2012) showed an error of 2.9 kg m^{-2} (R^2=0.47) using all environmental covariates and an error of 3.4 kg m^{-2} (R^2=0.33) using just the highest ranked environmental covariates after the repeated cross-validation. The highest five ranked environmental covariates of this model were: the DEM and the topographic wetness index, the vegetation seasonality (standard deviation) and vegetation productivity (mean) from the EVI and mean monthly precipitation.

3.3 Simulated annealing

The simulated annealing framework confirmed the explanatory power of land surface temperature and precipitable water vapor, because these variables were consistently ranked as the highest environmental covariates in the three models (1991-2010, 1991-2000, 2001-2010). For the three models, the simulated annealing framework revealed that mean annual precipitation and/or the total annual precipitation were also important predictors for the SOC dataset against a cross-validation strategy (5 times repeated, 5-fold). The error estimates from this algorithm were similar compared with the previous analysis (see section 3.2), but with higher levels of explained variance for years 1991-

2010 (2.2 kg m^{-2} ; R^2=0.41), years 1991-2000 (2.1 kg m^{-2} ; R^2=0.31), and years 2001-2010 (2.3 kg m^{-2} ; R^2=0.46).

The simulated annealing analysis on the independent datasets showed that a MODIS surface reflectance variable (M06MOD4, Supplementary Table S1) becomes one of the first five important variables for predicting SOC. Other highest ranked environmental covariates for the independent datasets were also consistent with our previous results: the DEM, mean monthly precipitation, the standard deviation of the EVI and precipitable water vapor. Modeling errors and explained variances (R^2) on this model were also similar, with a mean error of 3.1 kg m^{-2} (R^2= 0.42) using only the highest ranked environmental covariates.

Figure 4 Variogram analysis applied to residuals of SOC models. The variogram of residuals against independent datasets (A). The residual variance from the different pedotransfer functions used to calculate SOC stocks (B). The combined (independent models and pedotransfer functions) residuals for the periods 1991-2000 (C) and 2001-2010 (D). The numbers in the circles indicate the available pairs of points at a given distance. Variogram parameters are shown as insets on each plot: Sph = spherical model, Ste = Stein model parameterization (and its associated Kappa value).

3.4 SOC residual analysis

We obtained six different SOC stocks (and mean errors) from the six pedotransfer functions used to estimate BD values (Supplementary Figure S3). The equation provided by Drew (Drew 1973: BD = 1 / (0.6268 + 0.0361 * OM)) was the best correlated function with our SOC prediction (1991-2010, r=0.4). The residual variance of our predictions and the multiple SOC estimates derived from different BD pedotransfer functions had a standard deviation of 3.5, a median of 1.0, and mean variance of 1.2 kg m^{-2} . The residual variance of our predictions (1991-2010) against predictions from the independent model (RaCA- Mexican Forest Service; n=9239) showed a standard deviation of 2.7, a median of 2.0, and a mean value of 2.5 kg m^{-2} . We report a moderate spatial structure (nugget:sill ratio of 0.25) for the residuals of our models and independent SOC estimates (Figure 4a). However, the residual variance of our predictions against the multiple SOC stocks calculated from different BD pedotransfer functions showed a strong spatial structure (nugget:sill ratio <0.1) across both countries (Figure 4b).

When combining the residual variance from different BD pedotransfer functions and the residuals of the independent validation, we detected a significant increase of the nugget:sill ratio from 0.08 for the years 1991-2000 (Figure 4c) to a nugget:sill ratio of 0.25 for the period between 2001-2010 (Figure 4d). Thus, there was >100% increase of uncorrelated spatial variation of SOC data (nugget:sill ratio increased from 0.08 to 0.25) from the model using 1991-2000 data to the model using 2001-2010 data. This increase of uncorrelated variation (increase of the nugget:sill ratio of >100%) was found in the combined residuals against independent data sets and against the BD pedotransfer function. These differences in the nugget:sill ratio are associated with inconsistencies in data sampling strategies and multiple collection periods of SOC data (Figure 4).

Figure 5 Predicted SOC across CONUS and Mexico. Prediction using data for years 1991-2010 (a); Predictions with data for years 1991-2001 (b); and Predictions with data for years 2001-2010 (c).

3.5 SOC stocks

We estimated a total SOC stock (1991-2010) of 47 Pg (Figure 5a) that varies from 41 to 55 Pg of SOC for the models 1991-2000 (Figure 5b) and 2001-2010 (Figure 5c), respectively. For the years 1991-2010, the residual error map suggested 10.4±5.1 Pg of SOC variance associated with the use of multiple pedotransfer functions for BD and consequently calculating SOC stocks. The larger variance of associated with BD was found across the surroundings of the Great Lakes, in the states of Vermont, New York and borders between Pennsylvania and Ohio, in CONUS (Figure 6a). The residual error map of our models against two fully independent datasets (RaCA and Mexican Forest Service), suggested a higher value of 28.8 ± 9.1 Pg of SOC variance. The large variance associated with the independent datasets was found also across the surroundings of the Great Lakes, but in the states of Wisconsin and Minnesota (Figure 6b). Another large variance from the independent validation was found across the state of Florida, specifically across the south section in the everglades area where there limited observations for the training dataset (Figure 1)

Figure 6 Residual error maps interpolated using Ordinary Kriging. The residual error map of pedotransfer BD residuals (a). The residual error map of our models against independent validation datasets (b).

The estimated SOC stock after applying the same modeling strategy to the external datasets was 46 Pg of SOC (combined RaCA- Mexican Forest Service, Figure 7a), varying ±1 Pg of SOC with the model 1991-2010 (ISCN plus INEGI, 1991-2010, Figure 5a). A linear model of our predictions against the independent datasets (RaCA-Mexican Forest Service) showed a mean error of 1.0 kg m^{-2} (R^2=0.43) (Supplementary Figure S4).

The best correlation between SOC predictions was found between models 1991-2000 and 1991-2010 (r=0.8) with ±5 Pg of difference in the predicted SOC stocks. In contrast the model for the year 2001-2010 was better correlated with the RaCA- Mexican Forest Service combined predictions (r=0.6), but the SOC stocks varied for ±7 Pg of SOC. The correlation between the model 1991-2010 and the independent analysis was lower (r=0.3) but the SOC stocks showed less variation (±1 Pg of SOC).

The variance among all models based on INEGI and ISCN data was ±7.6 Pg of SOC (Figure 7b). This value increased up to ±12 Pg of SOC by adding the variance of the models based on the independent analysis (Figure 7c). Thus, we provide a SOC stock for both countries between 46 and 47 Pg of SOC with a total modeling variance of ±12 Pg.

Figure 7 Prediction of SOC generated using the independent datasets (Aa). Model variance for predictions 1991-2000 and 2001-2010 using the INEGI and ISCN available data (b). Variance of all SOC predictions (INEGI-ISCN, RaCA-Mexican Forest Service datasets) (c).

3.6 Quantile conditional distribution of residuals

The quantile conditional distribution (used to identify the model prediction limits) for the residuals against fully independent datasets suggest a maximum possible SOC variance of ±73 Pg of SOC. This is the SOC variance from the full quantile conditional response of these residuals to the highest ranked environmental covariates. From the BD pedotransfer function variance, the full conditional response to the highest ranked environmental covariates showed a lower value of ±20 Pg of SOC. Thus, less uncertainty was found from the use of multiple pedotransfer functions than from validating against fully independent datasets. These results highlight the large variance of possible SOC predictions given the use of multiple training data sources constrained to a relatively short (i.e., two decade) period of time. The model-based uncertainty results are shown in Figure 8. For the BD pedotransfer function variance, the larger range of model based uncertainty was found across the Great Lakes of northern CONUS and the border with Canada (Figure 8A), The model-based uncertainty using independent residuals shows the largest values across, Florida, the east coast and the surroundings of the Great Lakes in CONUS, as well as some areas of southeast Mexico (Figure 8B).

Figure 8 Conditional quantile distribution of SOC residuals to the highest ranked environmental covariates from the BD residual variance (a) and for the residual variance against models generated with fully independent datasets (b).

4 Discussion

We predicted SOC (across Mexico and CONUS at 0-30 cm of soil depth) and generated gridded estimates at 250m spatial resolution for the period 1991-2010. We provided predictions of SOC using multiple inputs of data and a feature selection framework (recursive elimination of predictors and the simulated annealing algorithms) that allowed us to identify the most informative SOC environmental covariates to determine SOC stocks between 1991-2010. We calculated SOC stocks for both countries (46-47 ±12 Pg) and these values were >30% below previous global estimates such as the SoilGrids system or the Harmonized World Soil Database. We have highlighted large discrepancies between modeling outputs based on multiple data collection periods (Figure 5) and between global SOC products such as the state of the art SoilGrids250m (Supplementary Figure S5). Furthermore, our results have implications for the use and interpretation of SOC legacy data or aggregated SOC information. Specifically, we found a large difference for predicted SOC stocks (from 41 to 55 Pg of SOC) between 1991-2000 and 2001-2010 that cannot be fully attributed to SOC dynamics, but also to inconsistencies in the spatial configuration of available datasets, the use of different SOC calculation methods, and the different periods of data collection. These results open new research questions about the interpretation of apparent changes in SOC stocks across time and future studies should determine: a) if regional-to-global differences are due to active management practices/land use change, or b) if apparent changes are overshadow by large uncertainty estimates due to inconsistencies in methods and modelling variability.

4.1 Highest ranked environmental predictors

Our results suggest that using a few informative environmental predictors (e.g., the DEM and terrain derivatives, the EVI or precipitation patterns) for the spatial variability of SOC have a similar performance (~50% of explained variance) to a high-dimensional covariate space (the 150 environmental covariates reported in Supplementary Table 1) across Mexico and CONUS and using the available datasets between 1991-2010. The use of a high-dimensional covariate space to predict SOC (and other soil properties) may be needed to maximize prediction accuracy at the global scale (Hengl *et al.* 2017); however, for local to regional applications some environmental covariates for SOC may be statistically redundant and lead to unnecessarily increases of the computing resources required for prediction purposes. Reducing the statistical redundancy of environmental covariates in SOC models will simplify computing requirements and model complexity (Guo *et al.* 2019). Reducing the complexity of SOC models would be appealing in further applications of SOC spatial information (e.g., land carbon uptake modeling, climate system modeling, niche modeling) which require similar predictors as for SOC.

Our simulated annealing analysis highlights SOC relationships (positive and negative) with climate variables (precipitation and land surface temperature), elevation (and terrain derivatives) and vegetation greenness (productivity and seasonality) that are consistent with previous literature describing SOC drivers across diverse environmental conditions (Hobley, *et al*. 2015, Evans et al. 2011). In addition, when applying the simulated annealing framework to the independent datasets, a MODIS surface reflectance short wave infrared band (M06MOD4, Supplementary Table S1) was one of the first five important variables predicting SOC, which is consistent with the infrared based methods used by the USDA for developing of the RaCA dataset (Wijewardane *et al.* 2016). The prediction capability of the infrared spectra (e.g., near infrared, mid infrared) for SOC can be attributed to the strong spectral absorption characteristics of soil organic matter and BD (the main components of SOC) in the infrared spectral bands (Guo *et al*. 2019). Thus, the main relationships driving our SOC predictions can be interpreted and associated with the use of different data inputs and specific environmental covariates (e.g., the DEM and terrain parameters, the EVI and the MODIS surface reflectance infrared data , precipitation and temperature gridded surfaces) that can be periodically acquired from remote sensing at the global scale.

4.2 Uncertainty quantification

4.2.1 BD pedotransfer functions

We report ~10Pg of SOC variance associated with SOC calculation inputs. The combined and quality-controlled dataset we used have been processed following international standards for increasing precision and accuracy (Harden *et al*. 2017, Batjes *et al*. 2017)**.** However, the major limitation of these datasets is arguably the low availability of BD and CF data. Inconsistencies in BD and CF data could explain the large variance found (Figure 6A) across the highly productive landscapes of the north east of CONUS. It has been discussed that SOC stocks are systematically overestimated by misuse of the BD and CF content parameters (Poeplau *et al.* 2017), although for some areas, significant underestimations of SOC have been reported (Chen *et al*. 2018). Thus, correction of BD is fundamental to achieve realistic SOC estimates and to reduce the potential overestimation of SOC stocks (Köchy *et al.* 2015). The lack of accurate BD and CF data and the large variance of the global SOC values are key issues that could explain the discrepancy between country-to-global SOC estimates (Tifafi et al. 2018). Thus, our results provide a spatially explicit measure of SOC variance derived from six conventional BD pedotransfer functions that can be used to explain discrepancies between national, regional and global SOC estimates.

4.2.2 Spatial and temporal variations of available data

We found differences in the spatial structure (i.e., autocorrelation) of modeling residuals from multiple models and periods of time (1991-2000 and 2001-2010), that result in large differences on predicted stocks (from 41 to 55 Pg of SOC) from these defined periods of time. This period of time (1991-2010) have experienced intensive land use and environmental changes across both countries and our results could be used for identifying sensitive areas of SOC changes or areas that require further research (Figure 5). However, a previous study suggested that SOC could increase under reforestation conditions around 2 Pg per century in the topsoil (Nave et al. 2018), so our "decadal" modeling results may be overestimating the SOC gain between those time-periods. Here, we discuss our results under this consideration.

While we detected a reduction of SOC across most of Mexico, we detected a larger gain of SOC mainly across higher latitudes of CONUS (when comparing models between 1991-2000 and 2001-2010). Recent efforts have shown multiple agricultural practices that can lead to substantial SOC gains (Singh *et al.* 2018) and SOC gains have been reported on previous studies across higher latitudes of CONUS in response to agricultural practices (Adhikari and Hartemink. 2017). For example, alpine forest have been recognized as important SOC sinks under warming conditions (Ding *et al*. 2017). Recent reports have shown that some land carbon uptake models tend to project increases in high latitude SOC that are inconsistent with empirical studies that indicate significant losses of SOC with predicted climate change across these areas (Lajtha *et al.* 2018). The uncertainty of current SOC available information is one limiting factor for increasing the agreement and explaining the aforementioned inconsistencies of SOC models (Crowther *et al.* 2017). When applying the analysis independently on the specific decades (1991-2000 and 2001-2010) we were forced to remove large amounts of data across large geographical areas; consequently, these areas were not equally represented (in terms of data information) on these models. We argue that the spatial distribution and statistical differences on data available for SOC models can explain discrepancies of SOC trends, as previously shown at the global scale (van Gestel *et al*. 2018). Reducing the amount of training data increases modeling errors and the uncertainty of SOC predictions (Lagacherie *et al.* 2019). Thus, we highlight that caution must be taken when limited amount of information is used to predict SOC stocks across large geographical areas and then use that information to quantify apparent changes in SOC stocks without considering uncertainty. In this study, rather than reporting a SOC change between decades, we postulate that a better practice is to use all available data (1991-2010) to increase spatial representation.

Thus, we were able to model SOC spatial variability and compared results with two fully independent datasets to determine the most probable SOC stock estimate (46 to 47 Pg of SOC) for the period around 1991 and 2010.

4.2.3 Quantile response of residual variance

Model prediction limits from the full quantile response of independent model residuals indicated a larger SOC variance across both countries (up to 73 Pg of SOC variance) than the full response of the residual variance associated with the BD pedotransfer functions (20 Pg of SOC). These results are useful for benchmarking SOC models and represent a valuable complement for the uncertainty indicators of the predicted SOC spatial variability (Lagacherie *et al.* 2019). This variance relies on a non-parametric and accurate way of estimating conditional quantiles and the overall reliability of tree-based ensembles such as Random Forests (Meinshausen, 2006). This approach has been used for analyzing the uncertainty on soil mapping applications and larger uncertainties have been reported when reducing the data availability in numerical experiments (Lagacherie *et al.* 2019, Vaysse and Lagacherie, 2017). Thus, we provide multiple uncertainty indicators as they are useful to better interpret model limitations associated with available datasets and complement (across unsampled areas) our cross-validation and independent validation results. We propose that the results of this quantile analysis applied to SOC modeling residuals could be used for identifying areas that require higher sampling effort due larger discrepancies of multiple SOC model predictions.

4.3 SOC stocks across CONUS and Mexico

The estimated SOC stock across both countries (46-47 Pg of SOC) could be used for quantifying the contribution of SOC to the regional (e.g., North America) carbon cycling for the analyzed period of time (1991-2010). Our predicted SOC stock is lower when compared to values obtained from global estimates such as the re-gridded HWSD (Wieder *et al.* 2014) or the SoilGrids250m system (Hengl *et al.* 2017), where this value increases to ~71 Pg and ~92 Pg of SOC, respectively. High discrepancy of these two global products has been reported earlier at global- (Tifati et al. 2017), country-, or region-specific scales (Guevara *et al*. 2018, Chen *et al.* 2018, Vitharana *et al*. 2019). Moreover, our results showed discrepancies comparing country-specific studies reporting SOC stocks in CONUS (29.3 Pg of SOC, Bliss *et al.* 2014) and Mexico (9.15 Pg, Cruz-Gaistardo and Paz-Pellat, 2014, Paz Pellat, *et al.* 2016), as we report ~39 Pg of SOC for CONUS and ~7 Pg of SOC for Mexico in the first 30cm of soil. Our results highlight the need to provide country-to-region specific estimates using the best available datasets, to improve global SOC estimates by developing analytical frameworks for optimizing multiple SOC modeling efforts and sampling strategies (Guevara *et al.* 2018).

We report a density of SOC across CONUS (4.98 kg m^{-2}) that was relatively higher than the soils of Mexico (4.22 kg m^{-2}). Globally, the soil carbon pool (at 1m depth) is estimated to have around 1500-2400 Pg (Sato *et al*. 2015), while the SOC pool in the upper 30 cm is estimated to be 755 ± 119 Pg (Batjes, 2016). Our results suggest that Mexico represents ~1 % and CONUS ~5 % of the global SOC pool at 30 cm depth. Recent revisions highlight that the SOC stock at 30 cm soil depth remains unclear (Lajtha *et al.* 2018), and this study provides new insights for interpreting the discrepancies around the topsoil SOC pool across CONUS and Mexico.

Our SOC estimates across forested areas are comparable to those reported on studies (Domke *et al*. 2017, Bolaños *et al.* 2017). However, our results show high uncertainty across areas dominated with high SOC values (e.g., >1 gr cm^{-2}, some northern and tropical forests, peatlands, other black soils dominated areas) and across higher latitudes (Figure 6), as documented in previous studies (Tian, *et al.* 2015). Unfortunately, these areas are poorly represented in the available datasets (<10% of available data with values >1 gr cm^{-2}) and we encourage future monitoring efforts to increase their representativeness.

4.3.1 SOC across land cover classes

Our study confirms the presence of important SOC stocks across both forest and agricultural soils. Across both countries, we found higher SOC stocks in croplands, representing 26% of total SOC within the upper 30cm of soil. Across Mexico, we found that 42% of SOC was stored in forest soils and 24% in agricultural soils, while 31% of SOC across CONUS is stored in forest soils and 27% in agricultural soils. While organic matter-rich and deep soils dominate most agricultural areas across CONUS (Adhikari & Hartemink, 2017), most agricultural soils in Mexico tend to be shallow (Guerrero *et al.* 2014, ~30cm depth); consequently, we emphasize that carbon management, monitoring and conservation strategies must be developed from a country-specific approach considering country-specific land cover classes.

We found that tropical or sub-tropical broadleaf evergreen forests is the natural vegetation class with the highest SOC pool across Mexico (1.22 Pg), while temperate broadleaf deciduous forests had the highest SOC pool across CONUS (6.41 Pg). Grasslands and shrublands are also important SOC reservoirs, as they store around 37.7% of SOC across Mexico and 34.9% of SOC across CONUS (Supplementary Table S2). Such estimates are relevant for public policy around SOC conservation efforts (e.g., FAO, 2017) because grasslands and shrublands transitions are increasingly vulnerable to global warming and the increase of aridity conditions which would result in a decrease of SOC stocks (Petrie *et al.* 2014, FAO 2017). Thus, accurately quantifying the spatial variability of SOC across grasslands and shrublands will be an important component for enhancing SOC sequestration by better informing conservation efforts of soil ecosystem functions across North America.

Accurate SOC estimates represent a key variable to quantify human induced disturbances to the carbon cycle across land cover classes. We report that temperate forests of CONUS contain the larger SOC reserves while tropical or sub-tropical broadleaf evergreen forest and wetlands are the land cover classes with higher SOC in Mexico than CONUS (Supplementary Table S2). Respectively, the tropical or sub-tropical broadleaf evergreen forests are the most productive ecosystems of Mexico (Murray-Tortarolo *et al.* 2016). The wetlands category, with high carbon sequestration potential, includes mangroves (and other coastal wetlands), which have been recognized as the ecosystems with higher carbon storage capacity from the site-specific to the global scales (Vázquez-Lule, *et al.* 2019, Adame *et al.* 2015, Atwood *et al.* 2017). Our results represent benchmarks for SOC monitoring across these land cover classes. Thus, the spatial predictions of SOC at 250m allows for the interpretation of SOC spatial patterns across land cover classes of Mexico and CONUS accounting for sensitivities associated with the use of multiple data inputs.

4.4 Final remarks

Optimizing future SOC sampling strategies while reducing modeling variance and increasing model agreement against model independent datasets collected under different circumstances (e.g., logistics, design, main purpose, SOC estimation methods) are large challenges for enabling SOC carbon mapping and monitoring systems. New and better SOC parameters are required for reducing the current discrepancy between multiple sources of SOC data (Tifafi *et al.* 2017, Guevara *et al.* 2018) and enabling SOC monitoring systems (Viscarra-Rossel *et al. 2014*). The lack of accurate SOC spatial information and the combination of multiple SOC data sources could result in large variance estimates across the two countries (e.g., red areas of Figure 6). We propose that areas with high variance suggest that these regions require higher SOC sampling efforts.

We provide high spatial resolution (e.g., 250m pixels) SOC estimates that account for model uncertainty. Such estimates are needed for identifying regions that should be targets for SOC protection (Lagacherie & McBratney, 2006). Soil carbon protection is increasingly important to restore the current negative imbalance in our soil carbon budget due, for example, to the development of agricultural systems and croplands (Sanderman *et al.* 2017). Accurate SOC estimates at the relevant (local) scale for farmers and landowners (e.g., spatial resolution of 250m or less) would be an important component to reduce land degradation and improve the efficiency of current efforts for sequestering SOC (Bonfatti *et al.* 2016; Malone *et al.* 2017). Thus, our results provide insights for identifying and delineating land-areas with high potential for SOC stocks that account for model sensitivity to multiple data inputs and sources. Finally, this research is timely because there is high discrepancy between SOC global estimates that needs to be solved in order to better quantify SOC dynamics (Tifati *et al.* 2017). Consequently, this discrepancy can influence the estimates of SOC warming response (Karhu *et al.* 2010) and the carbon–climate feedback that could accelerate climate change (Crowther *et al.* 2016). We hope that this study motivates an increase in country-specific soil surveys, data sharing, and modeling of SOC estimates at higher spatial resolution with a better quantification of uncertainty.

Acknowledgements

MG acknowledges support from a scholarship from the Consejo Nacional de Ciencia y Tecnología (CONACyT) of Mexico (382790). RV acknowledges support from NASA Carbon Monitoring Systems (80NSSC18K0173) and USDA (2014-67003-22070).

Data availability: All modeling output is available through The Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) for biogeochemical dynamics of the NASA-Earth Observing System Data and Information System (DOI: https://doi.org/10.3334/ORNLDAAC/1737). Public data to parameterize the model are available from the Rapid Carbon Assessment Project (https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/?cid=nrcs142p2_054164), the International Soil Carbon Network (https://iscn.fluxdata.org/), the Instituto Nacional de Estadística y Geografía (INEGI, SERIES 1 & 2; n >65 000 pedons available, Krasilnikov *et al.* 2013), and covariates from the SoilGrids250m project (https://soilgrids.org/#!/?layer=ORCDRC_M_sl2_250m&vector=1; Hengl *et al.* 2017).

References

Adame MF, Santini NS, Tovilla C, Vázquez-Lule A, Castro L (2015) Carbon stocks and soil sequestration rates of riverine mangroves and freshwater wetlands. *Biogeosciences Discussions*, **12**, 1015–1045.

Adams, W. (1973). The effect of organic matter on the bulk and true densities of some uncultivated podzolic soils. European Journal of Soil Science, 24(1):10–17.

Adhikari, K., Hartemink, A. E. (2017). Soil organic carbon increases under intensive agriculture in the Central Sands, Wisconsin, USA. *Geoderma Regional*, *10*, 115–125. https://doi.org/10.1016/j.geodrs.2017.07.003

Adhikari K, Hartemink AE (2017) Soil organic carbon increases under intensive agriculture in the Central Sands, Wisconsin, USA. *Geoderma Regional*, **10**, 115–125.

Ad-hoc-AG-Boden. 2005. Bodenkundliche Kartieranleitung – 5. Auflage. Hannover, Germany. 438 pp.

Arrouays D, Grundy MG, Hartemink AE, Hempel JW, Heuvelink GBM, Hong SY, et al. 2014. Chapter Three - GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties. In: Donald LS, editor. Advances in Agronomy [Internet]. Academic Press; p. 93–134.

Arrouays D, Lagacherie P, Hartemink AE (2017) Digital soil mapping across the globe. *Geoderma Regional*, **9**, 1–4.

Atwood, Trisha B., et al. "Global patterns in mangrove soil carbon stocks and losses." Nat. Clim. Change, vol. 7, no. 7, 26 June 2017, p. 523, doi:10.1038/nclimate3326.

Batjes NH, Ribeiro E, van Oostrum A, Leenaars J, Hengl T, Mendes de Jesus J (2017) WoSIS: providing standardised soil profile data for the world. *Earth System Science Data*, **9**, 1–14.

Banwart, Steve S., Helaina B. Black, Zucong Z. Cai, Patrick G. Gicheru, Hans J. Joosten, Reynaldo V. Victoria, Eleanor E. Milne, et al. 2014. "Benefits of Soil Carbon: Report on the Outcomes of an International Scientific Committee on Problems of the Environment Rapid Assessment Workshop." *Carbon Management* 5 (2): 185–92.

Batjes NH (2016) Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. *Geoderma*, **269**, 61–68.

Bond-Lamberty, B., Bailey, V.L., Chen, M., Gough, C.M. and Vargas, R., 2018. Globally rising soil heterotrophic respiration over recent decades. Nature, 560(7716), p.80.

Bishop TFA, McBratney AB, Laslett GM (1999) Modelling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma*, **91**, 27–45.

Bliss NB, Waltman SW, West LT, Neale A, Mehaffey M (2014) *Distribution of soil organic carbon in the conterminous United States*.

Bolaños González, Y., Bolaños González, M. A., Paz Pellat, F., Ponce Pulido, J. I., Bolaños González, Y., Bolaños González, M. A., ...Ponce Pulido, J. I. (2017). Estimación de carbono almacenado en bosques de oyamel y ciprés en Texcoco, Estado de México. Terra Latinoamericana, 35(1), 73–86. Retrieved from http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-57792017000100073

Bonfatti BR, Hartemink AE, Giasson E, Tornquist CG, Adhikari K (2016) Digital mapping of soil carbon in a viticultural region of Southern Brazil. *Geoderma*, **261**, 204–221.

Breiman L (2001) Random Forests. *Machine Learning*, **45**, 5–32.

Ciais P, Dolman AJ, Bombelli A et al. (2014) Current systematic carbon-cycle observations and the need for implementing a policy-relevant carbon observing system. *Biogeosciences*, **11**, 3547–3602.

Chen, S., & Arrouays, D. (2018). Soil carbon stocks are underestimated in mountainous regions. Geoderma, 320, 146–148. doi: 10.1016/j.geoderma.2018.01.029

Cortes C, Vapnik V (1995) Support-Vector Networks. *Machine Learning*, **20**, 273–297.

Crowther TW, Todd-Brown KEO, Rowe CW et al. (2016) Quantifying global soil carbon losses in response to warming. *Nature*, **540**, 104–108.

Cruz-Cárdenas G, López-Mata L, Ortiz-Solorio CA, Villaseñor JL, Ortiz E, Silva JT, Estrada-Godoy F (2014) Interpolation of Mexican soil properties at a scale of 1:1,000,000. *Geoderma*, **213**, 29–35.

Cruz-Gaistardo, C. and Paz-Pellat, F.: Mapa de carbono orgánico de los suelos de la República Mexicana, in: EstadoActual del Conocimiento del Ciclo del Carbono y sus Interacciones en México: Síntesis a 2013, edited by: Paz-Pellat, F., Wong-González, J., Bazan, M., and Saynes, V., Programa Mexicano del Carbono. Texcoco, Estado de México, México, ISBN 978-607-96490-1-2, 702, 2014.

Delgado-Baquerizo M, Eldridge DJ, Maestre FT, Karunaratne SB, Trivedi P, Reich PB, Singh BK (2017) Climate legacies drive global soil carbon stocks in terrestrial ecosystems. *Science Advances*, **3**, e1602008.

de Gruijter JJ, McBratney AB, Minasny B, Wheeler I, Malone BP, Stockmann U (2016) Farm-scale soil carbon auditing. *Geoderma*, **265**, 120–130.

Ding, J., Chen, L., Ji, C., Hugelius, G., Li, Y., Liu, L., et al. (2017). Decadal soil carbon accumulation across Tibetan permafrost regions. Nature Geoscience, 10(6), 420–424. https://doi.org/10.1038/ngeo2945

Domke GM, Perry CH, Walters BF, Nave LE, Woodall CW, Swanston CW (2017) Toward inventory-based estimates of soil organic carbon in forests of the United States. *Ecological Applications*, **27**, 1223–1235.

Drew, L. A.: Bulk Density Estimation Based on Organic Matter Content of Some Minnesota Soils, St. Paul, Minn., School of Forestry, University of Minnesota, Digital Conservancy, available at: http://hdl.handle.net/11299/58293 (last access: 16 July 2018), 1973

Evans, S. E., Burke, I. C., & Lauenroth, W. K. (2011). Controls on soil organic carbon and nitrogen in Inner Mongolia, China: A cross- continental comparison of temperate grasslands. Global Biogeochem. Cycles, 25(3). doi: 10.1029/2010GB003945

FAO 2017. Soil Organic Carbon: the hidden potential. Food and Agriculture Organization of the United Nations, Rome, Italy

FAO Guidelines for soil description Fourth edition, Rome 2006 109 pp. FAO and ITPS. 2018. Global Soil Organic Carbon Map (GSOCmap) Technical Report. Rome. 162 pp.

FAO and ITPS. 2018. Global Soil Organic Carbon Map (GSOCmap) Technical Report. Rome. 162pp.

Fernández-Delgado M, Cernadas E, Barro S, Amorim D (2014) Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, **15**, 3133–3181.

Finke PA (2012) On digital soil assessment with models and the Pedometrics agenda. *Geoderma*, **171–172**, 3–15.

Folberth C, Skalský R, Moltchanova E, Balkovič J, Azevedo LB, Obersteiner M, van der Velde M (2016) Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations. *Nature Communications*, **7**, 11872.

Grunwald S (2009) Multi-criteria characterization of recent digital soil mapping and modeling approaches. *Geoderma*, **152**, 195–207.

Grigal, D., Brovold, S., Nord, W., and Ohmann, L. (1989). Bulk density of surface soils and peat in the north central united states. Canadian Journal of Soil Science, 69(4):895–900.

Guerrero E, Pérez A, Arroyo C, Equihua J, Guevara M (2014) Building a national framework for pedometric mapping: Soil depth as an example from Mexico. In: *GlobalSoilMap* (Eds. Arrouays D, McKenzie N, Hempel J, de Forges A, McBratney A), pp. 103–108. CRC Press.

Guevara, M., Olmedo, G. F., Stell, E., Yigini, Y., Aguilar Duarte, Y., Arellano Hernández, C., Arévalo, G. E., Arroyo-Cruz, C. E., Bolivar, A., Bunning, S., Bustamante Cañas, N., Cruz-Gaistardo, C. O., Davila, F., Dell Acqua, M., Encina, A., Figueredo Tacona, H., Fontes, F., Hernández Herrera, J. A., Ibelles Navarro, A. R., Loayza, V., Manueles, A. M., Mendoza Jara, F., Olivera, C., Osorio Hermosilla, R., Pereira, G., Prieto, P., Ramos, I. A., Rey Brina, J. C., Rivera, R., Rodríguez-Rodríguez, J., Roopnarine, R., Rosales Ibarra, A., Rosales Riveiro, K. A., Schulz, G. A., Spence, A., Vasques, G. M., Vargas, R. R., and Vargas, R.: No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America, SOIL, 4, 173-193, https://doi.org/10.5194/soil-4-173-2018, 2018.

Guo, L., Zhang, H., Shi, T., Chen, Y., Jiang, Q., & Linderman, M. (2019). Prediction of soil organic carbon stock by laboratory spectral data and airborne hyperspectral images. Geoderma, 337, 32–41. doi: 10.1016/j.geoderma.2018.09.003

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. doi: 10.1016/j.geoderma.2019.01.023

Harden JW, Hugelius G, Ahlström A et al. (2017) Networking our science to characterize the state, vulnerabilities, and management opportunities of soil organic matter. *Global Change Biology*.

Hechenbichler K, Schliep K (2006) Weighted k-nearest-neighbor techniques and ordinal classification. In: *Discussion Paper 399, SFB 386*.

Hengl T, de Jesus JM, MacMillan RA et al. (2014) SoilGrids1km — Global Soil Information Based on Automated Mapping. *PLoS ONE*, **9**, e105992.

Hengl T. (2017). GSIF: Global Soil Information Facilities. R package version 0.5-1. https://CRAN.R-project.org/package=GSIF

Hengl T, Mendes J, Heuvelink GBM, et al. (2017) SoilGrids250m: global gridded soil information based on Machine Learning. PLoS ONE 12(2): e0169748.https://doi.org/10.1371/journal.pone.0169748

Hengl, T., Heuvelink, G. B. M., & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma, 120(1), 75–93. doi:10.1016/j.geoderma.2003.08.018

Heuvelink G.B.M. Uncertainty Quantification of GlobalSoilMap Products. (2014). D. Arrouays, McKenzie NJ, J.W. Hempel, A.C. Richer de Forges, A.B. McBratney (Eds.), GlobalSoilMap. Basis of the Global Soil Information System, Taylor & Francis, CRC press, Oxon, pp. 335–340

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hobley, E., Wilson, B., Wilkie, A., Gray, J., & Koen, T. (2015). Drivers of soil organic carbon storage and vertical distribution in Eastern Australia. Plant Soil, 390(1-2), 111–127. doi: 10.1007/s11104-015-2380-1

Honeysett, J. and Ratkowsky, D. (1989). The use of ignition loss to estimate bulk density of forest soils. European Journal of Soil Science, 40(2):299–308.

INEGI-Instituto Nagional de Geografia y Estadistica (2014) Diccionario de Datos Edafológicos escala 1:250000 (version 3) 55pp.
http://www.inegi.org.mx/geo/contenidos/recnat/edafologia/doc/dd_edafologicos_v3_250k.pdf

Jackson RB, Lajtha K, Crow SE, Hugelius G, Kramer MG, Piñeiro G (2017) The Ecology of Soil Carbon: Pools, Vulnerabilities, and Biotic and Abiotic Controls. *Annual Review of Ecology, Evolution, and Systematics*, **48**.

James G, Witten D, Hastie T, Tibshirani R (2013) *An Introduction to Statistical Learning*, Vol. 103. Springer New York, New York, NY.

Jenny H (1941) *Factors of soil formation: a system of quantitative pedology*. McGraw-Hill, 304 pp.

Jones C, Falloon P (2009) Sources of uncertainty in global modelling of future soil organic carbon storage. In: *Uncertainties in Environmental Modelling and Consequences for Policy Making* (eds Baveye PC, Laba M, Mysiak J), pp. 283–315. Springer Netherlands, Dordrecht.

Jeffrey, D. (1970). A note on the use of ignition loss as a means for the approximate estimation of soil bulk density. The Journal of Ecology, pages 297–299.

Jones C, McConnell C, Coleman K, Cox P, Falloon P, Jenkinson D, Powlson D (2005) Global climate change and soil carbon stocks; predictions from two contrasting models for the turnover of organic carbon in soil. *Global Change Biology*, **11**, 154–166.

Karhu K, Fritze H, Hämäläinen K et al. (2010) Temperature sensitivity of soil carbon fractions in boreal forest soil. *Ecology*, **91**, 370–376.

Khlopenkov, K., & Trichtchenko, A. (2008). Implementation and evaluation of concurrent gradient search method for reprojection of MODIS level 1B imagery. IEEE Transaction on Geoscience and Remote Sensing, 46:2016–2027.

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

Kuhn M (2008) Building Predictive Models in R Using the caret Package. J Stat Softw 28: 1–26. https://doi.org/10.18637/jss.v028.i05

Krasilnikov P, Gutiérrez-Castorena M del C, Ahrens RJ, Cruz-Gaistardo CO, Sedov S, Solleiro-Rebolledo E (2013) *The Soils of Mexico*. Springer Netherlands, Dordrecht.

Lagacherie P, McBratney AB (2006) Chapter 1 Spatial Soil Information Systems and Spatial Soil Inference Systems: Perspectives for Digital Soil Mapping. In: *Developments in Soil Science*, Vol. 31, pp. 3–22. Elsevier.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Martin, M., & Saby, N. P. A. (2019). How far can the uncertainty on a Digital Soil Map be known?: A numerical experiment using pseudo values of clay content obtained from Vis-SWIR hyperspectral imagery. Geoderma, 337, 1320–1328. doi: 10.1016/j.geoderma.2018.08.024

Lajtha, K., V. L. Bailey, K. McFarlane, K. Paustian, D. Bachelet, R . Abramoff, D. Angers, S. A. Billings, D. Cerkowniak, Y. G. Dialynas, A. Finzi, N. H. F. French, S. Frey, N. P. Gurwick, J. Harden, J. M. F. Johnson, K. Johnson, J. Lehmann,S. Liu, B. McConkey, U. Mishra, S. Ollinger, D. Paré, F. Paz Pellat, D. deB. Richter, S. M. Schaeffer, J. Schimel, C. Shaw, J. Tang, K. Todd-Brown, C. Trettin, M. Waldrop, T. Whitman, and K. Wickland, 2018: Chapter 12: Soils. In Second State of the Carbon Cycle Report (SOCCR2): A Sustained Assessment Report. Cavallaro, N., G. Shrestha, R . Birdsey, M. A. Mayes, R . G. Najjar, S. C. Reed, P. Romero-Lankao, and Z. Zhu (eds.). U.S. Global Change Research Program, Washington, DC, USA, pp. 469-506, https://doi.org/10.7930/SOCCR2.2018.Ch12.

Lal R (2004) Soil Carbon Sequestration Impacts on Global Climate Change and Food Security. *Science*, **304**, 1623–1627.

Malone BP, McBratney AB, Minasny B, Laslett GM (2009) Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma*, **154**, 138–152.

Malone BP, Styc Q, Minasny B, McBratney AB (2017) Digital soil mapping of soil carbon at the farm scale: A spatial downscaling approach in consideration of measured and uncertain data. *Geoderma*, **290**, 91–99.

McBratney A., Mendonça Santos M., Minasny B (2003) On digital soil mapping. *Geoderma*, **117**, 3–52.

McBratney A, Field DJ, Koch A (2014) The dimensions of soil security. *Geoderma*, **213**, 203–213.

Meinshausen N (2006) Quantile Regression Forests. *J. Mach. Learn. Res.*, **7**, 983–999.

Minasny B, McBratney AB, Lark RM (2008) Digital Soil Mapping Technologies for Countries with Sparse Data Infrastructures. In: *Digital Soil Mapping with Limited Data* (eds Hartemink AE, McBratney A, Mendonça-Santos M de L), pp. 15–30. Springer Netherlands, Dordrecht.

Minasny B, McBratney AB, Malone BP, Wheeler I (2013) Digital Mapping of Soil Carbon. In: *Advances in Agronomy*, Vol. 118, pp. 1–47. Elsevier.

Minasny B, Malone BP, McBratney AB et al. (2017) Soil carbon 4 per mille. *Geoderma*, **292**, 59–86.

Murray-Tortarolo G, Friedlingstein P, Sitch S et al. (2016) The carbon cycle in Mexico: past, present and future of C stocks and fluxes. *Biogeosciences*, **13**, 223–238.

Nave L and Johnson K, van Ingen C, Agarwal D, Humphrey M, Beekwilder N. 2017. International Soil Carbon Network (ISCN) Database, Version 3. Database Report: Calculations and Quality Assessment. Accessed 2 February 2017.

Nelson, D.W., and L.E. Sommers (1982) Total carbon, organic carbon, and organic matter. p. 539-580. In A.L. Page et al. (ed.) Methods of soil Analysis. Part 2. 2nd ed. Agron. Monogr. 9. ASA and SSSA, Madison, WI.

North American Land Cover at 250 m spatial resolution. Produced by Natural Resources Canada/Canadian Center for Remote Sensing (NRCan/CCRS), United States Geological Survey (USGS); Insituto Nacional de Estadística y Geografía (INEGI), Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO) and Comisión Nacional Forestal (CONAFOR) (2010).

Naipal, V., Ciais, P., Wang, Y., Lauerwald, R., Guenet, B., and Van Oost, K. (2018) Global soil organic carbon removal by water erosion under climate change and land use change during AD1980–2005, Biogeosciences, 15, 4459–4480, https://doi.org/10.5194/bg

Ogle SM, Breidt FJ, Easter M, Williams S, Killian K, Paustian K (2010) Scale and uncertainty in modeled soil organic carbon stock changes for US croplands using a process-based model. *Global Change Biology*, **16**, 810–822.

O'Rourke SM, Angers DA, Holden NM, McBratney AB (2015) Soil organic carbon across scales. *Global Change Biology*, **21**, 3561–3574.

Oliver, M. A. and R. Webster. "A tutorial guide to geostatistics: Computing and modelling variograms and kriging." CATENA, vol. 113, 1 Feb. 2014, pp. 56-69, doi:10.1016/j.catena.2013.09.006.

Padarian J, Minasny B, McBratney AB (2015) Using Google's cloud-based platform for digital soil mapping. *Computers & Geosciences*, **83**, 80–88.

Paz Pellat, F., J. Argumedo Espinoza, C. O. Cruz Gaistardo, J. D. Etchevers, B., and B. de Jong, 2016: Distribución especial y temporal del carbono orgánico del suelo en los ecosistemas terrestres. Terra Latinoamericana, 34 (3), 289-310.

Petrie MD, Collins SL, Swann AM, Ford PL, Litvak ME (2015) Grassland to shrubland state transitions enhance carbon sequestration in the northern Chihuahuan Desert. *Global Change Biology*, **21**, 1226–1235.

Pike RJ, Evans IS, Hengl T (2009) Chapter 1 Geomorphometry: A Brief Guide. In: *Developments in Soil Science*, Vol. 33, pp. 3–30. Elsevier.

Poeplau C, Vos C, Don A (2017) Soil organic carbon stocks are systematically overestimated by misuse of the parameters bulk density and rock fragment content. *SOIL*, **3**, 61–66.

Ramcharan, A., T. Hengl, T. Nauman, C. Brungard, S. Waltman, S. Wills, and J. Thompson (2018) Soil Property and Class Maps of the Conterminous United States at 100-Meter Spatial Resolution. Soil Sci. Soc. Am. J. 82:186-201. doi:10.2136/sssaj2017.04.0122

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reuter HI, Hengl T. Worldgrids—a public repository of global soil covariates. In: Miasny B, Malone BP, McBratney AB editors. Digital Soil Assessments and Beyond—Proceedings of the 5th Global Workshop on Digital Soil Mapping. Sydney: CRC Press; 2012. pp. 287–292. Available: https://doi.org/10.1201/b12728-57

Saini, G. (1966). Organic matter as a measure of bulk density of soil. Nature, 210(5042):1295–1296.
Sanchez PA, Ahamed S, Carré F, Hartemink AE, Hempel J, Huising J, et al. (2009) Digital soil map of the world. *Science*, **325** (5941), 680–1.

Sanderman J, Hengl T, Fiske GJ (2017) Soil carbon debt of 12,000 years of human land use. *Proceedings of the National Academy of Sciences*, **114**, 9575–9580.

Scharlemann JP, Tanner EV, Hiederer R, Kapos V (2014) Global soil carbon: understanding and managing the largest terrestrial carbon pool. *Carbon Management*, **5**, 81–91.

Siebe C, Jahn R, Stahr K Manual para la descripción y evaluación ecológica de suelos en el campo, 2nd edition (2006) Sociedad Mexicana de la Ciencia del Suelo A. C., Publicación Especial.

Singh, B. P., Setia, R., Wiesmeier, M., & Kunhikrishnan, A. (2018). Agricultural Management Practices and Soil Organic Carbon Storage. *Soil Carbon Storage*, 207–244. https://doi.org/10.1016/b978-0-12-812766-7.00007-x

Soil Survey Staff and T. Loecke. 2016. Rapid Carbon Assessment: Methodology, Sampling, and Summary. S. Wills (ed.). U.S. Department of Agriculture, Natural Resources Conservation Service.

Soil Survey Staff. 1999. Soil taxonomy: A basic system of soil classification for making and interpreting soil surveys. 2nd edition. Natural Resources Conservation Service. U.S. Department of Agriculture Handbook 436 pp.

Soil Survey Staff. 2014. Soil Survey Field and Laboratory Methods Manual. Soil Survey Investigations Report No. 51, Version 2.0. R. Burt and Soil Survey Staff (ed.). U.S. Department of Agriculture, Natural Resources Conservation Service. 487 pp.

Stockmann U, Adams MA, Crawford JW et al. (2013) The knowns, known unknowns and unknowns of sequestration of soil organic carbon. *Agriculture, Ecosystems & Environment*, **164**, 80–99.

Stockmann U, Padarian J, McBratney A et al. (2015) Global soil organic carbon assessment. *Global Food Security*, **6**, 9–16.

Stoorvogel JJ, Bakkenes M, Temme AJAM, Batjes NH, ten Brink BJE (2017) S-World: A Global Soil Map for Environmental Modelling. *Land Degradation & Development*, **28**, 22–33.

Tank, S.E., J.B. Fellman, E. Hood, E.S. KritzbergBeyond respiration: controls on lateral carbon fluxes across the terrestrial- aquatic interface Limnol. Oceanogr. Lett., 3 (3) (2018), pp. 76-88

Tian H, Lu C, Yang J et al. (2015) Global patterns and controls of soil organic carbon dynamics as simulated by multiple terrestrial biosphere models: Current status and future directions: *Global Biogeochemical Cycles*, **29**, 775–792.

Tifafi M, Guenet B, Hatté C Large Differences in Global and Regional Total Soil Carbon Stock Estimates Based on SoilGrids, HWSD, and NCSCD: Intercomparison and Evaluation Based on Field Data From USA, England, Wales, and France. (2017) *Global Biogeochemical Cycles*, 2017GB005678.

Vázquez-Lule, A., Colditz, R., Herrera-Silveira, J., Guevara, M., Rodríguez-Zúñiga, M.T., Cruz, I., Ressl, R. and Vargas, R., 2019. Greenness trends and carbon stocks of mangroves across Mexico. Environmental Research Letters, 14(7), p.075010.

Vaysse K, Lagacherie P (2017) Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma*, **291**, 55–64.

Vargas R, Alcaraz-Segura D, Birdsey R et al. (2017) Enhancing interoperability to facilitate implementation of REDD+: case study of Mexico. *Carbon Management*, **8**, 57–65.
Vázquez-Selem L, Heine K (2004) Late Quaternary glaciation of México. In: *Developments in Quaternary Sciences*, Vol. 2 (eds Ehlers J, Gibbard PL), pp. 233–242. Elsevier.

Villarreal S., Guevara M., Alcaraz-Segura D., Brunsell N., Hayes D., Loescher H. and Vargas R. (2018). Ecosystem functional diversity and the representativeness of environmental networks across the conterminous United States. Agric. For. Meteorol., 262, 423–433. doi: 10.1016/j.agrformet.2018.07.016

Viscarra-Rossel RA, Webster R, Bui EN, Baldock JA (2014) Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. *Global Change Biology*, **20**, 2953–2970.

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

Wieder WR, Cleveland CC, Smith WK, Todd-Brown K (2015) Future productivity and carbon storage limited by terrestrial nutrient availability. *Nature Geoscience*, **8**, 441–444.

Wijewardane NK, Ge Y, Wills S, Loecke T (2016) Prediction of Soil Carbon in the Conterminous United States: Visible and Near Infrared Reflectance Spectroscopy Analysis of the Rapid Carbon Assessment Project. *Soil Science Society of America Journal*, **80**, 973.

Wilson JP (2012) Digital terrain modeling. *Geomorphology*, **137**, 107–121.

Yigini, Y., Olmedo, G.F., Reiter, S., Baritz, R., Viatkin, K. and Vargas, R. (eds). 2018. Soil Organic Carbon Mapping Cookbook 2nd edition. Rome, FAO. 220 pp.

Supplementary Tables:

Supplementary Table S1 Detailed description of the SOC covariates used on for generating predictions across 250m grids. This table includes a code, units and source of each SOC prediction factor included in the model predictions.

Supplementary Table S2 Estimated Soil Organic Carbon (SOC) stocks in petagrams (Pg) for the different land cover classes reported by the North American Land Change Monitoring System. This table shows the different SOC stocks estimated for the different data/periods across the combined area of CONUS and Mexico, the SOC stock across land cover classes of CONUS and land cover classes of Mexico.

Supplementary Figures:

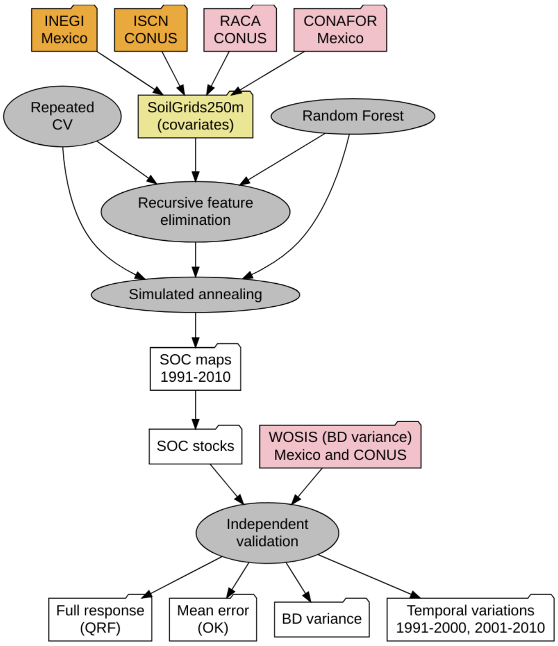
Supplementary Figure S1. Selection of prediction algorithm based on cross validated accuracy metrics. Random Forest (rf) generates the lowest error and the highest explained variance. (qrf=quantile regression forest, dnn=deep neural network, pls=partial least squared regression, bagEarth=multivariate adaptive regression splines, svmRadial=radial kernel support vector machines, kknn=kernel weighted nearest neighbors). These results are derived from repeated 5-fold-cross-validation. These methods were implemented using the R package caret. Highest explained variance (R2, 0-1), lowest mean absolute error (MAE, gr cm-2) or root mean squared errors (RMSE, gr cm-2) were achieved with rf. Accuracy indicator represents the individual values for each metric (e.g., R2, MAE, RMSE)

Supplementary Figure S2 Independent datasets (RaCA and the Mexican Forest Service). This combined dataset was collected between 2009-2012.

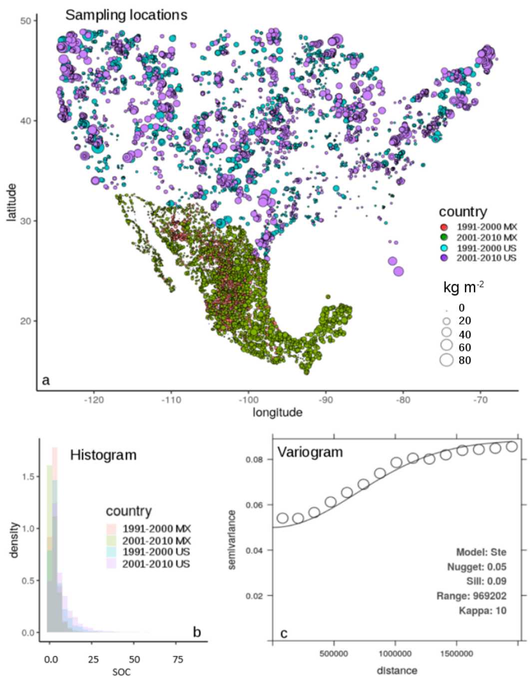
Supplementary Figure S3 Results from the different pedotransfer functions applied to BD for calculating SOC stocks and report the variance of the calculation method. We show the distribution of values in our predictions (1991-2010 first 3 boxplots), the prediction with independent datasets (2010-2012) and the residuals against each of the six pedotransfer functions for BD (A). The estimated BD data from the six pedotransfer functions (B). The mean absolute errors for each SOC estimated value based on Truncate Taylor series analysis (C). The horizontal line is an arbitrary reference for comparing the values in the three panels.

Supplementary Figure S4 Linear ensemble predicting nearly 50% of SOC variability (Rsquared) using our models as explanatory variables for the independent datasets. These results were cross-validated (5 repeats five folds). The mean absolute error (MAE) and the root mean squared error (RMSE) are between 0.1 and 0.2 gr cm -2, the lowest prediction error obtained for SOC data. This ensemble was performed using a random forest (rf) models and a kernel based model (kknn), since were the best approaches predicting SOC data from supplementary Figure S1. Accuracy indicator represents the individual values for each metric (e.g., R2, MAE, RMSE)

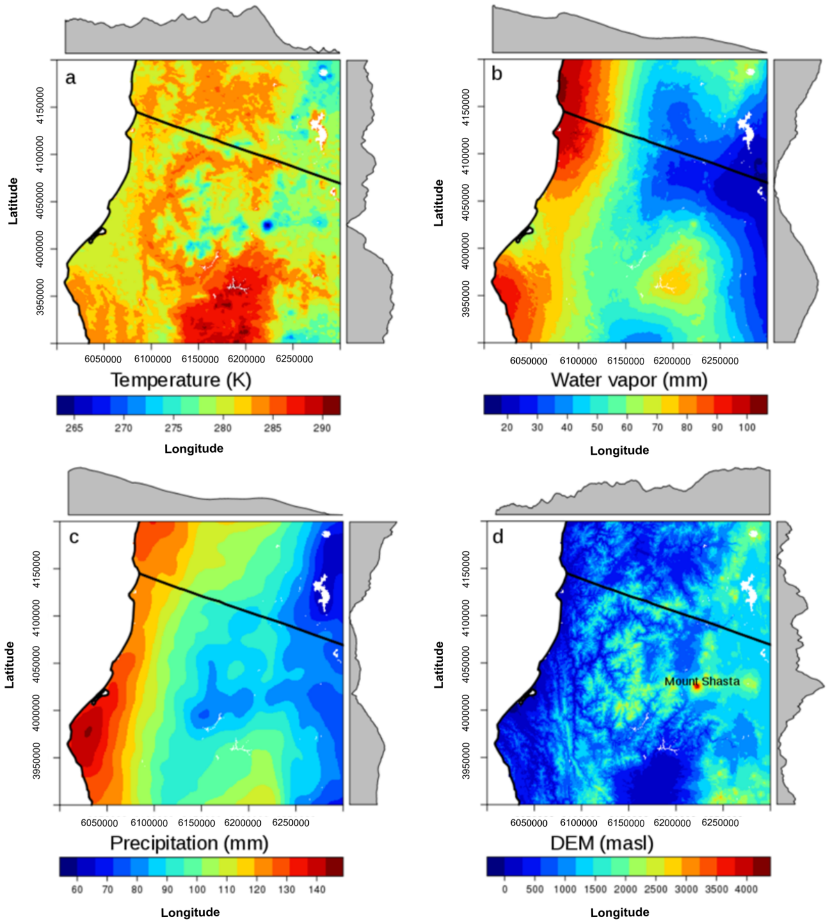
Supplementary Figure S5 Change vector map between our product (1991-2010) and SoilGrids250m (our map - SoilGrids250m). Areas in blue indicate areas where our model is predicting higher values than SoilGrids250m. Areas in red are showing areas where SoilGrids250m is predicting higher SOC carbon values.



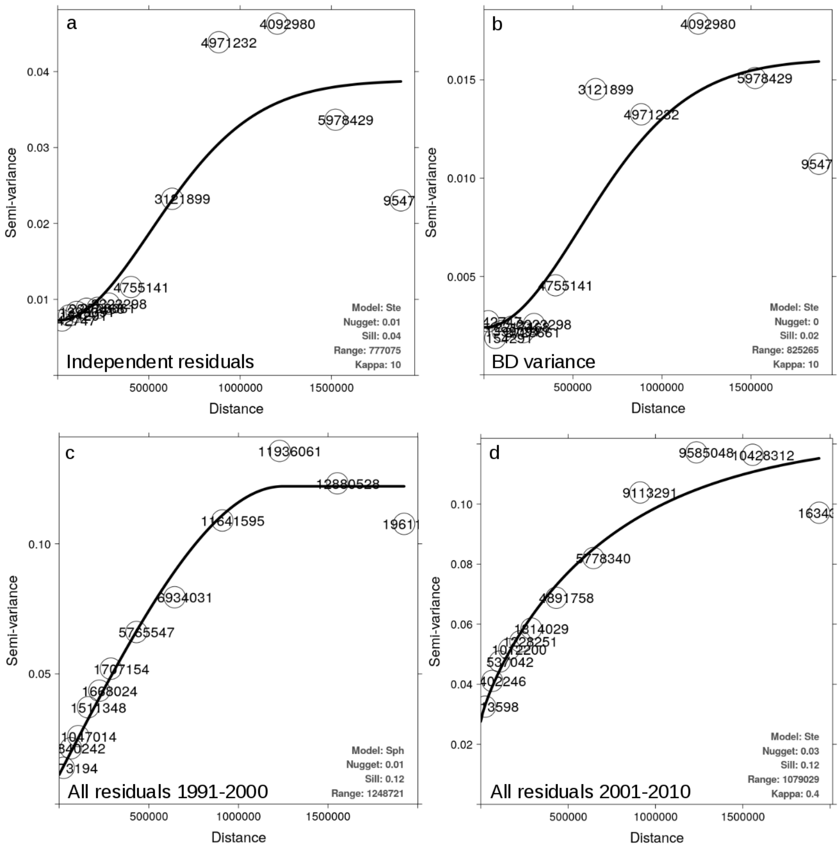
2019GB006219-f01-z-.tif



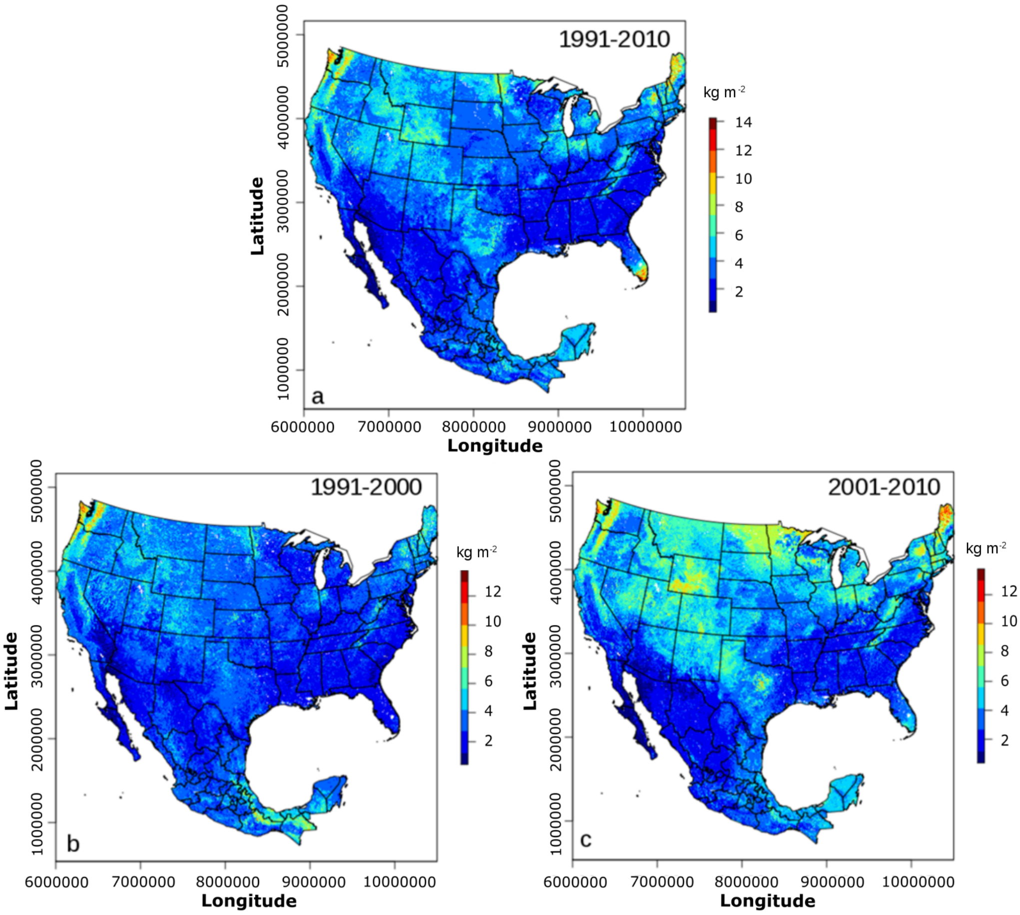
2019GB006219-f02-z-.tif



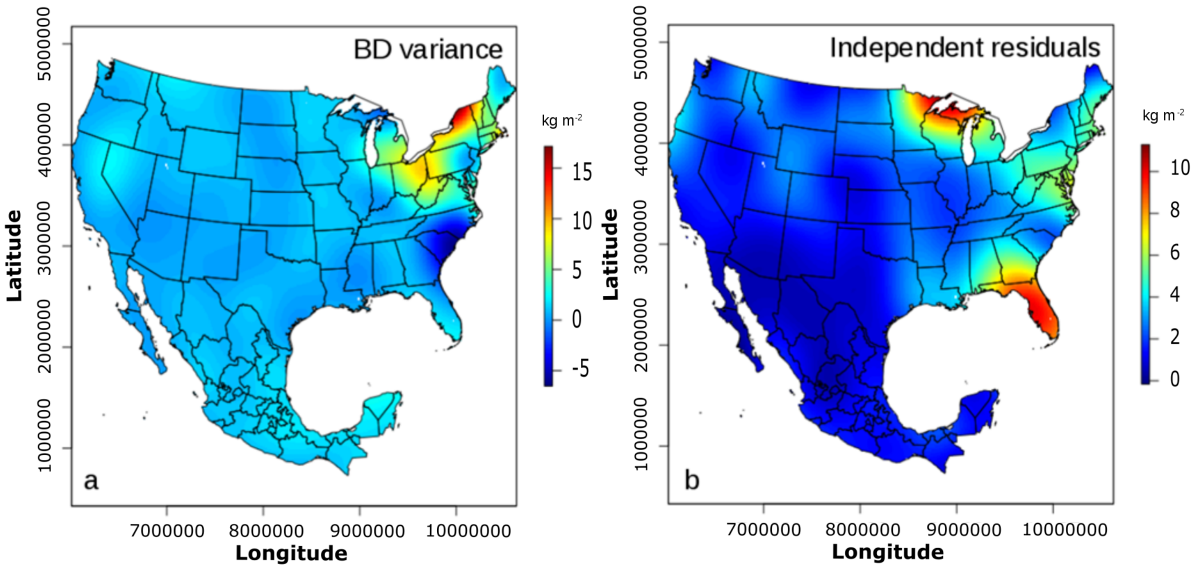
2019GB006219-f03-z-.tif



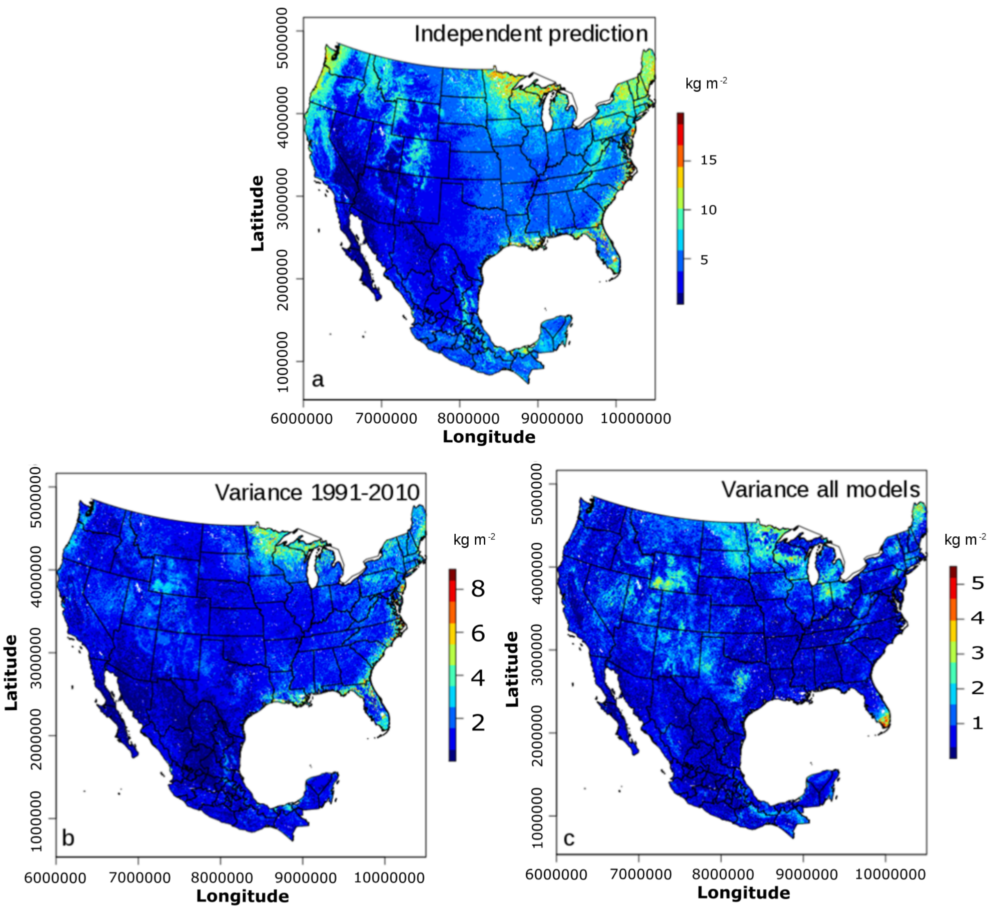
2019GB006219-f04-z-.tif



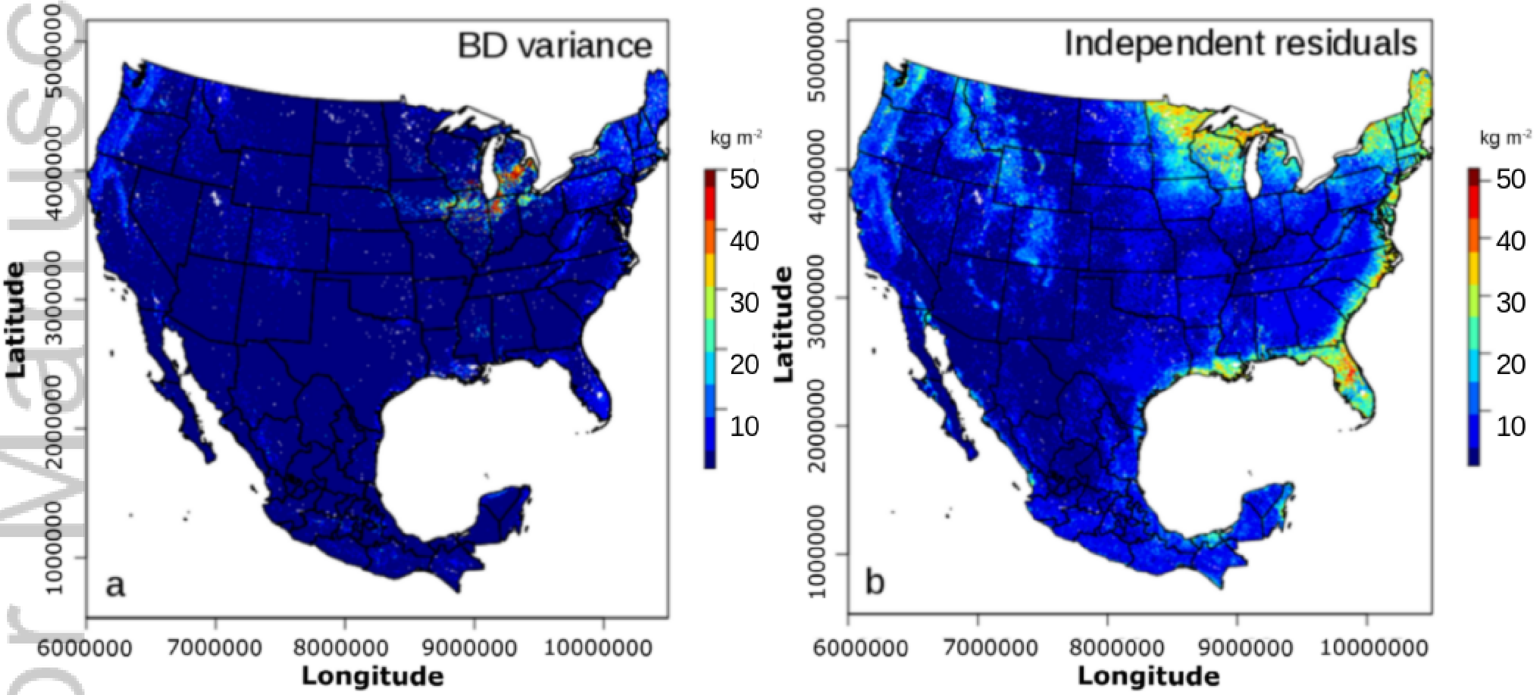
2019GB006219-f05-z-.tif



2019GB006219-f06-z-.tif



2019GB006219-f07-z-.tif



2019GB006219-f08-z-.tif