Cross-regional Digital Soil Carbon Modeling in Two Contrasting Soil-Ecological Regions in the U.S.

B. Cao, S. Grunwald & X. Xiong
University of Florida, Gainesville, United States of America

ABSTRACT: The implications of transposing regional digital soil models to continental and global scales are still poorly understood. This paper presents a 'controlled landscape experiment' where soil organic carbon (SOC) stocks were predicted using a standardized set of nationally available environmental predictor variables and sampling density of soil observations. Our specific objective was to compare the prediction performance of SOC in two contrasting regions in the United States. We used soil samples in the topsoil (0-20 cm) depth across Colorado and Florida from the U.S. National Soil Survey Database (Natural Resource and Conservation Service, NRCS). Environmental covariate sets were assembled representing a subset of STEP-AWBH factors (S: soils, T: topography, E: ecology, P: parent material, A: atmosphere, W: water, B: biota, and H: human). We used single regression trees (RT) and support vector machines (SVMs) and various error metrics to assess the prediction performance of SOC stocks. Our results demonstrate that in ecologically contrasting states, both RT and SVM could produce moderate good predictions with R^2 of 0.76 (SVMs) in Florida and R^2 of 0.62 (RT) in Colorado. The differences in model results elucidate on the contrasting relationships between SOC and environmental predictors that were climate, soil and vegetation driven in Colorado and soil and vegetation driven in Florida. These findings have implications for upscaling of regional digital soil models to continental and global scales, specifically for SOC modeling across the U.S.

1 INTRODUCTION

Soil organic carbon (SOC) sequestration has received much attention recently as the concentration of CO_2 rises in the atmosphere, intensifying climate change. Considering global issues of significance, such as food security, land degradation, and global climate change, more emphasis is needed on maintaining soil's natural condition under the impact of disturbances and human activities (Grunwald et al., 2011). Characterizing organic carbon (C) pools across large regions is critical to understanding the dynamics of soil C in the context of climate change.

Research on soil C has mainly focused on regional and local scales, while continental scale studies at a spatial resolution that matches the underlying soil C variability is lacking. Comparison of regional digital soil models is hampered by the fact that studies differ in terms of soil C measurement techniques, sampling densities, sample protocols, environmental co-variates (predictor variables), and statistical and geostatistical methods used to predict soil properties. But all these factors may play a role when continental and global digital soil models are created (Grunwald et al., 2011). To assess differences in the strength and magnitude in soil prediction models developed in regions that show contrasting topography, ecology, parent material, soils, land use and/or climate, soil and environmental data inputs and model setup need to be standardized. Understanding the effects of regional soil C predictions may then provide guidance to upscale to larger scales (e.g., whole U.S. or even larger extent).

Our specific objective was to compare the prediction performance of soil C models in two contrasting regions in the United States. These two regions differed in terms of topography, parent material, soils, geology, climate, and land use.

2 STUDY AREAS

Two states in the U.S. were selected: Colorado and Florida, which are extremely different in terms of ecoregion, climate zone, topography and other ecological conditions.

2.1 *Colorado*

Colorado is a U.S. state that encompasses most of the Southern Rocky Mountains as well as the northeastern portion of the Colorado Plateau and the western edge of the Great Plains (Fig. 1). Colorado is the 8th most extensive (size: 269,601 km^2). Colorado is noted for its vivid landscape of mountains, forests, high plains, mesas, canyons, plateaus, rivers, and desert lands. The climate of Colorado is quite complex compared to most of the United States due to the complex orography (1,011~4401.2 m). Northeast, east, and southeast Colorado are mostly the high plains, while Northern Colorado is a mix of high plains, foothills, and mountains. Northwest and west Colorado are predominantly mountainous, with some desert lands mixed in. Southwest and southern Colorado are a complex mixture of desert and mountain areas.

2.2 *Florida*

Florida is a state in the southeastern U.S., located on the southern coastal plain. It is bordered to the west

by the Gulf of Mexico, to the north by Alabama and Georgia and to the east by the Atlantic Ocean. Florida is the 22nd most extensive (size: 170,304 km^2). Its geography is marked by a coastline, by the omnipresence of water and the threat of hurricanes. The topography consists of gentle slopes varying from 0 to 5% in almost the whole state, with moderate slopes of 5 to 19% occurring in <1% of the state. Land use and land cover consist mainly of wetlands (28%), pinelands (18%), and urban and barren lands (15%). Agriculture, rangelands, and improved pasture occupy 9, 9, and 8% of Florida, respectively (Florida Fish and Wildlife Conservation Commission, 2003). North of Lake Okeechobee, the prevalent climate is humid subtropical, while coastal areas south of the lake (including the Florida Keys) have a tropical climate.

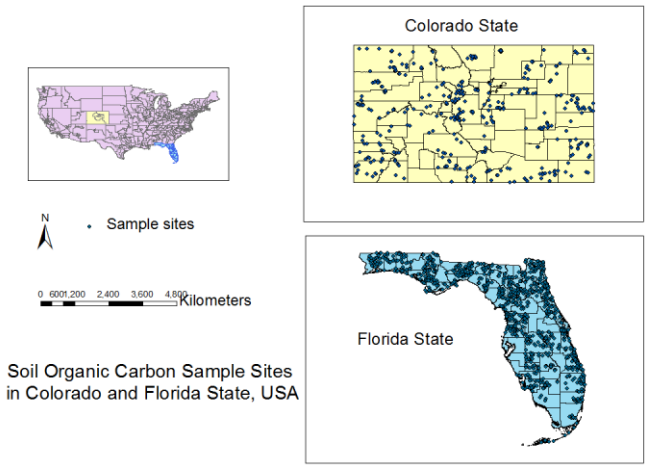

Figure 1. The study areas – Colorado and Florida States – and overview map that shows their geographic location within the U.S..

3 DATA AND METHODS

3.1 *Data*

Soil sample data in Colorado and Florida were derived from the U.S. National Soil Survey Database (Natural Resources Conservation Service, NRCS), This national set includes data from the Florida Soil Characterization Database (Grunwald et al., 2007). We retrieved SOC data (measured in Walkley Black) from the database and standardized them to 0-20 cm depth.

Each of the environmental variable data were from the same sources for both Colorado and Florida. Table 1 below lists detailed information of the retrieved environmental data that represent STEP-AWBH variables (Grunwald et al., 2011) expected to relate to SOC. These variables are available as grids or polygons for the whole U.S. and allow soil prediction modeling not only in Florida and Colorado, but potentially in other states in the U.S.

Since the area size of Colorado is almost twice as that of Florida, we standardized the sample observations retrieved from the database which contained geo-referenced 437 pedons for Colorado and 1,233 pedons for Florida. Thus, we randomly selected 227 samples as subset out of 1,233 samples collected in Florida to equal the sample density of Colorado (0.0016 pedons/km^2). Figure 1 shows the geophysical location of the study areas and sample sites in each one.

3.2 *Methods*

3.2.1 *Pedotransfer functions*

The term pedotransfer function (PTF) was coined by Bouma (1989) as translating data we have into what we need. These functions fill the gap between the available soil data and the properties which are more useful or required for a particular model or quality assessment. And it is common to have missing values in modeling for large regions. According to Heuscher et al. (2005) regression equations are a feasible alternative for bulk-density estimations. A PTF was developed to estimate bulk density (BD) values to derive SOC stocks if measured BD were not available from the national soil database. Five predictors of 565 samples with bulk density data in Colorado were used to develop the PTF: predicted BD (g cm^{-3}) = 1.7 - 0.22SOC$^{1/2}$ (square root of SOC, %) -0.012water content (%)+ 0.011total clay content (%) -0.0027total silt content (%) - 0.00077average horizon depth (cm), and R^2=0.58, Root mean square error, RMSE=0.14 g cm^{-3}.

3.2.2 *Regression trees*

We used single Regression Trees (RT) to identify the relationships between SOC stocks and environmental variables. Regression trees (Breiman et al., 1984) are a nonparametric data mining technique that has been incorporated in soil science (Brown et al., 2006). They have no assumptions about the distribution of the data and can identify nonlinear relationships in the data, offering an alternative to linear methods to analyze soil properties.

3.2.3 *Support vector machines*

Another method employed to relate SOC stocks to environmental predictor variables was support vector machines (SVMs). SVMs apply a simple linear method to the data (i.e., constructing a hyperplane) in a high-dimensional feature space non-linearly related to the input space (Karatzoglou et al., 2006). The best generalization performance is obtained when the hyperplane is as far as possible from all the data points. This generalization is with more complex surfaces by extending the measurement space, so that it includes transformations of the raw data.

Table 1. Datasets of environmental variables used for study areas.

Environmental Predictors	Variables	Data source[1]	Original Resolution	Year
Soil	Soil order	SSURGO 2.2, NRCS	1:24,000	N/A
	Soil suborder			
	Soil group			
	Soil subgroup			
	Annual minimum water table depth			
	Annual minimum water table depth from April to June			
	Plant available water holding capacity top 25 cm			
	Plant available water holding capacity top 50 cm			
	Plant available water holding capacity top 100 cm			
	Plant available water holding capacity top 150 cm			
Climate	Annual Precipitation	PRISM Climate Group	800m	1971-2000
	Average Max temperature			
	Average Min temperature			
Organism	Land use	NLCD	30m	2001
	Normalized difference vegetation index (NDVI) Annual average	MODIS	250m	2005
	NDVI Annual max			
	NDVI Annual min			
	Existing vegetation type	LANDFIRE	30m	2002
	Biophysical settings			
	Existing vegetation height			
	Existing vegetation cover			
Relief	Elevation	NED, USGS	1 arc sec	N/A
	Slope			
	Flow accumulation			
Parent material	Rock type*	Mineral Resources Program, USGS	1:500,000	2005

[1] SSURGO: Soil Survey Geographic Database; NRCS: Natural Resources Conservation Service; NLCD: National land cover database; MODIS: Moderate Resolution Imaging Spectroradiometer; LANDFIRE: Landscape Fire and Resource Management Planning Tools Project; NED: National elevation dataset; USGS: U.S. Geological Survey. * The predominant lithology found in the formation.

Then it is possible to develop a linear decision surface that perfectly separates the data in this enhanced space. SVMs had been successfully used to model and interpret soil properties such as soil moisture and soil diffuse reflectance spectra (Lamorski et al., 2008).

3.2.4 *Model assessment*

We tested the performance of different modeling techniques – RT and SVM – to identify relationships between SOC and environmental variables using a repeated 9-fold cross-validation in Colorado and Florida. The training data were separated into nine subgroups, where one subgroup is used as the validation data for testing, and the remaining eight subgroups were used as training data. This cross-validation process was repeated eight times so that each of the nine subgroups has been used once as the validation data. The coefficient of determination (R^2) was used to compare the model fit, and error statistics, including the RMSE and the residual prediction deviation (RPD).

4 RESULTS AND DISCUSSION

The descriptive statistics of SOC observed in the two States are shown in Table 2. Both the whole set and subset soil samples in Florida had a larger mean and median value than that of Colorado. And the maximum SOC stock in Florida was much higher than that of Colorado.

Table 2. Descriptive statistics of measured soil organic carbon stock (kg m^{-2}) in topsoil (0-20 cm depth) in two study areas.

Statistic	Colorado	Florida	
	(n: 437)	Whole set (n: 1,233)	Subset (n: 227)
Mean	3.1	4.7	5.0
Median	2.3	2.7	2.6
SD*	3.0	6.7	7.9
Minimum	0.4	0.4	0.5
Maximum	26.8	68.1	68.1

*SD: standard deviation.

Among the four predicted models, SVM for Florida showed the highest relationships, while SVM for

Colorado the lowest. For both RT and SVM methods, the RMSE was higher for Florida than Colorado. This might be explained by the smaller sample size in Florida or by the much larger maximum SOC stock observations (68 kg C m^{-2}) that occurred in Florida when compared to Colorado. The RPD for all four models was <2.0 suggesting moderate performance of SOC predictions. Although the R^2 and RPD values were higher in Florida than in Colorado suggesting better performing models, the RMSEs were substantially larger in Florida.

Table 3. Summary of model performances to estimate soil organic carbon stocks (kg C m^{-2}) in the topsoil using regression trees (RT) and support vector machines (SVMs) in Colorado and Florida.

Model	Colorado			Florida		
	R^2	RMSE	RPD	R^2	RMSE	RPD
RT	0.62	1.83	1.62	0.67	4.49	1.75
SVMs	0.54	2.22	1.34	0.76	4.10	1.92

Plots of predicted and observed SOC stocks, derived using 9-fold cross-validation are presented in Figure 2. The plot shows trends of underestimation of high values for both Florida and Colorado using SVM. The estimated SOC values smaller than 10 kg C m^{-2} using RT showed reasonable approximation to the 1:1 line of equal values; however, substantial spread between predicted and observed SOC stocks occurred for values larger than about 10 kg C m^{-2}.

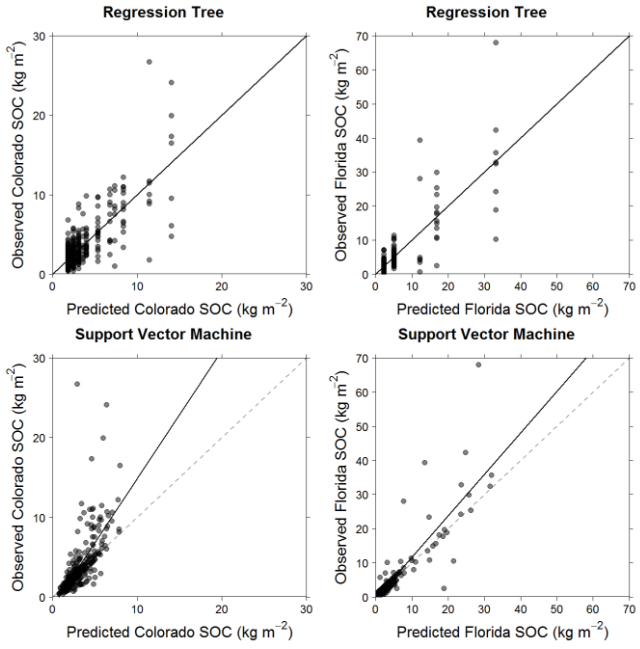

Figure 2. Predicted vs. observed soil organic carbon stock values in the topsoil derived from regression trees (RT) and support vector machines (SVMs) for Colorado and Florida.

5 CONCLUSIONS

Our modeling study indicated that both, RT and SVMs, had moderate capability to predict SOC stocks. Interestingly, although the soil forming factors, as represented by the STEP-AWBH environmental variables, differed profoundly in the two regions, successful prediction models could be build. In hindsight of developing SOC prediction models for the whole U.S. these are encouraging findings.

This controlled landscape experiment, where sampling density of SOC and environmental predictors were standardized, allowed direct comparison of model performances. The differences in model results elucidate on the contrasting relationships between SOC and environmental predictors that were climate, soil and vegetation driven in Colorado and soil and vegetation driven in Florida.

6 ACKNOWLEDGEMENTS

We would like to thank NRCS for soil database support specifically Henry Ferguson and C. Wade Ross, An-Min Wu, and Brandon Hoover for the help on data processing. Financial support was provided by the School of Natural Resources and Environment (SNRE), University of Florida and Chinese Scholarship Council Program.

7 REFERENCES

Bouma, J., 1989. Using soil survey data for quantitative land evaluation, 1989. Adv. Soil Sci. 9: 177-213.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. Classifi cation and regression trees. Wadsworth Stat./Probab. Ser. Wadsworth Int., Belmont, CA.

Brown, D.J., K.D. Shepherd, M.G. Walsh, M. Dewayne Mays, and T.G. Reinsch. 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132:273–290.

Florida Fish and Wildlife Conservation Commission. 2003. Florida vegetation and land cover data derived from Landsat ETM+ imagery. Raster layer. Spatial resolution: 30 m. FFWCC, Tallahassee, FL.

Grunwald, S. and Harris, W.G. 2007. Florida Soil Characterization Data (http://soils.ifas.ufl.edu/flsoils/).

Grunwald, S., J.A. Thompson and J.L. Boettinger. 2011. Digital soil mapping and modeling at continental scales – finding solutions for global issues. Soil Sci. Soc. Am. J. (SSSA 75th Anniversary Special Paper) 75: 1201-1213.

Heuscher, S.A., Brandt, C.C., Jardine, P.M. 2005. Using soil physical and chemical properties to estimate bulk density. Soil Sci. Soc. Am. J. 69: 51-56.

Holmes, K.W., Kyriakidis, P.C., Chadwick, O.A., Soares, J.V., Roberts D.A. 2004. Muti-scale variability in tropical soil nutrients following land-cover change. Biogeochemistry 74:173-203.

Karatzoglou A., D. Meyer, and K. Hornik. 2006. Support vector machines in R. J. of Stat. Software15(9): 1-28.

Lamorski, K., Y. Pachepsky, C. Slawihski, and R.T. Walczak. 2008. Using support vector machines to develop pedotransfer functions for water retention of soils in Poland. Soil Sci. Soc. Am. J. 72:1243–1247.