



No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America

Mario Guevara¹, Guillermo Federico Olmedo^{2,3}, Emma Stell¹, Yusuf Yigini³, Yameli Aguilar Duarte⁴, Carlos Arellano Hernández⁵, Gloria E. Arévalo⁶, Carlos Eduardo Arroyo-Cruz⁷, Adriana Bolívar⁸, Sally Bunning⁹, Nelson Bustamante Cañas¹⁰, Carlos Omar Cruz-Gaistardo⁵, Fabian Davila¹¹, Martin Dell Acqua¹¹, Arnulfo Encina¹², Hernán Figueredo Tacona¹³, Fernando Fontes¹¹, José Antonio Hernández Herrera¹⁴, Alejandro Roberto Ibelles Navarro⁵, Veronica Loayza¹⁵, Alexandra M. Manueles⁶, Fernando Mendoza Jara¹⁶, Carolina Olivera¹⁷, Rodrigo Osorio Herмосilla¹⁰, Gonzalo Pereira¹¹, Pablo Prieto¹¹, Iván Alexis Ramos¹⁸, Juan Carlos Rey Brina¹⁹, Rafael Rivera²⁰, Javier Rodríguez-Rodríguez⁷, Ronald Roopnarine^{21,22}, Albán Rosales Ibarra²³, Kenset Amaury Rosales Riveiro²⁴, Guillermo Andrés Schulz²⁵, Adrian Spence²⁶, Gustavo M. Vasques²⁷, Ronald R. Vargas³, and Rodrigo Vargas¹

¹University of Delaware, Department of Plant and Soil Sciences, Newark, DE, USA

²INTA EEA Mendoza, San Martín 3853, Luján de Cuyo, Mendoza, Argentina

³FAO, Vialle de Terme di Caracalla, Rome, Italy

⁴Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, Mérida, Mexico

⁵Instituto Nacional de Estadística y Geografía, Aguascalientes, Mexico

⁶Zamorano University of Honduras and Asociación Hondureña de la Ciencia del Suelo, Tegucigalpa, Honduras

⁷National Commission for the Knowledge and Use of Biodiversity, Mexico City, Mexico

⁸Subdirección Agrología, Instituto Geográfico Agustín Codazzi, Bogotá, Colombia

⁹Oficina Regional de la FAO para América Latina y el Caribe, Santiago de Chile, Chile

¹⁰Servicio Agrícola y Ganadero, Santiago de Chile, Chile

¹¹Dirección General de Recursos Naturales, Ministerio de Ganadería, Agricultura y Pesca, Montevideo, Uruguay

¹²Facultad de Ciencias Agrarias de la Universidad Nacional de Asunción, Asunción, Paraguay

¹³Land Viceministry, Ministry of Rural Development and Land, La Paz, Bolivia

¹⁴Universidad Autónoma Agraria Antonio Narro Unidad Laguna, Torreón, Mexico

¹⁵Ministerio de Agricultura y Ganadería, Quito, Ecuador

¹⁶Universidad Nacional Agraria, Managua, Nicaragua

¹⁷Oficina Regional de la FAO para América Latina y el Caribe, Bogotá, Colombia

¹⁸Instituto de Investigación Agropecuaria de Panamá, Panamá, Panama

¹⁹Sociedad Venezolana de la Ciencia del Suelo, Caracas, Venezuela

²⁰Ministerio de Medio Ambiente, Santo Domingo, Dominican Republic

²¹Department of Natural and Life Sciences, COSTAATT, Port of Spain, Trinidad and Tobago

²²University of the West Indies, St. Augustine Campus, St. Augustine, Trinidad and Tobago

²³Instituto de Innovación en Transferencia y Tecnología Agropecuaria, San José, Costa Rica

²⁴Ministerio de Ambiente y Recursos Naturales de Guatemala, Ciudad Guatemala, Guatemala

²⁵INTA CNIA, Buenos Aires, Argentina

²⁶International Centre for Environmental and Nuclear Sciences, University of the West Indies, Kingston, Jamaica

²⁷Embrapa Solos, Rio de Janeiro, Brazil

Correspondence: Rodrigo Vargas (rvargas@udel.edu)

Received: 15 December 2017 – Discussion started: 25 January 2018

Revised: 15 June 2018 – Accepted: 24 June 2018 – Published: 1 August 2018

Abstract. Country-specific soil organic carbon (SOC) estimates are the baseline for the Global SOC Map of the Global Soil Partnership (GSOCmap-GSP). This endeavor is key to explaining the uncertainty of global SOC estimates but requires harmonizing heterogeneous datasets and building country-specific capacities for digital soil mapping (DSM). We identified country-specific predictors for SOC and tested the performance of five predictive algorithms for mapping SOC across Latin America. The algorithms included support vector machines (SVMs), random forest (RF), kernel-weighted nearest neighbors (KK), partial least squares regression (PL), and regression kriging based on stepwise multiple linear models (RK). Country-specific training data and SOC predictors (5×5 km pixel resolution) were obtained from ISRIC – World Soil Information. Temperature, soil type, vegetation indices, and topographic constraints were the best predictors for SOC, but country-specific predictors and their respective weights varied across Latin America. We compared a large diversity of country-specific datasets and models, and were able to explain SOC variability in a range between ~ 1 and $\sim 60\%$, with no universal predictive algorithm among countries. A regional ($n = 11\,268$ SOC estimates) ensemble of these five algorithms was able to explain $\sim 39\%$ of SOC variability from repeated 5-fold cross-validation. We report a combined SOC stock of 77.8 ± 43.6 Pg (uncertainty represented by the full conditional response of independent model residuals) across Latin America. SOC stocks were higher in tropical forests (30 ± 16.5 Pg) and croplands (13 ± 8.1 Pg). Country-specific and regional ensembles revealed spatial discrepancies across geopolitical borders, higher elevations, and coastal plains, but provided similar regional stocks (77.8 ± 42.2 and 76.8 ± 45.1 Pg, respectively). These results are conservative compared to global estimates (e.g., SoilGrids250m 185.8 Pg, the Harmonized World Soil Database 138.4 Pg, or the GSOCmap-GSP 99.7 Pg). Countries with large area (i.e., Brazil, Bolivia, Mexico, Peru) and large spatial SOC heterogeneity had lower SOC stocks per unit area and larger uncertainty in their predictions. We highlight that expert opinion is needed to set boundary prediction limits to avoid unrealistically high modeling estimates. For maximizing explained variance while minimizing prediction bias, the selection of predictive algorithms for SOC mapping should consider density of available data and variability of country-specific environmental gradients. This study highlights the large degree of spatial uncertainty in SOC estimates across Latin America. We provide a framework for improving country-specific mapping efforts and reducing current discrepancy of global, regional, and country-specific SOC estimates.

1 Introduction

Soils store around 1500 Pg of carbon and represent the largest terrestrial carbon pool (Jackson et al., 2017); thus, it is critical to accurately quantify the variability of soil organic carbon (SOC) from local to global scales. During the fourth session of the Global Soil Partnership (GSP) Plenary Assembly held in May 2016 in Rome, it was agreed to develop a Global Soil Organic Carbon Map (GSOCmap) (FAO, 2017). The overarching goal is that a Global SOC Map of the Global Soil Partnership (GSOCmap-GSP) will be developed using a distributed approach relying on country-specific SOC maps. Country-specific maps represent a valuable source of information to explain the high discrepancy of current global SOC estimates such as the SoilGrids250m system and the Harmonized World Soil Database (Tifafi et al., 2018). The Food and Agriculture Organization (FAO) recently compiled how different statistical methods (e.g., regression kriging and machine learning) could be used to generate country-specific SOC maps and calculate uncertainty (Yigini et al., 2018). All these approaches consider the reference framework of the Soils, Climate, Organisms, Parent material, Age and (N) space or spatial position (SCORPAN) model for digital soil mapping (DSM) (McBratney et al., 2003). In the SCORPAN reference framework, a soil attribute (e.g., SOC) can be pre-

dicted as a function of the soil-forming environment, in correspondence with soil-forming factors from the Dokuchaev hypothesis and Jenny's soil-forming equation based on climate, organisms, relief, parent material, and elapsed time of soil formation (Florinsky, 2012). The SCORPAN reference framework is an empirical approach that can be expressed as in Eq. (1):

$$S_{a[x;y;t]} = f(S_{[x;y;t]}, C_{[x;y;t]}, O_{[x;y;t]}, R_{[x;y;t]}, P_{[x;y;t]}, A_{[x;y;t]}), \quad (1)$$

where S_a is the soil attribute of interest at a specific location N (represented by the spatial coordinates of field observations $x; y$) and at a specific period of time (t); S is the soil or other soil properties that are correlated with the soil attribute of interest (S_a); C is the climate or climatic properties of the environment; O is the organisms, vegetation, fauna, or human activity; R is topography or landscape attributes; P is parent material or lithology; and A is the substrate age or the time factor. To generate predictions of S_a across places where no soil data are available, N should be explicit for the information layers representing the soil-forming factors. These predictions will be representative of a specific period of time (t) when soil available data were collected. Therefore, the prediction factors ideally should represent the conditions of the soil-forming environment for the same period

of time (as much as possible) when soil available data were collected. In Eq. (1), the left side is usually represented by the available geospatial soil observational data (e.g., from legacy soil profile collections) and the right side of the equation is represented by the soil prediction factors. These prediction factors are normally derived from four main sources of information: (a) thematic maps (i.e., soil type, rock type, land use type); (b) remote sensing (i.e., active and passive sensors); (c) climate surfaces and meteorological data; and (d) digital terrain analysis or geomorphometry. The SCORPAN reference framework is widely used, but one critical challenge is to quantify the relative importance of the soil-forming factors (i.e., prediction factors) that could explain the underlying soil processes controlling the spatial variability of a specific soil attribute (i.e., SOC).

Arguably, there are two approaches for statistical modeling (Breiman, 2001) that influence the predictions of the spatial variability of SOC. One assumes that the variability of observations can be reproduced by a given stochastic data model (e.g., with hypotheses about the spatial structure of the variable). The other approach uses algorithms and treats as unknown the mechanisms generating the structure of values in available datasets (e.g., with hypothesis about the statistical distribution and moments of the variable). For SOC modeling, the accuracies of global models compared with country-specific estimates have not been systematically evaluated on detail. While globally available SOC predictions rely on large and complex multivariate spaces to represent the soil-forming environment, local (i.e., more simple) models may be useful for validation purposes and required to measure the bias of global SOC estimates, specifically, at particular sites/countries (well represented by available data), where SOC drivers may be easier to identify due to a smaller range of SOC variance. In addition, the assumptions of global models compared with local efforts may be different, and local datasets could complement global information sources. Because different mapping approaches use available information (i.e., training data and predictors) in different ways, comparing several approaches and methods is useful to quantify the relative importance of prediction factors across data configurations and distributional properties. We argue that a systematic analysis of predictive algorithms and consequently selection of predictors (by each one of the algorithms) could provide insights about the underlying factors that control the spatial variability of SOC.

The last decade has seen an increasing diversity of approaches for DSM. Data mining techniques have been successfully used to model and predict the spatial variability of soil properties (Rossel and Behrens, 2010; Hengl et al., 2017; Shangquan et al., 2017) and generate site-specific and country-specific SOC maps (Viscarra Rossel et al., 2014; Adhikari et al., 2014). The combination of regression modeling approaches with geostatistics of independent model residuals (i.e., regression kriging) is a combined strategy that has been widely used to map SOC (Hengl et al., 2004; Mishra

et al., 2009; Marchetti et al., 2012; Kumar et al., 2012; Peng et al., 2013; Adhikari et al., 2014; Yigini and Panagos, 2016; Nussbaum et al., 2014; Mondal et al., 2017). Machine learning algorithms such as random forests or support vector machines have also been used to increase statistical accuracy of soil carbon models (Martin et al., 2011; Hashimoto et al., 2017; Hengl et al., 2017) including applications for SOC mapping (Grimm et al., 2008; Sreenivas et al., 2016; Yang et al., 2016; Hengl et al., 2017; Delgado-Baquerizo et al., 2017; Ließ et al., 2016; Viscarra Rossel et al., 2014). Machine learning methods do not necessarily allow to extract information about the main effects of prediction factors in the response variable (e.g., SOC); consequently, a variable selection strategy is always useful to increase the interpretability of machine learning algorithms. With this diversity of approaches, one constant question is if there is a method that systematically improves the prediction capacity of the others aiming to predict SOC across large geographic areas (e.g., Latin America). We postulate that probably there is no universal method (i.e., silver bullet) for DSM, but both global and country-specific efforts are needed to test a variety of predictive algorithms including variable and parameter selection strategies for maximizing explained variance while minimizing prediction bias.

To minimize bias in SOC predictions, it is required to build baseline reference estimates to quantify SOC stocks and contribute to better parameterization for projections of SOC under future soil weathering conditions and land degradation scenarios. Therefore, SOC estimates based on statistical predictions should be ideally based on all available information for specific countries or regions of interest, from both national and global information sources. However, the availability of public SOC information is limited across large areas of Latin America and large discrepancies exist in current global SOC estimates (Tifafi et al., 2018). Thus, there is a pressing need to validate the accuracy of global SOC estimates, improve interoperability (Vargas et al., 2017) and contribute to the capacity of countries to meet the Global-SoilMap specifications (Arrouays et al., 2017) to inform policy decisions around climate change mitigation strategies.

This study focuses on Latin America, where site- or region-specific modeling efforts report high explained variance when mapping SOC (Reyes-Rojas et al., 2018). Accurate SOC maps are required to identify areas with the potential for soil carbon sequestration, and distinguish them from areas with high SOC. However, site-specific efforts to map SOC across Latin America highlight the challenge of predicting pedologically sound soil maps due to the complexity of SOC spatial variability (Angelini et al., 2016), including the inconsistencies of using simple linear approaches to explain soil and depth interrelationships (Angelini et al., 2017). Site-specific SOC mapping efforts across Latin America also suggest that variable selection and the spatial detail of SOC prediction factors also contribute to discrepancies of SOC predictions (Samuel-Rosa et al., 2015). To increase the accuracy

of SOC predictions, the use of high-performance computing through open-source platforms (i.e., Google Earth) represents a valuable resource to make and continuously update (as new and better data become available) country-specific SOC maps (Padarian et al., 2017). The constant challenge is how to increase SOC prediction accuracy while also reducing the uncertainty and granularity of SOC grids.

The overarching goal of this study is to compare different predictive algorithms across 19 data/country scenarios with publicly available information to support the development of country-specific SOC maps to be included in the GSOCmap-GSP. Currently, SOC information across Latin America has been derived from global models such as the SoilGrids system or the Harmonized World Soil Database (Hengl et al., 2017; Köchy et al., 2015), which lack quantification of uncertainty and where large areas remain parameterized with limited country-specific information. This challenge is not unique for Latin America as many regions around the world (e.g., Africa, Siberia) have limited SOC information to parameterize models to estimate the SOC pool. To inform future SOC mapping efforts, this study addresses two specific questions: (a) which environmental variables (derived from publicly available information) have the highest correlations with country-specific SOC information, and (b) which method (i.e., predictive algorithm) is best to represent SOC across Latin America and within each country. We assumed that methods could inform each other as they may explain different aspects of SOC variability. The ultimate aim of this study is to empower capacities for digital SOC mapping across Latin America and to contribute to the discussion about the importance of integrating country-specific information for representing and predicting soil-related variables (e.g., SOC) to improve regional-to-global SOC predictions.

2 Methods

We based our methodological approach on public sources of information and methods implemented in open-source platforms for statistical computing. Thus, our framework for modeling SOC stocks (Fig. 1) could be reproduced across the world for comparative purposes between country-specific and global estimates.

2.1 SOC observations

Soil organic carbon information was extracted from the World Soil Information Service (WoSIS) soil profile database. This dataset represents a great harmonization effort in which a large number of national legacy datasets have been compiled. It includes local-to-national soil profile collections with a sampling strategy generally based on morphological soil attributes (Batjes et al., 2017). The goal of the GSOCmap-GSP is to produce global information for the first 30 cm; thus, we generated synthetic horizons for this depth using a mass-preserving spline approach (Bishop et al.,

1999). We applied a pedotransfer function based on organic matter (OM) if the bulk density (BLD) information was missing: $BLD = 1/(0.6268 + 0.0361 \times OM)$ (Yigini et al., 2018). We decided to use this equation because it showed less extreme values than other available pedotransfer functions during preliminary discussion and training exercises (data not shown). Another reason is that there is not a single pedotransfer function applicable to all conditions across Latin America. This equation is representative for soils with organic matter content between 0.17 and 13.5 % (Drew, 1973). For coarse fragments (CRFVOL), a value of 0 % was used for missing information prior to the mass-preservative spline modeling. SOC estimates (0 to 30 cm) were derived following a standardized SOC calculation method (Nelson and Sommers, 1982) (Eq. 2):

$$SOC_{stock} = \frac{ORCDR}{1000} \times \frac{H}{100} \times BLD \times \frac{(100 - CRFVOL)}{100}, \quad (2)$$

where ORCDR is SOC density ($g\ kg^{-1}$) and H is soil depth (30 cm).

Because of the limitations and uncertainty in the available BD and CRFVOL data, we also included an error approximation of SOC estimates. This error was derived using Global Soil Information Facilities (GSIF; Hengl, 2017) as explained in the next section.

2.2 SOC error estimates

The GSIF approach for estimating SOC (function OC-SKGM) includes an approximate error which we used to quantify the reliability of SOC estimates (Hengl et al., 2017). This error was approximated using the Taylor series method, by a truncated Taylor series centered by the means explained previously (Heuvelink, 2018). We mapped the error trend of SOC estimates by interpolating the values on a per country basis using the generic framework for predictive modeling based on machine learning and buffer (geographical) distances (Hengl et al., 2018). We followed this method to provide a spatial explicit measure of the SOC estimation error. We used this method because it can be implemented without prediction factors (e.g., only buffer distances) and because it is practically free of assumptions but considers the geographical proximity to and composition of the sampling location points as explained by its developers (Hengl et al., 2018). SOC error estimates represent a component of uncertainty of the overall quality of country-specific input data.

2.3 SOC training data and exploratory analysis

Each country-specific SOC dataset was transformed to its natural logarithm to reduce the right-skewed distribution of SOC values and because exploratory analysis showed that this transformation can improve the prediction capacity of further modeling methods. To analyze the statistical distribution of SOC values, a probability distribution function was

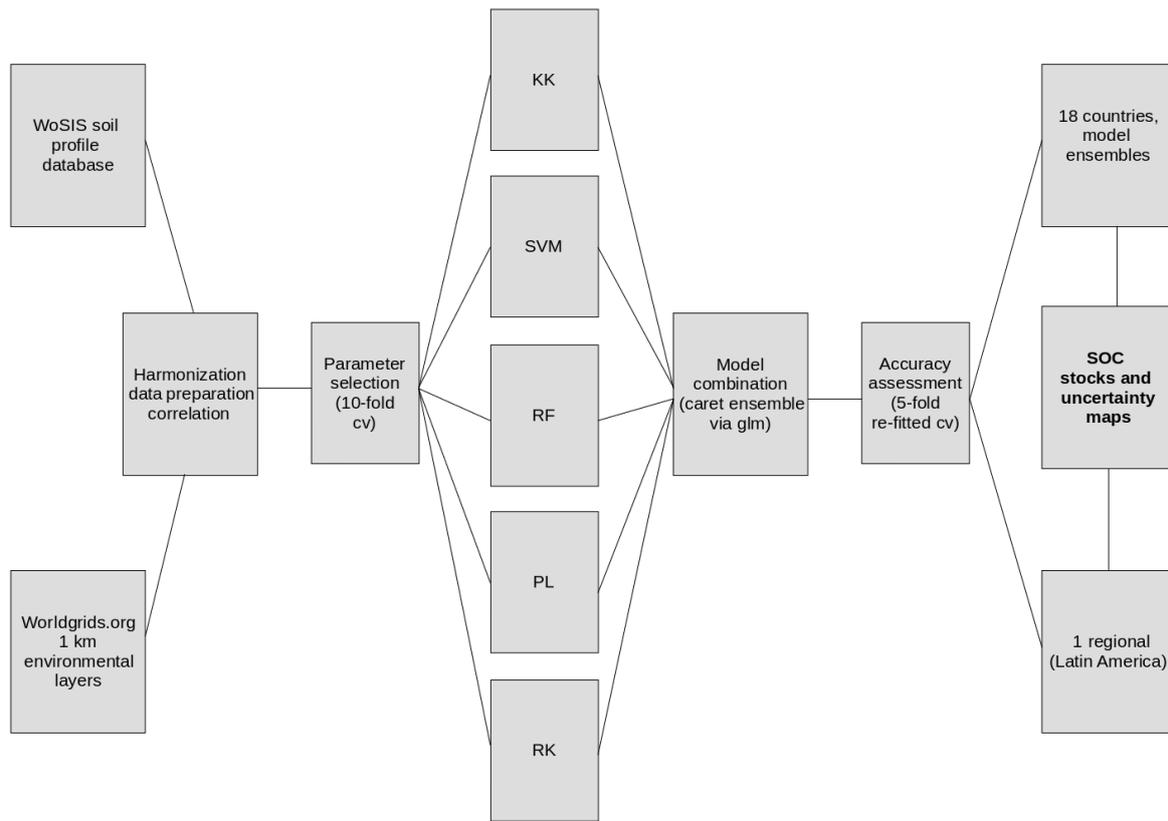


Figure 1. Flow diagram of the main methodological steps that we performed in order to generate country-specific and regional SOC predictions. The World Soil Information Service (WoSIS) dataset was harmonized with the <http://worldgrids.org> (last access: 20 February 2018) environmental data using 5×5 km grids. SOC stocks were calculated at points and correlated predictors identified. Five methods were parameterized and we created an ensemble of using a generalized linear approach. Accuracy of models and the ensembles was assessed with repeated cross-validation. Country-specific and regional (Latin America) ensembles were compared with global models. KK is kernel-weighted nearest neighbors, SVM is support vector machines, RF is random forests, PL is partial least squares regression, and RK is regression kriging.

plotted and a Shapiro–Wilk test of normality was conducted on each dataset. The units of the SOC estimates are kg m^{-2} . Our global (Latin America) dataset of 11 268 SOC estimates was divided using a simple bootstrapping technique (Kuhn et al., 2017) and 25 % of data were used for independent validation purposes, and the remaining 75 % of data for training prediction models. We coupled this information with a public source of prediction factors; see Sect. 2.4.

2.4 Soils prediction factors

We used environmental information from WorldGrids (worldgrids.org), which is an initiative of ISRIC-World Soil Information. We downloaded and masked 118 environmental layers (i.e., prediction factors) for each country to quantitatively represent the soil-forming environment (<http://worldgrids.org/doku.php/wiki:layers>, last access: 20 February 2018). The prediction factors were harmonized into a 1×1 km global grid by the WorldGrids project from three main information sources: remote sensing, climate surfaces, and digital terrain analysis. Additional terrain parameters

(e.g., terrain slope, aspect, catchment area, channel network base level, terrain curvature, topographic wetness index, and length–slope factor) from elevation data were calculated in the System for Automated Geoscientific Analyses geographic information system (SAGA GIS) for each country following the standard implementation for basic terrain parameters (Conrad et al., 2015). We resampled the prediction factors into a 5×5 km pixel size grid to reduce the computational demand required to make predictions and facilitate the reproducibility of this DSM framework without the need for high-performance computing.

2.5 Prediction of SOC

We made predictions on a country-specific and on a regional (Latin American) basis. We based our prediction framework on the following six steps:

- First, the relationship between SOC and prediction factors was explored using simple correlation analysis.

- Second, the 10 prediction factors with highest correlations with SOC data were identified for each country and used for further analyses.
- Third, we explored, parameterized, and compared five statistical methods with different assumptions to model SOC variability across Latin America: regression kriging (based on a multiple linear regression model (RK) and partial least squares (PLS) regression, support vector machines (SVMs), random forests (RF), and kernel-weighted nearest neighbors (KK). A brief explanation for each modeling approach is provided in Appendix A.
- Fourth, we used a five times repeated 5-fold cross-validation strategy of the aforementioned models to estimate the RMSE. Then, we used the caretEnsemble tools for stacking the five predictions (Deane-Mayer and Knowles, 2016; Kuhn et al., 2017). The caretEnsemble approach uses the RMSE to weight and create ensembles of regression models under a generalized approach to create a linear blend of predictions.
- Fifth, we calculated independent model residuals (by predicting the 25 % of data not used for model parameterization). For each 5×5 km pixel, we estimated the full conditional response of these residuals to the SOC prediction factors following the quantile regression method available within the quantregForest modeling framework (Meinshausen, 2017, 2006). We used this map as a surrogate of model uncertainty complementary to the approximated error trend of SOC estimates.
- Sixth, we used all Latin American data in the WoSIS system to repeat the fourth and fifth steps of our modeling framework, generating regional predictions of SOC and comparing with country-specific results and global SOC estimates. We also evaluated the prediction capacity of these models.

2.6 Model evaluation and accuracy

First, we selected the optimal parameters for each model/country by the means of a 10-fold cross-validation strategy following a generic recommendation (Borra and Di Ciaccio, 2010) (see parameter description in Appendix A). For each model, the train function of the caret package (Kuhn et al., 2017) included simple resampling techniques for automatic model parameter selection. Thus, we obtained unbiased residuals for each model/country that we compared using Taylor diagrams (Carslaw and Ropkins, 2012). A Taylor diagram summarizes multiple aspects of model performance, such as the agreement and variance between observed and predicted values (Taylor, 2001). In a Taylor diagram, each model is represented by a point in the

plot describing how well the patterns of observed and modeled values match each other. Two models have a similar predictive capacity if they overlap across the intersection of an error vector, a variance ratio, and a correlation vector.

We analyzed the overall ratio (EC_r) between model errors (RMSE) and the correlation between observed and predicted values (corr) for each model across all countries. We propose this ratio EC_r as an approach to better understand the agreement between the correlation (calculated by the means of cross-validation) and the RMSE (derived from the unbiased residuals of cross-validation). Before calculating the RMSE / correlation ratio, the RMSE and the correlation between observed and predicted values were standardized (by its maximum and minimum values) to a range between 0 and 1 using

$$RMSE_{SD} = \frac{RMSE_i - \min(RMSE)}{\text{range}(RMSE)} \quad (3)$$

$$\text{corr}_{SD} = \frac{\text{corr}_i - \min(\text{corr})}{\text{range}(\text{corr})} \quad (4)$$

$$EC_r = \frac{RMSE_{SD}}{\text{corr}_{SD}}, \quad (5)$$

where EC_r is the proposed ratio between errors and correlation between observed and predicted values; $RMSE_i$ is the observed RMSE for the i th model; $\min(RMSE)$ is the minimum observed value of RMSE, and $\text{range}(RMSE)$ is the difference between the maximum and minimum observed values of RMSE; corr_i is the observed correlation for the i th model; $\min(\text{corr})$ is the minimum observed value of correlation, and $\text{range}(\text{corr})$ is the difference between the maximum and minimum observed values of correlation

If the value of the EC_r was close to 0, then there was a stronger agreement between high RMSE and low correlation, or low RMSE and high correlation. If this value deviated from 0 (up to 1 or more), then the RMSE would tend to be high while the correlation was also high, suggesting that the method represented the variability of SOC but with high bias.

Model accuracy (also represented by the RMSE and R^2) was assessed for the model ensembles with a more strict (but computationally expensive) 5-fold and five times repeated cross-validation strategy. This model refitting allowed more stable accuracy results with the ultimate goal of comparing country-specific and regional (Latin America) estimates. Repeated 10- and 5-fold cross-validation have been used to compare both machine learning and geostatistical approaches for mapping soil properties from book examples to real applications at the global scale (Hengl et al., 2018, 2017). In addition, independent model residuals were also obtained from the 25 % of data not used for the country-specific and regional ensembles to estimate a spatially explicit measure of uncertainty (as explained in step five of our prediction framework).

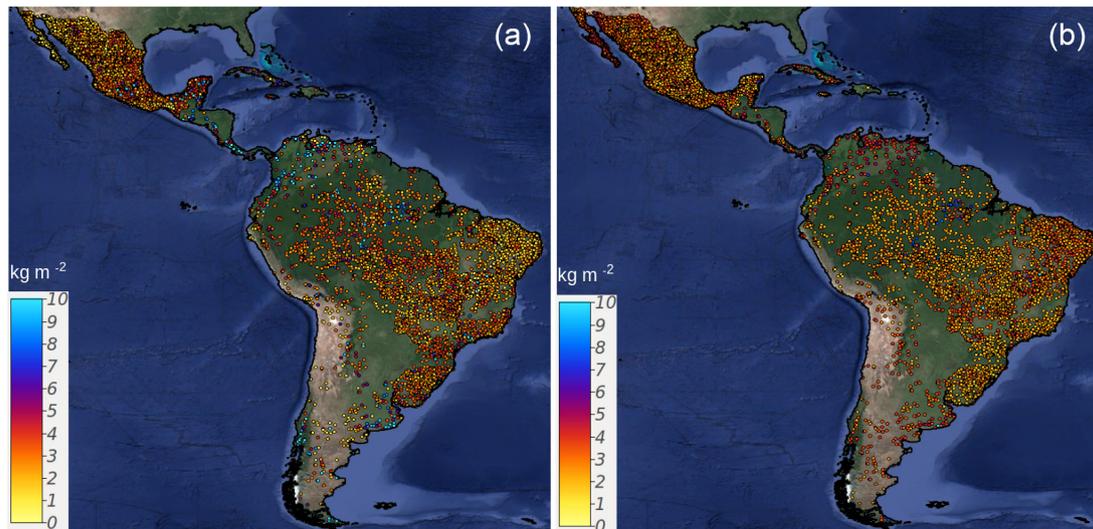


Figure 2. Spatial distribution of available SOC in WoSIS for Latin America. SOC estimates are calculated for each point using Eq. (2) (a). The approximated error is based on Taylor series as implemented in the R-GSIF package, as is explained in Heuvelink (2018) (b). Thus, panel (b) represents the uncertainty of SOC estimates at each point. The values of this map could be associated with data limitations and missing information for BLD and CRFVOL.

2.7 SOC stocks

First, we analyzed the influence of the maximum allowed prediction limits for each prediction algorithm. We harmonized the units of our SOC estimates with global datasets in Mg/ha (megagrams per hectare at 30 cm depth). The sensitivity of the total SOC stock to the model prediction limit was tested by increasing (every 10 Mg/ha) the maximum prediction limit from 0.5 Mg/ha, until finding a stable rate. Geopolitical limits were obtained from the Global Administrative Areas project (<https://gadm.org/>, last access: 16 July 2018). Using these country limits we report our country-specific and Latin American SOC estimates. For comparative purposes, we also extracted for each country the global SOC estimates from the SoilGrids system (Hengl et al., 2017), the Harmonized World Soil Database (Köchy et al., 2015), and the GSOCmap-GSP (see <http://54.229.242.119/apps/GSOCmap.html>, last access: 16 July 2018). We also report stocks across the land cover classes derived from the Latin American Network for Monitoring and Studying of Natural Resources, a product with an estimated accuracy of 84 % (Blanco et al., 2013). We report the overall uncertainty of these stocks with the independent model residuals map and the approximated error trend of the SOC estimates. Some countries with no data were filled with the average of the surrounding extent of the SOC predictions. All analyses were performed using the R software (R Core Team, 2017).

3 Results

3.1 Descriptive statistics

SOC across different countries showed a wide diversity of data scenarios (Table 1). Costa Rica (with a mean of 11.05 kg m^{-2}), Chile (with a mean of 9.88 kg m^{-2}), and Colombia (with a mean of 8.15 kg m^{-2}) are the countries with the highest SOC mean values. Brazil ($n = 5616$) and Mexico ($n = 4321$) were the countries with highest data availability. In contrast, Honduras ($n = 11$), Guatemala ($n = 20$), and Belize ($n = 21$) were the countries with lowest density of SOC estimated values (Table 1). With the original (untransformed) dataset, the only countries that showed a normal distribution after the Shapiro–Wilk test of normality with an α of 0.05 were Belize, Guatemala, Honduras, and Suriname.

3.2 Spatial distribution and point error estimates

There were large areas of Latin America with no available SOC observational data in the WoSIS system (e.g., the south of Chile, Argentina, or across large areas of Central America). We found substantial error estimates across large areas with high density of SOC data but low carbon contents, such as northern Mexico or the Brazilian semiarid savanna located at the eastern side of that country (Fig. 2).

3.3 Correlation of SOC and its predictors

Best correlated predictors were not the same across countries. We found higher correlations with the original datasets transformed to their natural logarithm, as data had a right-

Table 1. Descriptive statistics of SOC estimates (in kg m⁻²) and total land area for each analyzed country. *n* is the number of observations. We provide quantiles, median, mean, and the standard deviation of SOC data. The columns *p* and *p* log represent the probability values derived from the Shapiro–Wilk test of normality before *p* and after *p* log the log transformation of SOC values. When *p* is larger than *p* log, the log transformation of the data did not increase the probability of normality in the dataset. For comparative purposes, we provide (Fig. S1 in the Supplement) the probability distribution functions of available data before and after the log transformations. ARG is Argentina, BLZ is Belize, BOL is Bolivia, BRA is Brazil, CHL is Chile, COL is Colombia, CRI is Costa Rica, CUB is Cuba, ECU is Ecuador, GTM is Guatemala, HND is Honduras, JAM is Jamaica, MEX is Mexico, NIC is Nicaragua, PAN is Panama, PER is Peru, SUR is Suriname, SLV is El Salvador, URY is Uruguay, and VEN is Venezuela.

Country	<i>n</i>	Land area (km ²)	Min.	First <i>Q</i>	Med.	Mean	Third <i>Q</i>	Max.	SD	<i>p</i> / <i>p</i> log
ARG	231	2 736 690	0.34	1.88	3.21	5.65	5.96	86.85	9.33	< 0.001/0.03
BLZ	21	22 970	1.84	4.49	6.72	7.71	9.99	19.48	4.32	0.08/0.99
BOL	76	1 083 301	0.64	1.83	2.56	2.64	3.20	7.65	1.21	< 0.001/0.08
BRA	5616	8 358 140	0.07	1.99	2.67	3.23	3.34	573.76	9.18	< 0.001/ < 0.001
CHL	44	743 812	0.43	3.58	5.19	9.88	16.52	31.87	8.86	< 0.001/0.01
COL	166	1 038 700	0.66	3.44	5.78	8.15	9.95	52.62	7.35	< 0.001/0.96
CRI	43	51 060	2.27	4.07	7.23	11.05	10.85	82.57	14.90	< 0.001/0.001
CUB	48	109 820	0.36	2.85	3.61	4.32	5.73	10.98	2.23	0.004/ < 0.001
ECU	77	276 841	0.99	2.37	3.65	5.15	4.36	24.36	5.15	< 0.001/ < 0.001
GTM	20	107 159	2.60	5.66	8.48	7.73	9.75	12.41	3.11	0.14/0.007
HND	11	111 890	2.69	5.25	6.48	6.71	8.32	12.38	2.78	0.72/0.39
JAM	76	10 831	1.29	3.01	3.99	4.35	4.83	12.90	1.99	< 0.001/0.72
MEX	4321	1 943 945	0.00	1.73	2.49	2.56	3.25	35.55	1.49	< 0.001/ < 0.001
NIC	26	119 990	2.93	3.94	7.31	7.50	9.04	15.91	3.78	0.05/0.09
PAN	25	74 177	3.39	4.90	7.53	7.59	9.13	19.89	3.76	0.003/0.49
PER	145	1 279 996	0.19	1.89	2.93	2.92	3.55	8.35	1.42	0.005/ < 0.001
SUR	27	156 000	1.38	2.60	3.35	3.37	4.07	6.01	1.20	0.69/0.51
URY	130	175 015	0.82	2.70	3.38	4.34	3.90	46.54	4.67	< 0.001/ < 0.001
VEN	164	882 050	0.31	2.58	4.14	5.92	6.57	44.35	6.37	< 0.001/0.11

skewed distribution and did not follow a normal distribution (i.e., log normal). Highest correlations of available SOC data and their environmental predictors were associated with temperature-related variables across Honduras, Costa Rica, Peru, Chile, Guatemala, and Suriname (the r^2 varied from 0.35 to 0.58). However, there were a low number of available SOC observations across these countries in the WoSIS system (between 11 to 34). Similarly, across countries with high data availability (e.g., Mexico and Brazil), the strongest correlations between SOC and prediction factors were associated with temperature-related variables (Table 2). In all cases, the relationship between SOC and temperature-related variables was negative. In contrast, SOC had a positive relationship with elevation-derived terrain parameters (r^2 varied from 0.43 to 0.59) such as terrain curvature, potential incoming solar radiation, and slope of terrain.

Lower correlations of SOC data with prediction factors were found across Brazil, Bolivia, Uruguay, Cuba, Panama, Venezuela, and Argentina (e.g., $r^2 < 0.2$). The correlation analysis was useful to formulate a working hypothesis about the major drivers of the spatial variability of SOC across countries based on our DSM conceptual framework (e.g., $\text{SOC}_{\text{ARG}} = f[\text{px4wcl3a} + \text{px3wcl3a} + \text{evmmod3a} + \text{i07igb3a} + \text{px2wcl3a} + \dots]$). For example, the best correlated predictors with SOC for Argentina were

precipitation-related variables (px4wcl3a, px3wcl3a, px2wcl3a), remote-sensing-based vegetation indexes (evmmod3a), and a probability-based shrubland map (i07igb3a) (Table 2) (see sources of these maps in <http://worldgrids.org/doku.php/wiki:layers>, last access: 20 February 2018).

3.4 SOC-related properties

Correlations between SOC density (ORCDR) and prediction factors were higher with maximum and mean nighttime temperature, where Costa Rica and Chile had the highest correlations (r^2 varied from 0.61 to 0.71). The best correlated variables with BLD were terrain parameters: relative slope position, vertical distance to channel network, flow accumulation areas, and potential incoming solar radiation. These correlations were stronger across Guatemala, Belize, and Panama (r^2 varied from 0.52 to 0.67). We found that terrain slope and the standard deviation of temperature were the variables with highest correlations with CRFVOL, where Nicaragua, Honduras, and Argentina had the highest correlations (r^2 varied from 0.40 to 0.55). We did not find a dominant algorithm to predict ORCDR, BLD, and CRFVOL. Slightly higher correlations between observed and predicted values were achieved with RF, but in most cases different methods showed similar

Table 2. Best correlated predictors and their frequency across the analyzed data country scenarios, given available data in the WoSIS system; see predictor codes in <http://worldgrids.org/doku.php/wiki:layers> (last access: 20 February 2018). ARG is Argentina, BLZ is Belize, BOL is Bolivia, BRA is Brazil, CHL is Chile, COL is Colombia, CRI is Costa Rica, CUB is Cuba, DOM is Dominican Republic, ECU is Ecuador, GTM is Guatemala, HND is Honduras, JAM is Jamaica, MEX is Mexico, NIC is Nicaragua, PAN is Panama, PER is Peru, SUR is Suriname, SLV is El Salvador, URY is Uruguay, and VEN is Venezuela.

Var	Factor	Subfactor	Freq.	Country
gachws3a	Soil	Soil type	2	CUB, SUR
garhws3a	Soil	Soil type	2	PER, URY
ghshws3a	Soil	Soil type	2	BLZ, URY
gphws3a	Soil	Soil type	2	CUB, JAM
gplhws3a	Soil	Soil type	2	BLZ, BOL
gyrhws3a	Soil	Soil type	2	JAM, URY
tdmmod3a	Climate	Temperature	11	ARG, BOL, BRA, CHL, COL, CRI, CUB, ECU, MEX, PER, VEN
tx1mod3a	Climate	Temperature	10	ARG, BOL, BRA, COL, CUB, ECU, JAM, NIC, PER, URY
tx4mod3a	Climate	Temperature	10	BRA, CHL, CRI, CUB, ECU, GTM, JAM, MEX, PER, VEN
tx5mod3a	Climate	Temperature	9	BOL, BRA, CHL, CUB, ECU, JAM, MEX, PER, VEN
tx6mod3a	Climate	Temperature	9	ARG, BOL, BRA, CHL, COL, CRI, ECU, MEX, VEN
tnhmod3a	Climate	Temperature	8	BLZ, COL, CRI, GTM, HND, JAM, PAN, VEN
tnmmod3a	Climate	Temperature	8	BLZ, COL, CRI, GTM, HND, PAN, URY, VEN
tx3mod3a	Climate	Temperature	7	BRA, CHL, CUB, ECU, PAN, PER, VEN
tdhmod3a	Climate	Temperature	6	ARG, CUB, ECU, JAM, MEX, URY
tdlmod3a	Climate	Temperature	6	BRA, CHL, COL, ECU, GTM, JAM
tnsmod3a	Climate	Temperature	5	ARG, MEX, NIC, PAN, SUR
tx2mod3a	Climate	Temperature	4	ARG, ECU, PER, URY
tdsmod3a	Climate	Temperature	3	MEX, PAN, SUR
tnlmod3a	Climate	Temperature	3	BLZ, COL, GTM
px2wcl3a	Climate	Precipitation	2	BOL, PAN
px3wcl3a	Climate	Precipitation	2	CHL, MEX
px4wcl3a	Climate	Precipitation	2	BRA, CHL
etmmts3a	Climate	ET	2	ARG, MEX
evmmod3a	Organism	Vegetation	5	ARG, ECU, HND, MEX, VEN
107igb3a	Organism	Vegetation	2	ARG, CHL
DEMSRE3a	Topography		5	COL, CRI, GTM, HND, SUR
twisre3a	Topography		5	BRA, JAM, NIC, PAN, SUR
ChannNetworkBLevel	Topography		4	COL, HND, PAN, SUR
l3pobi3b	Topography		4	COL, CRI, PAN, VEN
inssre3a	Topography		3	BLZ, HND, SUR
opisre3a	Topography		3	CRI, NIC, SUR
SLPSRT3a	Topography		3	CRI, NIC, SUR
AnalyticalHillshading	Topography		2	BLZ, CUB
Aspect	Topography		2	BLZ, BOL
CovergenceIndex	Topography		2	BOL, HND
inmsre3a	Topography		2	CRI, GTM
ValleyDepth	Topography		2	BLZ, JAM
geaisg3a	Age		3	CHL, NIC, SUR

prediction capacity. The highest prediction error was found with RK for CRFVOL, but for all other output variables all prediction algorithms had a similar range of errors (Fig. 3). The PLS and SVM had the lowest variance for prediction of each one of the four soil properties. The r^2 values for predicting the combined SOC-related properties (i.e., ORCDR, CRFVOL, and BLD) for each prediction algorithm were RK (r^2

0.67 to 0.76), RF (r^2 0.56 to 0.74), SVM (r^2 0.32 to 0.71), PL (r^2 0.46 to 0.69), and KK (r^2 0.19 to 0.64). Across countries with lower data availability and sparse distribution, SVM and RK algorithms resulted in lower model performance.

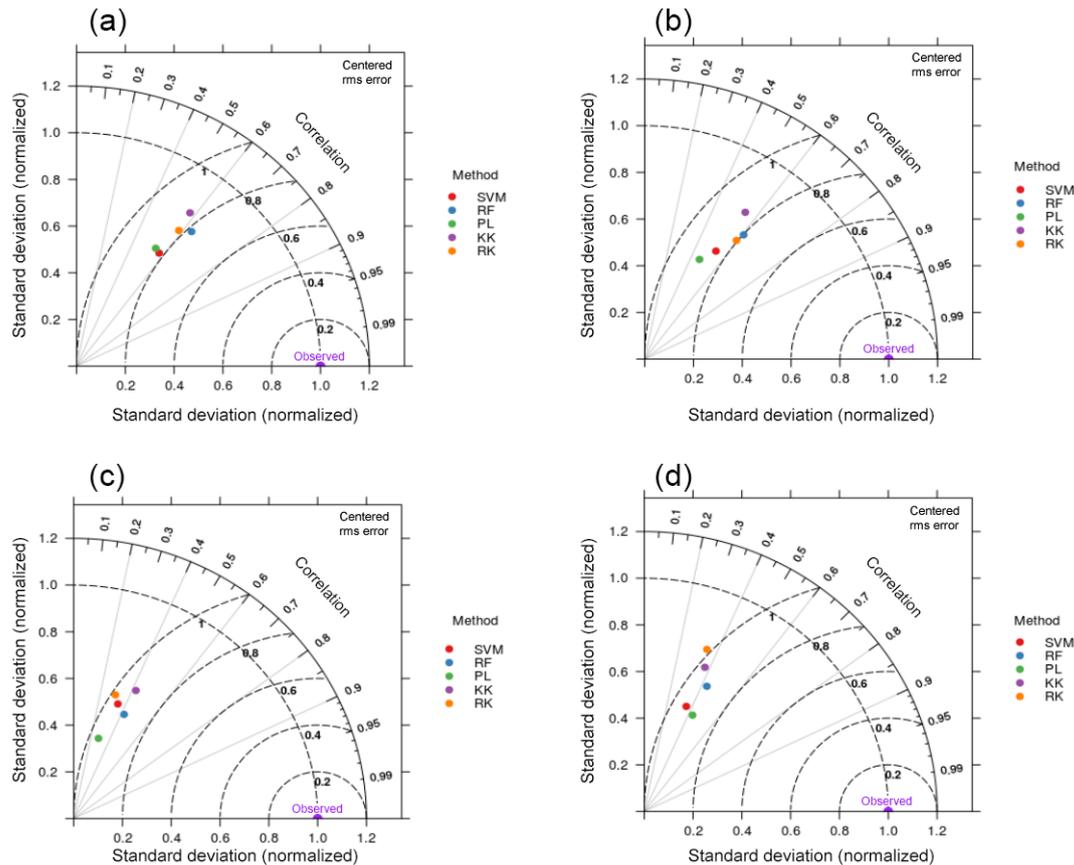


Figure 3. Taylor diagrams showing the performance of the five models evaluated. SOC stock (a), ORCDR (b), BLD (c), and CRFVOL (d). This analysis is based on all available data across Latin America. Although RF tends to generate higher correlation, it also shows high variance in predictions. The points are close to each other and the differences in accuracy between them generally fall within the same intersection of error, variance, and correlation, suggesting a similar prediction capacity by the implemented approaches.

3.5 Country-specific SOC predictions

We did not find a dominant algorithm to predict SOC on a country-specific basis (Fig. 4). Overall, machine learning prediction algorithms generated similar results. Higher agreement of machine learning prediction algorithms was found in small countries where environmental conditions and land cover/use characteristics tend to be more homogeneous (e.g., Jamaica, Suriname). RK showed higher discrepancies in countries where data distribution was sparse (e.g., Suriname, Chile, Guatemala) but effective across countries with higher and/or well-distributed data availability (e.g., Mexico, Brazil). Machine learning SOC predictions were conservative compared with RK (RK generated the higher density of extreme and unreliable SOC values). PL had comparable results with machine learning algorithms (i.e., KK, SVM, RF). From the cross-validation strategy, higher r^2 values between observed and predicted data were found for Costa Rica (0.58; $n = 21$) using SVM, while the lowest error was found in Suriname (0.36 kg m^{-2} ; $n = 37$) using PL. In contrast, algorithms had lower prediction capacity for countries with large areas (e.g., Brazil, Mexico) despite the large data availability.

The simple correlation (main effect) between the r^2 and RMSE for RF, PL, KK, and RK was positive (0.18, 0.35, 0.32, and 0.1, respectively). In contrast, this correlation was stronger for SVM (but negative; -0.65) where increasing the explained variance resulted in a lower error. Thus, we found a low level of agreement between these two information criteria (r^2 and RMSE) commonly used in DSM to assess performance of prediction algorithms.

Agreement between the RMSE and r^2 was found only in 12 of the 19 countries, resulting in country-specific “recommended” prediction algorithms. Here, we list the prediction algorithms that generated the best correlation and the best RMSE for each country: ARG (RK, RK), BLZ (RF, RK), BOL (SVM, KK), BRA (RF, RF), CHL (PL, PL), COL (RF, RF), CRI (SVM, SVM), CUB (PL, PL), ECU (RK, RK), GTM (KK, RF), HND (SVM, KK), JAM (RF, RF), MEX (RK, RK), NIC (RF, RF), PAN (PL, KK), PER (KK, KK), SUR (SVM, PL), URY (RF, RK), and VEN (RK, RK) (see country codes in Table 1). Brazil and Mexico had the highest number of observations (nearly 80 % of the total) and the same method yielded the highest r^2 and the lowest RMSE.

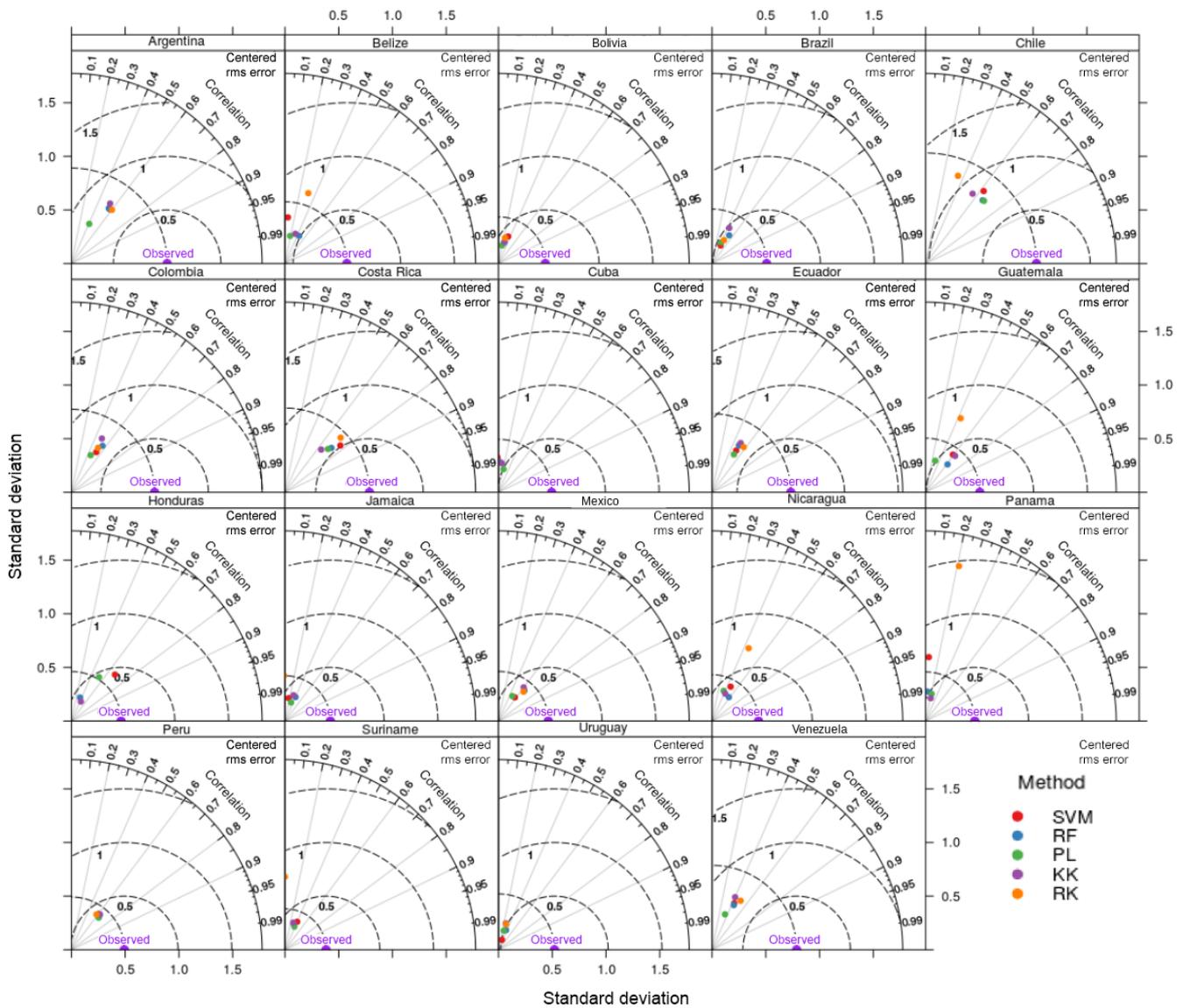


Figure 4. Taylor diagrams showing the performance of the five models evaluated for country-specific SOC estimates across Latin America. The position of each point/method varies from each dataset to another, suggesting that the predictive capacity changes when data characteristics are different.

We clarify that the best within-country method was not the same for every country. The higher EC_r was found with PL (0.96), followed by RF (0.54) and KK (0.43), informing that these predictive algorithms did not minimize prediction bias while increasing the explained variance. SVM (with 0.008) and RK (with 0.003) had the lowest EC_r , as they maximize the explained variance while minimizing prediction bias.

3.6 Model ensembles and SOC maps

High discrepancy was found among country-specific SOC predictions and between country-specific and regional SOC predictions. Although both maps predict SOC following a similar general pattern, the country-specific ensemble

showed a higher density of unrealistic patterns across Guatemala, Venezuela, northern Brazil, and the surroundings of Uruguay (Fig. 5a). These areas correspond to areas where we report both higher SOC calculation errors and model uncertainty (Fig. 6).

Compared with the country-specific ensemble, the regional model showed spatial differences predicting higher SOC across the highlands of the Southern Andes and boundaries of the Amazon Basin (Fig. 5b). As expected, the country-specific model showed spatial artifacts associated with country geopolitical borders. Based on the repeated 5-fold cross-validation, we report a $r^2 = 0.39$ for the regional model and r^2 values for the country-specific approach that vary from 0.01 to 0.55.

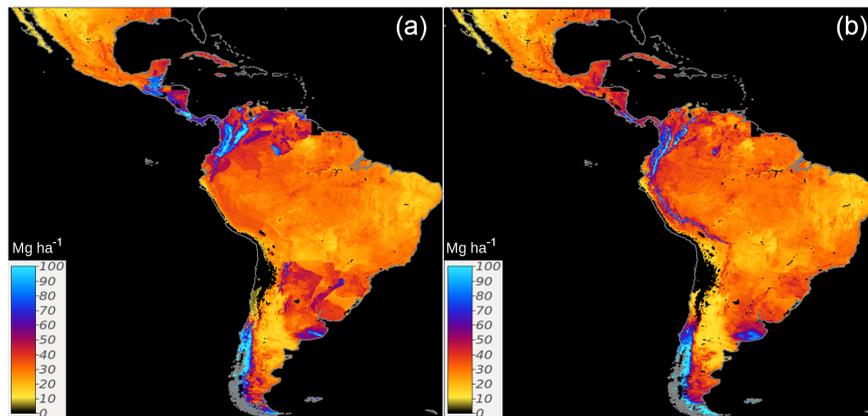


Figure 5. Country-specific (a) and regional (Latin America) (b) predictions of SOC based on a linear ensemble of methods. We present the units as Mg ha^{-1} for visualization purposes. These units were used to reduce the digits of the value range and highlight larger differences between SOC maps.

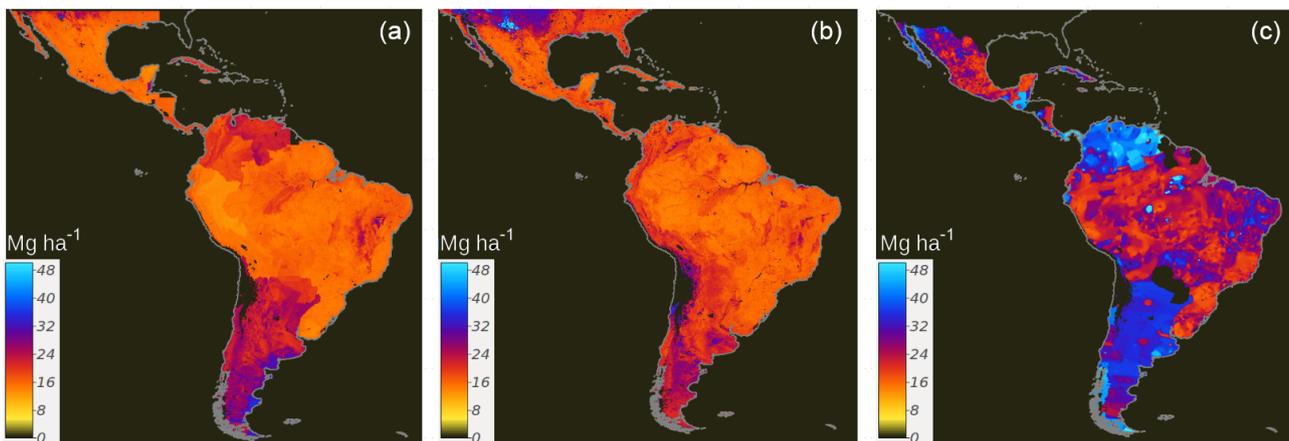


Figure 6. The full conditional response of residuals to the prediction factors on a country-specific basis (a). The full conditional response of residuals to the SOC prediction factors in the regional (Latin America) model (b). The trend of the approximated error of SOC estimates is derived from buffer distances and the random forest spatial framework (c).

High uncertainty in our modeling framework was found across tropical, arid, and semiarid regions of Latin America (Fig. 6a, b). Residual uncertainty from independent validation in the country-specific ensemble showed higher errors across geopolitical borders (in Chile, Argentina, Colombia, Ecuador, Venezuela, and the Brazilian savanna), while the residual uncertainty map from the regional model had higher uncertainty across ecologically meaningful transitions, with no evident effect of geopolitical borders. The trend of the mean approximated error suggests high uncertainty in the SOC calculation method (Fig. 6c). We used this map just to visualize the general trend of error estimates based only on geographical buffer distances.

Primarily, the Pacific coastal plains, the delta of the Amazon river, some closed watersheds and wetlands across Mexico, and some sparse points across Central America showed the higher discrepancies. Mexico and Brazil, with higher

density of SOC data, were the countries with less discrepancy between country and global models (Fig. 7a). We report that the geographical areas where country-specific models tend to predict higher SOC values than the regional ensemble (Fig. 7b). However, we report a similar SOC stock from both modeling approaches (country-specific and global) as we explain in Sect. 3.7.

3.7 SOC stocks and model uncertainties

For comparative purposes with previous reports (i.e., the SoilGrids system and the Harmonized World Soil Database), we harmonized the units of our maps to Mg ha^{-1} , which was also useful for visualization purposes. For our models, the uncertainty of the maximum prediction limit was estimated to be $\pm 10 \text{ Pg}$, which was the variance of the SOC stock by increasing the prediction limit from 1 to 700 Mg ha^{-1} (Fig. 8).

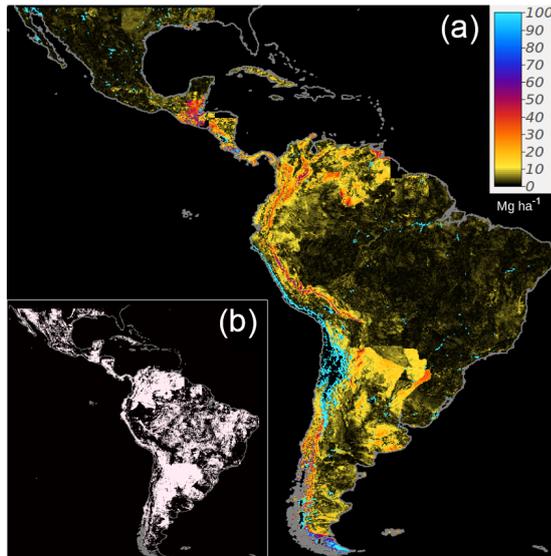


Figure 7. The absolute distance (Mg/ha) between the country-specific and the regional ensemble (a). The areas in white are areas where the country-specific modeling is predicting higher SOC than the regional estimate (i.e., country-specific is greater than regional) (b).

This relationship showed a stable (close to 0) trend after 200 Mg/ha. A larger density of extreme values was found with the regional model, and we calculated a maximum possible SOC stock of 83.62 Pg with this model.

Despite the spatial differences reported for the country-specific and regional ensembles, we report a similar stock between both approaches (77.8 ± 42.2 and 76.8 ± 45.1 Pg, respectively). We found that the global ensemble yields a slightly higher uncertainty. Our country-specific ensembles suggested that countries with highest SOC stocks were Brazil, Argentina, Colombia, Mexico, Peru, and Venezuela (Table 3).

Consistently, all models showed that tropical broadleaf evergreen forests, croplands, and temperate shrublands were the land cover classes that had higher SOC across all SOC available estimates (Table 4). However, using only the dataset contained in the WoSIS system, we predict nearly the half of SOC compared with previously reported SOC estimates such as the SoilGrids system (Table 3).

The model variance of predicted SOC reached values over 300 % for countries such as Mexico and Bolivia. In contrast, countries with higher SOC per unit area and relatively low prediction variances were Panama, Guatemala, Costa Rica, Nicaragua, and Belize. Overall, we found a median model prediction variance of 53 % across countries in Latin America. Areas with high uncertainty and model variance were across northern Mexico, Central America, limits between Colombia and Brazil, and the border between Chile and Argentina.

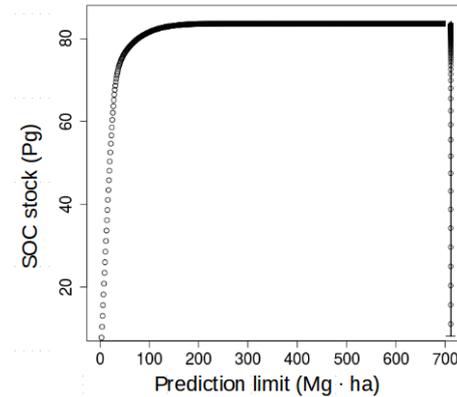


Figure 8. Relationship between the SOC stock and the prediction limit. The average breakdown points of this relationship are shown in the vertical line at the right of the plot.

4 Discussion

We developed a DSM framework to characterize the spatial variability of SOC across Latin America. Our results suggest that a multi-model approach was suitable to better understand modeling bias and uncertainty of SOC maps. We argue that uncertainty on SOC mapping can be associated with (a) the complexity of the property of interest (i.e., SOC), (b) the environmental heterogeneity within the area/country of interest, and (c) the characteristics of available data (e.g., data density, data quality, and data representativeness) to meet model-specific assumptions. Thus, when legacy soil profile collections that were collected for different purposes along long periods of time (i.e., decades), a multi-model approach (i.e., ensemble) would be convenient to maximize the predictive capacity considering the available information.

To maximize accuracy of our models, we used a generalized linear approach to combine single predictions, and at the continental scale we were able to explain 39 % of SOC variance using only information contained in the WoSIS system for Latin America. This result was within the range of the prediction capacity of country-specific models. Besides the low density of observation points, the performance could be partially affected by the generalization from the 1 : 1 scale of a soil profile (or field SOC observation) to a 5×5 km grid, representing an additional source of uncertainty. Higher discrepancy between country-specific and global efforts was evident across Brazil, the largest country, where our models tend to predict nearly half of SOC compared to previous efforts (e.g., the GSOCmap-GSP, the SoilGrids system, and the Harmonized World Soil Database). The SoilGrids system tends to predict the highest values, while our country-specific ensemble predicts the lowest. The GSOCmap-GSP and our ensembles predicted < 100 Pg of SOC across the analyzed countries, while all other products suggest higher stocks (see Tables 3 and 4).

Table 3. SOC stocks (Pg) at the contextual resolution of 5×5 km grids. The terms used are defined as follows: ens is country-specific, regional is Latin America ensemble, sg is the SoilGrids system, GSOCmap-GSP is country-specific 1 km, and hw is the Harmonized World Soil Database.

	Country	ens	regional	sg	GSOCmap-GSP	hw
1	Argentina	13.19	12.77	24.45	18.00	18.13
2	Belize	0.24	0.12	0.28	0.28	0.19
3	Bolivia	3.29	3.39	8.39	6.99	5.96
4	Brazil	26.82	27.16	68.45	42.79	47.20
5	Chile	6.31	7.20	15.15	1.93	8.28
6	Colombia	7.01	5.96	15.50	5.12	14.99
7	Costa Rica	0.56	0.34	0.83	0.83	0.71
8	Cuba	0.52	0.51	1.48	0.82	0.64
9	Ecuador	1.31	1.36	4.04	1.57	2.63
10	Guatemala	1.02	0.57	1.27	1.27	0.99
11	Jamaica	0.05	0.05	0.14	0.07	0.07
12	Mexico	5.98	6.12	14.43	9.04	17.59
13	Nicaragua	0.74	0.62	1.42	0.71	0.92
14	Panama	0.56	0.43	1.10	0.33	0.69
15	Peru	4.38	5.13	17.08	3.14	10.51
16	Suriname	0.56	0.51	1.20	0.45	1.33
17	Uruguay	0.92	0.88	1.99	0.84	2.27
18	Venezuela	4.71	3.77	9.39	5.28	5.64

Another source of discrepancy can be associated with the lack of available data to represent the SOC stock at the depth of interest (i.e., -30 cm of mineral soil). The predictive performance of the mass-preservative spline to continuously represent the SOC and depth relationships in some cases could be strongly influenced by the lack of observations across highly variable soil profiles. Some examples include SOC-rich agricultural soil profiles constantly transformed for food production purposes, or a volcanic setting. These high levels of missing data lead the trend map of approximated error (Fig. 6), which provides an idea of the uncertainty in the SOC estimates.

The GSOCmap-GSP, for example, was generated on a country basis, but the amount of SOC observations used for the countries to generate these maps was considerable higher than the available data in the WoSIS system ($> 1\,000\,000$ points). Both of our models predicted more conservative results than the GSOCmap-GSP, while at the same time, the GSOCmap-GSP predicted less SOC than the SoilGrids system and the Harmonized World Soil Database. Respectively, the SoilGrids system relies on a multivariate space suitable to represent the global soil-forming environment; however, a model would assume a similar relation of each covariate with the response across all land area in the world. The Harmonized World Soil Database may be a pedologically sound product, but large areas of Latin America have not been mapped at detailed scales (i.e., larger scales than 1:1 million) and this results in a polygon-based approach relying on wide generalizations.

Despite the aforementioned limitations, across Latin America, there is an increasing availability of relevant SOC

information across site- and country-specific regions (Reyes-Rojas et al., 2018; Vasques et al., 2016; Angelini et al., 2017; Samuel-Rosa et al., 2015; Angelini et al., 2016; Padarian et al., 2017), which could serve for validating and calibrating global SOC estimates. Thus, regional approaches considering multiple Latin American countries and SOC models could be a valuable resource to explain discrepancies between site- or country-specific and global SOC models.

Our results incorporate a multi-model perspective for quantifying/evaluating the spatial variability of SOC. The model with higher predictive capacity in terms of cross-validated r^2 was RF, an ensemble of regression trees based on bagging. However, this method yields high EC_r , and therefore it tends to capture the trend but with high bias. Taylor diagrams show that RF in any case yield the lower variance. SVM and RK were methods with higher agreement between RMSE and corr, and therefore lower EC_r . Large values of EC_r represent an accuracy limitation that was evident for RF, PL, and KK. To overcome these types of modeling biases, previous studies have suggested that the theory of ensemble learning applied to soil datasets could increase the accuracy of results (Finke, 2012; Nussbaum et al., 2018). Furthermore, recent studies highlight the applicability of selective ensembles across a large diversity of model algorithms useful for digital soil mapping purposes (Møller et al., 2018). Thus, our modeling approach included the combination of multiple predictions by using a linear stack of models as implemented in the caretEnsemble package of R (Deane-Mayer and Knowles, 2016), with the ultimate goal of reducing the uncertainty on SOC mapping efforts.

Table 4. SOC stocks at the contextual resolution of 5×5 km across land cover classes of Latin America for the 18 analyzed countries. The terms used are defined as follows: ens is country-specific, regional is Latin America ensemble, sg is the SoilGrids system, GSOCmap-GSP is country-specific 1 km, and hw is the Harmonized World Soil Database. These are the land cover classes described in Blanco et al. (2013). This land cover product was generated using 500 m grids and has 84 % of accuracy.

	Land cover	ens	GSOCmap-GSP	hw	sg	regional
1	Tropical broadleaf evergreen forest	30.39	40.30	59.15	80.44	29.73
2	Tropical broadleaf deciduous forest	0.43	0.65	1.00	1.09	0.42
3	Subtropical broadleaf evergreen forest	2.38	3.91	4.51	6.57	2.25
4	Subtropical broadleaf deciduous forest	1.42	2.04	1.87	2.55	1.07
5	Temperate broadleaf evergreen forest	3.32	1.26	4.97	6.91	3.56
6	Temperate broadleaf deciduous forest	0.48	0.52	1.02	1.21	0.63
7	Subtropical needleleaf forest	0.00	0.01	0.00	0.01	0.00
8	Temperate needleleaf forest	0.23	0.36	0.45	0.54	0.24
9	Mixed forest	0.67	1.08	1.34	1.66	0.66
10	Tropical shrubland	4.25	6.58	6.98	10.30	4.18
11	Subtropical shrubland	3.17	4.18	6.62	6.33	2.90
12	Temperate shrubland	4.56	5.08	7.33	9.97	5.32
13	Tropical grassland	3.01	2.48	3.56	5.46	2.45
14	Subtropical grassland	1.15	1.35	2.28	2.58	1.12
15	Temperate grassland	2.75	3.31	4.86	5.92	3.04
16	Inland water bodies	1.21	1.37	2.07	3.45	1.21
17	Urban area	0.24	0.31	0.45	0.55	0.22
18	Permanent ice and snow	0.14	0.08	0.14	0.38	0.17
19	Barren land	1.74	2.38	2.43	2.95	1.70
20	Cropland	12.95	19.33	21.89	27.94	12.42
21	Wetland	0.37	0.56	0.66	1.24	0.35
22	Salt flat	0.13	0.17	0.16	0.18	0.10
23	Coastal areas	1.59	1.39	2.23	4.31	1.78

Across Latin America, we did not find a common predictive algorithm for SOC. These results suggest that country-specific environmental predictors and available data influence the applicability of different approaches. This assessment is needed to address the requirements from the GSOCmap-GSP with the official mandate to generate and update country-specific soil information by the means of DSM. Thus, we argue that the DSM form of each country should assess and incorporate country-specific available data and environmental predictors to select the best prediction algorithm. The FAO SOC mapping cookbook explores possibilities to derive country-specific SOC maps from a variety of prediction algorithms (Yigini et al., 2018), and multiple resources have described the state of the art of modeling methods focused on DSM of soil carbon (Minasny et al., 2013; Malone et al., 2017) including geostatistics (Hengl, 2009, 2017). Thus, data characteristics (e.g., spatial structure, representativeness) are specifically important for developing a DSM framework as legacy soil profile collections, generated with long-term soil inventory purposes, will determine data availability and spatial distribution within a country.

This country-specific approach to map regional SOC results in artifacts across geopolitical borders. Therefore, data sharing, model validation, and calibration experiments among countries are required to better capture the spa-

tial variability of SOC. The use of a natural-defined prediction domain (e.g., ecoregional or physiographic map) could reduce the border effects. However, we understand that geopolitical borders are required for policy decisions around country-specific needs. We highlight that there is a lack of publicly available country-specific data that ultimately influence the performance of both country-specific to regional-to-global SOC estimates.

To achieve the highest possible accuracy of country-specific SOC estimates, the availability of point data sources for SOC modeling and mapping is an important consideration when selecting an efficient modeling strategy, especially when dealing with legacy SOC datasets. Our results highlight important uncertainty levels ($> 100\%$) across large areas of Latin America (Table 6). The data contained in WoSIS have a low-density distribution given the large area and environmental complexity of several countries analyzed. Thus, larger uncertainty dominates countries with larger SOC pools probably because available data do not capture the large spatial heterogeneity of SOC stocks. We highlight that the WoSIS dataset is a unique and invaluable effort that has proven to generate global SOC predictions (Hengl et al., 2017; Sanderman et al., 2017), but there is a global need to increase information and networking capabilities for SOC (Harden et al., 2017).

This study generated predictions of SOC across Latin America but also provided information about the main relationships driving the spatial distribution of SOC. Machine learning (i.e., data-driven) models have proven to be more efficient to model non-linear relationships of SOC (Hengl et al., 2015), but our results suggest that linear-based models (e.g., RK) could outperform machine learning methods under well-distributed and representative SOC data scenarios. Similar results were found across productive landscapes of Brazil (Bonfatti et al., 2016). We argue that our capacity to meet modeling assumptions will determine the most suitable prediction algorithm or ensemble methods (i.e., stack, blend, bucket of models). Machine learning models are usually conceived as black boxes and the influence of non-informative SOC prediction factors on machine learning-based SOC models has not been evaluated in detail. Therefore, we propose that the use of simple linear methods (i.e., correlation of available data and their predictors) can be a useful and parsimonious first step to inform data-driven approaches and enhance the interpretability of machine learning models to predict SOC. However, the simple selection of prediction factors based on simple correlation analysis does not prevent multi-collinearity, in which hypothesis-driven methods (e.g., RK) may be at risk to fail, but provides useful information about the main effects of the predictors on SOC. Thus, the use of machine learning and other statistical models (i.e., PL) is suitable to overcome the bias associated with the potential statistical redundancy of our simple variable selection approach based on simple correlation analysis. Furthermore, our data suggest that country-specific predictor factors are needed to better parameterize models but also could be useful for country-specific model interpretation. These results have important implications because it has been proposed that an extensive set of prediction factors is required to capture the large variance of the global SOC pool (Hengl et al., 2017). Thus, we propose that limited but informative country-specific prediction factors could be jointly explored to describe the local biophysical characteristics controlling SOC variability.

This study is expected to increase the capacity of Latin American institutions to provide accurate baseline estimates of SOC with a country-specific perspective following recommendations of GSOCmap-GSP. Ultimately, these efforts will enhance the development of new guidelines for measuring, mapping, reporting, verification, and monitoring SOC stocks (Vargas et al., 2013). Accurate country-specific DSM frameworks for SOC are required to facilitate interoperability and inform environmental policy across developing countries (Vargas et al., 2017). Our results highlight that attention is needed to better understand the influence of model prediction limits (e.g., the full conditional distribution) for the predicted SOC stocks. Setting an unreliable (excessive or low) prediction limit can have important effects (under- or overestimating) on the overall estimated stocks (Fig. 8). Therefore, we argue that data science systems for DSM focused on car-

bon assessments should be fundamentally based on SOC expert knowledge and informed by expert-based soil mapping systems.

5 Conclusions

We provided a multi-model comparison approach to map SOC stocks across Latin America and found that there is no dominant best prediction algorithm given the available data. The relative performance of the different methods varies from one place to another as well as the relative correlation of SOC with the prediction factors given the available data. We compared and combined hypothesis-driven approaches (e.g., linear geostatistics) and data-driven algorithms (e.g., machine learning), respectively, to generate interpretable and predictable models of SOC variability. We argue that models should not be conceived as competitors, because they have different assumptions (about the data themselves or about the empirical relationship between the response variable and its predictors) as different models will capture different portions of SOC variability. We highlight potential levels of uncertainty in SOC stocks associated with the maximum allowed prediction limit. Public data may not be representative across large areas, and we call for all countries to strengthen digital soil mapping capacity building initiatives, SOC research, and data sharing. The use of country-specific information and the use of different modeling approaches will enhance regional SOC mapping efforts and will provide insights to identify where and why different modeling approaches generate similar SOC estimates.

Code availability. The codes used for this work are available under the AGPL 3.0 license at <https://doi.org/10.5281/zenodo.1304392> (Guevara et al., 2018).

Working codes are also available at https://github.com/vargaslab/SoilCarbon_Latin_America (last access: 16 July 2018).

Data availability. The soil dataset can be downloaded from WoSIS at <http://www.isric.org/explore/wosis> (last access: 16 July 2018) and corresponds to the July 2016 version (Batjes et al., 2017). Soil covariates are available at <http://worldgrids.org> (last access: 20 February 2018). A list of the codes for the SOC prediction factors used here can be found at https://docs.google.com/spreadsheets/d/1yr09cPDoSVdoahN_fXcNLfgipQcCodRl66WCcj6hJ9A/edit?usp=sharing (last access: 16 July 2018).

Appendix A

A1 Brief description of implemented methods

RK is a hybrid model with both a deterministic and a stochastic component (Hengl et al., 2004). The regression part took the form of a stepwise (backward and forward) multiple linear regression to avoid statistical redundancy among the best prediction factors. The residual kriging was ordinary. The variogram parameters supporting the spatial interpolation were automatically fitted using the framework proposed by Hiemstra et al. (2008). RK was applied only to countries with 10 or more available observations.

PLS is a common method to deal with the presence of highly correlated predictors. The PLS algorithm integrates the compression and regression steps and it selects successive orthogonal factors that maximize the covariance between predictor and response variables (Wold, 1983; Viscarra Rossel et al., 2014). Most of its development and application are in the fields of chemometrics but it is used in several research areas to effectively solve regression and classification problems.

SVM applies a simple linear method to the data but in a high-dimensional feature space non-linearly related to the input space (Karatzoglou et al., 2006). It creates a hyperplane through n -dimensional spectral space. Then, SVM separates numerical data based on a kernel function and parameters (e.g., gamma and cost) that maximize the margin from the closest point to the hyperplane that divides data with the largest possible margin, being the support vectors the points which fall within (Heumann, 2011). Then, linear models are fitted to the support vectors. A radial general purpose kernel was found optimal after the cross-validation strategy for parameter selection.

RF is an ensemble of regression trees based on bagging (Breiman, 1996). This machine learning algorithm uses a different combination of prediction factors to train multiple regression trees. Each tree is generated using different subsets of available data (Breiman, 2001). The number of prediction factors to use on each tree is known as the m_{try} parameter. The final prediction is the weighted average of all individual trees.

KK is a pattern recognition technique which is based on the distances to training examples in the feature space (Silverman and Jones, 1989). The observations within the learning set, which are particularly close to the new observation (y, x) , should get a higher weight in the decision than such neighbors that are far away from (y, x) (Hechenbichler and Schliep, 2004). The parameter k determines the number of neighbors from which information will be considered for prediction, and a kernel function (e.g., triangular, Gaussian among others) converts distances into weights which will be used for regression problems. The Gaussian and (in less proportion) the triangular kernels were the optimal options for all countries.

The Supplement related to this article is available online at <https://doi.org/10.5194/soil-4-173-2018-supplement>.

Author contributions. All coauthors contributed to the planning of the study with support from the GSP secretariat to develop the GSOCmap. MG, GFO and RV designed the experiment. MG and GFO performed analyses. MG, ES and GFO prepared datasets. MG, GFO and RV wrote the manuscript with feedback from all coauthors.

Competing interests. The authors declare that they have no conflict of interest.

Special issue statement. This article is part of the special issue “Regional perspectives and challenges of soil organic carbon management and monitoring – a special issue from the Global Symposium on Soil Organic Carbon 2017”. It is a result of the Global Symposium on Soil Organic Carbon, Rome, Italy, 21–23 March 2017.

Acknowledgements. This work was supported by the Global Soil Partnership, the Central America, Caribbean and Mexico Soil Partnership, and the South America Soil Partnership in collaboration with the Department of Plant and Soil Sciences at the University of Delaware. Mario Guevara acknowledges support from a CONACYT fellowship. Guillermo Federico Olmedo is supported by the Argentinian government through the project INTA PNSUELO1134032. Rodrigo Vargas acknowledges support from NASA (80NSSC18K0173) and USDA (2014-67003-22070).

Edited by: Peter Finke

Reviewed by: Tomislav Hengl and one anonymous referee

References

- Adhikari, K., Hartemink, A. E., Minasny, B., Bou Kheir, R., Greve, M. B., and Greve, M. H.: Digital mapping of soil organic carbon contents and stocks in Denmark, *PLoS ONE*, 9, e105519, <https://doi.org/10.1371/journal.pone.0105519>, 2014.
- Angelini, M. E., Heuvelink, G. B., Kempen, B., and Morrás, H. J.: Mapping the soils of an Argentine Pampas region using structural equation modelling, *Geoderma*, 281, 102–118, <https://doi.org/10.1016/j.geoderma.2016.06.031>, 2016.
- Angelini, M. E., Heuvelink, G. B. M., and Kempen, B.: Multivariate mapping of soil with structural equation modelling, *Eur. J. Soil Sci.*, 68, 575–591, <https://doi.org/10.1111/ejss.12446>, 2017.
- Arrouays, D., Leenaars, J. G., de Forges, A. C. R., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J., Hengl, T., Heuvelink, G., Batjes, N., Carvalho, E., Hartemink, A., Hewitt, A., Hong, S.-Y., Krasilnikov, P., Lagacherie, P., Lelyk, G., Libohova, Z., Lilly, A., McBratney, A., McKenzie, N., Vasquez, G. M., Mulder, V. L., Minasny, B., Montanarella, L., Odeh, I., Padarian, J., Poggio, L., Roudier, P., Saby, N., Savin, I., Searle, R., Solbovoy, V., Thompson, J., Smith, S., Su-laeman, Y., Vintila, R., Rossel, R. V., Wilson, P., Zhang, G.-L., Swerts, M., Oorts, K., Karklins, A., Feng, L., Navarro, A. R. I., Levin, A., Laktionova, T., Dell’Acqua, M., Suvannang, N., Ruam, W., Prasad, J., Patil, N., Husnjak, S., Pásztor, L., Okx, J., Hallett, S., Keay, C., Farewell, T., Lilja, H., Juilleret, J., Marx, S., Takata, Y., Kazuyuki, Y., Mansuy, N., Panagos, P., Liedekerke, M. V., Skalsky, R., Sobocka, J., Kobza, J., Eftekhari, K., Alavipanah, S. K., Moussadek, R., Badraoui, M., Silva, M. D., Paterson, G., da Conceição Gonçalves, M., Theocharopoulos, S., Yemefack, M., Tedou, S., Vrscaj, B., Grob, U., Kozák, J., Boruvka, L., Dobos, E., Taboada, M., Moretti, L., and Rodriguez, D.: Soil legacy data rescue via GlobalSoilMap and other international and national initiatives, *GeoResJ*, 14, 1–19, <https://doi.org/10.1016/j.grj.2017.06.001>, 2017.
- Batjes, N. H., Ribeiro, E., van Oostrum, A., Leenaars, J., Hengl, T., and Mendes de Jesus, J.: WoSIS: providing standardised soil profile data for the world, *Earth Syst. Sci. Data*, 9, 1–14, <https://doi.org/10.5194/essd-9-1-2017>, 2017.
- Bishop, T., McBratney, A., and Laslett, G.: Modelling soil attribute depth functions with equal-area quadratic smoothing splines, *Geoderma*, 91, 27–45, [https://doi.org/10.1016/S0016-7061\(99\)00003-8](https://doi.org/10.1016/S0016-7061(99)00003-8), 1999.
- Blanco, P. D., Colditz, R. R., Saldaña, G. L., Hardtke, L. A., Llamas, R. M., Mari, N. A., Fischer, A., Caride, C., Aceñolaza, P. G., del Valle, H. F., Lillo-Saavedra, M., Coronato, F., Opazo, S. A., Morelli, F., Anaya, J. A., Sione, W. F., Zamboni, P., and Arroyo, V. B.: A land cover map of Latin America and the Caribbean in the framework of the SERENA project, *Remote Sens. Environ.*, 132, 13–31, <https://doi.org/10.1016/j.rse.2012.12.025>, 2013.
- Bonfatti, B. R., Hartemink, A. E., and Giasson, E.: Comparing Soil C Stocks from Soil Profile Data Using Four Different Methods, in: *Progress in Soil Science*, 315–329, Springer International Publishing, https://doi.org/10.1007/978-3-319-28295-4_20, 2016.
- Borra, S. and Di Ciaccio, A.: Measuring the Prediction Error. A Comparison of Cross-validation, Bootstrap and Covariance Penalty Methods, *Comput. Stat. Data An.*, 54, 2976–2989, <https://doi.org/10.1016/j.csda.2010.03.004>, 2010.
- Breiman, L.: Bagging Predictors, *Mach. Learn.*, 24, 123–140, <https://doi.org/10.1023/A:1018054314350>, 1996.
- Breiman, L.: Random forests, *Mach. Learn.*, 45, 5–32, <https://doi.org/10.1023/A:1010933404324>, 2001.
- Carslaw, D. C. and Ropkins, K.: openair – An R package for air quality data analysis, *Environ. Modell. Softw.*, 27–28, 52–61, <https://doi.org/10.1016/j.envsoft.2011.09.008>, 2012.
- Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J.: System for Automated Geoscientific Analyses (SAGA) v. 2.1.4, *Geosci. Model Dev.*, 8, 1991–2007, <https://doi.org/10.5194/gmd-8-1991-2015>, 2015.
- Deane-Mayer, Z. A. and Knowles, J. E.: caretEnsemble: Ensembles of Caret Models, available at: <https://CRAN.R-project.org/package=caretEnsemble> (last access: 16 July 2018), r package version 2.0.0, 2016.
- Delgado-Baquerizo, M., Eldridge, D. J., Maestre, F. T., Karunarathne, S. B., Trivedi, P., Reich, P. B., and Singh, B. K.: Climate legacies drive global soil carbon

- stocks in terrestrial ecosystems, *Sci. Adv.*, 3, e1602008, <https://doi.org/10.1126/sciadv.1602008>, 2017.
- 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.
- FAO: Fifth Meeting of the Global Soil Partnership Plenary Assembly, available at: <http://www.fao.org/3/a-bs973e.pdf> (last access: 16 July 2018), 2017.
- Finke, P. A.: On digital soil assessment with models and the Pedometrics agenda, *Geoderma*, 171–172, 3–15, <https://doi.org/10.1016/j.geoderma.2011.01.001>, 2012.
- Florinsky, I. V.: The Dokuchaev hypothesis as a basis for predictive digital soil mapping (on the 125th anniversary of its publication), *Eurasian Soil Sci.*, 45, 445–451, <https://doi.org/10.1134/S1064229312040047>, 2012.
- Grimm, R., Behrens, T., Märker, M., and Elsenbeer, H.: Soil organic carbon concentrations and stocks on Barro Colorado Island – Digital soil mapping using Random Forests analysis, *Geoderma*, 146, 102–113, <https://doi.org/10.1016/j.geoderma.2008.05.008>, 2008.
- Guevara, M., Olmedo, G. F., and Vargas, R.: DSM-LAC/NoSilverBulletOnDSM: No Silver Bullets – raw code, Zenodo, <https://doi.org/10.5281/zenodo.1304392>, 2018.
- Harden, J. W., Hugelius, G., Ahlström, A., Blankinship, J. C., Bond-Lamberty, B., Lawrence, C. R., Loisel, J., Malhotra, A., Jackson, R. B., Ogle, S., Phillips, C., Ryals, R., Todd-Brown, K., Vargas, R., Vergara, S. E., Cotrufo, M. F., Keiluweit, M., Heckman, K. A., Crow, S. E., Silver, W. L., DeLonge, M., and Nave, L. E.: Networking our science to characterize the state, vulnerabilities, and management opportunities of soil organic matter, *Glob. Change Biol.*, 24, e705–e718, <https://doi.org/10.1111/gcb.13896>, 2017.
- Hashimoto, S., Nanko, K., Ťupek, B., and Lehtonen, A.: Data-mining analysis of the global distribution of soil carbon in observational databases and Earth system models, *Geosci. Model Dev.*, 10, 1321–1337, <https://doi.org/10.5194/gmd-10-1321-2017>, 2017.
- Hechenbichler, K. and Schliep, K. P.: Weighted k -nearest-neighbor techniques and ordinal classification. Discussion paper 399, SFB 386, Ludwig-Maximilians University, Munich, available at: <http://www.stat.uni-muenchen.de/sfb386/papers/dsp/paper399.ps> (last access: 16 July 2018), 2004.
- Hengl, T.: A Practical Guide to Geostatistical Mapping, 2nd Edn., extended edition of the EUR 22904 EN Scientific and Technical Research series report published by 10 Office for Official Publications of the European Communities, Luxembourg, 293 pp., 2009.
- Hengl, T.: GSIF: Global Soil Information Facilities, available at: <https://CRAN.R-project.org/package=GSIF> (last access: 16 July 2018), r package version 0.5-4, 2017.
- Hengl, T., Heuvelink, G. B., and Stein, A.: A generic framework for spatial prediction of soil variables based on regression-kriging, *Geoderma*, 120, 75–93, <https://doi.org/10.1016/j.geoderma.2003.08.018>, 2004.
- Hengl, T., Heuvelink, G. B., Kempen, B., Leenaars, J. G., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., De Jesus, J. M., Tamene, L., and Tondoh, J. E.: Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions, *PLoS ONE*, 10, 1–26, <https://doi.org/10.1371/journal.pone.0125814>, 2015.
- Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, *PLoS ONE*, 12, e0169748, <https://doi.org/10.1371/journal.pone.0169748>, 2017.
- Hengl, T., Nussbaum, M., Wright, M. N., and Heuvelink, G. B.: Random Forest as a generic framework for predictive modeling of spatial and spatio-temporal variables, *PeerJ*, 6, e26693v1, <https://doi.org/10.7287/peerj.preprints.26693v1>, 2018.
- Heumann, B. W.: An object-based classification of mangroves using a hybrid decision tree-support vector machine approach, *Remote Sens.*, 3, 2440–2460, <https://doi.org/10.3390/rs3112440>, 2011.
- Heuvelink, G. B. M.: Uncertainty and Uncertainty Propagation in Soil Mapping and Modelling, Springer International Publishing, Cham, 439–461, https://doi.org/10.1007/978-3-319-63439-5_14, 2018.
- Hiemstra, P., Pebesma, E., Twenhöfel, C., and Heuvelink, G.: Real-time automatic interpolation of ambient gamma dose rates from the Dutch Radioactivity Monitoring Network, *Comput. Geosci.*, 35, 1711–1721, <https://doi.org/10.1016/j.cageo.2008.10.011>, 2008.
- Jackson, R. B., Lajtha, K., Crow, S. E., Hugelius, G., Kramer, M. G., and Piñeiro, G.: The Ecology of Soil Carbon: Pools, Vulnerabilities, and Biotic and Abiotic Controls, *Annu. Rev. Ecol. Evol. S.*, 48, 419–445, <https://doi.org/10.1146/annurev-ecolsys-112414-054234>, 2017.
- Karatzoglou, A., Meyer, D., and Hornik, K.: Support Vector Algorithm in R, *J. Stat. Softw.*, 15, 1–28, 2006.
- Köchy, M., Hiederer, R., and Freibauer, A.: Global distribution of soil organic carbon – Part I: Masses and frequency distributions of SOC stocks for the tropics, permafrost regions, wetlands, and the world, *SOIL*, 1, 351–365, <https://doi.org/10.5194/soil-1-351-2015>, 2015.
- Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., and Hunt, T.: caret: Classification and Regression Training, available at: <https://CRAN.R-project.org/package=caret> (last access: 16 July 2018), r package version 6.0-78, 2017.
- Kumar, S., Lal, R., and Liu, D.: A geographically weighted regression kriging approach for mapping soil organic carbon stock, *Geoderma*, 189–190, 627–634, <https://doi.org/10.1016/j.geoderma.2012.05.022>, 2012.
- Ließ, M., Schmidt, J., and Glaser, B.: Improving the spatial prediction of soil organic carbon stocks in a complex tropical mountain landscape by methodological specifications in machine learning approaches, *PLoS ONE*, 11, 1–22, <https://doi.org/10.1371/journal.pone.0153673>, 2016.
- Malone, B. P., Minasny, B., and McBratney, A. B.: Using R for Digital Soil Mapping, Springer International Publishing, <https://doi.org/10.1007/978-3-319-44327-0>, 2017.

- Marchetti, A., Piccini, C., Francaviglia, R., and Mabit, L.: Spatial Distribution of Soil Organic Matter Using Geostatistics: A Key Indicator to Assess Soil Degradation Status in Central Italy, *Pedosphere*, 22, 230–242, [https://doi.org/10.1016/S1002-0160\(12\)60010-1](https://doi.org/10.1016/S1002-0160(12)60010-1), 2012.
- Martin, M. P., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., and Arrouays, D.: Spatial distribution of soil organic carbon stocks in France, *Biogeosciences*, 8, 1053–1065, <https://doi.org/10.5194/bg-8-1053-2011>, 2011.
- McBratney, A., Santos, M. M., and Minasny, B.: On digital soil mapping, *Geoderma*, 117, 3–52, [https://doi.org/10.1016/S0016-7061\(03\)00223-4](https://doi.org/10.1016/S0016-7061(03)00223-4), 2003.
- Meinshausen, N.: Quantile Regression Forests, *J. Mach. Learn. Res.*, 7, 983–999, 2006.
- Meinshausen, N.: `quantregForest`: Quantile Regression Forests, available at: <https://CRAN.R-project.org/package=quantregForest> (last access: 16 July 2018), r package version 1.3-7, 2017.
- Minasny, B., McBratney, A. B., Malone, B. P., and Wheeler, I.: Digital Mapping of Soil Carbon, in: *Advances in Agronomy*, 1–47, Elsevier, <https://doi.org/10.1016/b978-0-12-405942-9.00001-3>, 2013.
- Mishra, U., Lal, R., Slater, B., Calhoun, F., Liu, D., and Van Meirvenne, M.: Predicting Soil Organic Carbon Stock Using Profile Depth Distribution Functions and Ordinary Kriging, *Soil Sci. Soc. Am. J.*, 73, 614–621, <https://doi.org/10.2136/sssaj2007.0410>, 2009.
- Møller, A. B., Beucher, A., Iversen, B. V., and Greve, M. H.: Predicting artificially drained areas by means of a selective model ensemble, *Geoderma*, 320, 30–42, <https://doi.org/10.1016/j.geoderma.2018.01.018>, 2018.
- Mondal, A., Khare, D., Kundu, S., Mondal, S., Mukherjee, S., and Mukhopadhyay, A.: Spatial soil organic carbon (SOC) prediction by regression kriging using remote sensing data, *Egyptian Journal of Remote Sensing and Space Science*, 20, 61–70, <https://doi.org/10.1016/j.ejrs.2016.06.004>, 2017.
- Nelson, D. W. and Sommers, L. E.: Total carbon, organic carbon and organic matter, in: *Methods of soil analysis. Part 2 Chemical and Microbiological Properties*, edited by: Page, A. L., Miller, R. H., and Keeney, D. R., 539–579, 1982.
- Nussbaum, M., Papritz, A., Baltensweiler, A., and Walthert, L.: Estimating soil organic carbon stocks of Swiss forest soils by robust external-drift kriging, *Geosci. Model Dev.*, 7, 1197–1210, <https://doi.org/10.5194/gmd-7-1197-2014>, 2014.
- Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaeppman, M. E., and Papritz, A.: Evaluation of digital soil mapping approaches with large sets of environmental covariates, *SOIL*, 4, 1–22, <https://doi.org/10.5194/soil-4-1-2018>, 2018.
- Padarian, J., Minasny, B., and McBratney, A.: Chile and the Chilean soil grid: A contribution to GlobalSoilMap, *Geoderma Regional*, 9, 17–28, <https://doi.org/10.1016/j.geodrs.2016.12.001>, 2017.
- Peng, G., Bing, W., Guangpo, G., and Guangan, Z.: Spatial distribution of soil organic carbon and total nitrogen based on GIS and geostatistics in a small watershed in a hilly area of northern China, *PLoS ONE*, 8, 1–9, <https://doi.org/10.1371/journal.pone.0083592>, 2013.
- R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, available at: <https://www.R-project.org/> (last access: 16 July 2018), 2017.
- Reyes-Rojas, L. A., Adhikari, K., and Ventura, S. J.: Projecting Soil Organic Carbon Distribution in Central Chile under Future Climate Scenarios, *J. Environ. Qual.*, 47, 735–745, <https://doi.org/10.2134/jeq2017.08.0329>, 2018.
- Rossel, R. V. and Behrens, T.: Using data mining to model and interpret soil diffuse reflectance spectra, *Geoderma*, 158, 46–54, <https://doi.org/10.1016/j.geoderma.2009.12.025>, 2010.
- Samuel-Rosa, A., Heuvelink, G., Vasques, G., and Anjos, L.: Do more detailed environmental covariates deliver more accurate soil maps?, *Geoderma*, 243–244, 214–227, <https://doi.org/10.1016/j.geoderma.2014.12.017>, 2015.
- Sanderman, J., Hengl, T., and Fiske, G. J.: Soil carbon debt of 12,000 years of human land use, *P. Natl. Acad. Sci. USA*, 114, 9575–9580, <https://doi.org/10.1073/pnas.1706103114>, 2017.
- Shangguan, W., Hengl, T., Mendes de Jesus, J., Yuan, H., and Dai, Y.: Mapping the global depth to bedrock for land surface modeling, *J. Adv. Model. Earth Sy.*, 9, 65–88, <https://doi.org/10.1002/2016MS000686>, 2017.
- Silverman, B. W. and Jones, M. C.: An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951), *Int. Stat. Rev.*, 57, 233–238, 1989.
- Sreenivas, K., Dadhwal, V. K., Kumar, S., Harsha, G. S., Mitrani, T., Sujatha, G., Suresh, G. J. R., Fyze, M. A., and Ravisankar, T.: Digital mapping of soil organic and inorganic carbon status in India, *Geoderma*, 269, 160–173, <https://doi.org/10.1016/j.geoderma.2016.02.002>, 2016.
- Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, *J. Geophys. Res.-Atmos.*, 106, 7183–7192, <https://doi.org/10.1029/2000JD900719>, 2001.
- Tifafi, M., Guenet, B., and 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, *Global Biogeochem. Cy.*, 32, 42–56, <https://doi.org/10.1002/2017GB005678>, 2018.
- Vargas, R., Paz, F., and de Jong, B.: Quantification of forest degradation and belowground carbon dynamics: ongoing challenges for monitoring, reporting and verification activities for REDD+, *Carbon Manag.*, 4, 579–582, <https://doi.org/10.4155/cmt.13.63>, 2013.
- Vargas, R., Alcaraz-Segura, D., Birdsey, R., Brunsell, N. A., Cruz-Gaistardo, C. O., de Jong, B., Etchevers, J., Guevara, M., Hayes, D. J., Johnson, K., Loescher, H. W., Paz, F., Ryu, Y., Sanchez-Mejia, Z., and Toledo-Gutierrez, K. P.: Enhancing interoperability to facilitate implementation of REDD+: case study of Mexico, *Carbon Manag.*, 8, 57–65, <https://doi.org/10.1080/17583004.2017.1285177>, 2017.
- Vasques, G. M., Coelho, M. A.-C. R., Dart, R. O., Oliveira, R. P., and Teixeira, W. G.: Mapping soil carbon, particle-size fractions, and water retention in tropical dry forest in Brazil, *Pesqui. Agropecu. Bras.*, 51, 1371–1385, 2016.
- Viscarra Rossel, R. A., Webster, R., Bui, E. N., and Baldock, J. A.: Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change, *Glob. Change Biol.*, 20, 2953–2970, <https://doi.org/10.1111/gcb.12569>, 2014.

- Wold, H.: Systems Analysis by Partial Least Squares, Iiasa collaborative paper, IIASA, Laxenburg, Austria, available at: <http://pure.iiasa.ac.at/2336/> (last access: 16 July 2018), 1983.
- Yang, R.-M., Zhang, G.-L., Yang, F., Zhi, J.-J., Yang, F., Liu, F., Zhao, Y.-G., and Li, D.-C.: Precise estimation of soil organic carbon stocks in the northeast Tibetan Plateau, *Sci. Rep.-UK*, 6, 21842, <https://doi.org/10.1038/srep21842>, 2016.
- Yigini, Y. and Panagos, P.: Assessment of soil organic carbon stocks under future climate and land cover changes in Europe, *Sci. Total Environ.*, 557–558, 838–850, <https://doi.org/10.1016/j.scitotenv.2016.03.085>, 2016.
- Yigini, Y., Olmedo, G. F., Reiter, S., Baritz, R., Viatkin, K., and Vargas, R. R. (Eds.): *Soil Organic Carbon Mapping Cookbook*, FAO, Rome, Italy, 2nd Edn., available at: <http://www.fao.org/documents/card/en/c/I8895EN> (last access: 16 July 2018), 2018.