COUNTRY TO GLOBAL PREDICTION OF SOIL ORGANIC CARBON AND SOIL MOISTURE USING DIGITAL SOIL MAPPING

by

Mario Guevara

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Plant and Soil Sciences

Winter 2020

COUNTRY TO GLOBAL PREDICTION OF SOIL ORGANIC CARBON AND SOIL MOISTURE USING DIGITAL SOIL MAPPING

by

Mario Guevara

Approved: _____
Erik H. Ervin, Ph.D.
Chair of the Department of Plant and Soil Sciences

Approved: _____
Mark Rieger, Ph.D.
Dean of the College of Agriculture and Natural Resources

Approved: _____
Douglas J. Doren, Ph.D.
Interim Vice Provost for Graduate and Professional Education and
Dean of the Graduate College

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Rodrigo Vargas, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Bruce Vasilas, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Michela Taufer, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Randy Wisser, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

I want to thank to my advisor Dr. Rodrigo Vargas for his guidance and support towards the development of this research. Thanks to Dr. José Sarukhán and to Biol. Jorge Larson for their guidance and support. Thanks to all my colleagues, family and friends. Thanks to my father, to my beautiful wife and daughter. This research is dedicated to the memory of Ana Santamaria Galvan.

This research was supported by the Mexican National Council for Science and Technology (Ph.D. fellowship 382790). This research was partially funded by the National Science Foundation (CIF21 DIBBs: PD: Cyberinfrastruture Tools for Precision Agriculture in the 21st Century, 1724847) and the National Aeronautics and Space Administration – Carbon Monitoring Systems Initiative (80NSSC18K0173).

"I am impressed both with the quantity and quality of this dissertation. It is an excellent example of how pedometrics can help assess key soil properties and functions for a sustainable world." Gerard B.M. Heuvelink, Special professor of Pedometrics and Digital Soil Mapping at Wageningen University & Research and ISRIC - World Soil Information.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The largest carbon pool in terrestrial ecosystems is contained in soils and it plays a key role regulating hydrological processes, such as the spatial variability of soil moisture dynamics. Specifically, soil moisture and soil organic carbon are variables directly linked to ecosystem services such as food production and water storage. However, there are important knowledge gaps in the spatial representation (e.g., maps) of soil moisture and soil organic information from the country specific to the global scales. There is a pressing need to update the spatial detail of soil moisture estimates and the accuracy of digital soil carbon maps for improved land management, improved Earth system modeling and improved strategies (i.e., public policy) to combat land degradation. From the country specific to the global scale, the overreaching goal of this PhD research is to develop a reproducible digital soil mapping framework to increase the statistical accuracy of spatially continuous information on soil moisture and soil organic carbon across different scales of data availability (e.g., country-specific, regional, global). Chapter 1 provides a general introduction. Chapters 2 and 3 are focused on up-scaling soil organic carbon from the country-specific scale to the continental scale. Chapter 2 provides a country-specific and multi-modeling approach for soil organic carbon mapping across Latin America, where I identify key predictors and conclude that there is no best modeling method in

a quantifiable basis across all the analyzed countries. In Chapter 3, I compare and test different methods and combinations of prediction factors to model the variability of soil organic carbon across Mexico and conterminous United States (CONUS). I describe soil organic carbon stocks across different land covers across the region, quantify the model uncertainty and discuss estimates derived from previous studies. Chapters 4 and 5 are devoted to improving the statistical detail and accuracy of satellite soil moisture from the country to the global scale. Chapter 4 describes how the machine learning fusion of satellite soil moisture with Geomorphometry increase the statistical accuracy and spatial detail of current soil moisture estimates across CONUS. Chapter 5 extends the previous chapter to the global scale and identifies global soil moisture trends. I provide a novel (gap-free) soil moisture global estimate that could be potentially used to predict the global feedback between primary productivity and long-term soil moisture trends. Chapters 4 and 5 reveal evidence of soil moisture decline across large areas of the world. Finally, chapter 6 summarizes the main findings of this research, the key conclusions, emergent challenges and future steps. The results of this research were useful to generate benchmarks against which to assess the impact of climate and land cover changes on soil organic carbon stocks and soil moisture trends. This research provides a framework (including high quality data and novel methodologies) to generate environmentally relevant science that can be used for the formulation of public policy around soil and water conservation efforts.

Chapter 1

INTRODUCTION

There is an increasing demand of updated soil data and soil information (e.g., soil property maps) to improve land management, to improve the spatial representation of Earth system models and to develop policy relevant research to reverse land degradation. Current global initiatives such as the Food and Agriculture Organization-Global Soil Partnership (Yigini. et al., 2018), the GlobalSoilMap.net consortium (Arrouays et al., 2018), the SoilGrids250m project (Hengl et al., 2017) or the World Soil Information Service (Batjes et al., 2019) are some examples evidencing the need of new and better soil information across the world. Continuous and updated soil data and soil information are required to quantify the response of soils to global environmental change and to generate land monitoring baselines for identifying regenerative (reversing land degradation) soil management practices. Soil data and information is key to develop food production strategies directly related to multiple aspects of human security (Amundson et al., 2015). However, soil information is challenging to obtain across large geographical areas (i.e., continents) because collecting soil samples, describing soil profiles, measuring and analyzing soil physical and chemical properties in laboratory is expensive and time consuming.

With a national perspective, soil samples are collected and analyzed along decades by multiple agencies and institutions with the official mandate to generate soil information on each country for soil (or natural resources) inventory purposes. Some examples of these institutions are the National institute of Statistics and Geography in Mexico, the Soil Survey Division of the United States Department of Agriculture, the Commonwealth Scientific and Industrial Research Organization in Australia, the National Institute of Agricultural Research in France, among many others. Soil data and soil information collected by these agencies are the basis for soil classification and mapping applications across national to global scales.

Soil classification and mapping are components of soil science research and 'digital soil mapping has evolved from traditional soil classification and mapping to the creation and population of spatial soil information systems by using field and laboratory observations coupled with environmental covariates' representing soil forming factors (Ma et al., 2019). Digital soil property (or digital soil classes) maps are increasingly needed for multiple soil information users (such as land managers, students, other scientists and policy makers) across multiple spatial scales (from the plot to the global scale). Up to date soil property maps are needed for characterizing soil spatial variability (i.e., physical, chemical and biological soil properties) across large geographical areas where no soil information has been collected in the past. However, the integration of multiple datasets to generate regional to global soil property maps is affected by many inconsistencies in methodological approaches used for soil classification and mapping from one place to another (Stell et al., 2019), or differences

in multiple data collection periods of time. Those inconsistencies are probably the reasons why is easy to find discrepancies on soil measurements, soil estimates or soil modeled data from multiple soil classification and mapping efforts. For example, soil carbon stocks and soil moisture are important soil indicators of land productivity and therefore spatial information (e.g., maps) on these indicators is required for monitoring land degradation and drought (Lorenz et al., 2019; van der Molen et al., 2011). One major issue is that there are large discrepancies in soil moisture and soil organic carbon spatial information across the world have been reported in recent work (Tifafi et al., 2018, Guevara et al., 2018, Gu et al., 2019). The discrepancies on these (soil moisture and carbon) estimates (e.g. different stocks or contrasting trends) increases the uncertainty forecasting the fate of land carbon uptake or detecting climate impact signals of soil production systems (Folberth et al., 2016; Walsh et al., 2017). Although national to global collaborative and networking efforts have increased the availability and access to soil data and soil information during the last decade (Batjes et al., 2019; Dorigo et al., 2011, 2017; Harden et al., 2018; Hengl et al., 2018; Malhotra et al., 2019; Pfeiffer et al., 2019; Samuel-Rosa et al., 2020; Bond-Lamberty and Thomson, 2010; Batjes et al., 2017; Stoorvogel et al., 2016), there are still large areas of the world with low available soil information for accurately representing the spatial variability of soil carbon and soil moisture patterns. To compare and test multiple modeling and prediction approaches (e.g., probabilistic or algorithmic) could be useful to reveal soil variability patterns across poorly represented areas with available soil spatial information.

To improve the spatial representation of soil functional attributes such as the spatial variability of soil organic carbon and soil moisture content is a pressing need that could be addressed using digital soil mapping. On digital soil mapping (McBratney et al., 2003) or pedometric mapping (Hengl, 2003), the targets are soil functional attributes or classes (y) for a specific soil depth. The available information on target soil functional attributes can be used to train models to solve regression (to predict continuous variables) or classification (to predict soil classes) problems. Model training data are the values of soil attributes or classes represented by soil profile or soil samples descriptions and/or results from laboratory analysis applied to soil samples to quantify soil physical and chemical properties and/or field direct soil observational data (e.g., from proximal soil sensors or electrochemical methods). The explanatory variables on these models are emergent sources of spatial information directly related to the soil forming environment (x). These explanatory variables are the basis to generate predictions ($f(x)=y$) and they are represented by multiple information sources including remotely sensed data (i.e., satellite based), digital terrain analysis, climate information and legacy or thematic maps (e.g., rock type, soil type or land use maps). The model (f) coefficients are then applied across the area of interest for deriving digital soil maps. Thus, the observational available soil data (quantitative or categorical) is the basis of a (statistical) learning process to find optimal model parameters (that allows to meet modeling assumptions). Multiple modeling approaches (i.e., from Geo-statistics to machine learning or combined approaches), including data driven and hypothesis driven prediction models or prediction algorithms (Breiman, 2001) are required for capturing

accurately linear and non-linear relationships between soils and the soil forming factors (environmental covariates). Multiple validation strategies (e.g., cross-validation, independent validation) are then required to assess the accuracy of digital soil maps. Accurate soil property maps lead the discovery of emergent patterns of soil variability from the plot to the continental scale. Information about soil variability is key to improve Earth system models (i.e., predicting the fate of land carbon uptake), soil-related policy making (i.e., for developing soil protection strategies), and land management practices.

Digital soil mapping is increasingly being adopted by several agencies or academic institutions (across multiple countries) with the mandate to generate soil information. Digital soil mapping is a conceptual framework to update soil information useful to support sustainable soil management strategies (e.g., United Nations' Sustainable Development Goals). In the United States, digital soil mapping is now a chapter of the United States Department of Agriculture's soil survey manual (Chapter 5, p. 295, Soil Science Division Staff, 2017), which is a global reference for soil taxonomy around the world. Many countries in the world have adopted soil taxonomy as the reference framework for soil inventory purposes. Digital soil mapping research can also be used for increasing the access to soil information across scales and multiple scenarios of data availability (from the country to the global scale). Consequently, there are emergent challenges such as the constant need of higher accuracy and higher resolution (spatial and temporal) of digital soil maps for its further use on digital soil

assessments (Carré et al., 2007; Greve and Seneviratne, 2015; Stell et al., 2019), ecological interpretations, or other environmental and Geo-scientific applications.

This digital soil mapping research is focused on soil organic carbon and soil moisture, which are major planetary resources supporting critical ecosystem services such as food and energy production and water storage. In this research, digital soil mapping is the basis for predicting soil carbon and soil moisture information from the plot, to the country-specific, to the global scale. Through this research, I compared and tested multiple digital soil mapping approaches to generate predictions of soil organic carbon and soil moisture including an evaluation of the reliability and uncertainty of these predictions compared with previous work. This research is timely because the contribution of soil organic carbon stocks and soil moisture trends to the global carbon cycling remains unknown across large areas of the world most likely due to the low availability of accurate estimates that has been reported on soil organic carbon and soil moisture datasets. Therefore, there is an increasing research opportunity to explore and evaluate soil organic carbon and soil moisture feedbacks in response to global environmental change. Using remote sensing data, open source platforms for statistical computing and digital terrain analysis, and multiple sources of soil-related information (climate datasets, geology maps, land use/cover maps), with this research I want to lead the discovery of new knowledge around soil carbon and soil moisture from the country to the global scale.

This research focuses on the coupling of soil science, Geomorphometry, remote sensing and machine learning to predict the spatial variability of soil organic carbon and

soil moisture by the means of digital soil mapping. The overreaching goal of my PhD research is to develop a digital soil mapping framework to increase the statistical accuracy of spatially continuous information on soil moisture and soil organic carbon across different scales of data availability (e.g., country-specific, regional, global).

Increasing the statistical accuracy and detail of spatially continuous information on soil organic carbon and soil moisture will benefit current research on the regional to-global carbon and water cycles, and the formulation of public policy around global societal issues such as land degradation and water scarcity. This research seeks to improve the statistical accuracy and spatial detail of currently soil organic carbon and soil moisture gridded estimates exploring the following interrelated research questions. 1) Which are the most effective prediction algorithms for predicting soil moisture and soil organic carbon from the country-specific to global case studies? 2) Which are the key prediction factors for modeling the spatial variability of soil moisture and soil organic carbon? 3) Can we improve the spatial detail and the statistical accuracy of satellite soil moisture using Geomorphometry and machine learning? and 4) How much we know of current trends soil moisture and soil organic? Developing this research, I will show the use of a digital soil mapping strategy applied to soil moisture and soil organic carbon using multiple forms of statistical learning across scales and spatial configurations of multiple soil information sources.

REFERENCES

Amundson, R., Berhe, A. A., Hopmans, J. W., Olson, C., Sztein, A. E. and Sparks, D. L.: Soil science. Soil and human security in the 21st century, Science, 348(6235), 1261071, doi:10.1126/science.1261071, 2015.

Arrouays, D., Richer-De-Forges, A., Mcbratney, A., Hartemink, A., Minasny, B., Savin, I., Grundy, M., Leenaars, J., Poggio, L., Roudier, P., Libohova, Z., Mckenzie, N., Bosch, H. van den, Kempen, B., Mulder, V., Lacoste, M., Chen, S., Saby, N., Martin, M., Dobarco, M. R., Cousin, I., Loiseau, T., Lehmann, S., Caubet, M., Lemercier, B., Walter, C., Vaudour, E., Gomez, C., Martelet, G., Krasilnikov, P. and Lagacherie, P.: The globalsoilmap project: past, present, future, and national examples from france, Dokuchaev Soil Bulletin, (95), 3–23, doi:10.19047/0136-1694-2018-95-3-23, 2018.

Batjes, N. H., Ribeiro, E. and Oostrum, A. van: Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019), Earth System Science Data Discussions, 1–46, doi:https://doi.org/10.5194/essd-2019-164, 2019.

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 System Science Data, 9(1), 1–14, doi:10.5194/essd-9-1-2017, 2017.

Bond-Lamberty, B. and Thomson, A.: A global database of soil respiration data, Biogeosciences, 7(6), 1915 – 1926, doi:10.5194/bg-7-1915-2010, 2010.

Carré, F., McBratney, A. B., Mayr, T. and Montanarella, L.: Digital soil assessments: Beyond DSM, Geoderma, 142(1), 69–79, doi:10.1016/j.geoderma.2007.08.015, 2007.

Dorigo, W., Oevelen, P. van, Wagner, W., Drusch, M., Mecklenburg, S., Robock, A. and Jackson, T.: A New International Network for in Situ Soil Moisture Data, Eos, Transactions American Geophysical Union, 92(17), 141–142, doi:10.1029/2011EO170001, 2011.

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A. and al, et: ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions, Remote Sensing of Environment, 203, 185–215, doi:10.1016/j.rse.2017.07.001, 2017.

Folberth, C., Skalsky, R., Moltchanova, E., Balkovic, J., Azevedo, L., Obersteiner, M. and van der Velde, M.: Uncertainty in soil data can outweigh climate impact signals in crop yield simulations, Nature Communications, 7, art.no.11872. DOI:10.1038/ncomms11872 https://doi.org/10.1038/ncomms11872, 2016.

Greve, P. and Seneviratne, S. I.: Assessment of future changes in water availability and aridity, Geophysical Research Letters, 42(13), 5493–5499, doi:10.1002/2015gl064127, 2015.

Guevara, M., Olmedo, G. F., Stell, E., Yigini, Y., Aguilar Duarte, Y., Arellano Hernández, C., Arévalo, G. E., Arroyo-Cruz, C. E., Bolivar, A., Bunning, S., Bustamante Cañas, N., Cruz-Gaistardo, C. O., Davila, F., Dell Acqua, M., Encina, A., Figueredo Tacona, H., Fontes, F., Hernández Herrera, J. A., Ibelles Navarro, A. R., Loayza, V., Manueles, A. M., Mendoza Jara, F., Olivera, C., Osorio Hermosilla, R., Pereira, G., Prieto, P., Ramos, I. A., Rey Brina, J. C., Rivera, R., Rodríguez-Rodríguez, J., Roopnarine, R., Rosales Ibarra, A., Rosales Riveiro, K. A., Schulz, G. A., Spence, A., Vasques, G. M., Vargas, R. R. and Vargas, R.: No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America, SOIL, 4(3), 173–193, doi:10.5194/soil-4-173-2018, 2018.

Gu, X., Li, J., Chen, Y. D., Kong, D. and Liu, J.: Consistency and Discrepancy of Global Surface Soil Moisture Changes From Multiple Model-Based Data Sets Against Satellite Observations, Journal of Geophysical Research: Atmospheres, 124(3), 1474–1495, doi:10.1029/2018JD029304, 2019.

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, Global Change Biology, 24(2), e705–e718, doi:10.1111/gcb.13896, 2018.

Hengl, Tomislav (2003). Pedometric mapping : bridging the gaps between conventional and pedometric approaches. Wageningen: s.n. ISBN 9789058088963.

Hengl, T., Jesus, J. M. de, Heuvelink, G. B. M., Gonzalez, M. R., 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(2), e0169748, doi:10.1371/journal.pone.0169748, 2017.

Hengl, T., Wheeler, I. and MacMillan, R. A.: A brief introduction to Open Data, Open Source Software and Collective Intelligence for environmental data creators and users, PeerJ Inc., 2018.

Lorenz, K., Lal, R. and Ehlers, K.: Soil organic carbon stock as an indicator for monitoring land and soil degradation in relation to United Nations' Sustainable Development Goals, Land Degradation & Development, 30(7), 824–838, doi:10.1002/ldr.3270, 2019.

Ma, Y., Minasny, B., Malone, B. P. and Mcbratney, A. B.: Pedology and digital soil mapping (DSM), European Journal of Soil Science, 70(2), 216–235, doi:10.1111/ejss.12790, 2019.

Malhotra, A., Todd-Brown, K., Nave, L. E., Batjes, N. H., Holmquist, J. R., Hoyt, A. M., Iversen, C. M., Jackson, R. B., Lajtha, K., Lawrence, C., Vindušková, O., Wieder, W., Williams, M., Hugelius, G. and Harden, J.: The landscape of soil carbon data: Emerging questions, synergies and databases:, Progress in Physical Geography: Earth and Environment, doi:10.1177/0309133319873309, 2019.

van der Molen, M. K., Dolman, A. J., Ciais, P., Eglin, T., Gobron, N., Law, B. E., Meir, P., Peters, W., Phillips, O. L., Reichstein, M. and al, et: Drought and ecosystem carbon cycling, Agricultural and Forest Meteorology, 151(7), 765–773, doi:10.1016/j.agrformet.2011.01.018, 2011.

Pfeiffer, M., Padarian, J., Osorio, R., Bustamante, N., Olmedo, G. F., Guevara, M., Aburto, F., Antilen, M., Araya, E., Arellano, E., Barret, M., Barrera, J., Boeckx, P., Briceño, M., Bunning, S., Cabrol, L., Casanova, M., Cornejo, P., Corradini, F., Curaqueo, G., Doetterl, S., Duran, P., Escudey, M., Espinoza, A., Francke, S., Fuentes, J. P., Fuentes, M., Gajardo, G., García, R., Gallaud, A., Galleguillos, M., Gomez, A., Hidalgo, M., Ivelic-Sáez, J., Mashalaba, L., Matus, F., Mora, M. de la L., Mora, J., Muñoz, C., Norambuena, P., Olivera, C., Ovalle, C., Panichini, M., Pauchard, A., Perez-Quezada, J. F., Radic, S., Ramirez, J., Riveras, N., Ruiz, G., Salazar, O., Salgado, I., Seguel, O., Sepúlveda, M., Sierra, C., Tapia, Y., Toledo, B., Torrico, J. M., Valle, S., Vargas, R., Wolff, M. and Zagal, E.: CHLSOC: The Chilean Soil Organic Carbon database, a multi-institutional collaborative effort, Earth System Science Data Discussions, 1–17, doi:https://doi.org/10.5194/essd-2019-161, 2019.

Samuel-Rosa, A., Dalmolin, R. S. D., Moura-Bueno, J. M., Teixeira, W. G., Alba, J. M. F., Samuel-Rosa, A., Dalmolin, R. S. D., Moura-Bueno, J. M., Teixeira, W. G. and Alba, J. M. F.: Open legacy soil survey data in Brazil: geospatial data quality and how to improve it, Scientia Agricola, 77(1), doi:10.1590/1678-992x-2017-0430, 2020.

Soil Science Division Staff. 2017. Soil survey manual. C. Ditzler, K. Scheffe, and H.C. Monger (eds.). USDA Handbook 18. Government Printing Office, Washington, D.C.

Stoorvogel, J. J., Bakkenes, M., Temme, A. J. A. M., Batjes, N. H. and Brink, B. J. E.: S-World: A Global Soil Map for Environmental Modelling, Land Degradation & Development, 28(1), 22–33, doi:10.1002/ldr.2656, 2016.

Stell, E., Guevara, M. and Vargas, R.: Soil swelling potential across Colorado: A digital soil mapping assessment, [online] Available from: https://pubag.nal.usda.gov/catalog/6474669 (Accessed 1 January 2020), 2019.

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 Biogeochemical Cycles, 32(1), 42–56, doi:10.1002/2017GB005678, 2018.

Walsh, B., Ciais, P., Janssens, I. A., Peñuelas, J., Riahi, K., Rydzak, F., van Vuuren, D. P. and Obersteiner, M.: Pathways for balancing CO2 emissions and sinks, Nat Commun, 8, 14856, doi:10.1038/ncomms14856, 2017.

Yigini Y., G. Olmedo, S. Reiter, K. Viatkin and R. Vargas: Soil Organic Carbon Mapping Cookbook 2nd edition, 2018.

Chapter 2

NO SILVER BULLET FOR DIGITAL SOIL MAPPING: COUNTRY-SPECIFIC SOIL ORGANIC CARBON ESTIMATES ACROSS LATIN AMERICA

Mario Guevara[1], Guillermo Federico Olmedo[2,3,] Emma Stell[1], Yusuf Yigini[3], Yameli Aguilar Duarte[4], Carlos Arellano Hernández[5], Gloria E. Arévalo[6], Carlos Eduardo Arroyo-Cruz[7], Adriana Bolivar[8], Sally Bunning[9], Nelson Bustamante Cañas[10], Carlos Omar Cruz-Gaistardo[5], Fabian Davila[11], Martin Dell Acqua[11], Arnulfo Encina[12], Hernán Figueredo Tacona[13], Fernando Fontes[11], José Antonio Hernández Herrera[14], Alejandro Roberto Ibelles Navarro[5], Veronica Loayza[15], Alexandra M. Manueles[6], Fernando Mendoza Jara[16], Carolina Olivera[17], Rodrigo Osorio Hermosilla[10], Gonzalo Pereira[11], Pablo Prieto[11], Iván Alexis Ramos[18], Juan Carlos Rey Brina[19], Rafael Rivera[20], Javier Rodríguez-Rodríguez[7], Ronald Roopnarine[21,22], Albán Rosales Ibarra[23], Kenset Amaury Rosales Riveiro[24], Guillermo Andrés Schulz[25], Adrian Spence[26], Gustavo M. Vasques[27], Ronald R. Vargas[3], and Rodrigo Vargas[1]

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

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

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

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

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

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

[7]National Commission for the Knowledge and Use of Biodiversity, Mexico City, Mexico
8Subdirección Agrología, Instituto Geográfico Agustín Codazzi, Bogotá, Colombia

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

[10]Servicio Agrícola y Ganadero, Santiago de Chile, Chile

[11]Direccion General de Recursos Naturales, Ministerio de Ganaderia, Agricultura y Pesca, Montevideo, Uruguay

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

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

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

15Ministerio de Agricultura y Ganaderia, Quito, Ecuador

[16]Universidad Nacional Agraria, Managua, Nicaragua

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

[18]Instituto de Investigación Agropecuaria de Panamá, Panamá, Panama

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

[20]Ministerio de Medio Ambiente, Santo Domingo, Dominican Republic

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

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

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

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

[25]INTA CNIA, Buenos Aires, Argentina

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

[27]Embrapa Solos, Rio de Janeiro, Brazil

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 60 %, with no universal predictive algorithm among countries. A regional (*n* 11 268 SOC estimates) ensemble of these five algorithms was able to explain 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.

2.1 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 or- ganic 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 predicted as a function of the soil-forming environment, in cor- respondence with soil-forming factors from the Dokuchaev hypothesis and Jenny's soil-forming equation based on cli- mate, 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):

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

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 at- tribute 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 (ideally) 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 mea- sure the bias of global SOC estimates, specifically, at particu- lar 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, com- paring several approaches and methods is useful to quantify the relative importance of prediction factors across data con- figurations 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; Shangguan et al., 2017) and generate site-specific and country-specific SOC maps (Viscarra Rossel et al., 2014; Ad- hikari 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.

2.2 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 (Figure 2.1) could be reproduced across the world for comparative purposes between country-specific and global estimates.

2.2.1 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 is 1/(0.6268 0.0361 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):

$$SOC_{stock} = (ORCDR/1000) * (H/100) * BDL * (100-CRFVOL/100) \quad (2)$$

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.

2.2.2 SOC error estimates

The GSIF approach for estimating SOC (function OCSKGM) 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.

2.2.3 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 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.

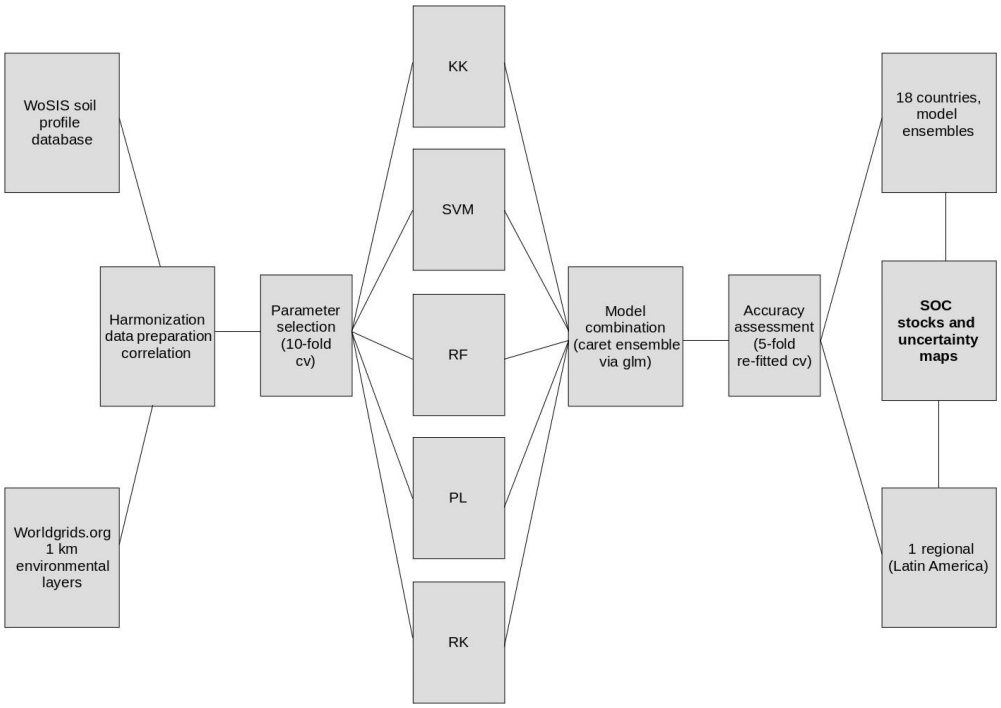

Figure 2.1. Flow diagram of the main methodological steps that we performed in order to generate country-specific and regional SOC predictions. The World Soil Information Service (WoSIS) dataset was harmonized with the http://worldgrids.org (last access: 20 February 2018) environmental data using 5 5 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 re- peated 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.

2.2.4 Soil 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. (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.

2.2.5 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 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 modelling 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.

2.2.6 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 (ECr) 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

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

$$corr_{SD} = corr_i - \min(corr)/range(corr) \quad (4)$$

$$EC = RMSE_{SD}/corr_{SD} \quad (5)$$

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 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(corr) is the minimum observed value of correlation, and range(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 R2) 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).

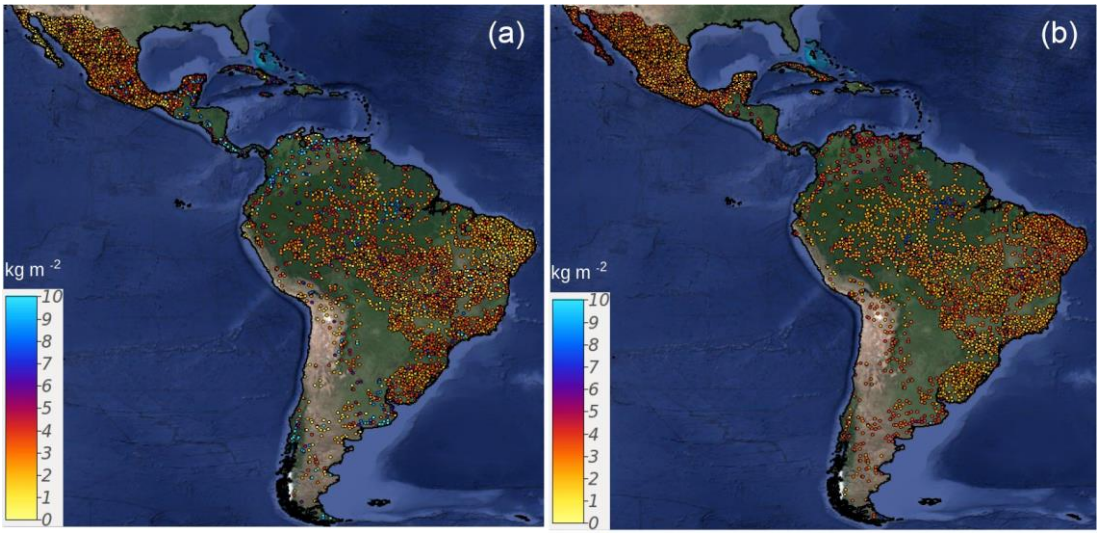

Figure 2.2. Spatial distribution of available SOC in WoSIS for Latin America. SOC estimates are calculated for each point using Eq. (2) **(a)**. The approximated error is based on Taylor series as implemented in the R-GSIF package, as is explained in Heuvelink (2018) **(b)**. Thus, panel **(b)** represents 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.

2.2.7 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).

2.3 Results

2.3.1 Descriptive statistics

SOC across different countries showed a wide diversity of data scenarios (Table 2.1). 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 2.1). 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.

2.3.2 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 (Figure 2.2).

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 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 2.2). 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., SOC_{ARG} f [px4wcl3a + px3wcl3a + evmmod3a + l07igb3a+ px2wcl3a …]).

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 (l07igb3a) (Table 2.2) (see sources of these maps in http://worldgrids.org/doku.php/wiki:layers, last access: 20 February 2018).

Table 2.1. 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	n	Land area (km2)	Min.	First Q	Med.	Mean	Third Q	Max.	SD	p/p 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

2.3.4 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 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 (Figure 2.3). 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.

Table 2.2. 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
gphhws3a	Soil	Soil type	2	CUB, JAM
gplhws3a	Soil	Soil type	2	BLZ, BOL
gvrhws3a	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
etmnts3a	Climate	ET	2	ARG, MEX
evmmod3a	Organism	Vegetation	5	ARG, ECU, HND, MEX, VEN
l07igb3a	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

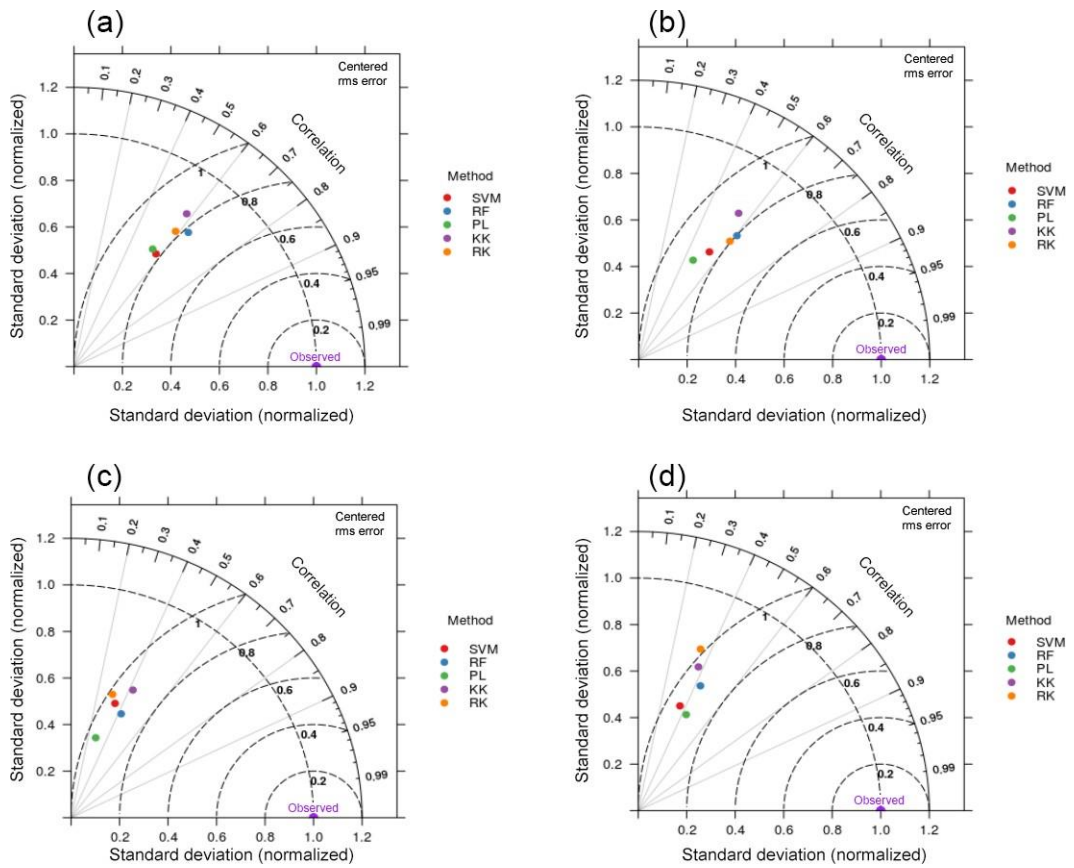

Figure 2.3. Taylor diagrams showing the performance of the five models evaluated. SOC stock **(a)**, ORCDR **(b)**, BLD **(c)**, and CRFVOL **(d)**. This analysis is based on all available data across Latin America. Although RF tends to generate higher correlation, it also shows high variance in predictions. 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.

2.3.5 Country-specific SOC predictions

We did not find a dominant algorithm to predict SOC on a country-specific basis (Figure 2.4). Overall, machine learn- ing 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., Suri- name, 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 2.1). 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.

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.

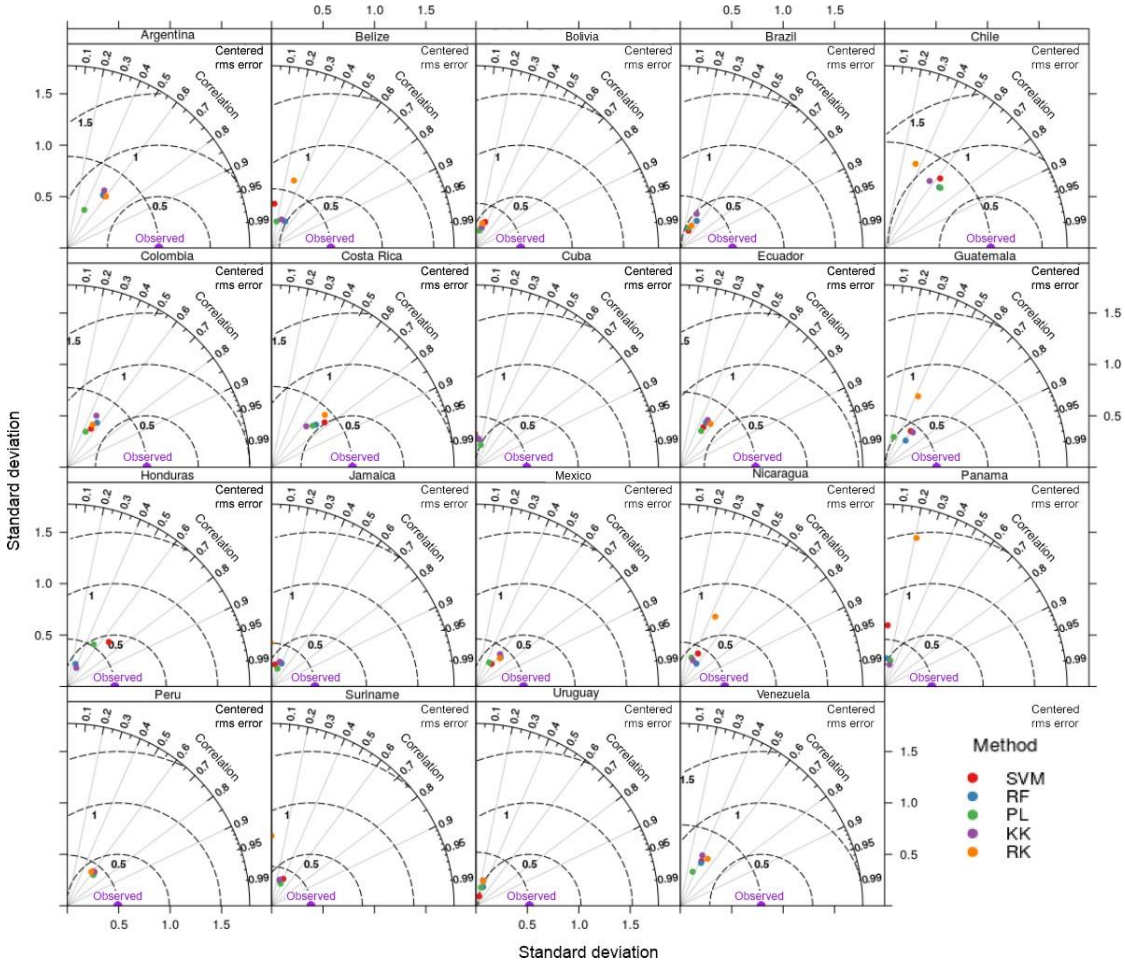

Figure 2.3 Taylor diagrams showing the performance of the five models evaluated for country-specific SOC estimates across Latin America. The position of each point/method varies from each dataset to another, suggesting that the predictive capacity changes when data characteristics are different.

2.4 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 (Figure 2.5a). These areas correspond to areas where we report both higher SOC calculation errors and model un- certainty (Figure 2.6).

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 (Figure 2.5b). 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 r2 0.39 for the regional model and r2 values for the country-specific approach that vary from 0.01 to 0.55.

High uncertainty in our modeling framework was found across tropical, arid, and semiarid regions of Latin America (Figure 2.6a, 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 (Figure 2.6c). 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 (Figure 2.7a). We report that the geographical areas where country-specific models tend to predict higher SOC values than the regional ensemble (Figure 2.7b). However, we report a similar SOC stock from both modeling approaches (country-specific and global) as we explain in Sect. 3.7.

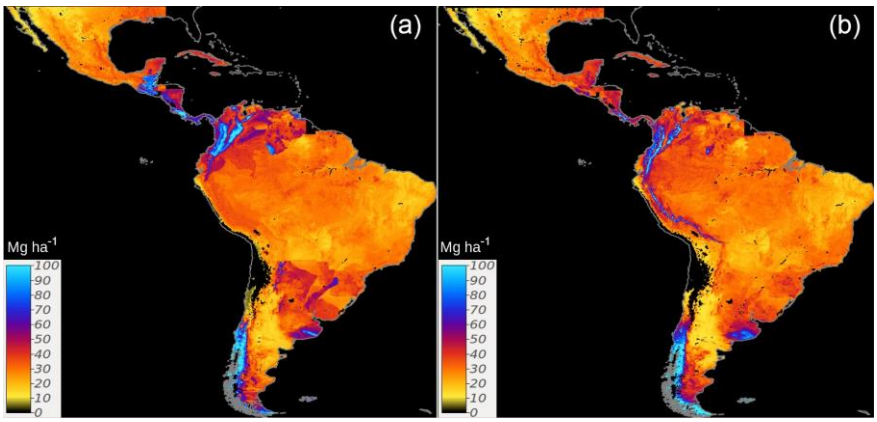

Figure 2.4. Country-specific **(a)** and regional (Latin America) **(b)** predictions of SOC based on a linear ensemble of methods. We present the units as Mg ha for visualization purposes. These units were used to reduce the digits of the value range and highlight larger differences between SOC maps.

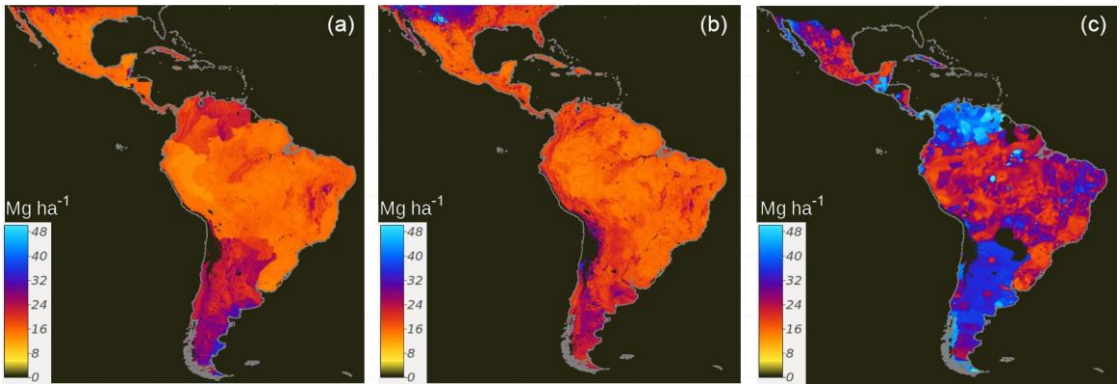

Figure 2.5. The full conditional response of residuals to the prediction factors on a country-specific basis **(a)**. The full conditional response of residuals to the SOC prediction factors in the regional (Latin America) model **(b)**. The trend of the approximated error of SOC estimates is derived from buffer distances and the random forest spatial framework **(c)**.

2.4.1 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, which was also useful for visualization purposes. For our models, the uncertainty of the maximum prediction limit was estimated to be ± 10 Pg, which was the variance of the SOC stock by increasing the prediction limit from 1 to 700 Mg ha (Figure 2.8).

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 2.3).

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 2.4). 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 2.3).

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.

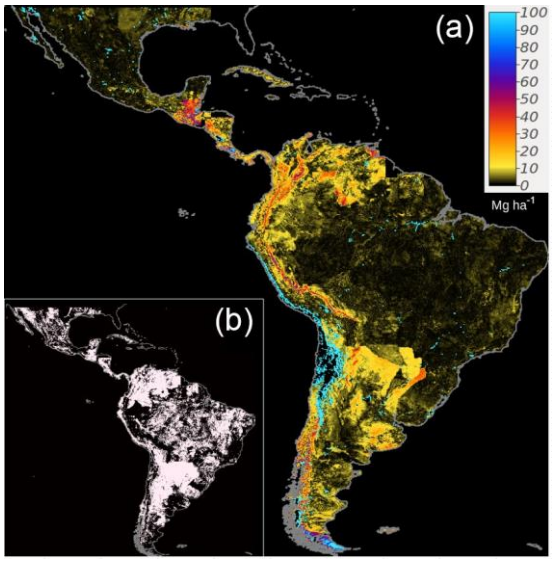

Figure 2.6. The absolute distance (Mg ha) between the country- specific and the regional ensemble **(a)**. The areas in white are areas where the country-specific modeling is predicting higher SOC than the regional estimate (i.e., country-specific is greater than regional) **(b)**.

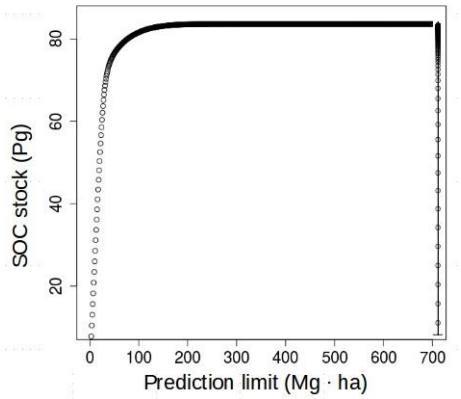

Figure 2.7. The absolute distance (Mg ha) between the country- specific and the regional ensemble **(a)**. The areas in white are areas where the country-specific modeling is predicting higher SOC than the regional estimate (i.e., country-specific is greater than regional) **(b)**.

2.5 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 2.3 and 2.4).

Table 2.3. SOC stocks (Pg) at the contextual resolution of 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	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	47.2
5	Chile	6.31	7.2	15.15	1.93	8.28
6	Colombia	7.01	5.96	15.5	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	1.1	0.33	0.69
15	Peru	4.38	5.13	17.08	3.14	10.51
16	Suriname	0.56	0.51	1.2	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

Table 2.4. SOC stocks (Pg) at the contextual resolution of 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

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 (Figure 2.6), 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 r2 was RF, an ensemble of regression trees based on bagging. However, this method yields high ECr, 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 ECr. Large values of ECr 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 ac-curacy 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.

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 spatial 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. 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; Sander- man 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 (Figure 2.8). Therefore, we argue that data science systems for DSM focused on carbon assessments should be fundamentally based on SOC expert knowledge and informed by expert-based soil mapping systems.

2.6 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 sev- eral 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 mtry 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.

Supplement. 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., Sulaeman, Y., Vintila, R., Rossel, R. V., Wilson, P., Zhang, G, 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, 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., Karunaratne, 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., Tˇupek, 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.uniuenchen.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., Blagotic´, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Gue-vara, 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 spatiotemporal 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., Twenh”ofel, 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 1: 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, 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, 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., Schaepman, 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 Guangcan, 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., Mitran, T., Sujatha, G., Suresh, G. J. R., Fyzee, 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 colaborative 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.

Chapter 3

SOIL ORGANIC CARBON ACROSS MEXICO AND THE CONTERMINOUS UNITED STATES (1991-2010)

Authors:

[1]Mario Guevara, [2]Carlos Arroyo, [3]Nathaniel Brunsell, [4]Carlos O. Cruz, [5]Grant Domke, [2]Julian Equihua, [6]Jorge Etchevers, [7]Daniel Hayes, [8]Tom Hengl, [4]Alejandro Ibelles, [5]Kris Johnson, [9]Ben de Jong, [10]Zamir Libohova, [1]Ricardo Llamas, [11]Lucas Nave, [4]Jose L. Ornelas, [7]Fernando Paz, [2]Rainer Ressl, [12]Anita Schwartz, [4]Arturo Victoria, [10]Skye Wills and [1]Rodrigo Vargas.

Affiliations:

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

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

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

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

[5]United States Department of Agriculture, Forest Service

[6] Colegio de Postgraduados, Texcoco, México

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

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

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

[10]National Soil Resource Conservation Service, Lincoln, Nebraska
[11]The International Soil Carbon Network-University of Michigan Biological Station, Michigan

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

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

Key points:

Multisource topsoil organic carbon prediction and prediction variance in Mexico and the conterminous

United States.

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

regression framework.

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

Abstract

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

3.1 Introduction

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

cycling requires a good understanding of different uncertainty sources around SOC datasets and SOC modeling efforts.

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

Global modelling outputs for SOC represent the only estimates of SOC across large areas of the world without in situ ground observations. These global estimates rely on large datasets that combine multiple SOC data collection periods and methods for calculating SOC stocks. These inconsistencies represent a known but unquantified bias for calculating SOC stocks (Poeplau et al., 2017). Thus, there is a need to test different modelling approaches across areas with high density of SOC observations to improve the accuracy, detail and reliability of global SOC estimates (Vitharana et al., 2019). The Harmonized World Soil Data Base (HWSD) or the harmonized soil property values for broadscale modeling (WISE30sec, Batjes, 2016), are probably the most commonly used datasets for spatially quantifying SOC stocks and its spatial variability patterns at the global scale (Köchy *et al.*2015; O'Rourke *et al.* 2015). The HWSD provides the most complete global soil description from synthesizing many regional or national soil maps,

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

Large discrepancies have been reported among global (Tifati *et al.* 2017) or country specific SOC estimates (Guevara *et al.* 2018). Consequently, reporting uncertainty and bias of SOC estimates will allow better parameterization of land surface models, improved local to regional monitoring baselines and informed policy and

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

North America is a region characterized by a long history of soil data collection that has produced unprecedented information of SOC. For example, SOC predictions and estimates across Mexico and the CONUS are available from a variety of methods

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

In this study, we quantified the spatial variability and associated uncertainty of SOC stocks across different land use categories of CONUS and Mexico; two countries with rich information of SOC measurements. Previous analyses have shown large discrepancies in SOC stocks (0-30 cm, ranging from ~38 to ~92 Pg of SOC) derived

from country-specific or global SOC esimates (Lajtha *et al.* 2018, Hengl *et al*, 2017, Paz-Pellat *et al.* 2016, Bliss *et al.* 2014, Wieder *et al.* 2014). Our main goal was to generate a spatial predictive model of SOC variability for the top 30 cm depth at 250 m spatial resolution across both countries with information collected between 1991 and 2010. We asked the following interrelated questions: 1) Which are the best SOC environmental covariates (i.e., prediction factors increasing SOC modeling accuracy) across Mexico and CONUS? 2) How much variation in SOC can machine learning methods (e.g., tree-based, kernel based, probabilistic-based) explain across this region using repeated cross-validation? 3) What is the SOC variance associated with multiple calculation methods for estimating SOC stocks and what is the variance associated to multiple model predictions? and 4) What are the sensitivities of these predictions associated to different training datasets (i.e., decadal information from different collection periods; 1991-2000 and 2001-2010)? The value of considering different collection periods is to explore the sensitivity of model uncertainty associated to different training datasets and provide insights for better interpretation of decadal changes in SOC stocks. In summary, this study provides benchmark information about how SOC spatial distribution is constrained by the soil forming environment (i.e., climate, biota, topography and geology), and quantifies the variance (spatial and temporal) from using multiple SOC observational datasets.

3.2 Datasets and methods

We followed a digital soil mapping strategy (Figure 3.1) for the prediction of the spatial variability of SOC across both countries. Digital soil maps are generated using

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

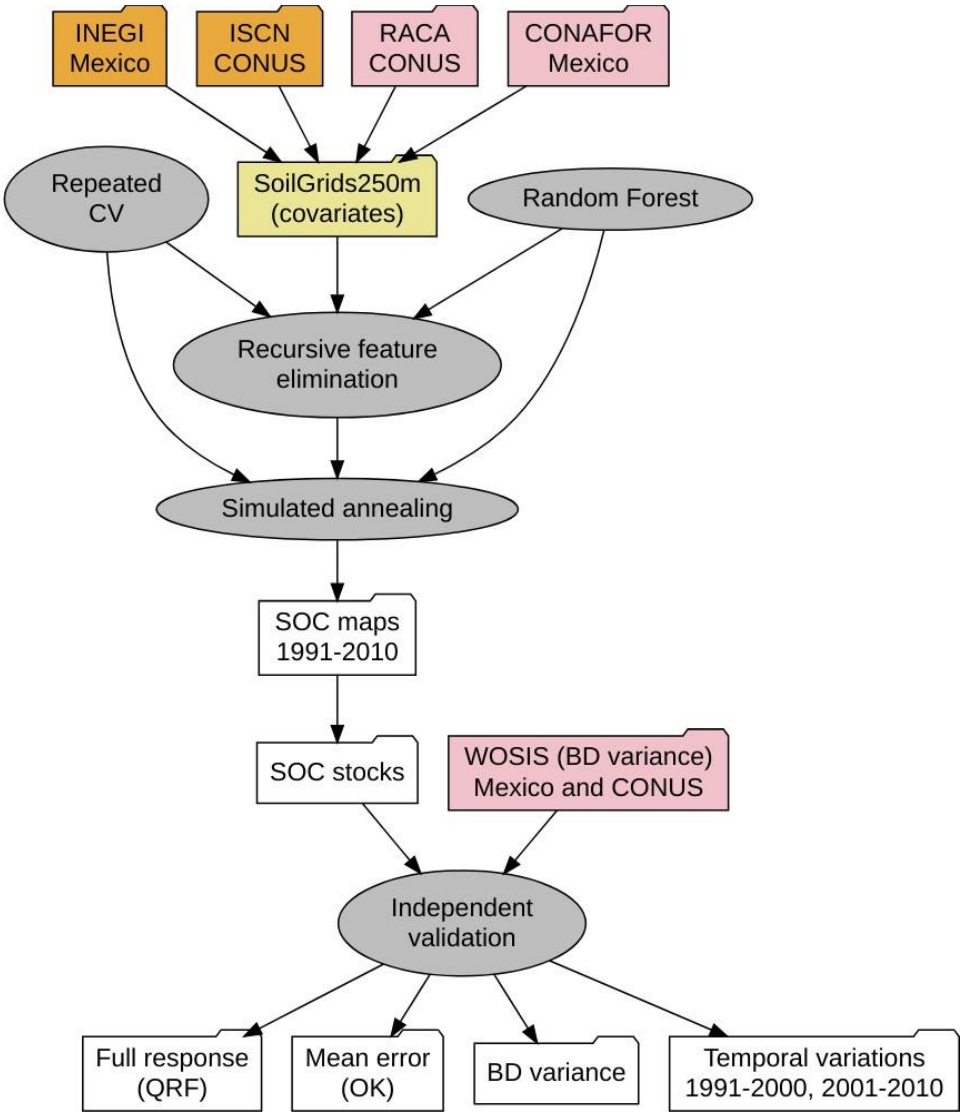

Figure 3.1. Flow diagram of the proposed methodology to predict the spatial variability of SOC stocks across Mexico and CONUS. Orange folders indicate the SOC data sources used for training and pink folders indicate SOC data sources used for validating. These sources were harmonized with the SoilGrids250m covariates. White folders indicate the main results of this methodology. Gray ovals indicate main methodological steps. CV: 5-fold cross validation, QRF: quantile regression forest, OK: Ordinary Kriging, BD: bulk density.

3.2.1 SOC observational data

Legacy SOC estimates across CONUS were obtained from the International Soil Carbon Network (ISCN latest version 2018, >18 000 pedons available, Harden *et al.* 2017). Data from Mexico was provided by the Instituto Nacional de Estadística y Geografía (INEGI, SERIES 1 & 2; n >65 000 pedons available, Krasilnikov *et al.* 2013). We used only the observations collected between 1991 and 2010 to minimize confounding factors (related to potential changes in the SOC pool; n = 10385, Figure 3.2). We considered all soil horizons containing upper and lower soil depth limit information. The combination of using soil depth continuous functions (Bishop *et al.* 1999; Malone *et al.* 2009) and deriving the weighted average (by depth) from the first sampled soil horizon at 0 cm depth to all soil horizons sampled within the first 30 cm of soil depth, allowed aggregating irregular soil horizons for calculating SOC stocks across both countries. The weights for calculating these stocks were selected defining the proportion of each horizon within this 0-30 cm interval of soil depth.

Most contributors (across CONUS) and INEGI (in Mexico) considered (or adapted) the United States Department of Agriculture Soil Taxonomy guidelines for interpreting soil surveys including SOC and other soil variables (Soil Survey Staff. 1999). For CONUS, the ISCN database provides a harmonized compilation from many contributors (e.g., Natural Resources Conservation Service, United States Geological Survey and site-specific research or academic groups; Harden *et al.* 2017). However, the largest contributor for this curated dataset is the United States Department of Agriculture Natural Resource Conservation Service, where the SOC concentration was

mainly obtained by the Walkley-Black technique (Soil Survey Staff. 2014). All samples for Mexico were systematically collected and analyzed by INEGI (INEGI, 2014, Krasilnikov *et al.* 2013) and SOC concentration was also measured using the Walkley-Black technique (IUSS-WRB-FAO, 2014). Potential error propagation from the use of different methods to calculate SOC using information collected over long periods of time (before 1991) is beyond the scope of this study. We only considered the sensitivity of SOC models (i.e., model outputs) to variations in training data and inputs derived from different pedotransfer functions for estimating bulk density (section 2.6).

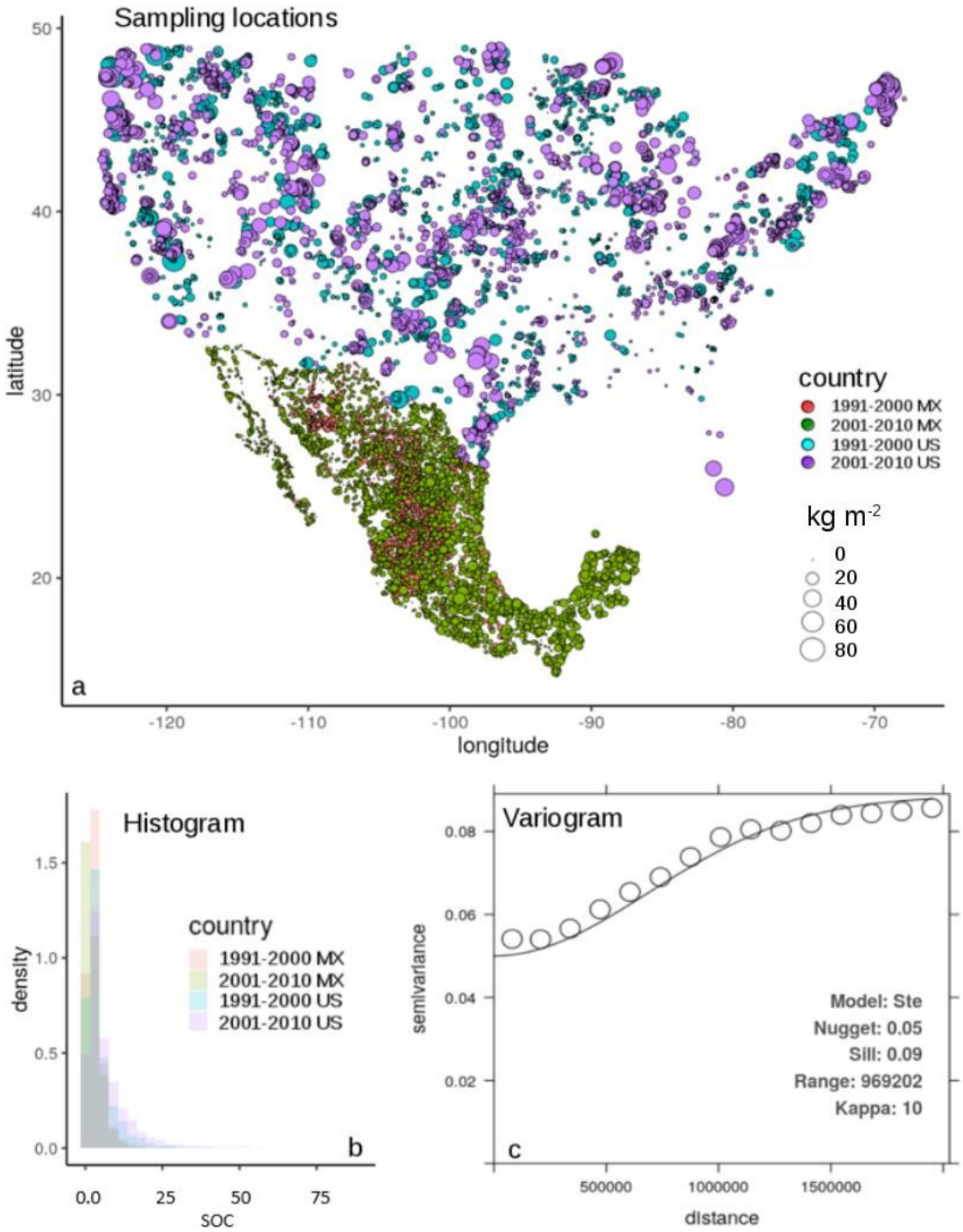
Sampling locations
latitude
longitude
country
1991-2000 MX
2001-2010 MX
1991-2000 US
2001-2010 US
kg m-2
0
20
40
60
80
a
Histogram
density
SOC
country
1991-2000 MX
2001-2010 MX
1991-2000 US
2001-2010 US
b
Variogram
semivariance
distance
Model: Ste
Nugget: 0.05
Sill: 0.09
Range: 969202
Kappa: 10
c

Figure 3.2. Distribution and descriptive statistics of available datasets. The point map shows the spatial sampling locations of available data for the period 1991-2010 (1991-2000 and 2001-2010) (a). The colored histograms are representing the statistical distribution of all datasets (i.e., combined CONUS and Mexico information) (b). The variogram (relation between distance and variance of observed values) and variogram parameters (nugget, sill and range) are representing the spatial structure captured with available data (c). *Ste* is a Stein model parameterization (and its associated Kappa value) for the covariance function between all pairs of points separated by distance units (range in meters) defining the spatial structure of SOC available datasets.

3.2.2 Calculation of SOC stocks

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

3.2.3 The environmental covariate space

For spatially representing soil forming factors (Jenny, 1941) we used environmental covariates from the SoilGrids250m system (Hengl *et al.* 2017). This

dataset represents over 150 variables of environmental gridded data including terrain derivatives from a digital elevation model (DEM), the enhanced vegetation index (EVI), climate (precipitation and land surface temperature) and other soil related gridded variables (Figure 3.3, Supplementary Table S1). This covariate space is representative for the analyzed period of time (1991-2010) and is described in detail by previous publications (Hengl, *et al.* 2017, Reuter and Hengl, 2012). We extracted this global information within the geographical limits of the NALCMS (North American Land Change Monitoring System, 77% of land use classification accuracy) at 250m spatial resolution (NRCan/CCRS-USGS-INEGI-CONABIO-CONAFOR, 2005).

3.2.4 Recursive feature elimination

We first performed a variable reduction strategy using of a recursive feature elimination technique (Kuhn *et al.* 2008) and multiple models were fitted repeatedly using all possible combinations of highly ranked predictors. Predictors were ranked using as indicator the cross-validated prediction error of a Random Forest tree ensemble. We selected the Random Forest as our overall accuracy indicator method because it showed the highest predictive capacity compared against different machine learning algorithms tested in our modeling selection strategy (Supplementary Figure S1). This method is based on bagging predictors and the combination of multiples regression trees derived from different random data subsets (Breiman, 2001). Each model grows with the number of trees for minimizing the prediction variance. Model parameters to define the number of predictors and subsets on each regression tree were automatically selected by the means of 10-fold cross validation (Figure 3.1). Cross

validation is a re-sampling technique that we used for maximizing the accuracy of results while obtaining a robust and stable prediction error estimate used for further selecting the most informative predictors. In addition, Random Forests uses an out-of-bag cross-validation form for assessing the relevance of each predictor in the model. Thus, multiple lists of the "best" predictors are generated from each Random Forest model realization in the recursive feature elimination framework. This provides a probabilistic assessment to determine the best predictors to retain at the end of the algorithm (Kuhn *et al.* 2008). After a 5-times repeated 5-fold cross validation for the recursive feature elimination technique (to account for the model sensitivity to data variations and reduce overfitting), we selected the first 25 environmental covariates for SOC.

3.2.5 Simulated annealing

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

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

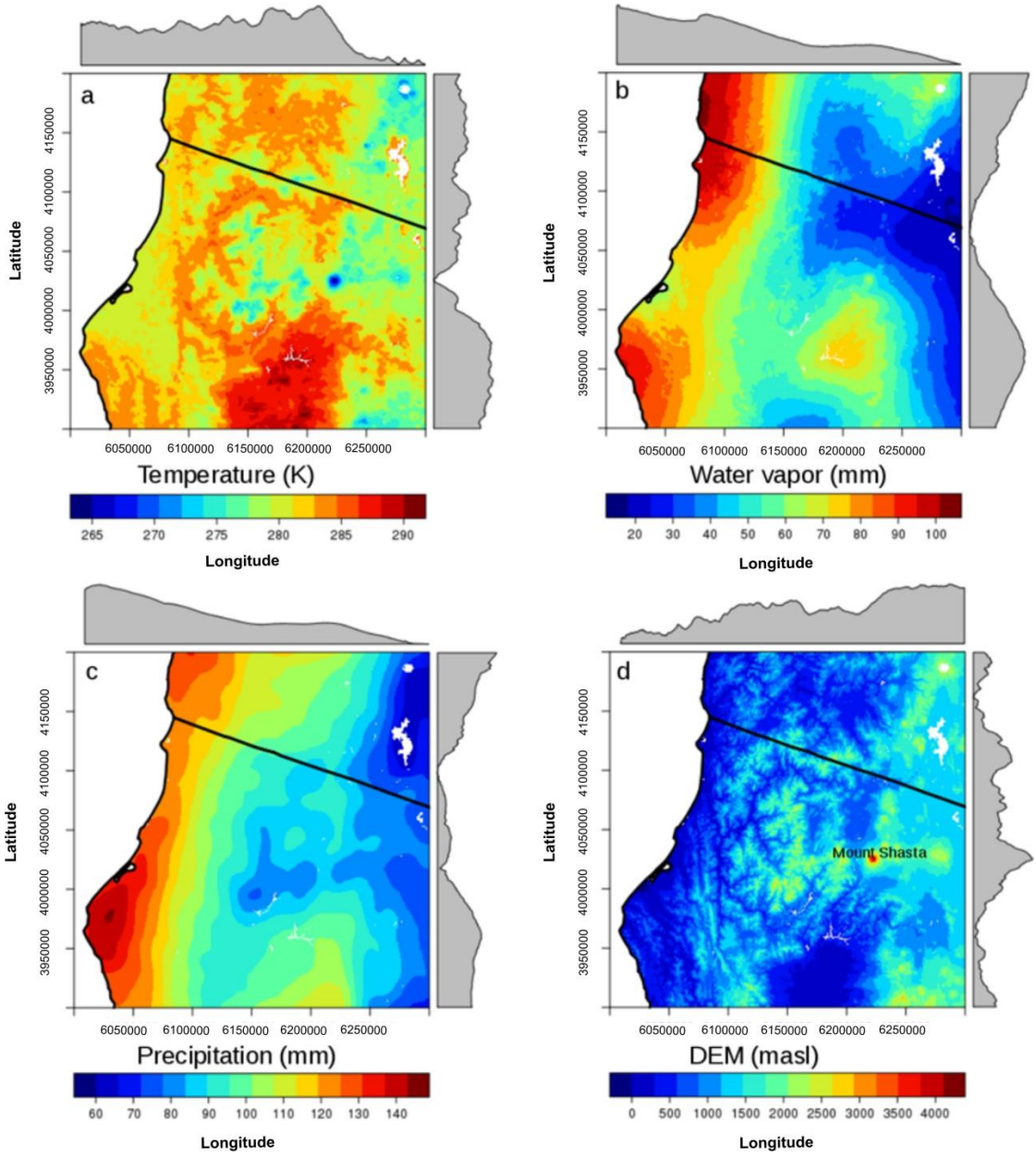
a
b
c
d
Latitude
Longitude
6050000 6100000 6150000 6200000 6250000
3950000 4000000 4050000 4100000 4150000
Temperature (K)
265 270 275 280 285 290
Water vapor (mm)
20 30 40 50 60 70 80 90 100
Precipitation (mm)
60 70 80 90 100 110 120 130 140
DEM (masl)
0 500 1000 1500 2000 2500 3000 3500 4000
Mount Shasta

Figure 3.3. Visualization of covariates across the political boundaries between California and Oregon in western CONUS. Land surface temperature (a); precipitation (b); precipitable water vapor (c); and a digital elevation model (d); see Supplementary Table S1 for detailed description and sources of these variables. Gray histograms represent variation across latitude or longitude.

3.2.6 Uncertainty analysis

We represented uncertainty of our modeling approach as the sensitivity of prediction models to multiple data inputs. We first explored the residual variance of our SOC training data against six SOC stocks calculated using six BD pedotransfer functions. Then, we analyzed the spatial structure of these residuals using Geostatistics, and computed a model residual error against fully independent SOC datasets. Finally, we computed the modeling SOC prediction variance and the full quantile response of residuals (from independent datasets and from the BD variance) to the highest ranked environmental covariates (Figure 3.1). We postulate that estimates of SOC fall within a range of errors, and it is therefore important to account for variation in model inputs and model outputs. Our main goal was to quantify the variability range around predicted SOC stocks using multiple uncertainty indicators.

3.2.6.1 Pedotransfer functions for bulk density variance

We predicted SOC stocks at the pedon locations available in the WoSIS system (Batjes et al. 2017). We calculated the residual variance of our predictions and independent SOC stocks. These independent stocks were calculated using the WoSIS

SOC concentration data (%), and six conventional pedotransfer functions for estimating BD. This resulted in six different SOC stocks estimates from the following pedotransfer functions:

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

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

3.2.6.2 Independent datasets for model prediction

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

3.2.7 Spatial autocorrelation of model residuals

We compared the spatial structure (i.e., spatial autocorrelation) of model residuals using linear geostatistics. The spatial structure accounts for the variance of values as a response of the geographical distance (e.g., meters) between SOC sampling points (Figure 3.2a). The spatial structure of a soil property can be quantified using variograms (a graphical method for modeling the relationship between distance between points and the variance of their values, Figure 3.2c) and the variogram parameters: nugget (uncorrelated variance), sill (spatially-autocorrelated variance) and range (distance to the maximum variance) as explained previously (Oliver and Webster, 2014). We used automated variogram fitting (Hiemstra, *et al.* 2008) for calculating the nugget:sill ratio. The nugget is an uncorrelated component of soil variation that cannot

be explained by our data, it depends on the calculation methods, the sampling resolution, and the spatial variability of SOC. The sill is the distance between the nugget and the variance stabilization (y - axis) point while increasing distance (x – axis). The range is the distance of the variance stabilization point. As in previous studies (Cruz-Cárdenas *et al*. 2014), a nugget:sill of <0.25 was considered evidence of a strong spatial dependence, a relationship between 0.25 and 0.75 was considered a moderate spatial dependence, and a relationship > 0.75 was considered a weak spatial dependence (Cambardella *et al.* 1994). We then used these variogram parameters to generate error maps by the means of Ordinary Kriging (Oliver and Webster, 2014), as explained earlier (Hengl *et al.*, 2004), accounting for the potential spatially autocorrelation of the model residuals.

3.2.8 Model residual limits

For analyzing the model-based uncertainty, we estimated the quantile conditional response of the aforementioned modeling residuals to the 'best' environmental covariates identified by our simulated nnealing framework aiming to estimate model prediction limits. The main purpose of estimating model prediction limits is to identify the variance from the most probable predicted SOC stock for each pixel across the 250m grids. For this purpose, we used the quantile regression forest approach, which is a variant of Random Forests. This method is able to: a) maintain the value of all observations in each node for each tree, not just their mean (as is the case of Random Forests) and b) assesses the quantile conditional distribution at each predicted

location (pixel). This method has the assumption that the full conditional estimated response is not different from the mean of the training dataset (Meinshausen 2006). This method allowed us to quantify the maximum possible range of SOC prediction limits (e.g., 95%) given available data and available environmental covariates.

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

3.3 Results

3.3.1 Descriptive statistics

The harmonized SOC training dataset across both countries (n = 10 385, 1991-2010, Figure 3.2a) showed a right skewed distribution and most estimates were between 0 and 10 kg m^{-2} up to a maximum of 87.9 kg m^{-2}. While the Mexican dataset dominates the 0-10 kg m^{-2} range, the CONUS dataset has larger SOC values (>10 kg m^{-2}) (Figure 3.2b). We used a logarithmic transformation (i.e., log(1+x)) of the combined (Mexico-CONUS) dataset to reduce the skewed distribution for further analysis. The combined dataset shows a nugget:sill ratio of 0.55, suggesting a moderate spatial autocorrelation of its values, a sill of 0.9 and a nugget of 0.5 (units in log(kg m^{-2} + 1), Figure 3.2c).

3.3.2 Recursive feature elimination

The five times repeated 5-fold cross validation (applied to the recursive variable elimination framework) showed errors of 1.7 (R^2 = 0.30), 2.0 (R^2= 0.27) and 2.6 (R^2=0.34) kg m^{-2} for the models 445 1991-2010 (n=10385), 1991 to 2000 (n=4877) and 2001-2010 (n=5508), respectively. For the years 1991-2010, the highest ranked environmental covariates for predicting SOC were: the DEM and topographic terrain attributes (the wetness index, the valley bottom flatness index and the valley depth index) and a remotely sensed precipitable water vapor estimate. For the years 1991-2000, the highest ranked environmental covariates were: the land surface temperature, the DEM, the wetness index, the valley bottom flatness index and the standard deviation of the EVI (surrogate of vegetation seasonality). Finally, for the years 2001-2010, the highest ranked environmental covariates were: valley bottom flatness index, mean value of the EVI (surrogate of vegetation productivity), the valley depth index, the wetness index and the night-time land surface temperature. These results showed consistency on the highest ranked environmental covariates such as the DEM and other terrain derivatives, considering the three recursive feature elimination models and years.

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

wetness index, the vegetation seasonality (standard deviation) and vegetation productivity (mean) from the EVI and mean monthly precipitation.

3.3.3 Simulated annealing

The simulated annealing framework confirmed the explanatory power of land surface temperature and precipitable water vapor, because these variables were consistently ranked as the highest environmental covariates in the three models (1991-2010, 1991-2000, 2001-2010). For the three models, the simulated annealing framework revealed that mean annual precipitation and/or the total annual precipitation were also important predictors for the SOC dataset against a cross-validation strategy (5 times repeated, 5-fold). The error estimates from this algorithm were similar compared with the previous analysis (see section 3.2), but with higher levels of explained variance for years 1991-2010 (2.2 kg m^{-2} ; R^2=0.41), years 1991-2000 (2.1 kg m^{-2} ; R^2=0.31), and years 2001-2010 (2.3 kg m^{-2} ; 473 R^2=0.46).

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

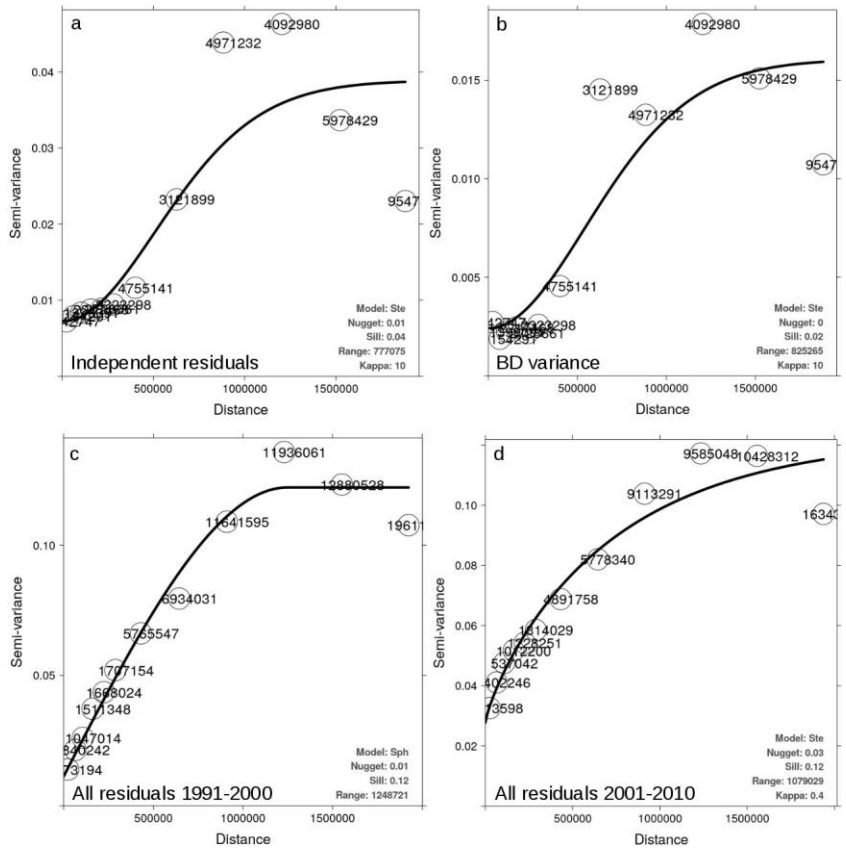

Figure 3.4. Variogram analysis applied to residuals of SOC models. The variogram of residuals against independent datasets (A). The residual variance from the different pedotransfer functions used to calculate SOC stocks (B). The combined (independent models and pedotransfer functions) residuals for the periods 1991-2000 (C) and 2001-2010 (D). The numbers in the circles indicate the available pairs of points at a given distance. Variogram parameters are shown as insets on each plot: Sph = spherical model, Ste = Stein model parameterization (and its associated Kappa value).

3.3.4 SOC residual analysis

We obtained six different SOC stocks (and mean errors) from the six pedotransfer functions used to estimate BD values (Supplementary Figure S3). The equation provided by Drew (Drew 1973: BD = 1 / (0.6268 + 0.0361 * OM)) was the best correlated function with our SOC prediction (1991-2010, *r*=0.4). The residual variance of our predictions and the multiple SOC estimates derived from different BD pedotransfer functions had a standard deviation of 3.5, a median of 1.0, and mean variance of 1.2 kg m^{-2} . The residual variance of our predictions (1991-2010) against predictions from the independent model (RaCA- Mexican Forest Service; n=9239) showed a standard deviation of 2.7, a median of 2.0, and a mean value of 2.5 kg m^{-2} . We report a moderate spatial structure (nugget:sill ratio of 0.25) for the residuals of our models and independent SOC estimates (Figure 3.4a). However, the residual variance of our predictions against the multiple SOC stocks calculated from different BD pedotransfer functions showed a strong spatial structure (nugget:sill ratio <0.1) across both countries (Figure 3.4b).

When combining the residual variance from different BD pedotransfer functions and the residuals of the independent validation, we detected a significant increase of the nugget:sill ratio from 0.08 for the years 1991-2000 (Figure 3.4c) to a nugget:sill ratio of 0.25 for the period between 2001-2010 (Figure 3.4d). Thus, there was >100% increase of uncorrelated spatial variation of SOC data (nugget:sill ratio increased from 0.08 to

0.25) from the model using 1991-2000 data to the model using 2001-2010 data. This increase of uncorrelated variation (increase of the nugget:sill ratio of >100%) was found in the combined residuals against independent data sets and against the BD pedotransfer function. These differences in the nugget:sill ratio are associated with inconsistencies in data sampling strategies and multiple collection periods of SOC data (Figure 3.4).

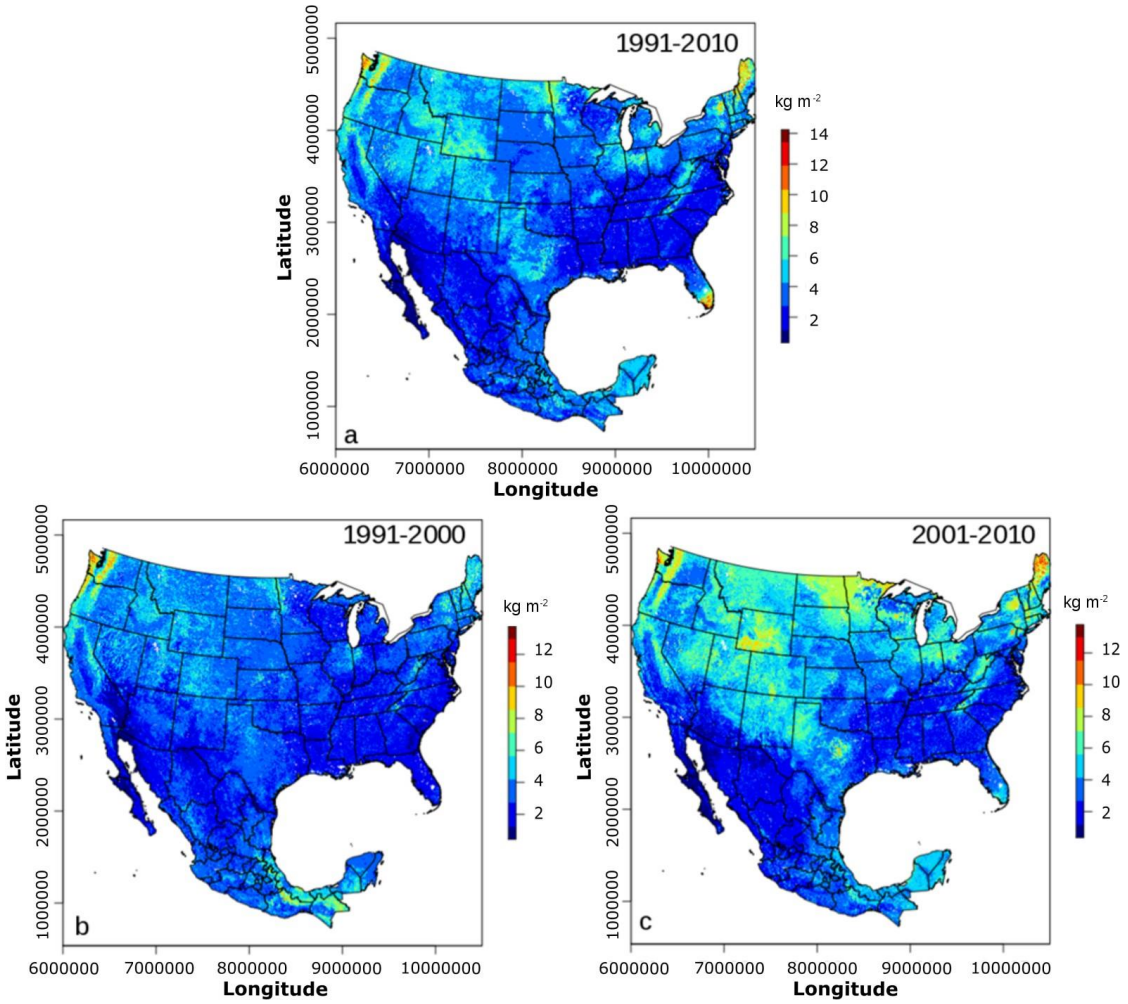

Figure 3.5. Predicted SOC across CONUS and Mexico. Prediction using data for years 1991-2010 (a); Predictions with data for years 1991-2001 (b); and Predictions with data for years 2001-2010 (c).

3.3.5 SOC stocks

We estimated a total SOC stock (1991-2010) of 47 Pg (Figure 3.5a) that varies from 41 to 55 Pg of SOC for the models 1991-2000 (Figure 3.5b) and 2001-2010 (Figure 3.5c), respectively. For the years 1991-2010, the residual error map suggested 10.4±5.1 Pg of SOC variance associated with the use of multiple pedotransfer functions for BD and consequently calculating SOC stocks. The larger variance of associated with BD was found across the surroundings of the Great Lakes, in the states of Vermont, New York and borders between Pennsylvania and Ohio, in CONUS (Figure 3.6a). The residual error map of our models against two fully independent datasets (RaCA and Mexican Forest Service), suggested a higher value of 28.8 ± 9.1 Pg of SOC variance. The large variance associated with the independent datasets was found also across the surroundings of the Great Lakes, but in the states of Wisconsin and Minnesota (Figure 3.6b). Another large variance from the independent validation was found across the state of Florida, specifically across the south section in the Everglades area where there limited observations for the training dataset (Figure 3.1).

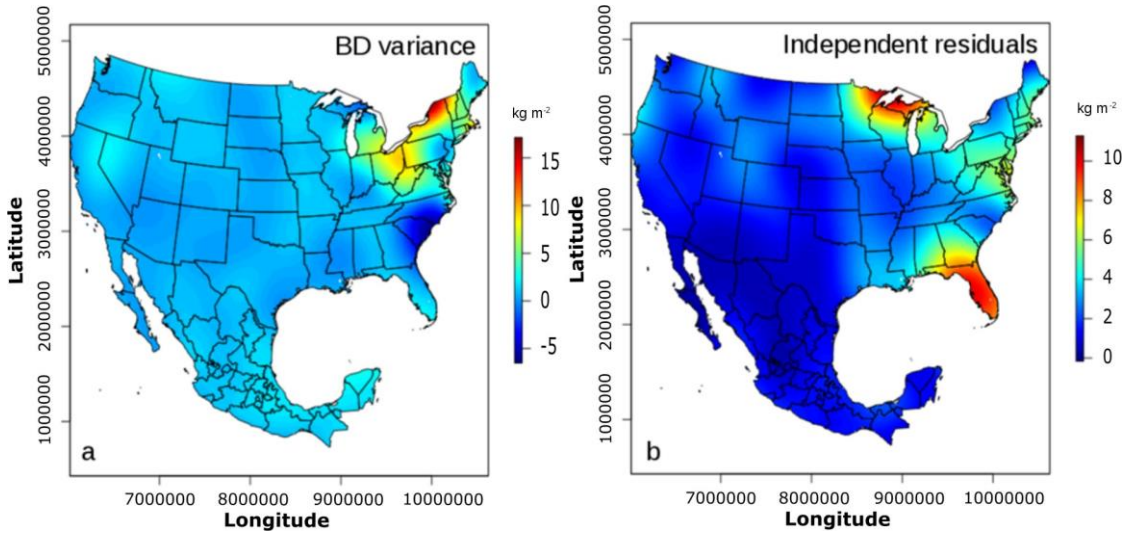

Figure 3.6. Residual error maps interpolated using Ordinary Kriging. The residual error map of pedotransfer BD residuals (a). The residual error map of our models against independent validation datasets (b).

The estimated SOC stock after applying the same modeling strategy to the external datasets was 46 Pg of SOC (combined RaCA- Mexican Forest Service, Figure 3.7a), varying ±1 Pg of SOC with the model 1991-2010 (ISCN plus INEGI, 1991-2010, Figure 3.5a). A linear model of our predictions against the independent datasets (RaCA-Mexican Forest Service) showed a mean error of 1.0 kg m^{-2} (R^2=0.43) (Supplementary Figure S4). The best correlation between SOC predictions was found between models 1991-2000 and 1991- 2010 (*r*=0.8) with ±5 Pg of difference in the predicted SOC stocks. In contrast the model for the year 2001-2010 was better correlated with the RaCA- Mexican Forest Service combined predictions (*r*=0.6), but the SOC stocks varied for ±7 Pg of SOC. The correlation between the model 1991-2010 and the

independent analysis was lower (r=0.3) but the SOC stocks showed less variation (±1 Pg of SOC).

The variance among all models based on INEGI and ISCN data was ±7.6 Pg of SOC (Figure 3.7b). This value increased up to ±12 Pg of SOC by adding the variance of the models based on the independent analysis (Figure 3.7c). Thus, we provide a SOC stock for both countries between 46 and 47 Pg of SOC with a total modeling variance of ±12 Pg.

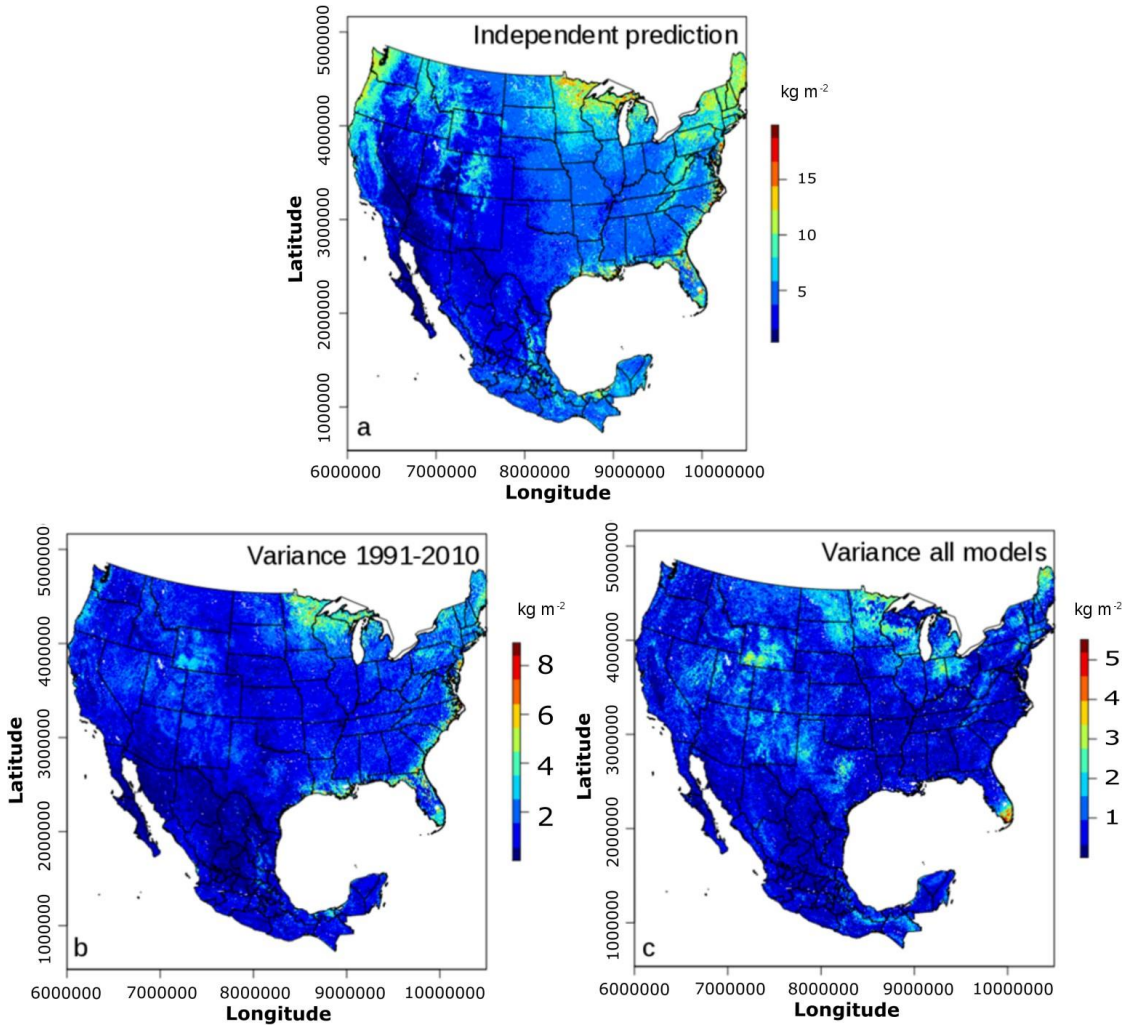

Figure 3.7. Prediction of SOC generated using the independent datasets (a). Model variance for predictions 1991-2000 and 2001-2010 using the INEGI and ISCN available data (b). Variance of all SOC predictions (INEGI-ISCN, RaCA-Mexican Forest Service datasets) (c).

3.3.6 Quantile conditional distribution of residuals

The quantile conditional distribution (used to identify the model prediction limits) for the residuals against fully independent datasets suggest a maximum possible SOC variance of ±73 Pg of SOC. This is the SOC variance from the full quantile conditional response of these residuals to the highest ranked environmental covariates. From the BD pedotransfer function variance, the full conditional response to the highest ranked environmental covariates showed a lower value of ±20 Pg of SOC. Thus, less uncertainty was found from the use of multiple pedotransfer functions than from validating against fully independent datasets. These results highlight the large variance of possible SOC predictions given the use of multiple training data sources constrained to a relatively short (i.e., two decade) period of time. The model-based uncertainty results are shown in Figure 3.8. For the BD pedotransfer function variance, the larger range of model based uncertainty was found across the Great Lakes of northern CONUS and the border with Canada (Figure 3.8A), The model-based uncertainty using independent residuals shows the largest values across, Florida, the east coast and the surroundings of the Great Lakes in CONUS, as well as some areas of southeast Mexico (Figure 3.8B).

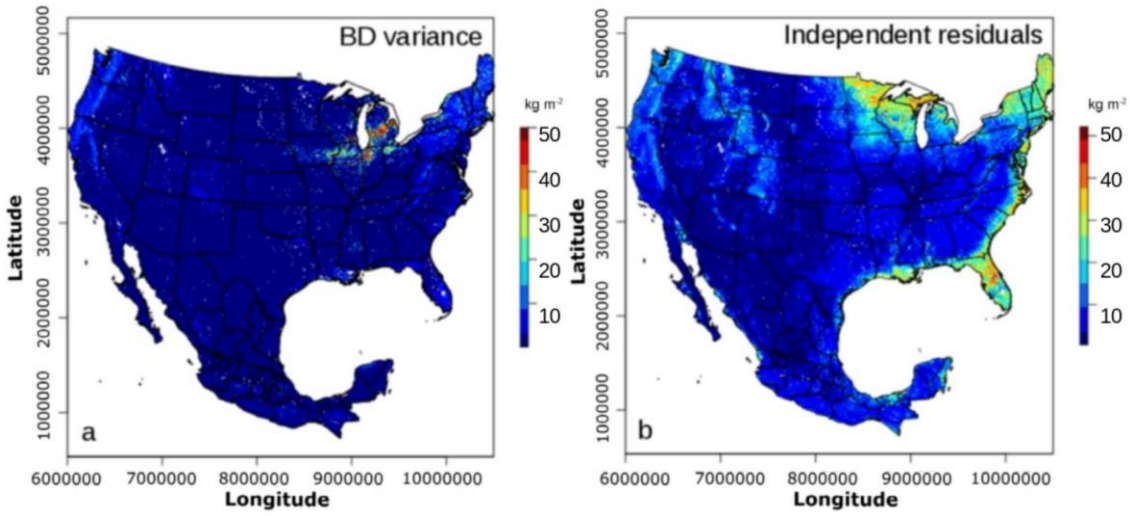

Figure 3.8. Conditional quantile distribution of SOC residuals to the highest ranked environmental covariates from the BD residual variance (a) and for the residual variance against models generated with fully independent datasets (b).

3.4 Discussion

We predicted SOC (across Mexico and CONUS at 0-30 cm of soil depth) and generated gridded estimates at 250m spatial resolution for the period 1991-2010. We provided predictions of SOC using multiple inputs of data and a feature selection framework (recursive elimination of predictors and the simulated annealing algorithms) that allowed us to identify the most informative SOC environmental covariates to determine SOC stocks between 1991-2010. We calculated SOC stocks for both countries (46-47 ±12 Pg) and these values were >30% below previous global estimates such as the SoilGrids system or the Harmonized World Soil Database. We have highlighted large discrepancies between modeling outputs based on multiple data collection periods (Figure 3.5) and between global SOC products such as the state of the art SoilGrids250m (Supplementary Figure S5). Furthermore, our results have implications for the use and interpretation of SOC legacy data or aggregated SOC information. Specifically, we found a large difference for predicted SOC stocks (from 41 to 55 Pg of SOC) between 1991-2000 and 2001-2010 that cannot be fully attributed to SOC dynamics, but also to inconsistencies in the spatial configuration of available datasets, the use of different SOC calculation methods, and the different periods of data collection. These results open new research questions about the interpretation of apparent changes in SOC stocks across time and future studies should determine: a) if regional-to-global differences are due to active management practices/land use change, or b) if apparent changes are overshadow by large uncertainty estimates due to inconsistencies in methods and modelling variability.

3.4.1 Highest ranked environmental predictors

Our results suggest that using a few informative environmental predictors (e.g., the DEM and terrain derivatives, the EVI or precipitation patterns) for the spatial variability of SOC have a similar performance (~50% of explained variance) to a high-dimensional covariate space (the 150 environmental covariates reported in Supplementary Table S1) across Mexico and CONUS and using the available datasets between 1991-2010. The use of a high-dimensional covariate space to predict SOC (and other soil properties) may be needed to maximize prediction accuracy at the global scale (Hengl *et al.* 2017); however, for local to regional applications some environmental covariates for SOC may be statistically redundant and lead to unnecessarily increases of the computing resources required for prediction purposes. Reducing the statistical redundancy of environmental covariates in SOC models will simplify computing requirements and model complexity (Guo *et al.* 2019. Reducing the complexity of SOC models would be appealing in further applications of SOC spatial information (e.g., land carbon uptake modeling, climate system modeling, niche modeling) which require similar predictors as for SOC.

Our simulated annealing analysis highlights SOC relationships (positive and negative) with climate variables (precipitation and land surface temperature), elevation (and terrain derivatives) and vegetation greenness (productivity and seasonality) that are consistent with previous literature describing SOC drivers across diverse environmental conditions (Hobley, *et al*. 2015, Evans et al., 2011). In addition, when applying the simulated annealing framework to the independent datasets, a MODIS surface

reflectance short wave infrared band (M06MOD4, Supplementary Table S1) was one of the first five important variables predicting SOC, which is consistent with the infrared based methods used by the USDA for developing of the RaCA dataset (Wijewardane *et al.* 2016). The prediction capability of the infrared spectra (e.g., near infrared, mid infrared) for SOC can be attributed to the strong spectral absorption characteristics of soil organic matter and BD (the main components of SOC) in the infrared spectral bands (Guo *et al*. 2019). Thus, the main relationships driving our SOC predictions can be interpreted and associated with the use of different data inputs and specific environmental covariates (e.g., the DEM and terrain parameters, the EVI and the MODIS surface reflectance infrared data , precipitation and temperature gridded surfaces) that can be periodically acquired from remote sensing at the global scale.

3.4.2 Uncertainty quantification

3.4.2.1 BD pedotransfer functions

We report ~10Pg of SOC variance associated with SOC calculation inputs. The combined and quality-controlled dataset we used have been processed following international standards for increasing precision and accuracy (Harden ***et al.*** 2017, Batjes *et al*. 2017)**.** However, the major limitation of these datasets is arguably the low availability of BD and CF data. Inconsistencies in BD and CF data could explain the large variance found (Figure 3.6A) across the highly productive landscapes of the north east of CONUS. It has been discussed that SOC stocks are systematically overestimated by misuse of the BD and CF content parameters (Poeplau *et al.*, 2017), although for

some areas, significant underestimations of SOC have been reported (Chen *et al*. 2018). Thus, correction of BD is fundamental to achieve realistic SOC estimates and to reduce the potential overestimation of SOC stocks (Köchy *et al.* 2015). The lack of accurate BD and CF data and the large variance of the global SOC values are key issues that could explain the discrepancy between country-to-global SOC estimates (Tifafi et al. 2018). Thus, our results provide a spatially explicit measure of SOC variance derived from six conventional BD pedotransfer functions that can be used to explain discrepancies between national, regional and global SOC estimates.

3.4.2.2 Spatial and temporal variations of available data

We found differences in the spatial structure (i.e., autocorrelation) of modeling residuals from multiple models and periods of time (1991-2000 and 2001-2010), that result in large differences on predicted stocks (from 41 to 55 Pg of SOC) from these defined periods of time. This period of time (1991-2010) have experienced intensive land use and environmental changes across both countries and our results could be used for identifying sensitive areas of SOC changes or areas that require further research (Figure 3.5). However, a previous study suggested that SOC could increase under reforestation conditions around 2 Pg per century in the topsoil (Nave et al. 2018), so our "decadal" modeling results may be overestimating the SOC gain between those time-periods. Here, we discuss our results under this consideration.

While we detected a reduction of SOC across most of Mexico, we detected a larger gain of SOC mainly across higher latitudes of CONUS (when comparing models

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

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

3.4.2.3 Quantile response of residual variance

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

that require higher sampling effort due larger discrepancies of multiple SOC model predictions.

3.4.3 SOC stocks across CONUS and Mexico

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

We report a density of SOC across CONUS (4.98 kg m^{-2}) that was relatively higher than the soils of Mexico (4.22 kg m^{-2}). Globally, the soil carbon pool (at 1m

depth) is estimated to have around 1500-2400 Pg (Sato *et al*. 2015), while the SOC pool in the upper 30 cm is estimated to be 755 ± 119 Pg (Batjes, 2016). Our results suggest that Mexico represents ~1 % and CONUS ~5 % of the global SOC pool at 30 cm depth. Recent revisions highlight that the SOC stock at 30 cm soil depth remains unclear (Lajtha *et al.* 2018), and this study provides new insights for interpreting the discrepancies around the topsoil SOC pool across CONUS and Mexico.

Our SOC estimates across forested areas are comparable to those reported on studies (Domke *et al*. 2017, Bolaños *et al.* 2017). However, our results show high uncertainty across areas dominated with high SOC values (e.g., >1 gr cm^{-2}, some northern and tropical forests, peatlands, other black soils dominated areas) and across higher latitudes (Figure 3.6), as documented in previous studies (Tian, *et al.* 2015). Unfortunately, these areas are poorly represented in the available datasets (<10% of available data with values >1 gr cm^{-2}) and we encourage future monitoring efforts to increase their representativeness.

3.4.3.1 SOC across land cover classes

Our study confirms the presence of important SOC stocks across both forest and agricultural soils. Across both countries, we found higher SOC stocks in croplands, representing 26% of total SOC within the upper 30cm of soil. Across Mexico, we found that 42% of SOC was stored in forest soils and 24% in agricultural soils, while 31% of SOC across CONUS is stored in forest soils and 27% in agricultural soils. While organic matter-rich and deep soils dominate most agricultural areas across CONUS

(Adhikari & Hartemink, 2017), most agricultural soils in Mexico tend to be shallow (Guerrero *et al.* 2014, ~30cm depth); consequently, we emphasize that carbon management, monitoring and conservation strategies must be developed from a country-specific approach considering country-specific land cover classes.

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

Accurate SOC estimates represent a key variable to quantify human induced disturbances to the carbon cycle across land cover classes. We report that temperate forests of CONUS contain the larger SOC reserves while tropical or sub-tropical broadleaf evergreen forest and wetlands are the land cover classes with higher SOC in Mexico than CONUS (Supplementary Table S2). Respectively, the tropical or sub-tropical broadleaf evergreen forests are the most productive ecosystems of Mexico

(Murray-Tortarolo et al. 2016). The wetlands category, with high carbon sequestration potential, includes mangroves (and other coastal wetlands), which have been recognized as the ecosystems with higher carbon storage capacity from the site-specific to the global scales (Vázquez-Lule, et al., 2019, Adame *et al.* 2015, Atwood *et al.*, 2017). Our results represent benchmarks for SOC monitoring across these land cover classes. Thus, the spatial predictions of SOC at 250m allows for the interpretation of SOC spatial patterns across land cover classes of Mexico and CONUS accounting for sensitivities associated with the use of multiple data inputs.

3.4.4 Final remarks

Optimizing future SOC sampling strategies while reducing modeling variance and increasing model agreement against model independent datasets collected under different circumstances (e.g., logistics, design, main purpose, SOC estimation methods) are large challenges for enabling SOC carbon mapping and monitoring systems. New and better SOC parameters are required for reducing the current discrepancy between multiple sources of SOC data (Tifafi *et al.* 2017, Guevara *et al.* 2018) and enabling SOC monitoring systems (Viscarra-Rossel *et al. 2014*). The lack of accurate SOC spatial information and the combination of multiple SOC data sources could result in large variance estimates across the two countries (e.g., red areas of Figure 3.6). We propose that areas with high variance suggest that these regions require higher SOC sampling efforts.

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

Consequently, this discrepancy can influence the estimates of SOC warming response (Karhu *et al.*,2010) and the carbon–climate feedback that could accelerate climate change (Crowther *et al.*, 2016). We hope that this study motivates an increase in country-specific soil surveys, data sharing, and modeling of SOC estimates at higher spatial resolution with a better quantification of uncertainty.

Acknowledgements

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

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

REFERENCES

Adame MF, Santini NS, Tovilla C, Vázquez-Lule A, Castro L (2015) Carbon stocks and soil sequestration rates of riverine mangroves and freshwater wetlands. *Biogeosciences Discussions*, **12**, 817 1015–1045.

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

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

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

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

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

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

Atwood, Trisha B., et al. "Global patterns in mangrove soil carbon stocks and losses." Nat. Clim. 839Change, vol. 7, no. 7, 26 June 2017, p. 523, doi:10.1038/nclimate3326.

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

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

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

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

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

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

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

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

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

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

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

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

Crowther TW, Todd-Brown KEO, Rowe CW et al. (2016) Quantifying global soil carbon losses in 881 response to warming. *Nature*, **540**, 104–108.

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

Cruz-Gaistardo, C. and Paz-Pellat, F.: Mapa de carbono orgánico de los suelos de la República Mexicana, in: EstadoActual del Conocimiento del Ciclo del Carbono y sus Interacciones en México: Síntesis a 2013, edited by: Paz-Pellat, F., Wong-González, J., Bazan, M., and Saynes, V., Programa 889 Mexicano del Carbono. Texcoco, Estado de México, México, ISBN 978-607-96490-1-2, 702, 2014.

Delgado-Baquerizo M, Eldridge DJ, Maestre FT, Karunaratne SB, Trivedi P, Reich PB, Singh BK (2017) Climate legacies drive global soil carbon stocks in terrestrial ecosystems. *Science Advances*, **3**, e1602008.de Gruijter JJ, McBratney AB, Minasny B, Wheeler I, Malone BP, Stockmann U (2016) Farm-scale soil carbon auditing. *Geoderma*, **265**, 120–130.

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

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

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

FAO 2017. Soil Organic Carbon: the hidden potential. Food and Agriculture Organization of the United Nations, Rome, Italy FAO Guidelines for soil description Fourth edition, Rome 2006 109 pp. FAO and ITPS. 2018. Global Soil Organic Carbon Map (GSOCmap) Technical Report. Rome. 162 pp.

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

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

Finke PA (2012) On digital soil assessment with models and the Pedometrics agenda. *Geoderma*, **171**– 927 **172**, 3–15.

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

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

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

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

Guevara, M., Olmedo, G. F., Stell, E., Yigini, Y., Aguilar Duarte, Y., Arellano Hernández, C., Arévalo, G. E., Arroyo-Cruz, C. E., Bolivar, A., Bunning, S., Bustamante Cañas, N., Cruz-Gaistardo, C. O., Davila, F., Dell Acqua, M., Encina, A., Figueredo Tacona, H., Fontes, F., Hernández Herrera, J. A., Ibelles Navarro, A. R., Loayza, V., Manueles, A. M., Mendoza Jara, F., Olivera, C., Osorio Hermosilla, R., Pereira, G., Prieto, P., Ramos, I. A., Rey Brina, J. C., Rivera, R., Rodríguez-Rodríguez, J., Roopnarine, R., Rosales Ibarra, A., Rosales Riveiro, K. A., Schulz, G. A., Spence, A., Vasques, G. M., Vargas, R. R., and Vargas, R.: No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America, SOIL, 4, 173-193, https://doi.org/10.5194/soil-4-173-2018, 951 2018.

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

Guo, Z., Adhikari, K., Chellasamy, M., Greve, M. B., Owens, P. R., & Greve, M. H. (2019). Selection of terrain attributes and its scale dependency on soil organic carbon prediction. Geoderma, 340, 303– 959312. doi: 10.1016/j.geoderma.2019.01.023 960

Harden JW, Hugelius G, Ahlström A et al. (2017) Networking our science to characterize the state, 962 vulnerabilities, and management opportunities of soil organic matter. *Global Change Biology*.

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

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

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

Hengl T, Mendes J, Heuvelink GBM, et al. (2017) SoilGrids250m: global gridded soil information based on Machine Learning. PLoS ONE 12(2): e0169748.https://doi.org/10.1371/journal.pone.0169748 975

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

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

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

Hobley, E., Wilson, B., Wilkie, A., Gray, J., & Koen, T. (2015). Drivers of soil organic carbon storage and vertical distribution in Eastern Australia. Plant Soil, 390(1-2), 111–127. doi: 10.1007/s11104-015- 988 2380-1

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

McBratney A, Field DJ, Koch A (2014) The dimensions of soil security. *Geoderma*, **213**, 203–213. Meinshausen N (2006) Quantile Regression Forests. *J. Mach. Learn. Res.*, **7**, 983–999.

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

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

Minasny B, Malone BP, McBratney AB et al. (2017) Soil carbon 4 per mille. *Geoderma*, **292**, 59–86. Murray-Tortarolo G, Friedlingstein P, Sitch S et al. (2016) The carbon cycle in Mexico: past, present and future of C stocks and fluxes. *Biogeosciences*, **13**, 223–238.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Vargas R, Alcaraz-Segura D, Birdsey R et al. (2017) Enhancing interoperability to facilitate implementation of REDD+: case study of Mexico. *Carbon Management*, **8**, 57–65.

Vázquez-Selem L, Heine K (2004) Late Quaternary glaciation of México. In: *Developments in Quaternary Sciences*, Vol. 2 (eds Ehlers J, Gibbard PL), pp. 233–242. Elsevier.

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

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

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

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

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

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

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

Supplementary Tables:

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

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

Supplementary Figures:

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

Supplementary material in:

https://github.com/marioguevara/utilityCodes/tree/master/SOC_reports_ch_4

Chapter 4

DOWNSCALING SATELLITE SOIL MOISTURE USING GEOMORPHOM AND MACHINE LEARNING

Authors:

Mario Guevara1 & Rodrigo Vargas1

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

Abstract

Annual soil moisture estimates are useful to characterize trends in the climate system, in the capacity of soils to retain water and for predicting land and atmosphere interactions. The main source of soil moisture spatial information across large areas (e.g., continents) is satellite-based microwave remote sensing. However, satellite soil moisture datasets have coarse spatial resolution (e.g., 25–50 km grids); and large areas from regional-to-global scales have spatial information gaps. We provide an alternative approach to predict soil moisture spatial patterns (and associated uncertainty) with higher spatial resolution across areas where no information is otherwise available. This approach relies on geomorphometry derived terrain parameters and machine learning models to improve the statistical accuracy and the spatial resolution (from 27km to 1km grids) of satellite soil moisture information across the conterminous United States on an annual basis (1991–2016). We derived 15 primary and secondary terrain parameters from

a digital elevation model. We trained a machine learning algorithm (i.e., kernel weighted nearest neighbors) for each year. Terrain parameters were used as predictors and annual satellite soil moisture estimates were used to train the models. The explained variance for all models-years was >70% (10-fold cross- validation). The 1km soil moisture grids (compared to the original satellite soil moisture estimates) had higher correlations (improving from $r^2 = 0.1$ to $r^2 = 0.46$) and lower bias (improving from 0.062 to 0.057 m3/m3) with field soil moisture observations from the North American Soil Moisture Database (n = 668 locations with available data between 1991– 2013; 0-5cm depth). We conclude that the fusion of geomorphometry methods and satellite soil moisture estimates is useful to increase the spatial resolution and accuracy of satellite- derived soil moisture. This approach can be applied to other satellite-derived soil moisture estimates and regions across the world.

4.1 Introduction

Continuous national to continental scale soil moisture information is increasingly needed to characterize spatial and temporal trends of terrestrial productivity patterns (e.g., production of food, fiber and energy). This is because soil moisture is a key variable regulating hydrological and biogeochemical cycles, and thus studying its spatial-temporal dynamics is crucial for assessing the potential impact of climate change on water resources [1-4]. Currently, the most feasible way to obtain national to continental soil moisture information is using remote sensing. Microwave remote sensing devices deployed on multiple earth observation satellites are able to quantify the dielectric constant of soil surface and retrieve soil

moisture estimates [5]. However, there are spatial gaps of satellite-based soil moisture information and its current spatial resolution (> 1km grids) limits its applicability at the ecosystem-to-landscape scales to address the ecological implications of soil moisture dynamics [5-8].

Satellite soil moisture records are an effective indicator for monitoring global soil conditions and forecasting climate impacts on terrestrial ecosystems, because soil moisture estimates are required for assessing feedbacks between water and biogeochemical cycles [9-12]. In addition, accurate soil moisture information is critical to predict terrestrial and atmospheric inter- actions such as water evapotranspiration or CO_2 emissions from soils [3, 13-15]. However, soil moisture information at spatial resolution of 1x1km pixels or less is not yet available across large areas of the world and the coarse pixel size (>1km pixels) of available satellite soil moisture records is limited for spatial analysis (i.e., hydrological, ecological) at small regional levels (e.g., county- to state). In addition, satellite soil moisture estimates are representative only of the first few 0–5 to 10 cm of top-soil surface [16]. Therefore, comparing multiple sources for satellite soil moisture and field soil moisture estimates is constantly required for precise interpretations of soil moisture spatial patterns [17-19].

There is an opportunity for exploring statistical relationships across different sources of remote sensing information (e.g., topography and soil moisture) and developing alternative soil moisture spatial datasets (i.e., grids) to improve the continental-to-global spatial resolution of current satellite soil

moisture estimates [7]. Spatially explicit soil moisture estimates can be obtained across large areas with a coarse spatial resolution (between 25–50 km grids) from radar-based microwave platforms deployed across different satellite soil moisture missions [20-21]. The availability of historical soil moisture records of these sources has increased during the last decade with unprecedented levels of temporal resolution (i.e., daily from years 1978-present) at the global scale. However, large areas constantly covered by snow, extremely dry regions or tropical rain forests (where there is a higher content of water above ground) lack of precise soil moisture satellite records due to sensor intrinsic limitations (e.g., saturation or noise) across these environmental conditions [22].

One valuable product that is affected by the aforementioned environmental conditions is the ESA-CCI (European Space Agency Climate Change Initiative) soil moisture product [20- 21]. The ESA-CCI mission makes rapidly available long-term soil moisture estimates with daily temporal resolution from the 1978s to date, and it represents the state-of-the-art knowledge tool for assessing long term trends in the climate system. Modeling, validation and calibration frameworks are required for improving the spatial representation of this important dataset, and for predicting soil moisture patterns across areas where no satellite estimates are available.

Currently, there is an increasing availability of fine-gridded information sources and modeling approaches that could be used for increasing the spatial resolution (hereinafter downscaling) of the ESA-CCI satellite soil moisture

estimates (e.g., soil moisture predictions across <1x1km grids). Downscaling (and subsequently gap-filling) satellite soil moisture estimates has been the objective of empirical modeling approaches based on sub-grids of soil moisture related information such as soil texture [23]. Other approaches followed environ- mental correlation methods and generated soil moisture predictions for satellite soil moisture estimates using both data-driven or hypothesis driven models and multiple sub-grids of ancillary information [24-26]. These sub-grids of information usually include vegetation related optical remote sensing imagery, gridded soil information, land cover classes and landforms [27-30]. Most of these approaches have been tested for specific study sites. Other studies have focused on applying a digital soil mapping approach (a reference framework for understanding the spatial distribution of soil variability [31]) and multiple upscaling methods for predicting soil moisture patterns at the continental scale [26, 32]. An overview of multiple approaches for downscaling satellite soil moisture (e.g., empirically based, physically based) has been previously discussed [33]. Here, we propose that digital terrain analysis (i.e., geomorphometry) can also be applied for empirically downscaling soil moisture satellite-based information across continental-to-global spatial scales.

Geomorphometry is an emergent discipline in earth sciences dedicated to the quantitative analysis of land surface characteristics and topography [34-35]. For geomorphometry, the analysis of topography includes the generation of a diversity of hydrologically meaningful terrain parameters (i.e., slope, aspect, curvature,

valley depth index, topographic wetness index) that aim to represent how the landscape physically constrains water inputs (e.g., rainwater, irrigation, overland flow) that reaches the soil surface [35-36]. These terrain parameters are referred as "digital" because they are usually derived from digital elevation models using Geographic Information Systems (GIS). These digital terrain parameters are hydrologically meaningful because at the landscape scale, soil moisture is partially controlled by topography- related factors (i.e., slope, aspect, curvature) that physically constrain soil water inputs and soil hydraulic properties (e.g., soil texture, structure). Based on these geomorphometry principles [35-39], we propose that it is possible to determine which terrain parameters are the strongest predictors of the spatial variability of satellite soil moisture. Statistically coupling the spatial variability of satellite soil moisture with hydrologically meaningful terrain parameters could be an alternative way to improve the spatial resolution and accuracy of satellite soil moisture esti- mates across scales. This is possible because topography (represented by digital terrain parameters) directly affects: 1) the angle of the satellite microwave signal at the soil surface; and 2) the overall distribution of water in the landscape.

Topography is a major driver for soil moisture and topography surrogates (e.g., land form or elevation map) have been combined with other variables (e.g., climate, soils, vegetation and land use) for downscaling satellite soil moisture estimates [33]. Furthermore, previous studies have shown that realistic soil moisture patterns can be obtained using topographic information across site

specific and catchment scales including physically based and empirical approaches [40-41], However, the exclusive use of geomorphometry derived products (i.e., digital terrain parameters) for downscaling satellite soil moisture estimates has not yet been explored in detail from national-to-continental scales. This approach is relevant to avoid statistical redundancies and potential spurious correlations when downscaled soil moisture is further used or analyzed with vegetation- or climate-related variables (when these aforementioned variables were used for downscaling of satellite derived soil moisture). In this study, we show the potential of a soil moisture prediction framework purely based on digital terrain parameters.

Our main objective is to generate a soil moisture prediction framework by coupling satellite soil moisture estimates with digital terrain parameters as prediction factors. Coupling the complexity of topographic gradients and the multi-temporal nature of satellite soil moisture requires an approach that should account for non-linear relationships. Machine learning approaches could account for non-linearity based on probability and the ability of computer systems to reproduce and 'learn' (i.e., decide the best solution after multiple model realizations) from multiple modeling outputs (i.e., varying model parameters of combinations of training and testing random samples) [42]. Furthermore, machine learning is now a common component of geoscientific research leading the discovery of new knowledge in the earth system [43] including mapping of soil organic carbon [44], soil greenhouse gas fluxes [8, 45] and soil moisture estimates [46].

We postulate that the data fusion between satellite soil moisture with hydrologically meaningful terrain parameters can enhance the spatial resolution, representativeness and quality (i.e., accuracy) of current coarse satellite soil moisture grids. We focus on the conterminous United States (CONUS) given the large availability of soil moisture records for validation purposes from the North American Soil Moisture Database (NASMD) [47]. The novelty of this research relies on proposing an alternative approach for obtaining soil moisture gridded estimates with no gaps and at high spatial resolution (i.e., 1km) determined by topographic pre- diction factors. This study is based on public sources of satellite information (derived from ESA-CCI soil moisture product) and a data-driven framework that could be reproduced and applied across the world.

4.2 Materials and methods

Our downscaling approach relied on a Digital Elevation Model (DEM), and satellite soil moisture records. Soil moisture information was acquired from the ESA-CCI [20-21]. The development and reliability (i.e., validation) of this remote sensing soil moisture product has been documented by previous studies [20-21, 48]. Our framework includes prediction factors for soil moisture from digital terrain analysis. These terrain predictors were derived across CONUS using 1km grids. Machine learning was used for generating soil moisture predictions (annual, 1991–2016) and the satellite soil moisture estimates provided by the ESA-CCI were used as training data. Field soil moisture observations from the North American Soil

Moisture dataset were used for validating the soil moisture predictions based on digital terrain analysis (Figure 4.1).

4.2.1 Datasets and data preparation

The downscaled dataset were obtained from the ESA-CCI satellite soil moisture estimates between 1991 and 2016, and the validation dataset were field soil moisture measurements from the NASMD (S1 Fig). The downscaling framework is explained in the following sections. The ESA-CCI soil moisture product has a daily temporal coverage from 1978 to 2016 and a spatial resolution of ~27 km (S2 Fig). Among several remotely sensed soil moisture products [16, 49-53], we decided to use the ESA-CCI soil moisture product because it covers a larger period of time compared with other satellite soil moisture products (e.g., NASA SMAP). We highlight that satellite soil moisture information is used for training a machine learning model for each year, and independent field soil moisture records area only used for validating the downscaled soil moisture predictions.

For externally validating, we used the NASMD because it has been curated following a strict quality control calibrated for CONUS [44] (S1 Fig). This data collection effort consists of a harmonized and quality-controlled soil moisture dataset with contributions from over 2000 meteorological stations across CONUS described in previous studies [47]. The NASMD also includes records of soil moisture registered in the International Soil Moisture Network (ISMN) [47, 54]. The NASMD provides processed data from each station location in each net- work following a standardization framework focused in North America [47-19]. We used

soil moisture records at 5 cm of depth (n = 5541 daily measurements) from 668 stations with avail- able (from the aforementioned sources) soil moisture estimates at this depth because radar- based soil moisture estimates are representative for these first few centimeters of topsoil sur- face [16].

As prediction factors for soil moisture, we calculated hydrologically meaningful terrain parameters for CONUS (S1 Table) using information from a radar-based DEM [55-56]. These terrain parameters are quantitative spatial grids representing the topographic variability that directly influence the water distribution across the landscape [35], which supports the physical link between soil moisture and topography. These parameters were the basis for downscaling satellite soil moisture records to 1km grids. This spatial resolution captures the major variability of topographic features across CONUS and is commonly used on large-scale ecosystem studies and soil mapping efforts [56-57].

For the calculation of soil moisture prediction factors, we used automated digital terrain analysis using the System for Automated Geographical Analysis-Geographical Information System (SAGA-GIS) [36]. The automated implementation of SAGA-GIS for Geomorphometry (module for basic terrain analysis) includes a preprocessing stage to remove spurious sinks and reduce the presence of other artifacts in the elevation gridded surface (e.g., false pikes or flat areas). After preprocessing the DEM, fifteen hydrologically meaningful terrain parameters were generated for the CONUS from elevation data including primary (i.e., slope, aspect) and secondary parameters (i.e., cross-sectional curvature,

longitudinal curvature, analytical hill- shading, convergence index, closed depressions, catchment area, topographic wetness index, length-slope factor, channel network base level, vertical distance to channel network, and valley depth index; Figure 4.2).

Primary terrain parameters are direct descriptions of elevation data, for example slope and aspect, which are respectively the first and second derivative of elevation data. Secondary parameters are generated by combining a primary terrain parameter with a mathematical formulation to describe process controlled by topography that are directly related to soil moisture variability [34-35, 37], such as the overland flow accumulation. The topographic wetness index for example, indicates areas where water tend to accumulate by effect of topography and is a secondary index derived by the combination upslope area draining through a certain point per unit contour length and the local terrain slope [34-35]. The values of these terrain parameters (S1 Table) were extracted for each one of the ESA-CCI soil moisture pixels using as reference the central coordinates of each ESA-CCI pixel (Figure 4.1*inputs*).

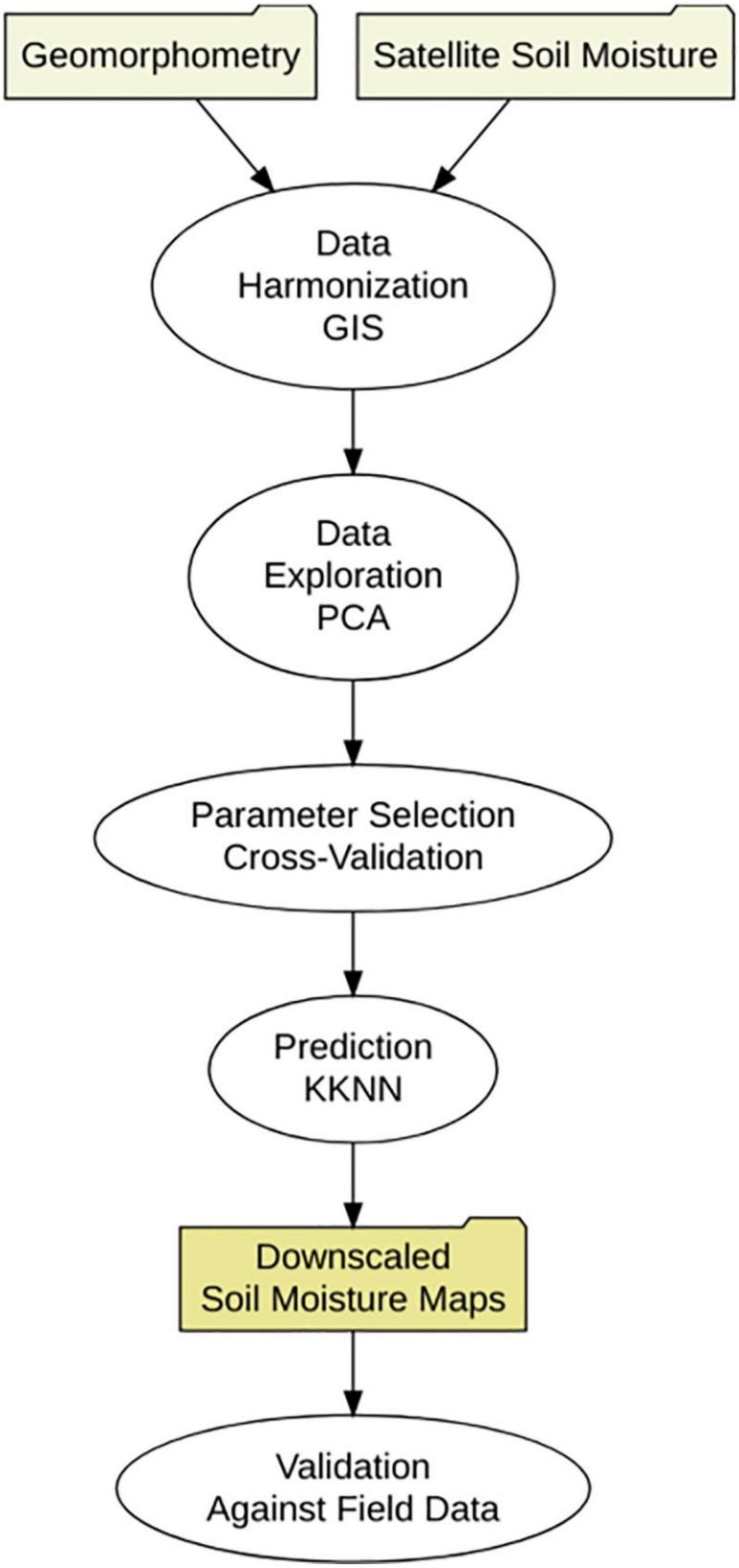
Geomorphometry
Satellite Soil Moisture
Data
Harmonization
GIS
Data
Exploration
PCA
Parameter Selection
Cross-Validation
Prediction
KKNN
Downscaled
Soil Moisture Maps
Validation
Against Field Data

Figure 4.1. Soil moisture prediction framework. The folders are the inputs and outputs and the ovals are methods for data preparation (data bases harmonization), modeling (for prediction) and validation (for assessing the reliability of soil moisture maps). The field data from the North American Soil Moisture Database (NASMD) was only used for validation purposes (i.e., not for training the model).

4.2.2 Data exploration

We used a principal component analysis (PCA) prior to modeling for data exploration and description of general relationships between soil moisture values and topography (represented by the aforementioned terrain parameters). The purpose was to simplify the dimensionality of the data set to identify the main relationships (between soil moisture and topographic parameters) driving our downscaling framework (Figure 4.1*methods*). The PCA was implemented as in previous work [58], based on a reference value representing the 0.95-quantile of the variability obtained by randomly simulating 300 data tables of equivalent size on the basis of a normal distribution. This analysis was applied to the terrain parameters at the locations of the field stations in order to compare the relationship of the first PCA and the values of soil moisture from the ESA-CCI grids and from the field data.

4.2.3 Model building

For this study we built a model for each annual mean of satellite soil moisture grids. We used a machine learning kernel-based model (kernel weighted nearest neighbors, kknn) [59-60] to generate predictions and downscale satellite

soil moisture grids (Figure 4.1$methods$). The training dataset for each model/year were the annual mean values of the ESA-CCI soil moisture product. The kknn model has two main model parameters: the optimum number of neighbors (k) and the optimal kernel function (okf). First, we defined k, which is the number of neighbors to be considered for the prediction. Second, we selected the okf, which is a reference (e.g., triangular, epanechnikov, Gaussian, optimal) for the probability density function of the variable to be predicted. The okf is used to convert distances (i.e., Minkowski distance) into weights used to calculate the k-weighted average. These kknn model parameters (k and okf) were selected by the means of 10-fold cross validation as previously recommended [61]. Cross-validation is a well-known re-sampling technique that divides data into 10 roughly equal subsets. For every possible parameter value (e.g., k from 1 to 50 and okf [triangular, epanechnikov, Gaussian, optimal]), 10 different models are generated, each using 90% of the data then being evaluated on the remaining 10%. To predict soil moisture information at 1 km of spatial resolution for each year (between 1991 and 2016), we selected the combination of optimal k and okf that lead to the highest correlation (between observed and predicted data) with the lowest root mean squared error (RMSE) after the cross-validation strategy. Thus, for each year we were able to predict soil moisture across 1x1 km grids (Figure 4.1$outputs$).

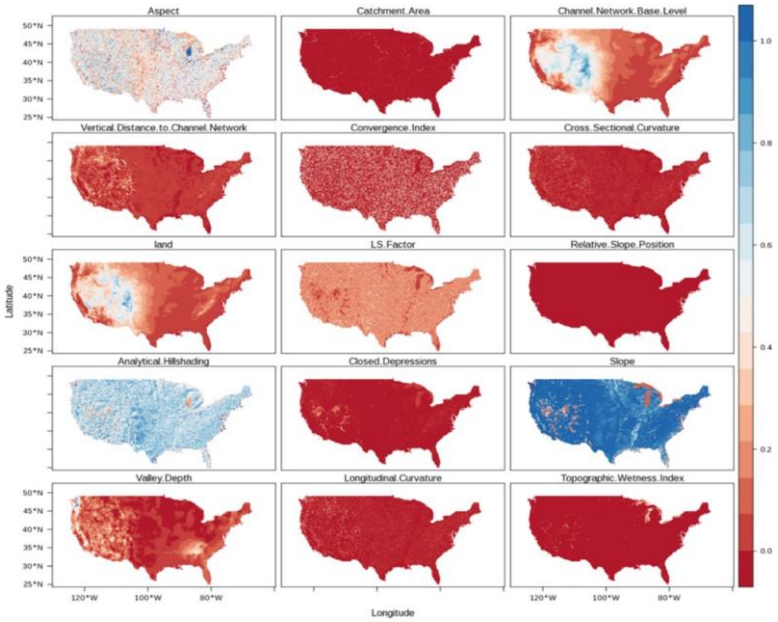

Figure 4.2. Elevation and hydrologically meaningful terrain parameters at 1x1km of spatial resolution derived using the standard SAGA-GIS basic terrain parameters module. These maps were normalized (between 0–1) and then used as prediction factors to downscale soil moisture across CONUS

4.2.4 Validation using field observations across CONUS

Downscaled soil moisture grids were compared against field measurements and we computed the explained variance (r^2) using a linear fit (observed *vs* predicted) for each field soil moisture location. Given the relatively low density and sparse spatial distribution of field data for validating (S1 Fig), we bootstrapped the independent validation using different sample sizes (from 10 to 100% of data with increments each 10%) to avoid systematic bias associated with the spatial distribution and density of field soil moisture information. We sampled (n = 1000) repeatedly the original and the downscaled soil moisture. We also computed the

spatial structure (spatial autocorrelation) of the explained variance (correlation between geographical distance and variance of r^2 values) for estimating an r^2 map using an interpolation technique known in geostatistics as Ordinary Kriging [62]. Ordinary Kriging is a well-known method for spatial interpolation based on the spatial structure or spatial autocorrelation of the variable of interest (the r^2 values between the field observations and the predicted soil moisture values). The spatial autocorrelation is defined by the relationship between geographical distances and variance of values at a given distance, and it is commonly characterized using variograms. We followed an automated variogram parameterization (the optimal selection for the variogram parameters nugget, sill and range required to perform Ordinary Kriging) proposed in previous work [63].

As implemented in the automap package of R [63], the initial sill is estimated as the mean of the maximum and the median values of the semi-variance. The semi-variance is defined by the variance within multiple distance intervals. For modeling the spatial autocorrelation this algorithm iterates over multiple variogram model parameters selecting the model (e.g., spherical, exponential, Gaussian) that has the smallest residual sum of squares with the sample vario- gram. The initial range is defined as 0.10 times the diagonal of the bounding box of the data. The initial nugget is defined as the minimum value of the semi-variance. Thus, the parameters used for obtaining a continuous map showing spatial trends in the r^2 were: a Gaussian (nor- mal) model form, a nugget value of 0.06 m^3 m^{-3}, a sill of 0.08 m^3 m^{-3} and an approximate range of 428.7 km. This map was generated

because it could provide insights about overall sources of modeling errors (e.g., environmental similarities in multiple areas showing low or high explained variance) and their spatial distribution. All analyzes were performed in R [64] using public sources of data. Our protocol and R code used for generating the soil moisture predictions at 1km grids is available online (http://dx.doi.org/10.17504/protocols.io.6cahase) for reproducibility of this research [65].grids aiming to identify their correlation with the aforementioned validation dataset (i.e., observed vs predicted).

4.3 Results

The exploratory PCA showed that the first two PCs explained 33% of the total dataset variability (S3A Fig), where the first PC explained 18% of total variability and at least five PCs were needed to explain 70% of total variability. The first PC was best correlated with elevation (r = 0.82) and with the vertical distance to channel network (r = 0.88). Elevation varied negatively with soil moisture, as well as other secondary terrain parameters such as the base level channel network elevation (distance from each pixel to the closer highest point), while the valley depth index varied positively with soil moisture (S3B Fig). The relative slope position (indicating the dominance of flat or complex terrain) and the topographic wetness index (which indicates areas where water tends to accumulate) were also correlated with soil moisture across the first 5 PCs. Thus, multiple terrain parameters varied positively and negatively with soil moisture values (S1 Appendix).

Our framework to predict soil moisture based on topography and remote sensing was able to explain, on average 79±0.1% of the variability of satellite soil moisture information as revealed by the cross-validation strategy. The root mean squared error (RMSE) derived from the cross-validation varied around 0.03 m^3/m^3, while the percentage of explained variance was in all cases above 70% (Table 4.1).

By applying the model coefficients to the topographic prediction factors across CONUS, we generated 26 cross-validated maps (for years 1991–2016) of mean annual soil moisture estimates within 1x1km grids (Figure 4.3). The downscaled product shows a higher level of spatial variability due the increased spatial detail achieved by downscaling soil moisture to 1x1km grids (S4 Fig). Our predictions reveal a clear bimodal distribution of soil moisture values (e.g., from the east to the west, Figure 4.4) which is also evident in the original estimate (S5 Fig). The statistical comparison between the original product and the downscaled product shows a high level of agreement with an r^2 value of 0.72.

We provided a visual comparison between the original satellite estimate and the down- scaled results including both median (Figure 4.4A and Figure 4.4.B) and standard deviation values (Figure 4.4C and Figure 4.4D). We also show the uncertainty of the original soil moisture product as reported by its developers (Figure 4.4E) and the r^2 map from the validation against field stations (Figure 4.4F). This r^2 map (Figure 4.4F) is based on the spatial autocorrelation found in the variogram estimation (e.g., sill > nugget) applied to the r^2 values of the data used for validating our approach (S1 Fig).

This r^2 map is indicating areas with similar soil moisture conditions regulating the spatial variability of soil moisture and affecting the performance of our models, around the sites of avail- able field data for validating our approach. The r^2 map shows the lowest values across the Central Plains of the US and the lower Mississippi basin. The lower values in the r^2 map are consistent with the high uncertainty values of the original satellite estimate (Figure 4.4E).

The r^2 map in Figure 4.4Fprovided insights about the relationship between soil moisture gridded surfaces and soil moisture field data. Higher r^2 values were found across the east coast, the Northern Plains and water-limited environments across the western states. We found that our soil moisture downscaled output better correlates (nearly 25% improvement) with NASMD field observations when compared to the original soil moisture satellite estimates (Figure 4.5).

This improvement was consistent after repeating it using random samples and different sample sizes (from 10 to 90% of available validation data) from the NASMD field observations (Figure 4.5). Consistently, we found a negative mean bias (surrogate of systematic error) in the ESA-CCI when compared against field stations of the NASMD (-0.051) that is slightly lower when comparing the downscaled soil moisture predictions against the NASMD (-0.048). The resulting RMSE of validating against field data was also slightly lower (0.057 m3/m3) for the downscaled estimate compared with the original product (0.062 m3/m3). However, there is a sparse distribution of validation data and large areas of CONUS lack of

field information for validating/calibrating soil moisture predictions (Figure 4.6). Considering the quality-controlled records available from the NASMD across CONUS and the coarse scale of the ESA-CCI soil moisture product, our approach suggests an improvement in the spatial resolution (from 27 to 1km grids) of soil moisture estimates while maintaining the integrity of the original satellite values.

The original satellite values, the downscaled product and the ISMN dataset showed a similar correlation with the terrain predictors. For example, the first PCA (represented by the distance to channel network and elevation), was negatively correlated with field soil moisture, the satellite original product and our soil moisture predictions. The correlation values were r = -0.17, r = -0.27, and r = -0.28 respectively. These relationships showed a similar pattern in the statistical space (Figure 4.7).

Table 4.1. The cross-validation results for each year. This table shows the correlation, root mean squared error (RMSE) in m^3/m^3, the number of training data available (*n*), the optimal kernel function (*okf*), and the optimal number of neighbors used for predicting to new data (*k*).

Model	Year	Correl ation	RMSE	*n*	*okf*	*k*
1	1991	0.85	0.03	18058	triangular	18
2	1992	0.89	0.03	18429	triangular	16
3	1993	0.88	0.03	18107	triangular	18
4	1994	0.9	0.03	18367	triangular	16
5	1995	0.88	0.03	18385	triangular	18
6	1996	0.9	0.03	18454	triangular	15
7	1997	0.88	0.03	18428	triangular	15
8	1998	0.88	0.03	18540	triangular	16
9	1999	0.89	0.03	18542	triangular	15
10	2000	0.9	0.03	18547	triangular	15
11	2001	0.9	0.03	18523	triangular	15
12	2002	0.9	0.03	19170	triangular	16
13	2003	0.89	0.03	19132	triangular	16
14	2004	0.89	0.03	18934	triangular	16
15	2005	0.89	0.03	19132	triangular	16
16	2006	0.9	0.03	19131	triangular	16
17	2007	0.88	0.03	19142	triangular	16
18	2008	0.9	0.03	19136	triangular	16
19	2009	0.9	0.03	19142	triangular	16
20	2010	0.88	0.03	19245	triangular	18
21	2011	0.9	0.03	19255	triangular	18
22	2012	0.9	0.03	19252	triangular	16
23	2013	0.89	0.03	19226	triangular	16
24	2014	0.89	0.03	19227	triangular	16
25	2015	0.88	0.03	19231	triangular	16
26	2016	0.88	0.03	19225	triangular	16

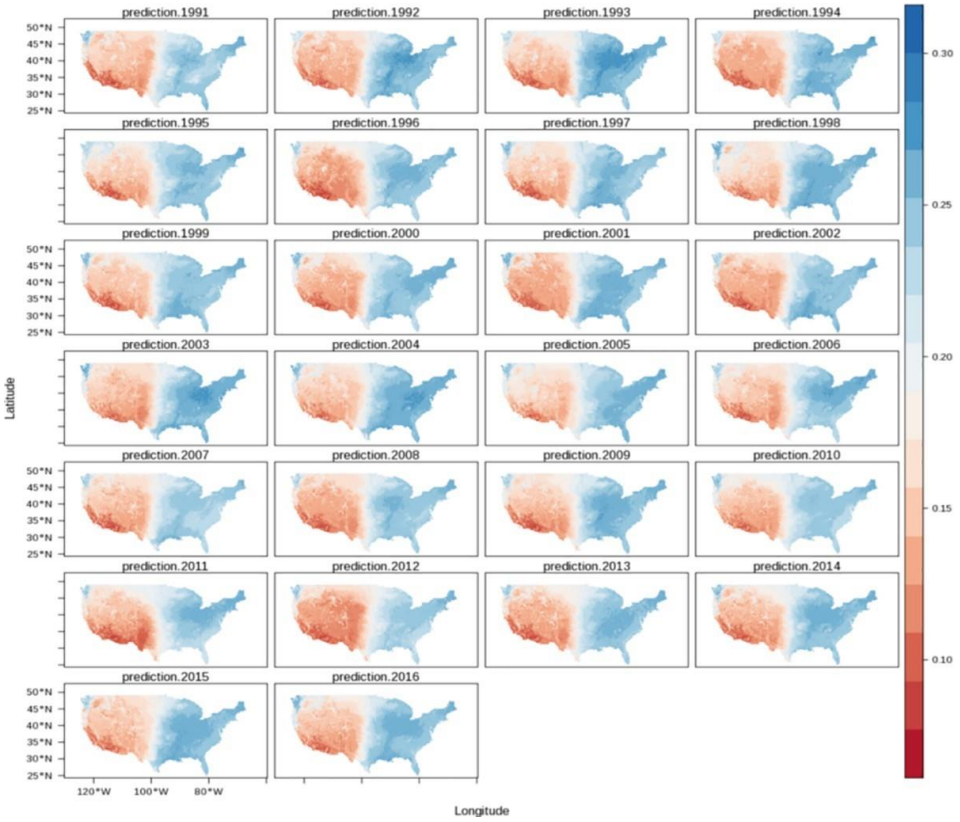

Figure 4.3. Annual means of soil moisture (1991–2016) downscaled to 1x1km grids across CONUS using terrain parameters as prediction factors.

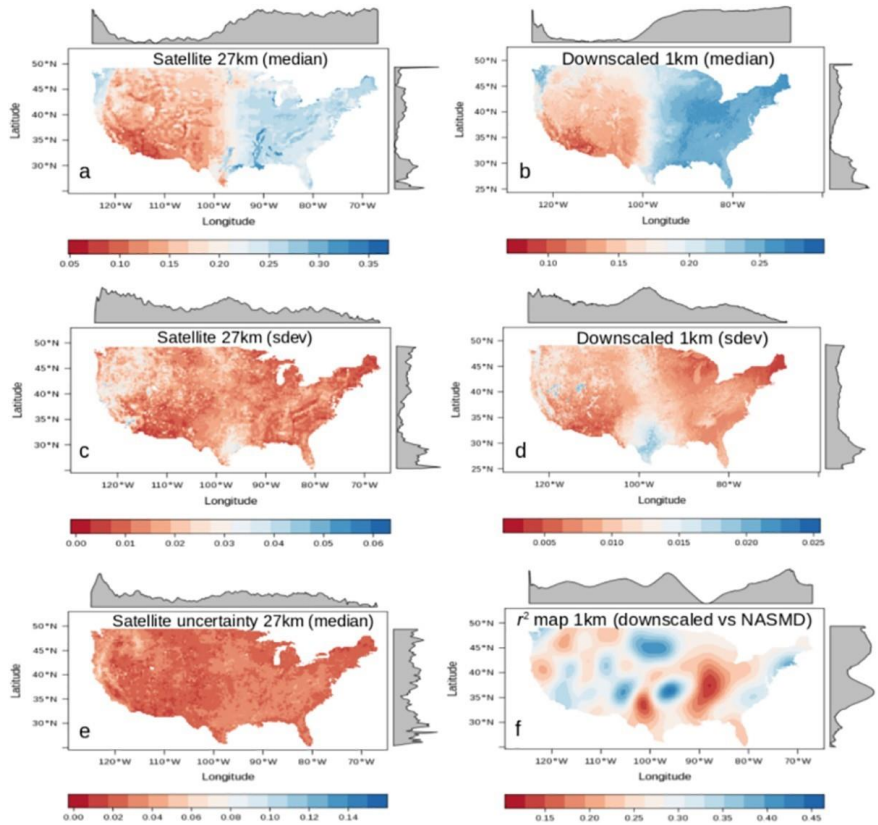

Figure 4.4. Comparison of the original (27km grids) and the downscaled (1km grids) soil moisture products. Median (a, b) and standard deviation (c, d, sdev) values of satellite soil moisture and downscaled soil moisture values (1991– 2016). Uncertainty reported by the ESA-CCI soil moisture (e) and the explained variance map (r^2) between field data and downscaled soil moisture (f).

4.4 Discussion

Our soil moisture downscaling framework was able to improve the spatial detail of ESA-CCI satellite soil moisture product and its agreement with field soil moisture records from the NASMD. It is well known that topography has a direct influence on the overall water distribution across the landscape [38-39] and in the angle between satellite retrieval and the Earth's surface. Thus, we demonstrated how a coarse scale satellite-based soil moisture product (27x27km of spatial resolution), in combination with hydrologically meaningful terrain parameters, can be coupled using machine learning algorithms to generate a fine-gridded and gap-free soil moisture product at the annual scale across CONUS.

We found a correlation between field soil moisture estimates and topography that is similar to the correlation between satellite estimates and topography (Figure 4.7), suggesting that topography can be an effective predictor for direct soil moisture measurements (i.e., from microwave remote sensing). Previous studies have confirmed this correlation between topographic variability and soil moisture conditions for downscaling soil moisture across multiple catchment scales and environmental conditions [40-41]. We recognize that our modeling approach based on digital terrain analysis does not directly account for local variations of evaporation, soil structure or vegetation (but indirectly). Our approach assumed that topography (described by the shape of multiple digital terrain parameters) is also capturing some variability associated to all factors affecting soil moisture. In contrast to previous downscaling efforts using vegetation and climate information [33, 66], we generated 26 annual soil

moisture predictions (1991–2016, 1x1 km of spatial resolution) that are independent of ecological data (i.e., vegetation greenness) and climate information, (i.e., precipitation and temperature). This topography-based approach is able to maintain the original satellite soil moisture statistical distribution (S5 Fig) and has the advantage that our modelling output could be further related to independent datasets of ecological or climate variables [67-68] and avoid subsequent spurious relationships.

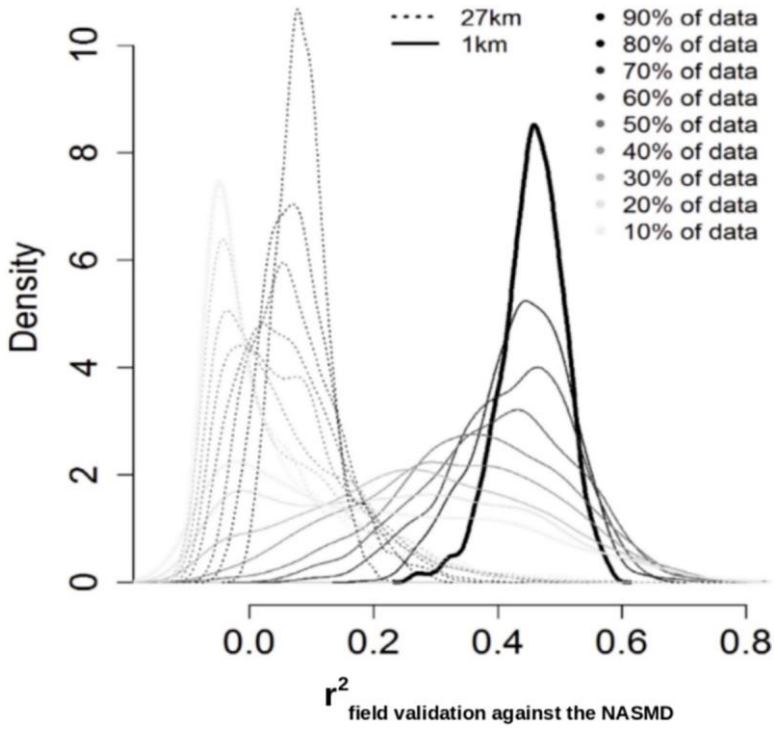

Figure 4.5. Validation of soil moisture gridded estimates (original 27 and 1km grids) against NASMD field observations. Dashed line represents the relationship of field stations and soil moisture gridded estimates at 27x27km, while black line represents the relationship between field stations and the downscaled 1x1km soil moisture product. In all cases (all sample sizes), the 1x1km product showed higher r2 with the NASMD than the ESA-CCI soil moisture estimates.

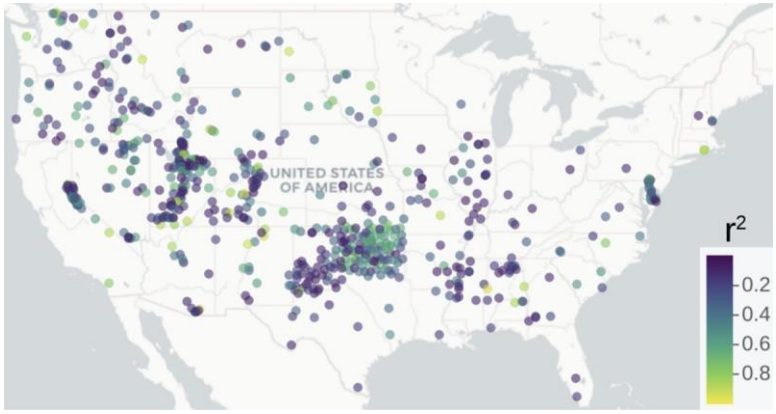

Figure 4.6. Explained variances computed for each meteorological station of the NASMD and the corresponding pixel of our soil moisture predictions based on geomorphometry.

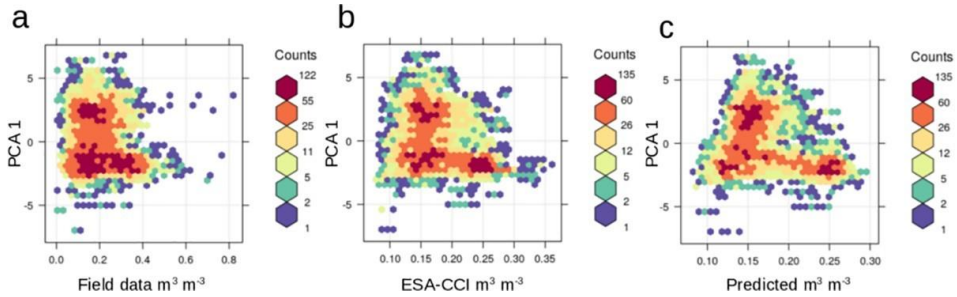

Figure 4.7. Relationships between the first PC of terrain parameters with soil moisture field data (a), with the ESA-CCI satellite product (b), and with the soil moisture predictions based on terrain parameters (c).

We showed that this topography-based approach is able to improve the correlation of the original estimate when compared against field data from the NASMD while maintaining the integrity of its values (i.e., same systematic error between the NASMD and the original or the downscaled product). Therefore, we provided a reliable (topography-based) approach to predict the satellite soil moisture patterns across finer spatial grids and in areas where no satellite soil moisture is available.

The downscaling process of satellite soil moisture from 27 to 1km grids across CONUS is supported on both internal (Table 4.1) and independent (Figure 4.5) validation frameworks to describe modeling performance. The accuracy of our modeling framework showed explained variances >70% and RMSE values considerably below (~0.03 m^3 m^{-3}) the satellite soil moisture mean of 0.22 m^3 m^{-3}, which is suitable for many applications [66], such as the detection of irrigation signals [69]. Similar results have been found recently for specific study sites [70].

Our results obtained by the cross-validation strategy and ground validation supports the application of a topography-based model to predict satellite soil moisture estimates (Figure 4.4).

Our results showed that higher soil moisture values could be found across lower elevations, areas with generally large and gentle slopes mainly across valley bottoms and across catchment areas where water tends to accumulate. This interpretation could explain the short distance in the multivariate analysis of satellite soil moisture estimates to elevation and derived terrain parameters such as

the vertical distance (of each pixel) to the nearest channel network, the valley depth index and the topographic wetness index. The multivariate analysis also suggested some degree of statistical redundancy between the topographic prediction factors (S1 Appen- dix) as they were derived from the digital elevation model by the means of geomorphometry [34-39]. For example, we found that the topographic wetness index is highly correlated with the length-slope factor (>0.80%), and this is because they are two secondary parameters that depend on slope (primary terrain attribute) [35]. Elevation and slope are respectively required for calculating secondary terrain parameters such as the valley depth index and the topographic wetness index [36] and these terrain parameters varied closely with soil moisture in the multivariate space (S1 Appendix). Thus, understanding the main relationships between topographic prediction factors and soil moisture can be useful for reducing modeling complexity while increasing our capacity to interpret modeling results.

The spatial detail of soil moisture estimates using 1km grids across the continental scale of CONUS is consistent with the variability of soil moisture patterns between the western and eastern United States. While drought scenarios have been recently reported for the western states [71] evidence of precipitation increase has been reported recently in the eastern states [72]. Our soil moisture downscaled estimates (Figure 4.3) revealed a soil moisture gradient across the Central Plains of CONUS and a clear separation of two major soil moisture data

populations (i.e., soil moisture values with a bimodality distribution) from the drier west, to the humid east (S5 Fig).

The original satellite soil moisture estimates also show this bimodal distribution but with a lesser extent (S2 Fig). The bimodal distribution of soil moisture could be explained by a negative soil moisture and precipitation feedback in the western CONUS and a positive soil moisture and precipitation feedback in the eastern CONUS [68]. Furthermore, areas with soil moisture bimodality have been recognized across global satellite observations and climate models [73]. We identified areas of low agreement between our soil moisture predictions and field stations (lower r^2 values) across the transitional ecosystems (Figure 4.4) from drier to humid soil moisture environments (i.e., Central Plains and lower Mississippi basin). It is likely that these transitional areas drive changes in water availability in surface and subsurface hydrological systems [74]. The lower Mississippi basin, specifically the area across the surroundings of the Mississippi delta, is an example of a transitional area experiencing aquifer depletion [75] where both flooding events and droughts tend to occur within shorter distances that are not captured by the original satellite soil moisture information. These are the type areas where we found lower values of agreement (r^2 values) between satellite and ground soil moisture observations. These low correlation values can be also explained by the use of multiple soil moisture networks with different types of sensors and measurement techniques [19]. Also, the imperfections of prediction factors used

for soil moisture spatial variability models represent a potential source of uncertainty.

As any downscaling effort dependent on covariates (i.e., terrain parameters), our approach is vulnerable to data quality limitations such as the presence of systematic errors on these covariates. Other errors are derived from input data imperfections and difficulties meeting modeling assumptions. These errors in soil moisture modeling inputs increase the risk of bias and uncertainty propagation to subsequent soil moisture modeling outputs and soil mapping applications [76-78]. For example, elevation data surfaces derived from remote sensing data (such as the global DEM used here) could show artifacts (i.e., false pikes or spurious sinks) due to data saturation or signal noise that can be propagated to final soil moisture predictions [79]. We minimized this issue by using SAGA-GIS [36] as it has adopted methods for preprocessing and perform DEM quality checks [80] before deriving the topographic prediction factors used in this study. Because input covariates could not be fully free of errors, we advocate for reporting information on bias and r^2 values to inform about accuracy (e.g., Table 4.1) as important components for interpreting soil moisture predictions.

Our results suggest that the original coarse scale soil moisture product and the values of soil moisture from the NASMD (Fig 5) are difficult to compare in terms of spatial variability, as is highlighted in previous studies [19]. This is because a satellite soil moisture pixel from the ESA-CCI product provides a value across a larger area (27x27km) than a field measurement at a specific sampling location

(defined by geographical coordinates). This scale dependent effect (27x27km vs 1:1 field scale) is reduced (>25%) with soil moisture predictions across finer grids (1km). The downscaled soil moisture maps showed a higher agreement with field soil moisture records from the NASMD (Figure 4.5), supporting the applicability of this soil moisture product for applications that required higher spatial resolution.

Our soil moisture predictions across 1km grids suggest that topography can be effectively used to improve the spatial detail and accuracy of satellite soil moisture estimates. Several studies have highlighted differences in spatial representativeness between ground-based observations and satellite soil moisture products [77, 81]. Other studies have shown that the spatial representativeness of the ESA-CCI soil moisture compared with field observations is higher from regional-to-continental scales than from ecosystem-to-landscape scales [82-83]. There- fore, large uncertainties of soil moisture spatial patterns (below 1km grids) needs to be resolved for assessing and better understanding the local variability of soil moisture trends.

We argue that currently there is an increasing availability of high-quality digital elevation data sources with high levels of spatial resolution (e.g., 1–2 to 30 to 90m grids) across large areas of the world [84-85] that can be used to derive reliable hydrologically meaningful terrain parameters for predicting soil moisture. The relationship of these digital terrain parameters and field soil moisture (i.e., meteorological stations) is similar to the relationship between terrain parameters and satellite soil moisture gridded estimates (Figure 4.7).

From a single information source (a remotely sensed DEM), we downscaled satellite records of soil moisture using a framework that theoretically is reproducible across multiple scales. The ultimate goal of reducing the multiple information sources for predicting soil moisture is to reduce the statistical redundancy in further modeling efforts (i.e., land carbon uptake models) and large-scale ecosystem studies (i.e., ecological niche modeling) that combine similar prediction factors for soil moisture (i.e., climate or vegetation indexes). These include models estimating water evapotranspiration trends [86] and process based global carbon mod- els that could also benefit from more accurate and independent soil moisture inputs [78]. To improve the spatial representativeness of satellite soil moisture estimates, the number of studies developing new downscaling approaches based on prediction factors is rapidly expanding [26, 28, 66, 87]. There is a pressing need to solve the current uncertainty of soil moisture estimates to accurately understand how soil moisture is limiting the primary productivity of terrestrial ecosystems [6]. Previous studies have identified a topographic signal in satellite soil moisture [88] supporting the reliability of our modeling approach. Therefore, our results pro- vide an alternative applicable to continental scales for downscaling satellite soil moisture estimates based on hydrologically meaningful terrain parameters.

The novelty of this approach is that it could be applicable to multiple temporal resolutions (e.g., monthly or daily) as it generates independent models for each period of interest and at multiple spatial scales as the availability of terrain

parameters for modeling purposes has increased substantially (i.e., meters) in the last decade. Increasing the temporal resolution of downscaled maps (i.e., from annual to monthly predictions) is beyond the scope of this study, will increase computational costs, but are theoretically possible following this approach. While monthly or weekly (or even daily soil moisture datasets) are valuable sources for large scale earth system modeling, annual averages are also valuable for detecting long term trends in the climate-land system. Rather than focusing on temporal variability of soil moisture, our results provide insights for improving the spatial variability and consequently the spatial representation of soil moisture gridded surfaces derived from satellite information.

4.5 Conclusion

Recent studies highlight the necessity of detailed soil moisture products to account for soil moisture limitation in terrestrial ecosystems. We developed a geomorphometry-based framework to couple satellite soil moisture records with hydrologically meaningful terrain parameters. This approach is useful to avoid statistical redundancies when downscaled soil moisture is further used or analyzed with vegetation- or climate-related variables not included in the downscaling framework. We predicted (i.e., downscaled) soil moisture using 1x1km grids across CONUS at an annual scale from 1991 to 2016. This gap-free soil moisture product improved the spatial detail of the original satellite soil moisture grids and the overall agreement (increased by $>20\%$) of these grids with the NASMD field soil moisture records. Our findings suggest that digital ter- rain analysis can be applied to elevation

data sources to derive hydrologically meaningful terrain parameters and use these parameters predict soil moisture spatial patterns. Our framework is reproducible across the world because it is based on publicly available DEMs, ground and satellite soil moisture data. Our protocol and R code is available online (http://dx.doi.org/10.17504/ protocols.io.6cahase) as well as input parameters and annual means of soil moisture at 1km grids (https://www.hydroshare.org/resource/b8f6eae9d89241cf8b5904033460af61/).

Supporting information available in:
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0219639#sec010

S1 Appendix. This file (AppendixS1.html) contains the results of the automated PCA report described in the methods section, for exploring relationships between soil moisture and topography.
(HTML)

S1 Fig. Spatial distribution of the NASMD validation set with information for the first 0- 5cm depth (n 668).

S2 Fig. ESA-CCI satellite soil moisture across CONUS (~27km grids).
(TIF)

S3 Fig. PCA map of the first and second principal components. The individual point cloud across the plane between the first and second PCA (a). The orthogonal relationship of the vari- ables with higher contribution to this plane (b). Soil moisture is represented by the dotted blue line. An interpretation of these can be found in S1 Appendix.
(TIF)

S4 Fig. Mean soil moisture for the year 2016. Comparison between the original satellite soil moisture (a) and soil moisture predicted at 1km grids (b).
(TIF)

S5 Fig. Bimodal distribution of satellite soil moisture (black) and the downscaled soil moisture estimates (red).
(TIF)

S1 Table. This table (S1 Table) contains a description of primary and secondary terrain parameters used in this study for generating soil moisture predictions.
(DOCX)

Author contributions

Conceptualization: Mario Guevara, Rodrigo Vargas.
Data curation: Mario Guevara.

Formal analysis: Mario Guevara.

Supervision: Rodrigo Vargas.

Writing – original draft: Mario Guevara.

Writing – review & editing: Rodrigo Vargas.

REFERENCES

1. Greve P., Gudmundsson L., Seneviratne S.I., 2018. Regional scaling of annual mean precipitation and water availability with global temperature change. Earth Syst. Dynam. 9, 227–240. https://doi.org/10. 5194/esd-9-227-2018

2. Seneviratne S.I., Corti T., Davin E.L., Hirschi M., Jaeger E.B., Lehner, et al. 2010. Investigating soil moisture–climate interactions in a changing climate: A review. Earth-Science Reviews 99, 125–161. https://doi.org/10.1016/j.earscirev.2010.02.004

3. Seneviratne S.I., Wilhelm M., Stanelle T., van den Hurk B., Hagemann S., Berg,et al. 2013. Impact of soil moisture-climate feedbacks on CMIP5 projections: First results from the GLACE-CMIP5 experi- ment: GLACE-CMIP5 EXPERIMENT. Geophysical Research Letters 40, 5212–5217. https://doi.org/ 10.1002/grl.50956

4. Western A.W., Grayson R.B., Blö schl G., Willgoose G.R., McMahon T.A., 1999. Observed spatial orga- nization of soil moisture and its relation to terrain indices. Water Resour. Res. 35, 797–810. https://doi. org/10.1029/1998WR900065

5. Dorigo W., Wagner W., Albergel C., Albrecht F., Balsamo G., Brocca, et al. 2017. ESA CCI Soil Mois- ture for improved Earth system understanding: State-of-the art and future directions. Remote Sensing of Environment, Earth Observation of Essential Climate Variables 203, 185–215. https://doi.org/10. 1016/j.rse.2017.07.001

6. Stocker B. D., Zscheischler J., Keenan T. F., Prentice I. C., Peñuelas J., & Seneviratne S. I. 2018. Quantifying soil moisture impacts on light use efficiency across biomes. New Phytol., 218(4), 1430– 1449. https://doi.org/10.1111/nph.15123 PMID: 29604221

7. Brocca L., Ciabatta L., Massari C., Camici S., & Tarpanelli A. 2017. Soil Moisture for Hydrological Appli- cations: Open Questions and New Opportunities. Water, 9(2), 140. https://doi.org/10.3390/w9020140

8. Vargas R., Sánchez-Cañete P., Serrano-Ortiz P., Curiel Yuste J., Domingo F., López- Ballesteros A. et al. 2018. Hot-moments of soil CO2 efflux in a water-limited grassland. Soil Systems, 2(3), p.47.

9. Asner G.P., Alencar A., 2010. Drought impacts on the Amazon forest: the remote sensing perspective. New phytologist.

10. Cook B.D., Corp L.A., Nelson R.F., Middleton E.M., Morton D.C., McCorkel J.T., et al. 2013. NASA Goddard's LiDAR, Hyperspectral and Thermal (G-LiHT) Airborne Imager. Remote Sensing 5, 4045– 4066. https://doi.org/10.3390/rs5084045

11. Dai A., 2011. Drought under global warming: a review. Wiley Interdisciplinary Reviews: Climate Change 2, 45–65. https://doi.org/10.1002/wcc.81

12. Samaniego L., Thober S., Kumar R., Wanders N., Rakovec O., Pan,et al. 2018. Anthropogenic warming exacerbates European soil moisture droughts. Nature Climate Change 8, 421–426. https://doi.org/10. 1038/s41558-018-0138-5

13. van der Molen M.K., Dolman A.J., Ciais P., Eglin T., Gobron N., Law B.E. et al. 2011. Drought and eco- system carbon cycling. Agricultural and Forest Meteorology 151, 765–773. https://doi.org/10.1016/j. agrformet.2011.01.018

14. Luo Y., Ahlström A., Allison S.D., Batjes N.H., Brovkin V., Carvalhais, et al., 2016. Toward more realistic projections of soil carbon dynamics by Earth system models. Global Biogeochemical Cycles 30, 40– 56. https://doi.org/10.1002/2015GB005239

15. Walsh B., Ciais P., Janssens I.A., Peñuelas J., Riahi K., Rydzak F., et al., 2017. Pathways for balancing CO2 emissions and sinks. Nature Communications 8, 14856. https://doi.org/10.1038/ncomms14856 PMID: 28406154

16. Owe M., Van de Griend A.A., 1998. Comparison of soil moisture penetration depths for several bare soils at two microwave frequencies and implications for remote sensing. Water Resources Research 34, 2319–2327. https://doi.org/10.1029/98WR01469

17. Entekhabi D., Yueh S., O'Neill P., Kellog K., Allen A., et al., 2014. SMAP handbook—Soil Moisture Active Passive: Mapping Soil Moisture and Freeze/Thaw From Space, Jet Propulsion Lab., California Inst. Technol., Pasadena, Calif.

18. Singh R.S., Reager J.T., Miller N.L., Famiglietti J.S., 2015. Toward hyper-resolution land-surface modeling: The effects of fine-scale topography and soil texture on CLM4.0 simulations over the South- western U.S.: Effects of fine-scale resolution on CLM4.0 in Southwest US. Water Resources Research 51, 2648–2667. https://doi.org/10.1002/2014WR015686

19. Dirmeyer P., Wu J., Norton H., Dorigo W., Quiring S., Trenton W., et al. 2016. "Confronting Weather and Climate Models with Observational Data from Soil Moisture Networks over the United States." Jour- nal of Hydrometeorology 17 (4): 1049–67. https://doi.org/10.1175/JHM-D-15-0196.1 PMID: 29645013

20. Liu Y.Y., Dorigo W.A., Parinussa R.M., de Jeu R.A.M., Wagner W., McCabe, et al. 2012. Trend-preserv- ing blending of passive and active microwave soil moisture retrievals. Remote Sensing of Environment 123, 280–297. https://doi.org/10.1016/j.rse.2012.03.014

21. Liu Y.Y., Parinussa R.M., Dorigo W.A., De Jeu R.A.M., Wagner W., van Dijk A.I.J.M., et al. 2011. Devel- oping an improved soil moisture dataset by blending passive and active microwave satellite-based retrievals. Hydrology and Earth System Sciences 15, 425–436. https://doi.org/10.5194/hess-15-425- 2011

22. McColl K.A., Alemohammad S.H., Akbar R., Konings A.G., Yueh S., Entekhabi D., 2017. The global dis- tribution and dynamics of surface soil moisture. Nature Geoscience 10, 100–104. https://doi.org/10.1038/ngeo2868

23. Montzka C., Rö tzer K., Bogena H.R., and Vereecken H. 2018. A new soil moisture downscaling approach for SMAP, SMOS and ASCAT by predicting sub-grid variability. Remote Sens. 10(3):427. https://doi.org/10.3390/rs10030427

24. Afshar M.H., Yilmaz M.T., 2017. The added utility of nonlinear methods compared to linear methods in rescaling soil moisture products. Remote Sensing of Environment 196, 224–237. https://doi.org/10.1016/j.rse.2017.05.017

25. Jin Y., Ge Y., Wang J., Heuvelink G.B.M., Wang L., 2018. Geographically Weighted Area-to-Point Regression Kriging for Spatial Downscaling in Remote Sensing. Remote Sensing 10, 579. https://doi.org/10.3390/rs10040579

26. Kearney M.R., Maino J.L., 2018. Can next-generation soil data products improve soil moisture model- ling at the continental scale? An assessment using a new microclimate package for the R programming environment. Journal of Hydrology 561, 662–673. https://doi.org/10.1016/j.jhydrol.2018.04.040

27. Piles M., Camps A., Vall-llossera M., Corbella I., Panciera R., Rudiger C., Kerr Y.H., et al., 2011. Down- scaling SMOS-Derived Soil Moisture Using MODIS Visible/Infrared Data. IEEE Transactions on Geo- science and Remote Sensing 49, 3156–3166. https://doi.org/10.1109/TGRS.2011.2120615

28. Ranney K.J., Niemann J.D., Lehman B.M., Green T.R., Jones A.S., 2015. A method to downscale soil moisture to fine resolutions using topographic, vegetation, and soil data. Advances in Water Resources 76, 81–96. https://doi.org/10.1016/j.advwatres.2014.12.003

29. Wang A., Zhang M., Shi J., Mu T., Gong H., Xie C., 2012. Space-time analysis on downscaled soil mois- ture data and parameters of plant growth. Transactions of the Chinese Society of Agricultural Engineer- ing 28, 164–169.

30. Yu G., Di L., Yang W., 2008. Downscaling of Global Soil Moisture using Auxiliary Data. IEEE, pp. III- 230–III–233. https://doi.org/10.1109/IGARSS.2008.4779325

31. McBratney A., Mendonça Santos M., Minasny B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

32. Bauer-Marschallinger B., Freeman V., Cao S., Paulik C., Schaufler S., Stachl T., et al. (2019). Toward Global Soil Moisture Monitoring With Sentinel-1: Harnessing Assets and Overcoming Obstacles. IEEE Trans. Geosci. Remote Sens., 57(1), 520–539. https://doi.org/10.1109/TGRS.2018.2858004

33. Morelo B., Merlin O., Malbeteau Y., Al Bitar A., Cabot F., Stefan V., et al. 2016. SMOS disaggregated soil moisture product at 1km resolution: Processor overview and first validation results. Remote Sensing of Environment 180, 361–376 https://doi.org/10.1016/j.rse.2016.02.045

34. Pike R.J., Evans I.S., T., 2009. Chapter 1 Geomorphometry: A Brief Guide, in: Developments in Soil Science. Elsevier, pp. 3–30.

35. Wilson J. P., & Gallant J. C. (2000). Digital terrain analysis. Terrain analysis: Principles and applications, 6(12), 1–27.

36. Conrad O., Bechtel B., Bock M., Dietrich H., Fischer E., Gerlitz L., et al. 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geoscientific Model Development 8, 1991–2007. https://doi. org/10.5194/gmd-8-1991-2015

37. Wilson, J.P., 2012. Digital terrain modeling. Geomorphology, Geospatial Technologies and Geomor- phological Mapping Proceedings of the 41st Annual Binghamton Geomorphology Symposium 137, 107–121. https://doi.org/10.1016/j.geomorph.2011.03.012

38. Florinsky I.V., 2016. Chapter 9—Influence of Topography on Soil Properties, in: Florinsky I.V. (Ed.), Dig- ital Terrain Analysis in Soil Science and Geology (Second Edition). Academic Press, pp. 265–270. https://doi.org/10.1016/B978-0-12-804632-6.00009-2

39. Florinsky I.V., 2012. The Dokuchaev hypothesis as a basis for predictive digital soil mapping (on the 125th anniversary of its publication). Eurasian Soil Science 45, 445–451. https://doi.org/10.1134/ S1064229312040047

40. Pellenq J, Kalma J, Boulet G, Saulnier G-M, Wooldridge S. et al. A disaggregation scheme for soil mois- ture based on topography and soil depth. Journal of Hydrology. 2003; 276: 112–127. https://doi.org/10. 1016/s0022-1694(03)00066-0

41. Busch FA, Niemann JD, Coleman M. Evaluation of an empirical orthogonal function-based method to downscale soil moisture patterns based on topographical attributes. Hydrological Processes. 2011; 26: 2696–2709. https://doi.org/10.1002/hyp.8363

42. Hengl T., MacMillan R.A., 2019. Predictive Soil Mapping with R. OpenGeoHub foundation, Wagenin- gen, the Netherlands, 370 pages, www.soilmapper.org, ISBN: 978-0-359-30635-0.

43. Reichstein M., Camps-Valls G., Stevens B., Jung M., Denzler J., Carvalhais N., Et al. 2019. Deep learn- ing and process understanding for data-driven Earth system science. Nature, 566(7743), 195. https://doi.org/10.1038/s41586-019-0912-1 PMID: 30760912

44. Guevara M, Olmedo GF, Stell E, Yigini Y, Aguilar Duarte Y, Arellano Hernández C, et al. No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America. SOIL. 2018; 4: 173–193. https://doi.org/10.5194/soil-4-173-2018

45. Warner D.L., Guevara M., Inamdar S. and Vargas R., 2019. Upscaling soil-atmosphere CO2 and CH4 fluxes across a topographically complex forested landscape. Agricultural and forest meteorology, 264, pp.80–91.

46. Coopersmith E. J., Cosh M. H., Bell J. E., & Boyles R. 2016. Using machine learning to produce near surface soil moisture estimates from deeper in situ records at U.S. Climate Reference Network (USCRN) locations: Analysis and applications to AMSR-E satellite validation. Adv. Water Resour., 98, 122–131. https://doi.org/10.1016/j.advwatres.2016.10.007

47. Quiring S.M., Ford T.W., Wang J.K., Khong A., Harris E., Lindgren T.,et al., 2016. The North American Soil Moisture Database: Development and Applications. Bulletin of the American Meteorological Soci- ety 97, 1441–1459. https://doi.org/10.1175/BAMS-D-13-00263.1

48. Dorigo W., Wagner W., Albergel C., Albrecht F., Balsamo G., Brocca, et al. 2017. ESA CCI Soil Mois- ture for improved Earth system understanding: State-of-the art and future directions. Remote Sensing of Environment, Earth Observation of Essential Climate Variables 203, 185–215. https://doi.org/10. 1016/j.rse.2017.07.001

49. Bindlish R., Jackson T., Cosh M., Tianjie Zhao O'Neill P., 2015. Global Soil Moisture From the Aquar- ius/SAC-D Satellite: Description and Initial Assessment. IEEE Geoscience and Remote Sensing Letters 12, 923–927. https://doi.org/10.1109/LGRS.2014.2364151

50. Entekhabi D., Njoku E., O'Neill P., Kellogg K., Crow W., Edelstein W., et al. 2010. The Soil Moisture Active Passive (SMAP) Mission. Proceedings of the IEEE 98, 704–716. https://doi.org/10.1109/ JPROC.2010.2043918

51. Naeimi V., Paulik C., Bartsch A., Wagner W., Kidd R., Park, et al. ASCAT Surface State Flag (SSF): Extracting Information on Surface Freeze/Thaw Conditions From Backscatter Data Using an Empirical Threshold-Analysis Algorithm. IEEE Transactions on Geoscience and Remote Sensing 50, 2566– 2582. https://doi.org/10.1109/TGRS.2011.2177667

52. Naeimi V., Scipal K., Bartalis Z., Hasenauer S., Wagner W., 2009. An Improved Soil Moisture Retrieval Algorithm for ERS and METOP Scatterometer Observations. IEEE Transactions on Geoscience and Remote Sensing 47, 1999–2013. https://doi.org/10.1109/TGRS.2008.2011617

53. Wagner W., Lemoine G., Rott H., 1999. A Method for Estimating Soil Moisture from ERS Scatterometer and Soil Data. Remote Sensing of Environment 70, 191–207. https://doi.org/10.1016/S0034-4257(99) 00036-X

54. Dorigo W.A., Wagner W., Hohensinn R., Hahn S., Paulik C., Xaver A., et al. 2011. The International Soil Moisture Network: a data hosting facility for global in situ soil moisture measurements. Hydrology and Earth System Sciences 15, 1675–1698. https://doi.org/10.5194/hess-15-1675-2011

55. Becker J.J., Sandwell D.T., Smith W.H.F., Braud J., Binder B., Depner J., et al. Global Bathymetry and Elevation Data at 30 Arc Seconds Resolution: SRTM30_PLUS. Marine Geodesy 32, 355–371. https://doi.org/10.1080/01490410903297766

56. Hengl T., de Jesus J., MacMillan R., Batjes N., Heuvelink GBM., Ribeiro E, et al. SoilGrids1km—Global Soil Information Based on Automated Mapping. Bond-Lamberty B, editor. PLoS ONE. 2014; 9: e105992. https://doi.org/10.1371/journal.pone.0105992 PMID: 25171179

57. Tuanmu M.-N., & Jetz W. A global 1-km consensus land-cover product for biodiversity and ecosystem modelling. Global Ecol. Biogeogr., 23(9), 1031–1045. 2014. https://doi.org/10.1111/geb.12182

58. Thuleau S, and Husson F. 2018. FactoInvestigate: Automatic Description of Factorial Analysis. R pack- age version 1.3. https://CRAN.R-project.org/package=FactoInvestigate

59. Hechenbichler, K., Schliep, K., 2006. Weighted k-nearest-neighbor techniques and ordinal classifica- tion, in: Discussion Paper 399, SFB 386.

60. Hechenbichler, K., Schliep, K., 2004. Weighted k-Nearest-Neighbor Techniques and Ordinal Classifica- tion [WWW Document]. URL https://epub.ub.uni-muenchen.de/1769/ (accessed 12.24.16).

61. Borra S., Di Ciaccio A., 2010. Measuring the prediction error. A comparison of cross-validation, boot- strap and covariance penalty methods. Computational Statistics & Data Analysis 54, 2976–2989. https://doi.org/10.1016/j.csda.2010.03.004

62. Oliver M. A., & Webster R. 2014. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. CATENA, 113, 56–69. https://doi.org/10.1016/j.catena.2013.09.006

63. Hiemstra P. H., Pebesma E. J., Twenhö fel C. J. W., & Heuvelink G. B. M. 2009. Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. Comput. Geosci., 35(8), 1711–1721. https://doi.org/10.1016/j.cageo.2008.10.011

64. R Core Team 2018. R: A language and environment for statistical computing. R Foundation for Statisti- cal Computing, Vienna, Austria. URL https://www.R-project.org/.

65. Guevara M. Vargas R. Protocol for Downscaling Satellite Soil Moisture Estimates using Geomorphome- try and Machine Learning. Protocols.io. protocols.io; 1970; https://doi.org/10.17504/protocols.io. 6cahase

66. Colliander A., Fisher J. B., Halverson G., Merlin O., Misra S., Bindlish, et al. 2017. Spatial Downscaling of SMAP Soil Moisture Using MODIS Land Surface Temperature and NDVI During SMAPVEX15. IEEE Geosci. Remote Sens. Lett., 14(11), 2107–2111. https://doi.org/10.1109/LGRS.2017.2753203

67. He L., Chen J. M., Liu J., Bélair S., & Luo X. 2017. Assessment of SMAP soil moisture for global simula- tion of gross primary production. J. Geophys. Res. Biogeosci., 122(7), 1549–1563. https://doi.org/10. 1002/2016JG003603

68. Tuttle S. & Salvucci G. 2016. Empirical evidence of contrasting soil moisture-precipitation feedbacks across the United States. Science 352, 825–828. https://doi.org/10.1126/science.aaa7185 PMID: 27174987

69. Lawston P. M., Santanello J. A., & Kumar S. V. 2017. Irrigation Signals Detected From SMAP Soil Mois- ture Retrievals. Geophys. Res. Lett., 44(23), 11,860–11,867. https://doi.org/10.1002/2017GL075733

70. Colliander A., Jackson T. J., Bindlish R., Chan S., Das N., Kim S. B.,et al. 2017. Validation of SMAP sur- face soil moisture products with core validation sites. Remote Sens. Environ., 191, 215–231. https://doi.org/10.1016/j.rse.2017.01.021

71. Diffenbaugh N. S., Swain D. L., & Touma D. 2015. Anthropogenic warming has increased drought risk in California. Proc. Natl. Acad. Sci. U.S.A., 112(13), 3931–3936. https://doi.org/10.1073/pnas. 1422385112 PMID: 25733875

72. Easterling D. R., Kunkel K. E., Arnold J. R., Knutson T., LeGrande A. N., Leung L. et al. 2017. Precipita- tion change in the United States. In Wuebbles D. J., Fahey D. W., Hibbard K. A., Dokken D. J., Stewart B. C., & Maycock T. K. (Eds.), Climate science special report: Fourth national climate assessment (Vol. I, pp. 207–230). Washington, DC: U.S. Global Change Research Program

73. Vilasa L, Miralles DG, de Jeu RAM, Dolman AJ. Global soil moisture bimodality in satellite observations and climate models. Journal of Geophysical Research: Atmospheres. 2017; 122: 4299–4311. https://doi.org/10.1002/2016jd026099

74. Dyer J., & Mercer A. 2013. Assessment of Spatial Rainfall Variability over the Lower Mississippi River Alluvial Valley. J. Hydrometeorol., 14(6), 1826–1843. https://doi.org/10.2307/24914344

75. Reba M. L., Massey J. H., Adviento-Borbe M. A., Leslie D., Yaeger M. A., Anders M., et al. 2017. Aquifer Depletion in the Lower Mississippi River Basin: Challenges and Solutions. Journal of Contemporary Water Research & Education, 162(1), 128–139. https://doi.org/10.1111/j.1936-704X.2017.03264.x

76. Heuvelink G.B. M., Millward A.A., 1999. Error propagation in environmental modelling with GIS. Carto- graphica 36, 69.

77. Miralles D.G., Crow W.T., Cosh M.H., 2010. Estimating Spatial Sampling Errors in Coarse-Scale Soil Moisture Estimates Derived from Point-Scale Observations. Journal of Hydrometeorology 11, 1423– 1429. https://doi.org/10.1175/2010JHM1285.1

78. Munguia-Flores F., Arndt S., Ganesan A. L., Murray-Tortarolo G., & Hornibrook E. R. C. 2018. Soil Methanotrophy Model (MeMo v1.0): a process-based model to quantify global uptake of atmospheric methane by soil. Geosci. Model Dev., 11(6), 2009–2032. https://doi.org/10.5194/gmd-11-2009-2018

79. Lindsay J.B, Creed I.F. Removal of artifact depressions from digital elevation models: towards a mini- mum impact approach. Hydrological Processes. 2005; 19: 3113–3126. https://doi.org/10.1002/hyp. 5835

80. Planchon O. & Darboux F. (2001): A fast, simple and versatile algorithm to fill the depressions of digital elevation models. Catena 46: 159–176

81. Gruber A., Dorigo W.A., Zwieback S., Xaver A., Wagner W., 2013. Characterizing Coarse-Scale Repre- sentativeness of in situ Soil Moisture Measurements from the International Soil Moisture Network. Vadose Zone Journal 12, 0. https://doi.org/10.2136/vzj2012.0170

82. Nicolai-Shaw N., Hirschi M., Mittelbach H., Seneviratne S.I., 2015. Spatial representativeness of soil moisture using in situ, remote sensing, and land reanalysis data: SPATIAL REPRESENTATIVENESS OF SOIL MOISTURE. Journal of Geophysical Research: Atmospheres 120, 9955–9964. https://doi. org/10.1002/2015JD023305

83. Vargas R., Sonnentag O., Abramowitz G., Carrara A., Chen J.M., Ciais P., et al. 2013. Drought influ- ences the accuracy of simulated ecosystem fluxes: a model-data meta-analysis for Mediterranean oak woodlands. Ecosystems, 16(5), pp.749–764.

84. Nelson A., Reuter H.I., Gessler P., 2009. Chapter 3 DEM Production Methods and Sources, in: Devel- opments in Soil Science. Elsevier, pp. 65–85.

85. Tadono T., Ishida H., Oda F., Naito S., Minakawa K., Iwamoto H., 2014. Precise Global DEM Genera- tion by ALOS PRISM. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sci- ences II–4, 71–76. https://doi.org/10.5194/isprsannals-II-4-71-2014

86. Schwingshackl C., Hirschi M., Seneviratne S. I., Schwingshackl C., Hirschi M., & Seneviratne S. I. 2017. Quantifying Spatiotemporal Variations of Soil Moisture Control on Surface Energy Balance and Near-Surface Air Temperature. J. Clim. Retrieved from https://journals.ametsoc.org/doi/full/10.1175/ JCLI-D-16-0727.1

87. Jin Y., Ge Y., Wang J., Heuvelink G.B.M., Wang L., 2018. Geographically Weighted Area-to-Point Regression Kriging for Spatial Downscaling in Remote Sensing. Remote Sensing 10, 579. https://doi. org/10.3390/rs10040579

88. Mason D. C., Garcia-Pintado J., Cloke H. L. and Dance S. L.: Evidence of a topographic signal in sur- face soil moisture derived from ENVISAT ASAR wide swath data, International Journal of Applied Earth Observation and Geoinformation, 45, 178–186, https://doi.org/10.1016/j.jag.2015.02.004, 2016.

Chapter 5

GAP-FREE GLOBAL SOIL MOISTURE: 15KM GRIDS FOR 1991-2016

Authors:

Mario Guevara[1], Michela Taufer[2], Rodrigo Vargas[1]

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

[2]Department of Electrical Engineering and Computer Science, The University of Tennessee, Knoxville. United States.

Correspondence to: Rodrigo Vargas (rvargas@udel.edu)

Abstract

Soil moisture is key for quantifying soil-atmosphere interactions and the ESA-CCI (European Space Agency Climate Change Initiative) provides historical (>30 years) satellite soil moisture gridded data at the global scale. We evaluate an alternative approach to increase the spatial resolution of the original ESA-CCI soil moisture measurements from 27km to 15km grids by coupling machine learning (ML) with information from digital terrain analysis at the global scale. We modeled mean annual ESA-CCI soil moisture values across 26 years of available data (1991-2016) using a ML kernel method and multiple terrain parameters (e.g., slope, wetness index) as prediction factors. We used ground information from the International Soil Moisture Network (ISMN, n=13376) for evaluating soil moisture predictions. We provide gap-free mean annual soil moisture predictions, which increase by nearly 50% the spatial

resolution of ESA-CCI soil moisture product. Our predictions showed a statistical accuracy varying 0.69-0.87% and 0.04 m^3/m^3 of cross-validated explained variance and root mean squared error (RMSE). We found no significant differences between the ESA-CCI and our predictions, but we found discrepancy between multiple evaluation metrics (e.g., bias vs efficiency) comparing the ESA-CCI with the ISMN. We found a negative bias (-0.01 to -0.08 m^3/m^3) between the values of ISMN when comparing with the ESA-CCI and our predictions across the analyzed years. A temporal analysis, using a robust trend detection strategy (i.e., Theil-Sen estimator), suggests a decline of soil moisture at the global scale that is consistent in both gridded datasets and field measurements of soil moisture varying from -0.7[-0.77, -0.62]% in the ESA-CCI product, -0.9[-1.01, -0.8]% in the downscaled predictions, and -1.6 [-1.7, -1.5]% in the ISMN. The soil moisture predictions provided here (Guevara, et al., 2019, https://doi.org/10.4211/hs.b940b704429244a99f902ff7cb30a31f) could be useful for quantifying soil moisture spatial and temporal dynamics across areas with low availability of soil moisture information in the original ESA-CCI soil moisture current and future versions.

5.1 Introduction

Assessing the reliability of currently available soil moisture datasets is fundamental for a comprehensive understanding of the global water cycle (Al-Yaari et al., 2019). Soil moisture datasets are useful to characterize hydrological patterns (Greve and Seneviratne, 2015), the influence of soil moisture on terrestrial carbon dynamics

(van der Molen et al., 2011), and global climate variability (Seneviratne et al., 2013). Soil moisture information is used for identifying trends in the water cycle and could be useful to better characterize the response of ecosystems productivity to soil moisture decline (Zhou et al., 2014). However, quantifying the response of ecosystems productivity to soil moisture decline is challenging due to a current lack of accurate and detailed long-term soil moisture datasets across large areas of the world. This lack of soil moisture information can affect our capacity to detect regional-to-global soil moisture trends and could be an important source of uncertainty in global models of land-atmosphere interactions (May et al., 2016).

Large-scale hydrological and ecological analyses (e.g., continental, global) and syntheses of global climate variability benefit from soil moisture information provided by multiple soil moisture monitoring networks (Dorigo et al., 2011a) and from satellite soil moisture measurements (Dorigo et al., 2017; Liu et al., 2011). Field soil moisture measurements (from soil moisture monitoring networks; Figure 5.1) and satellite soil moisture measurements are two main sources of continuous soil moisture information at the global scale used for quantifying regional-to-continental soil moisture patterns and trends.

Field soil moisture measurements are representative of small footprints within specific study sites, at specific soil depths (e.g., 0-5 cm) and the availability of field soil moisture measurements is sparse and limited across large areas of the world. On the other hand, satellite microwave radiometry using L-band (~ 1.4-1.427 GHz) and C-band (~4-8 GHz) for example, is optimal to estimate regional soil moisture (Mohanty et al.,

2017). The number of efforts for providing satellite soil moisture data at the global scale (based on microwave radiometry) have increased during the last decade (Al-Yaari et al., 2019).

Satellite soil moisture datasets are representative for the first few cm of soil depth (e.g., 0-5 cm) and they are provided in grids with spatial resolution varying between 9 and 25 km (Senanayake et al., 2019). This is a range of spatial resolution commonly used in global studies quantifying land and atmosphere interactions (Crow et al., 2012; Jung et al., 2010). However, there are also large areas of the world (across specific environmental conditions) where no satellite soil moisture data are available due to intrinsic sensor limitations (McColl et al., 2017). Collection of field soil moisture information across these missing areas is expensive, time consuming, and in many cases impossible due to logistical reasons. Consequently, many information gaps exist and modeling/validation efforts for predicting soil moisture values across unmeasured areas are required to increase the applicability of soil moisture datasets in studies at multiple scales (Singh et al., 2015).

Currently, the historical soil moisture product from the European Space Agency-Climate Change Initiative (ESA-CCI) provides soil moisture grids at the spatial resolution of ~27x27 km grids (Liu et al., 2011). The ESA-CCI soil moisture product is a synthesis from multiple soil moisture sources and it covers four decades (from the 1978 to 2018) of accurate soil moisture values at the global scale. The ESA-CCI soil moisture product contains uncertainty of measurements and it covers a longer period of time (from 1978 to 2018) compared with other satellite-derived soil moisture products

(Al-Yaari et al., 2019), and is suitable for applications in long-term ecological and hydrological studies (Dorigo et al., 2017).

Across large areas of the world, the ESA-CCI soil moisture product is being validated and calibrated against ground truth (i.e., field) soil moisture measurements (Al-Yaari et al., 2019; Dorigo et al., 2011a). Previous work has included a quality control framework to remove potentially wrong measurements from the ESA-CCI soil moisture product, with the ultimate goal of improving its spatial representativeness and reliability (Gruber et al., 2017). The ESA-CCI soil moisture product is constantly being improved in each released version and we highlight that versions 4.4 and 4.5 have substantial spatial gaps across the world (Figure. S1). Therefore, the development of alternative modeling and validation frameworks is needed to provide new datasets and information to complement the different versions of the ESA-CCI soil moisture product.

To improve the spatial resolution of satellite soil moisture gridded datasets, multiple efforts have used statistical learning methods to couple coarse satellite soil moisture datasets with multiple sources of environmental information at higher spatial resolutions. These sources of environmental information include vegetation indexes (from optical imagery) and climate information (Alemohammad et al., 2018). Chloropeth maps (i.e., land use, land forms) and soil information have also been used as prediction factors to improve the spatial resolution of soil moisture gridded datasets under statistical modeling (e.g., regression) frameworks (Peng et al., 2017). The sources of environmental information are the basis for downscaling satellite soil moisture grids because they are available at higher spatial resolution than the original satellite soil

moisture grids. For example, optical remote sensing (i.e., light detecting and ranging) is able to provide vegetation indexes with a spatial resolution of meters (Dubayah and Drake, 2000), as well as elevation data is now available at higher levels of spatial resolution (e.g., meters; Tadono et al., 2014). For this study, we focus on the potential of using elevation data to represent topographic variability and consequently represent the role of topography in the overall distribution of water across the landscape (Moeslund et al., 2013, Mason et al., 2016, Guevara and Vargas 2019).

Elevation data is the basis of digital terrain analysis to quantify topographic variability and land surface characteristics (e.g., terrain roughness, terrain slope, terrain convexity; Wilson, 2012). Previous studies have found evidence of a topographic signal in satellite soil moisture measurements (Mason et al., 2016). Other studies have highlighted the potential of using these land surface characteristics as prediction factors for soil moisture and developing soil moisture products with higher spatial resolution compared with the original satellite soil moisture measurements (Guevara and Vargas, 2019; Western et al., 2002). Digital terrain analysis are calculations of land surface characteristics that largely depend on topography, such as the terrain slope and aspect, or the topographic wetness index, which is a parameter that characterizes areas where soil moisture could increase by the effect of the overland flow accumulation. The overland flow as well as the potential incoming solar radiation are two important topographic drivers of the spatial distribution of soil moisture (Nicolai-Shaw et al., 2015), its memory after precipitation events (McColl et al., 2017), and its role as a dominant control of plant productivity (Forkel et al., 2015).

Our overarching goal is to test how the application of digital terrain parameters (e.g., terrain slope and aspect, or the topographic wetness index) as predictors of soil moisture information could increase the spatial resolution of global mean annual satellite soil moisture information. The specific objectives are: 1) improve the spatial resolution (from 27 to 15km grids; an improvement of about 50%) of the ESA-CCI soil moisture product (version 4.2 at the annual scale; years 1991- 2016); 2) test how this downscaled mean annual global estimate compares with the statistical distribution of the ESA-CCI soil moisture measurements; and 3) test the consistency of temporal analyses at the global scale using our downscaled product, the ESA-CCI product, and field soil moisture measurements around the world. In this study, we provide a new dataset of gap-free mean annual global soil moisture estimates at 15km resolution for the years 1991-2016 (Guevara, et al., 2019, https://doi.org/10.4211/hs.b940b704429244a99f902ff7cb30a31f).

5.2 Methods

We used a data-driven modelling approach including Geographical Information Systems (GIS) and platforms for geoscientific analysis and statistical computing for combining digital terrain parameters, soil moisture gridded estimates and soil moisture tabular datasets. Field information of soil moisture is used for validating our soil moisture predictions based on one data driven 115 model for each year of available satellite soil moisture data in the ESA-CCI.

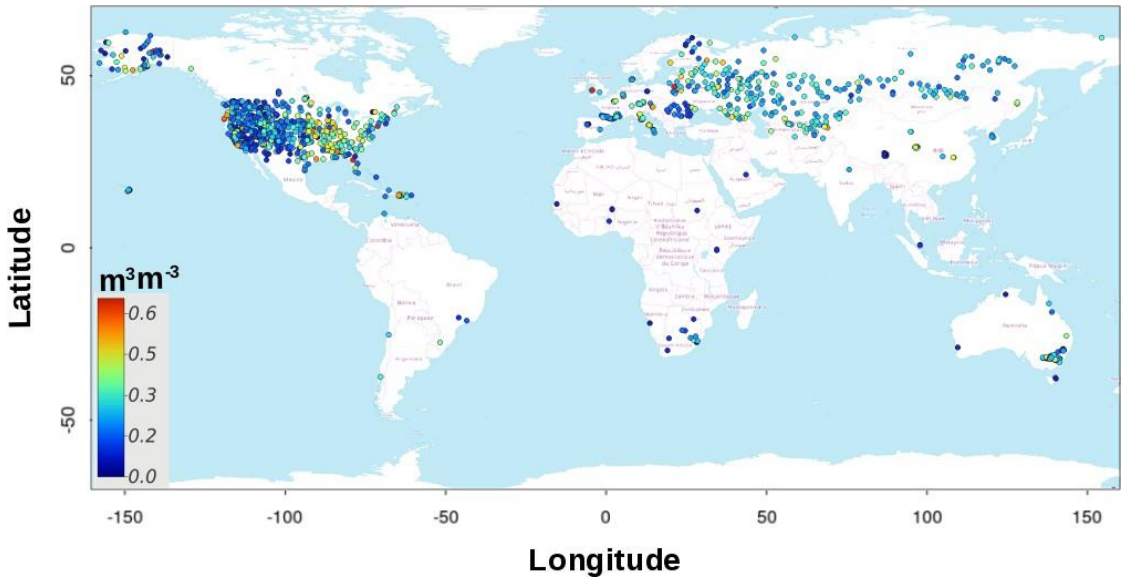

Figure 5.1. Data distribution of the ISMN dataset (n=13376) available for the period 1991-2016. Values represent overall mean values at each location for the period 1991-2016.

5.2.1 Datasets

This study is based on the analysis of ESA-CCI soil moisture measurements (version 4.2, years 1991-2016; Figure 5.2). We postulate that mean annual soil moisture (for any specific year) can be predicted using a regression model and digital terrain parameters (derived from elevation data) as prediction factors. In this regression framework, soil moisture is represented by the yearly mean of soil moisture values (for each calendar year) at the central coordinates (latitude and longitude) of pixels from the ESA-CCI soil moisture measurements. A similar approach was applied at the regional-scale within the conterminous United States providing an improvement in spatial

resolution and ground truth validation of satellite soil moisture derived from the ESA-CCI (Guevara and Vargas 2019).

The explanatory variables for soil moisture were represented with the values of the terrain parameters for the locations of the aforementioned central coordinates of the original ESA-CCI soil moisture measurements. Therefore, soil moisture for each model/year can be predicted at finer spatial resolutions defined by the spatial resolution of the digital terrain parameters (Guevara and Vargas, 2019). We first harmonized (e.g., same geo-spatial reference, same extent) the ESA-CCI soil moisture and topographic data (elevation and derived terrain parameters) using open source geographic information systems (Hijmans, 2019). The digital terrain parameters (Figure 5.3) were derived from elevation data in SAGA-GIS (System for Automated Geoscientific Analysis-GIS) (Conrad et al., 2015). The source of elevation data was a radar based digital elevation model (Becker et al., 2009) that we resampled to a spatial resolution of 15 km across the world. We recognize that this resolution is still too coarse to represent the local variability of soil moisture but this dataset has two advantages: 1) it is nearly a 50% improvement of spatial resolution when compared with the original ESA-CCI (~27 km grids) soil moisture product, and 2) our framework produced a gap-free global annual soil moisture estimate at 15 km resolution and it is theoretically applicable for predictions at higher temporal resolution (e.g., of months, weeks or days).

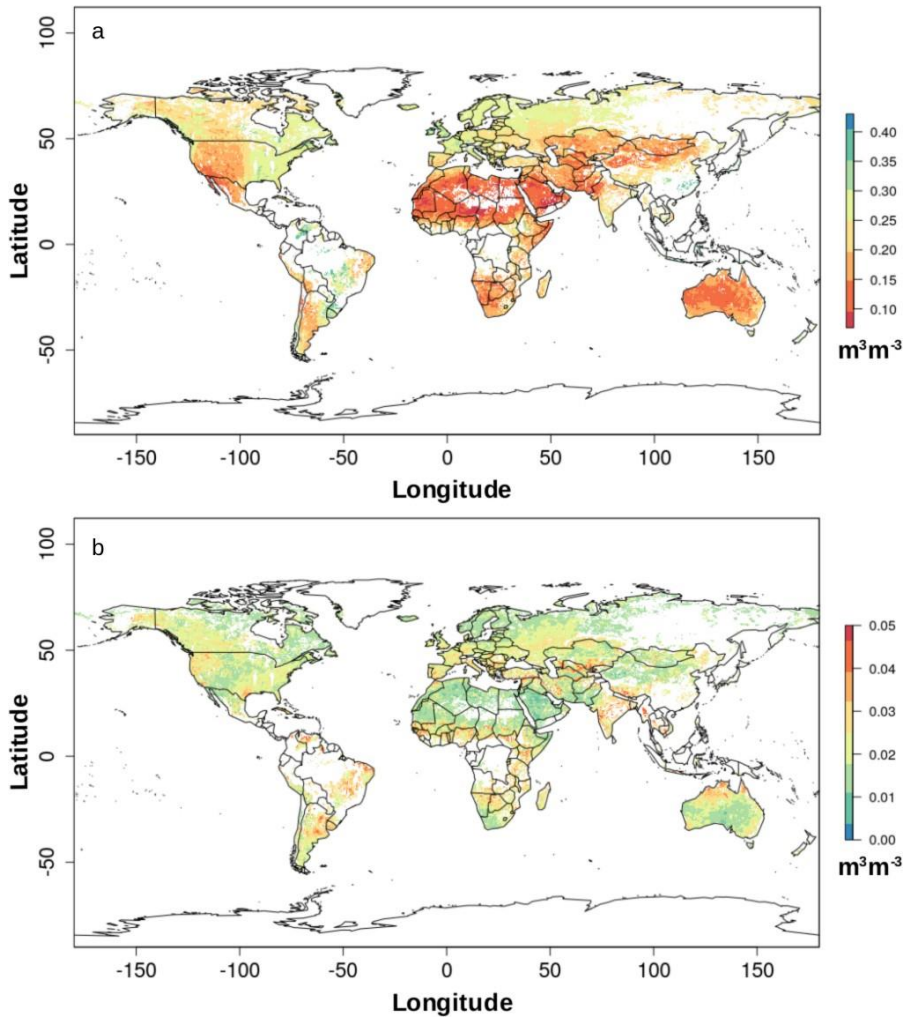

Figure 5.2. ESA-CCI soil moisture mean (a) and standard deviation (b) for the period 1991-2016 from ESA-CCI soil moisture version 4.2. White areas are areas where no complete information is available during the analyzed period. The black line shows geopolitical borders.

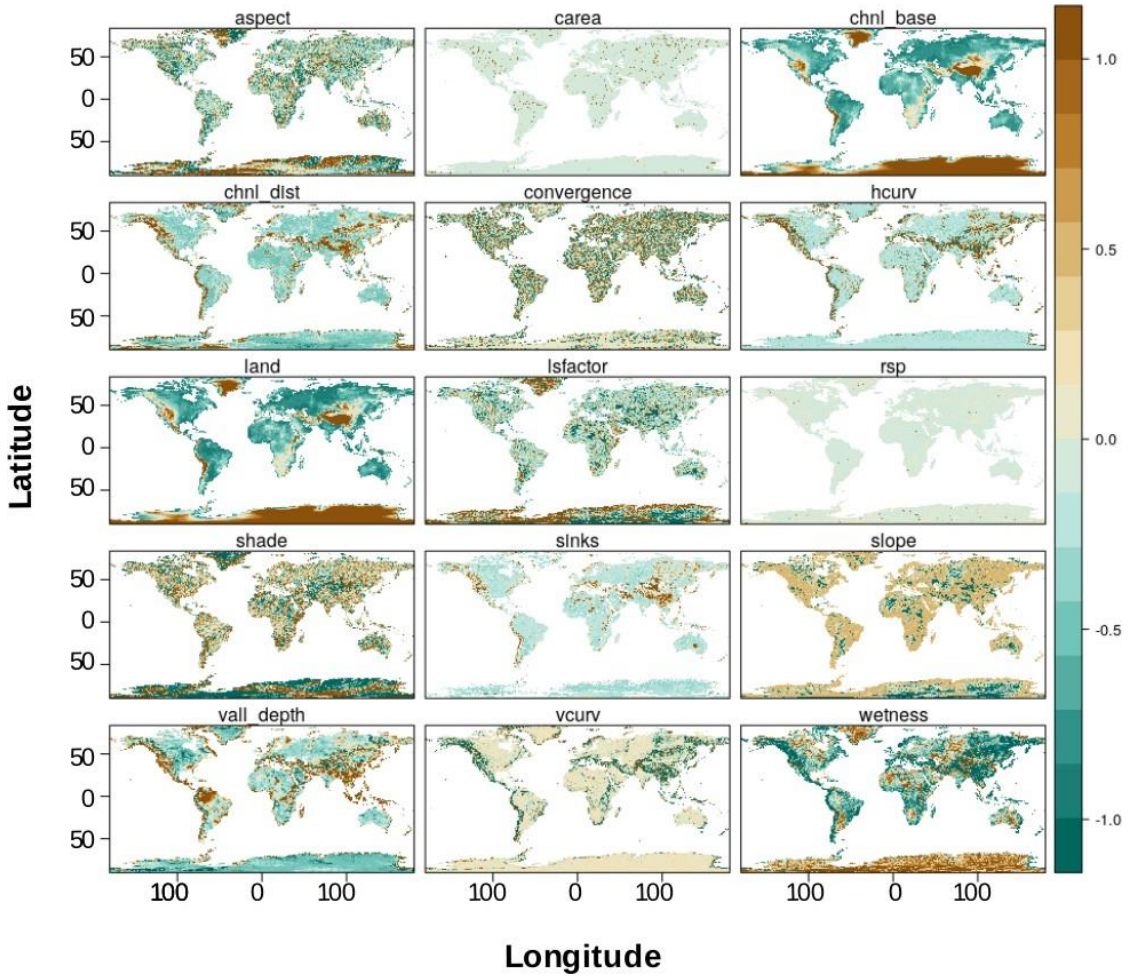

Figure 5.3. Topographic terrain parameters that were derived from the DEM using SAGA-GIS. These terrain parameters were used as prediction factors for the values of the ESA-CCI soil moisture product. These terrain parameters were standardized by centering their means in 0 by a variance unit for improving visualization purposes. a) aspect: terrain aspect, b) carea: specific catchment area, c) chnl base: channel network base level, d) chnl dist: distance to channel network, e) convergence: flow convergence index, f) hcurv: horizontal curvature, g) land: digital elevation model, h) lsfactor: length-slope factor, i) rsp: relative slope position, j) shade: analytical hillshading, k) sinks: smoothed elevation, l) slope: terrain slope, m) vall depth: valley depth index, n) vcurv: vertical curvature, o) wetness: topographic wetness index.

5.2.2 Refinement modeling

To predict soil moisture at a finer spatial resolution (15 km) than the original ESA-CCI product (27 km), we used a machine learning (ML) kernel-based method known as weighted Kernel Nearest Neighbors (KKNN; Hechenbichler and Schliep, 2004). This ML method was used to account for likely non-linear relationships between soil moisture and terrain parameters at the global scale. The KKNN method is a time efficient algorithm compared with more complex ML algorithms (i.e., tree- based, deep learning based). It is a pattern recognition technique based on multiple data neighbors to account also for variations in the relationship of soil moisture with its explanatory variables from one place to another. The KKNN algorithm has two main user defined parameters, the parameter *k* determines the number of neighbors from which information will be considered for prediction. The second parameter is a kernel function (e.g., triangular, epanechnikov or Gaussian among others) that allows to convert distances into weights (the farther the neighbor, the smaller the weight it will be assigned) that can then be used to take a weighted vote or a weighted average respectively for classification (i.e., predicting categorical variables such as soil type) of for regression problems (i.e., predicting continuous variables such as soil moisture).

5.2.3 Model parameter selection

For each yearly mean, each model was first parameterized (selection of optimal parameters for each model/year) using cross validation and folds of 10 % of available data out of each iteration. The cross-validation indicators (information criteria about model performance) were the Pearson correlation coefficient (r) and the root mean

squared error (RMSE) for each modeled yearly mean. Using the combination of *k* and kernel function of the model generating the lowest RMSE and highest r for each model/year, we predicted soil moisture at the resolution of the aforementioned terrain parameters (i.e., 15 km) at the global scale.

Then, the resulting soil moisture predictions and the ESA-CCI soil moisture product were validated against the available soil moisture data reported in the International Soil Moisture Network (ISMN; Dorigo et al., 2011a, 2017) for each annual mean (1991-2016). We extracted the values of soil moisture gridded measurements to the locations on the ISMN available data. A total of 8080 tables with soil moisture information with multiple data sizes (from multiple contributing networks) were extracted from the ISMN for the analyzed period of time (1991-2016). We also provided this information harmonized in an annual basis, including the values of the ESA-CCI and our predictions for the locations of the ISMN dataset (see section 5).

5.2.4 Assessment metrics

The ISMN data was used to calculate multiple model evaluation indicators (see Carslaw ,2015) for comparing the ESA-CCI soil moisture product and the soil moisture predictions based on digital terrain analysis approach. These evaluation statistics were performed using the *openair* package of the R software (Carslaw and Ropkins, 2012). These evaluation statistics included the number of complete pairs of data (n) and the Pearson correlation coefficient (r) between predicted and observed values as well as systematic error indicators:

- MB, the mean bias;

- MGE, the mean gross error;
- NMB, the normalized mean bias;
- NMGE, the normalized mean gross error;
- RMSE, the root mean squared error;

The fraction of predictions within a factor of two of the observed values (FAC2) was another evaluation indicator included in our analysis. The FAC2 is a robust metric for model evaluation because it is not overly influenced by outliers (Chang and Hanna, 2004). Other evaluation indicators that are not sensitive to outliers and extreme values were also included:

- The Coefficient of Efficiency (COE, based on Legates and McCabe, (1999) and Legates and McCabe (2013)). The COE has been used widely used to evaluate the performance of hydrological models. A perfect model has a COE =A value of COE= 0 or negative implies low prediction capacity;
- IOA, the Index of Agreement (Willmott, Robeson, and Matsuura 2012), which spans between -1 and 1 and with values approaching +1 representing better model performance. The IOA indicates the proportion of the sum of the error magnitudes in relation to the sum of the observed-deviation magnitudes.

By interpreting the aforementioned model evaluation indicators, a perfect model would have a FAC2, r, COE and IOA ~ 1.0, while all the others ~ 0. These metrics represent a valuable set of information criteria for comparing both the ESA-CCI soil moisture product and the soil moisture predictions based on digital terrain analysis against ground data from the ISMN.

5.2.5 Trend detection

We also performed a non-parametric trend detection test (i.e., Theil-Sen estimator) to compare soil moisture trends between the ESA-CCI soil moisture product and the downscaled soil moisture predictions based on digital terrain analysis (yearly means 1991-2016). The same trend detection test was applied to the field information contained in the ISMN dataset for comparative purposes. A pixel-wise trend detection test was also applied to search for changes in the regression relationship (i.e., soil moisture ~ years) using different regression parameters before and after any possible breakpoint. A minimum of four years are required between break points for detecting trends and segments between break points with less than eight observations are not considered. Therefore, this method is considered to provide a robust trend detection estimate (Forkel et al., 2013, 2015).

5.3 Results

We provided a dataset of gap-free downscaled mean annual soil moisture predictions at the global scale at 15 km grids for years 1991-2016. These predictions are based on ML and digital terrain parameters and they are provided in a generic raster format (Guevara, et al., 2019, https://doi.org/10.4211/hs.b940b704429244a99f902ff7cb30a31f). Supporting the reliability of our prediction framework, the original ESA-CCI and our downscaled soil moisture predictions based on digital terrain analysis showed consistently a global mean annual value of 0.19 m^3/m^3 and a standard deviation of 0.6 m^3/m^3 considering the mean value of all analyzed years. The downscaled soil moisture predictions showed a similar

bimodal distribution compared with the original ESA-CCI soil moisture product (Figure 5.4). The ISMN showed also a bimodal distribution but a larger range (>50%) of soil moisture values compared with both soil moisture gridded measurements (Figure 5. 4). The ISMN values (in an annual basis) showed a mean value of 0.24 and a standard deviation 0.12 m^3/m^3. Thus, the complete range of soil moisture variability in the ISMN is higher than the ESA-CCI satellite measurements and therefore is also higher than the downscaled soil moisture predictions based on digital terrain analysis (Figure 5. 4).

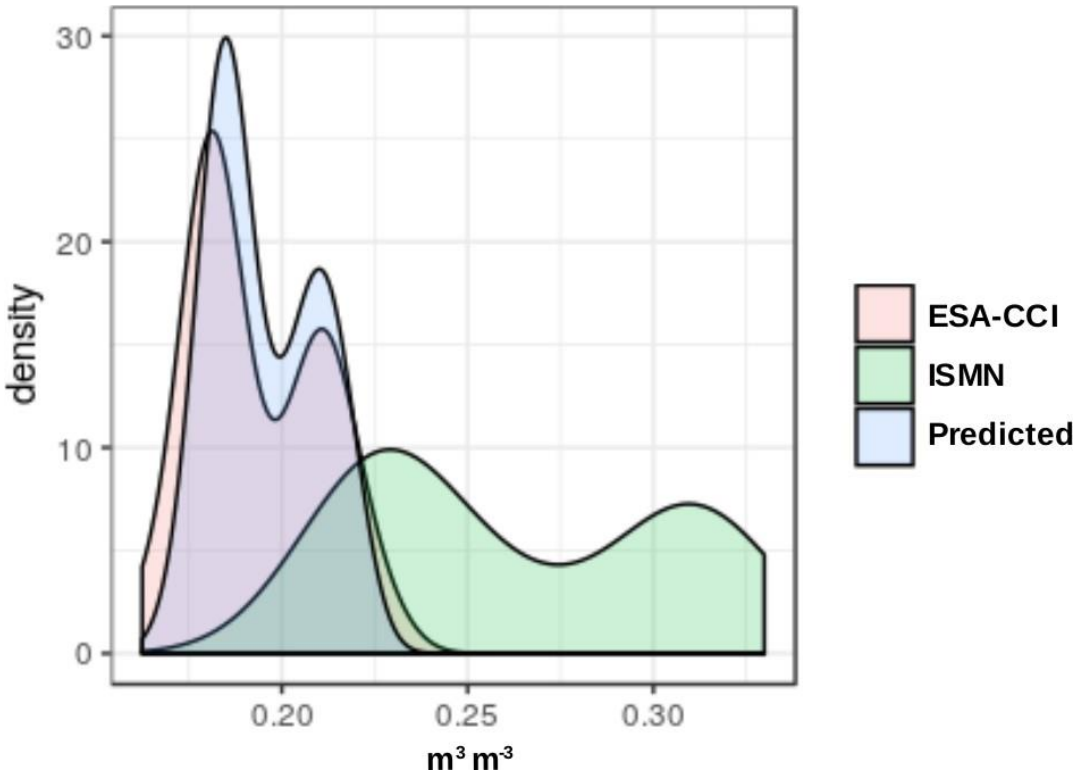

Figure 5.4. Probability distribution functions showing the statistical distribution of soil moisture gridded datasets (i.e., ESA-CCI, and our soil moisture product [Predicted]) and soil moisture observations from the ISMN.

5.3.1 Model parameter selection

The cross-validated Pearson correlation coefficient (r) between observed and predicted soil moisture across the analyzed years varied from 0.78 to 0.81, demonstrating a reliable prediction capacity. The RMSE varied from 0.03 to 0.04 m^3/m^3, in all cases below the first quartile of the ESA-CCI soil moisture training data distribution (0.15 m^3/m^3, Table 5.1).

Table 5.1 Cross validated correlation (r), RMSE, training data pixels (n), the kernel type, and the number of neighbors of the parameter k in the soil moisture prediction models for each year.

	year	**r**	**rmse**	**n**	**kernel**	**k**
1	1991	0.78	0.0448	195886	triangular	20
2	1992	0.81	0.0404	198624	triangular	19
3	1993	0.81	0.0405	198144	triangular	19
4	1994	0.81	0.0408	198458	triangular	19
5	1995	0.82	0.0400	198595	triangular	18
6	1996	0.81	0.0398	198752	triangular	18
7	1997	0.81	0.0398	198655	triangular	19
8	1998	0.81	0.0409	199076	triangular	18
9	1999	0.81	0.0407	199067	triangular	18
10	2000	0.81	0.0404	199089	triangular	19
11	2001	0.80	0.0422	199008	triangular	19
12	2002	0.81	0.0407	208167	triangular	19
13	2003	0.78	0.0441	178580	triangular	21
14	2004	0.83	0.0374	152696	triangular	19
15	2005	0.82	0.0398	173179	triangular	18
16	2006	0.80	0.0405	176761	triangular	19
17	2007	0.80	0.0398	208161	triangular	18
18	2008	0.81	0.0393	209543	triangular	18
19	2009	0.81	0.0391	209276	triangular	19
20	2010	0.79	0.0398	211717	triangular	18
21	2011	0.80	0.0391	211581	triangular	19
22	2012	0.80	0.0399	212042	triangular	18
23	2013	0.81	0.0400	209584	triangular	18
24	2014	0.80	0.0403	209573	triangular	18
25	2015	0.80	0.0402	209590	triangular	19

5.3.2 Evaluation against field data

The validation of the ESA-CCI soil moisture product and the ISMN (Table 5.2) showed relatively similar results compared to the validation of the downscaled soil moisture predictions based on digital terrain analysis and the ISMN (Table 5.3). In Table 5.2 and Table 5.3 the calculated evaluation statistics (n, FAC2, MB, MGE, NMB, RMSE, r, COE, and IOA) are provided.

Table 5.2 and Table 5.3 provide model evaluation statistics (agreement metrics between observed and modeled) for 26 different yearly models for the years 1991 to 2016. The number of complete pairs of data available for validation (n) increases with time. (i.e., respectively for Table 5.2 and Table 5.3, we counted in 102 and 104 available pairs of points for 1991 while for 2016 these numbers increased to 1165 and 1194 pairs of points). The higher data density (e.g., with >1000 spatial coordinates with soil moisture from the ISMN and pixel soil moisture values from the ESA-CCI) was found between 2011 and 2016 (Table 5.2). In all cases, the evaluation statistics are equal or better for the downscaled soil moisture predictions based on digital terrain analysis (Table 5.3) than the original ESA-CCI soil moisture product (Table 5.2).

Evaluation statistics such as FAC2, RMSE, r, COE and IOA showed variability across the analyzed years (Table 5.2 and Table 5.3). The FAC2 from 1991 to 2016 never falls below 0.70. This implies that no less than 70% of the predicted values were within a factor of 2 of the observed values. For the ESA-CCI soil moisture product, the lowest FAC2 value was 0.70 for 2012 and the highest value was 0.93 for 2000. These

numbers were consistent for the downscaled predictions based on digital terrain analysis (0.79 and 0.94 respectively for the same years). From both Table 5.2 and Table 5.3 the RMSE varied from 0.09 to 0.13 m^3/m^3. We found in all cases a negative mean bias, confirming that the ESA-CCI soil moisture product (and consequently our predictions) tend to underestimate the values of yearly soil moisture means from field measurements in the ISMN. The values of the mean bias varied from -0.01 to -0.1. Lowest bias was found for 2013 and highest bias for 1996. For the models after 1997, the r values range from 0.30 to 0.60. These values indicate a weak to moderate positive relationship between soil moisture gridded measurements and field soil moisture data. Before 1998, values on both tables indicate weak to no linear relationship. In all cases, we also found that the COE was less than 0 from 1991 to 2001. This suggests that the simple observed mean has a better agreement than the ESA-CCI with the values of the ISMN. For years 2001 onward, the models showed a relatively higher COE (Table 5.3). For the IOA values in both tables, all of the 26 values are above 0, and range from 0.10 to 0.57. Therefore, this indicates a low to moderate, positive correlation between gridded soil moisture measurements and field soil moisture data. From this evaluation section, our results show that the original ESA-CCI soil moisture product tends to underestimate the values of the ISMN. The downscaled predictions based on digital terrain analysis are not significantly different compared with the ESA-CCI soil moisture product when evaluated against observed field values from the ISMN, but they provide 1) gap free soil moisture-related information and 2) higher spatial resolution (from 27 to 15 km grids).

Table 5.2. Agreement metrics between the ISMN dataset and the ESA-CCI soil moisture product at the annual scale. n, the number of complete pairs of data; FAC2, fraction of predictions within a factor of two; MB, the mean bias; MGE, the mean gross error; NMB, the normalized mean bias; NMGE, the normalized mean gross error; RMSE, the root mean squared error; r, the Pearson correlation coefficient; COE, the Coefficient of Efficiency; IOA, the Index of Agreement.

year	n	FAC2	MB	MGE	NMB	NMGE	RMSE	r	COE	IOA
1991	102	0.892	-0.069	0.081	-0.241	0.282	0.105	0.091	-0.499	0.251
1992	123	0.837	-0.077	0.091	-0.269	0.316	0.115	0.165	-0.455	0.272
1993	113	0.903	-0.076	0.085	-0.251	0.282	0.108	0.234	-0.466	0.267
1994	113	0.823	-0.088	0.093	-0.292	0.308	0.119	0.055	-0.782	0.109
1995	94	0.851	-0.078	0.090	-0.272	0.317	0.110	0.251	-0.478	0.261
1996	118	0.847	-0.099	0.106	-0.332	0.357	0.128	0.252	-0.644	0.178
1997	121	0.909	-0.080	0.098	-0.274	0.339	0.115	0.319	-0.433	0.284
1998	119	0.899	-0.082	0.098	-0.282	0.334	0.117	0.357	-0.380	0.310
1999	72	0.903	-0.055	0.091	-0.203	0.337	0.103	0.589	-0.012	0.494
2000	82	0.927	-0.062	0.089	-0.228	0.328	0.104	0.522	-0.067	0.467
2001	133	0.880	-0.039	0.085	-0.163	0.357	0.102	0.465	0.077	0.538
2002	205	0.883	-0.038	0.078	-0.168	0.349	0.094	0.581	0.126	0.563
2003	299	0.796	-0.032	0.083	-0.148	0.385	0.102	0.470	0.108	0.554
2004	380	0.845	-0.047	0.081	-0.205	0.351	0.105	0.433	0.041	0.521
2005	469	0.806	-0.033	0.083	-0.156	0.390	0.109	0.271	0.020	0.510
2006	507	0.805	-0.033	0.080	-0.159	0.381	0.102	0.308	0.008	0.504
2007	573	0.859	-0.028	0.073	-0.133	0.350	0.090	0.452	0.059	0.529
2008	613	0.853	-0.041	0.081	-0.187	0.368	0.106	0.453	0.065	0.532
2009	696	0.876	-0.035	0.079	-0.157	0.358	0.104	0.468	0.089	0.544
2010	900	0.839	-0.029	0.078	-0.135	0.363	0.107	0.331	0.041	0.521
2011	1031	0.826	-0.032	0.080	-0.152	0.375	0.112	0.334	0.046	0.523
2012	1318	0.705	-0.045	0.091	-0.218	0.439	0.126	0.200	-0.055	0.472
2013	1242	0.812	-0.019	0.076	-0.099	0.387	0.107	0.433	0.133	0.566
2014	1237	0.840	-0.026	0.075	-0.126	0.368	0.103	0.469	0.129	0.565
2015	1253	0.812	-0.034	0.083	-0.161	0.388	0.119	0.415	0.126	0.563
2016	1194	0.845	-0.031	0.082	0.145	0.378	0.118	0.366	0.086	0.543

Table 5.3. Agreement metrics between the ISMN dataset and the downscaled soil moisture predictions based on digital terrain analysis. n, the number of complete pairs of data; FAC2, fraction of predictions within a factor of two; MB, the mean bias; MGE, the mean gross error; NMB, the normalized mean bias; NMGE, the normalized mean gross error; RMSE, the root mean squared error; r, the Pearson correlation coefficient; COE, the Coefficient of Efficiency; IOA, the Index of Agreement.

year	n	FAC2	MB	MGE	NMB	NMGE	RMSE	r	COE	IOA
1991	104	0.904	-0.078	0.084	-0.270	0.292	0.108	0.077	-0.561	0.220
1992	125	0.880	-0.084	0.094	-0.292	0.326	0.118	0.120	-0.509	0.245
1993	115	0.887	-0.085	0.093	-0.280	0.306	0.116	0.112	-0.597	0.202
1994	115	0.826	-0.092	0.095	-0.302	0.315	0.117	0.133	-0.806	0.097
1995	96	0.865	-0.085	0.094	-0.296	0.328	0.115	0.214	-0.528	0.236
1996	119	0.849	-0.099	0.106	-0.333	0.355	0.127	0.276	-0.638	0.181
1997	122	0.910	-0.081	0.098	-0.279	0.337	0.117	0.295	-0.427	0.287
1998	120	0.900	-0.082	0.096	-0.279	0.326	0.116	0.372	-0.349	0.325
1999	73	0.904	-0.051	0.092	-0.189	0.344	0.103	0.601	-0.015	0.492
2000	84	0.940	-0.056	0.089	-0.208	0.328	0.102	0.560	-0.047	0.477
2001	136	0.904	-0.031	0.083	-0.129	0.348	0.096	0.544	0.106	0.553
2002	211	0.872	-0.030	0.079	-0.134	0.355	0.094	0.568	0.127	0.564
2003	307	0.798	-0.027	0.084	-0.128	0.391	0.101	0.492	0.108	0.554
2004	389	0.871	-0.040	0.081	-0.173	0.354	0.103	0.411	0.045	0.522
2005	483	0.810	-0.032	0.083	-0.150	0.388	0.107	0.331	0.031	0.516
2006	522	0.830	-0.031	0.080	-0.148	0.376	0.101	0.353	0.035	0.517
2007	586	0.884	-0.026	0.073	-0.127	0.351	0.090	0.454	0.059	0.530
2008	626	0.874	-0.041	0.082	-0.187	0.371	0.107	0.456	0.056	0.528
2009	710	0.885	-0.034	0.081	-0.154	0.365	0.106	0.449	0.070	0.535
2010	911	0.856	-0.028	0.080	-0.129	0.368	0.106	0.333	0.030	0.515
2011	1051	0.829	-0.031	0.082	-0.143	0.384	0.112	0.301	0.024	0.512
2012	1341	0.786	-0.035	0.084	-0.166	0.405	0.115	0.299	0.028	0.514
2013	1273	0.807	-0.012	0.079	-0.061	0.402	0.108	0.380	0.096	0.548

2014	1268	0.834	-0.020	0.079	-0.098	0.382	0.106	0.380	0.091	0.546
2015	1283	0.812	-0.028	0.086	-0.129	0.401	0.121	0.341	0.093	0.547
2016	1194	0.845	-0.031	0.082	0.145	0.378	0.118	0.366	0.086	0.543

5.3.3 Trend detection results

At the locations with available field soil moisture information from the ISMN, we found a consistent soil moisture decline. This result was consistent when considering the values of the same locations for both tested gridded soil moisture datasets (i.e., ESA-CCI and the downscaled soil moisture predictions based on digital terrain analysis predictions). Overall, the trend detection test suggests a consistent decline of soil moisture at the global scale (Figure 5.6). The confidence intervals of the detected trends overlap between both gridded datasets, but both gridded datasets significantly underestimate the soil moisture decline detected from the field soil moisture measurements contained in the ISMN. These detected trends (using yearly means between 1991-2016) decreased from -1.6 [-1.7, -1.5]% in the field measurements of the ISMN (Figure 5.6a), to -0.9[-1.01, -0.8]% in the downscaled soil moisture predictions based on digital terrain analysis (Figure 5.6b), to -0.7[-0.77, -0.62]% in the ESA-CCI soil moisture product (Figure 5.6c). Thus, supporting the reliability of our prediction framework, we found that the trend calculated using the downscaled soil moisture predictions is closer to the trend of the ISMN data than the trend of the ESA-CCI soil moisture values (Figure 5.6).

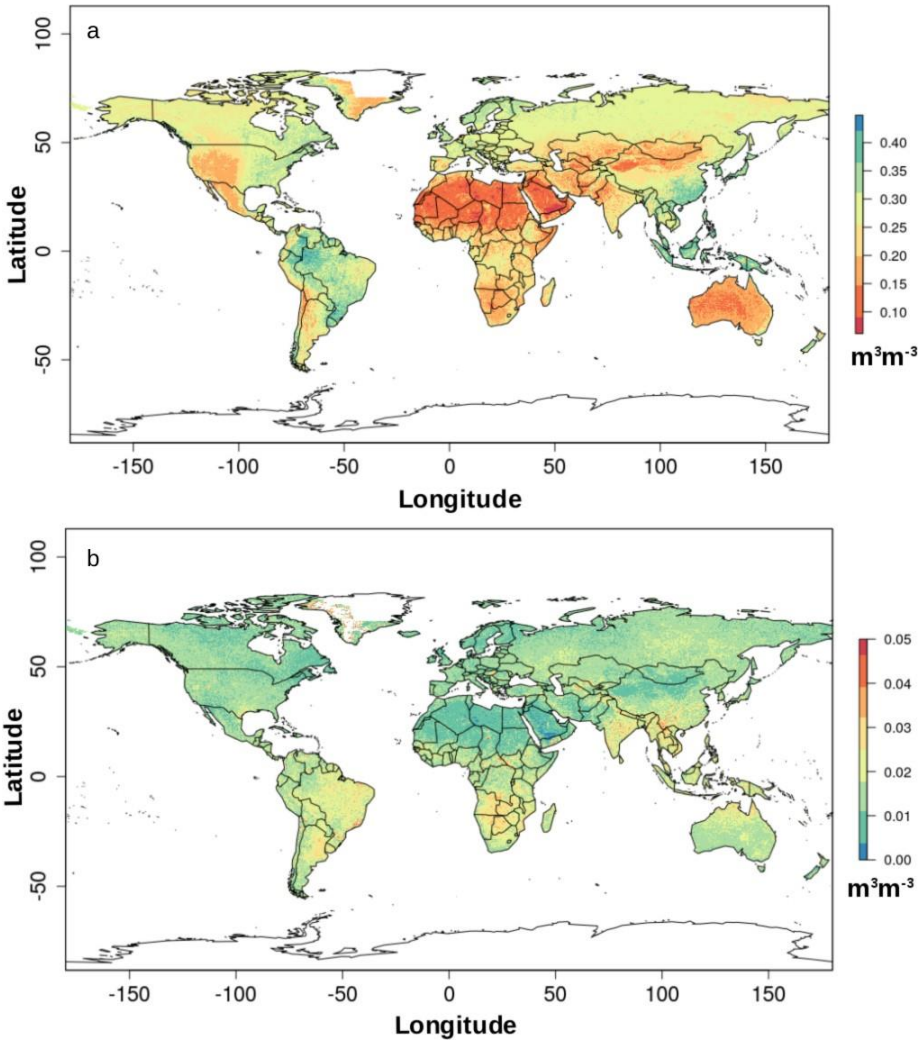

Figure 5.5. Predicted soil moisture mean based on digital terrain parameters (a) and standard deviation (b) for the period 1991-2016. The black line shows geopolitical borders

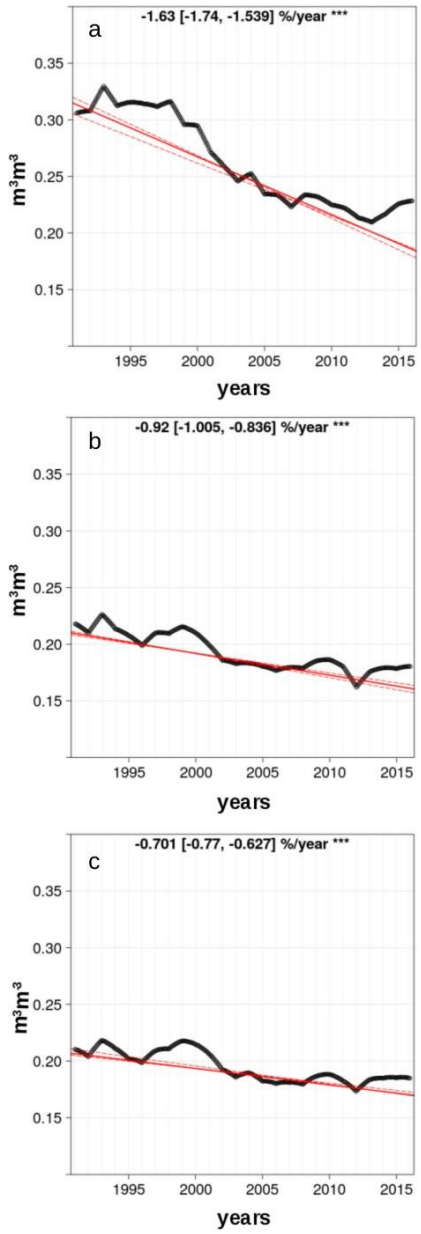

Figure 5.6. Trend detection results for soil moisture. Trends of the ISMN dataset (a), the soil moisture predictions based on digital terrain analysis (b) and the ESA-CCI soil moisture product (c).

Both soil moisture gridded products (ESA-CCI and the downscaled soil moisture predictions based on digital terrain analysis predictions) showed a variety of pixels showing both positive and negative trends during the analyzed period of time at the global scale (Figure 5.7). At the global scale, from the original ESA-CCI soil moisture product we found that 25.16% of pixels (53374 of 212141 total pixels of 27x27km) with available data during 1991 and 2016 showed significant positive and negative trends (using a probability threshold of 0.05). For the soil moisture predictions based on digital terrain analysis this value showed a similar percentage (26.16%, 368614 of 1409020 total pixels) of pixels showing positive and negative trends (same probability threshold of 0.05) during the analyzed period. The downscaled soil moisture product based on digital terrain analysis revealed a larger area of significant soil moisture decline (significant negative trend, probability threshold <0.05). The ESA-CCI soil moisture product showed 16635 pixels of 27 km with significant positive trend (12126915 km^2) and 36739 pixels (26782731 km^2) showing significant negative trend (Figure 5.7a). However, in the downscaled soil moisture product a total of 103863 pixels of 15 km of spatial resolution showed a significant positive trend (23369175 km^2), while 264751 pixels of the same dimensions (59568975 km^2) showed a significant negative trend (Figure 5.7b). Thus, the soil moisture decline detected using the downscaled product occupies an area >2 times larger than the area where soil moisture decline was detected using the ESA-CCI soil moisture product.

The downscaled soil moisture predictions based on digital terrain analysis are useful for quantifying soil moisture trends across areas where no soil moisture

information for long periods of time is otherwise available (Figure 5.8). These soil moisture predictions revealed a negative trend across tropical rain forests of the Amazon basin in Latin America, and the Congo region in Africa (Figure 5.7b). Across these areas, there are still large spatial gaps of information in the original ESA-CCI satellite product due to intrinsic sensor limitations and therefore trends and spatial patterns cannot be resolved across these areas using the ESA-CCI satellite product on its version 4.2 (Figure 5.7a) or 4.4 (Supplementary Figure S1).

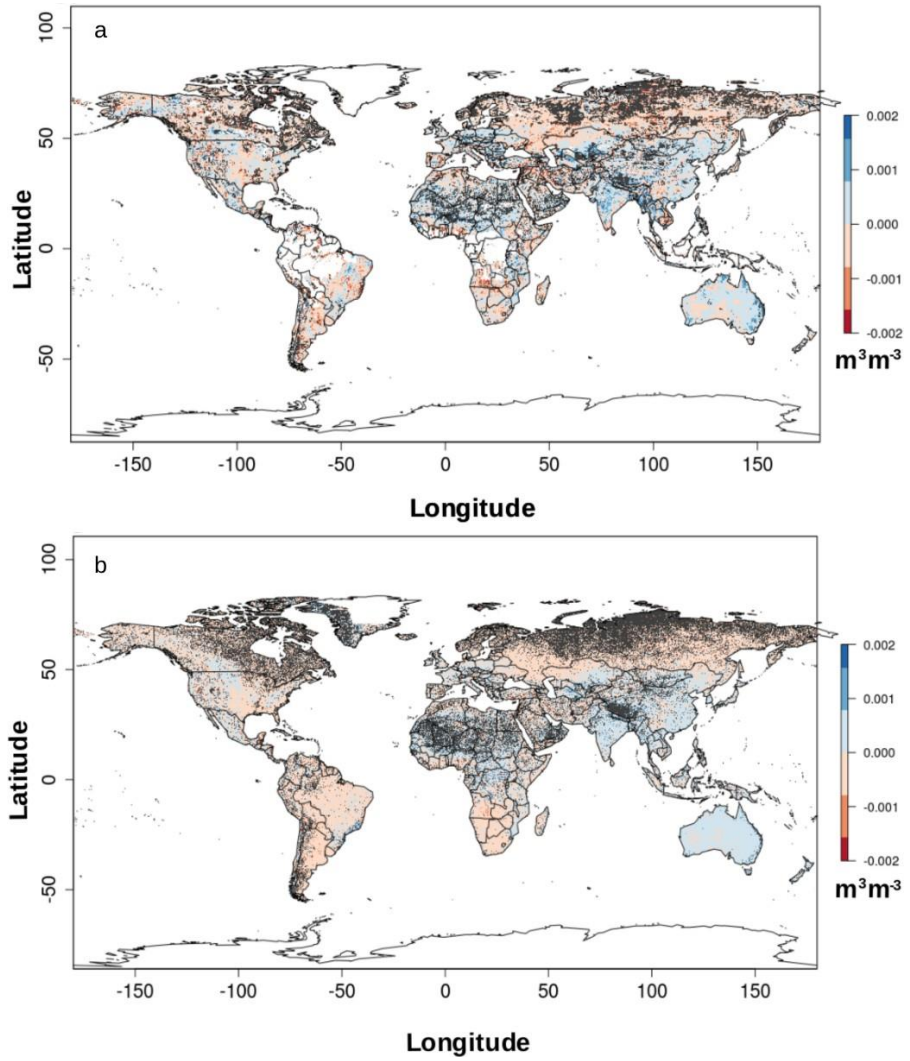

Figure 5.7. Pixel based soil moisture annual trend based on the ESA-CCI soil moisture product (a) and soil moisture annual trends from the soil moisture predictions based on digital terrain analysis (b) for the period 1991-2016. The black line shows geopolitical borders. Dark gray areas are areas where no significant trend (p-value >0.05) was detected.

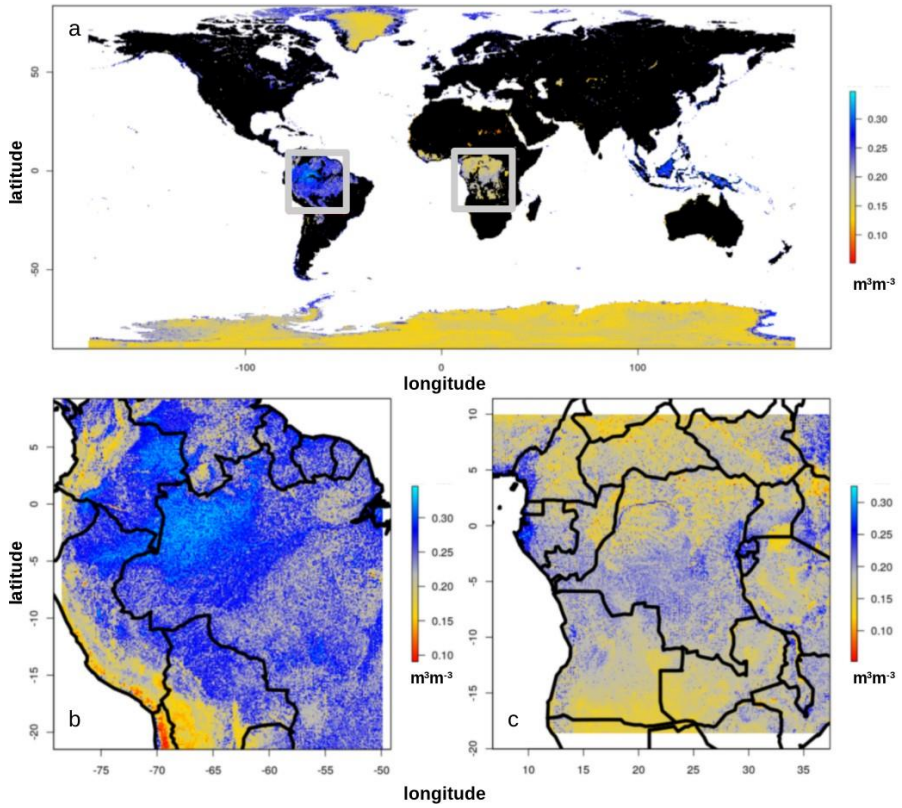

Figure 5.8. Soil moisture predicted across areas with gaps in the ESA-CCI soil moisture product. Black areas in the global map indicate the areas that are well covered by the ESA-CCI soil moisture product while the colored areas are the predictions of soil moisture based on geomorphometry and machine learning (a). We show mean annual soil moisture predictions across areas with no data in the ESA-CCI soil moisture product across the Amazon (b) and the Congo (c) basins.

5.4 Discussion

We developed a regression framework coupling ML and digital terrain analysis for improving (by nearly 50%) the spatial resolution of satellite soil moisture datasets from the ESA-CCI soil moisture product at the global scale. We provided a gap- free

annual mean soil moisture dataset at the global scale for years 1991-2016 using a spatial resolution of 15 km grids. This dataset could prove useful for analyzing the spatial variability of long-term drought scenarios associated with soil moisture decline (Berg and Sheffield, 2018) and its implications in global hydrological models (Zhuo et al., 2016), climate change predictions (Samaniego et al., 2018), carbon cycling models (Green et al., 2019), or for monitoring famine spatial relationships with water scarcity for crops and human use (Mishra et al., 2019).

We provided soil moisture information across areas where no information of the ESA-CCI soil moisture product is available. These results could contribute with the validation and calibration of current initiatives for improving the spatial representativeness and data quality of the ESA-CCI soil moisture product (Gruber et al., 2017). To predict soil moisture across these areas with gaps in the ESA-CCI soil moisture product (Figure 5.7a) we assumed that soil moisture (each annual mean soil moisture estimate) can be predicted as a function of topography (Guevara and Vargas 2019). This is based on physical principles because topography is a major control of the water distribution in the landscape and topography determines the angle between satellite soil moisture sensors and the earth surface. Thus, our results bring attention for the potential of using digital terrain parameters (i.e., surrogates of topography) for improving the spatial resolution and gap- filling of the current satellite-derived soil moisture products.

The accuracy of our downscaled soil moisture predictions (Table 5.3) is consistent with the accuracy of the original ESA-CCI soil moisture product (Table 5.2).

However, we found that the extreme values (minimum and maximum values) in the ISMN dataset in an annual basis are underestimated by the ESA-CCI soil moisture product (Figure 5.4). This subestimation, as explained in previous work (McColl et al., 2017, Liu et al., 2019), is because satellite soil moisture sensors are not able to provide accurate estimates across extremely dry conditions or across areas where water aboveground is higher than water belowground (e.g., extremely humid conditions). For a better calibration and understanding the main limitations of satellite soil moisture measurements across multiple environmental conditions, there is an increasing number of studies reporting validation performances across multiple scales (spatial and temporal) of available soil moisture satellite datasets (An et al., 2016; Colliander et al., 2017b; Dorigo et al., 2011b; Minet et al., 2012; Mohanty et al., 2017; Yee et al., 2016). Several reports have shown similar prediction bias (e.g., similar RMSE values) (Al-Yaari et al., 2019; Colliander et al., 2017a) as the results of this study. Thus, we contribute with a modeling framework supported by the direct correlation between topography and satellite soil moisture (Mason et al., 2016) that brings positive implications to increase the accuracy of satellite soil moisture measurements.

Increasing the accuracy of satellite soil moisture measurements is critical for improving interpretations about soil moisture spatial and temporal dynamics. Thus, multiple alternative modeling evaluation frameworks and model evaluation statistics are required (or could be useful) to better interpret the spatial variability and dynamics of global soil moisture (McColl et al., 2017). The use of multiple evaluation statistics is important because there is not a single best measure, and it is necessary to use a

combination of these performance measures for model evaluation purposes (Chang and Hanna, 2004). This is specifically important when working with limited spatial data for validation of global models based on coarse scale soil moisture gridded measurements.

Our model evaluation is based on a set of multiple evaluation statistics that allow a better understanding of the discrepancies and the sources of the discrepancies between the ISMN and both soil moisture gridded datasets (ISMN vs ESA-CCI and ISMN vs the downscaled soil moisture predictions). The sources of discrepancies can be associated with the spatial representation of multiple soil moisture products (e.g., points vs grids) and their spatial representativeness (Nicolai-Shaw et al., 2015). While the ESA-CCI product shows a coarser spatial resolution and large areas with no available data (Figure 5.2), then the ISMN shows a sparse distribution of available data. In addition, most of the ISMN datasets are located at higher latitudes, but large areas around the tropics and water-limited environments are not represented within the ISMN (Figure 5.1). Water limited environments across arid and semi-arid regions where drying trends are prevalent, large discrepancy has been found between satellite and model-based soil moisture estimates (Liu et al., 2019). The lack of field soil moisture information across these areas is a major limitation for interpreting the discrepancies between satellite and model-based soil moisture estimates. This lack of information is also a limitation for interpreting the accuracy of our prediction framework (and the accuracy of the ESA-CCI soil moisture product) across large areas where no field data for validation purposes are available in the ISMN.

Using the available data contained in the ISMN for validating the ESA-CCI soil moisture product and the soil moisture predictions generated here, we found consistent negative trends of soil moisture at the global scale (Figure 5. 6). The main attribution of this soil moisture drying at the global scale, which is consistent with recent soil moisture monitoring efforts (Gu et al., 2019a) is anthropogenic climate change (i.e., land use change, Gu et al., 2019b). This is consistent with previous reports that have found similar results reporting a dominance in previous decades of negative soil moisture trends across the world that were detected using field and satellite soil moisture measurements (Albergel et al., 2013). It has been shown how soil moisture decline can be intensified by land warming (Samaniego et al., 2018) or by land use change (Chen, et al., 2016, Garg et al., 2019) and agricultural practices (Bradford et al., 2017) or transformations to vegetation cover that directly affect primary productivity, evapotranspiration rates and drought (Stocker et al, 2019, Martens et al., 2018). Areas with high rates of primary productivity and the evapotranspiration rates such as the tropical rain forest of the Amazon or the Congo regions, are examples of areas where is challenging to accurately assess soil moisture trends due to limitations of historical soil moisture records (such as in the ESA-CCI). The ESA-CCI also lack spatial information in higher latitudes across areas with high density of small water bodies (i.e., northeast United States) or high latitude forested areas (i.e., boreal forest) constantly covered by snow (Reich et al., 2018). It has been shown that soil moisture regulates climate warming effects on forest tree species across the aforementioned areas (Reich et al., 2018). The response of vegetation productivity to long term soil moisture trends is

needed for improved land management and land surface modeling across all climate conditions and finer spatial grids. Therefore, we provide annual soil moisture values across the world (Figure 5.8) that can be used to monitor the long term response of vegetation to soil moisture decline based on geomorphometry and remote sensing of soil moisture.

Our prediction framework was useful to improve the spatial representation of ESA-CCI soil moisture product. Recent soil moisture remote sensing products (Entekhabi et al., 2010, Piles et al., 2019) are able to provide soil moisture information across areas with spatial gaps in the ESA-CCI and provide global estimates, however there are only recent records with full soil moisture coverage (e.g., 2010 to date,) and this represent a limitation for studying historical soil moisture and atmosphere interactions across longer periods of time. Soil moisture trends are crucial in future projections of the water cycle for identifying regions of strong land–atmosphere coupling (Lorenz et al., 2015). However higher resolution of soil moisture products are needed to precisely quantify the contribution of soil moisture on larnd-surface models (Singh et al., 2015). Although model- based estimates of soil moisture associated to global land data assimilation systems are available for studying historical soil moisture trends at the global scale (Fang et al., 2009, Liu et al., 2019), they remain represented with spatial resolutions > 15km grids. By increasing the spatial resolution of the ESA-CCI soil moisture product by nearly 50%, we demonstrate the potential of digital terrain analysis to predict satellite soil moisture spatial patterns and trends while increasing the agreement between satellite and field soil moisture records predicting historical soil

moisture trends (Fig 6). While our results are consistent with previous studies predicting global soil moisture decline (Jung et al., 2010, Albergel et al., 2013), they are also generalizable to specific spatial extents or higher spatial resolution (e.g., across the continental United States using 1x1km grids, Guevara and Vargas, 2019) to analyze spatial and temporal trends of soil moisture.

Future improvements for our approach could include predicting soil moisture patterns across finer pixel sizes (e.g., 1km or <1km) and higher temporal resolutions (e.g., monthly, daily). The current version of the downscaled soil moisture predictions based on digital terrain analysis is provided in an annual basis because is a temporal resolution useful for multiple ecological and hydrological studies related to climate change (Green et al., 2019). We recognize that there is an increasing need of soil moisture datasets with higher temporal resolutions to analyze the seasonal and short-term memory soil moisture effects after precipitation events (McColl et al., 2017). A spatial resolution of 15 km is still a coarse pixel size for detailed analysis of hydro-ecological patterns (e.g., at the hillslope scale), but the main focus of this study was to test the potential of digital terrain analysis for increasing the spatial resolution of the original ESA-CCI soil moisture product. Our decision for selecting a 15km pixel size was driven by the reproducibility or our framework by multiple groups without the need of high performance computing infrastructure.

In conclusion, to downscale (i.e., increase spatial resolution) coarse satellite soil moisture grids we used ML to combine satellite soil moisture data with terrain parameters (as surrogates of topographic variability). Our results support that digital

terrain analysis can be used for improving the spatial resolution (from 27 to 15 km grids) of available global satellite soil moisture datasets based on multiple evaluation metrics against field soil moisture data. We provide a new gap-free and annual soil moisture product (1991-2016) that can be used for further ecological and hydrological analysis. The current version of the new generated data set is composed by 26 annual soil moisture predictions provided across 15 km grids in an annual basis (1991-2016). These grids could be useful for identifying and characterizing long term patterns in the soil moisture content across 26 years of available satellite soil moisture datasets from the regional to the global scale. Future efforts could apply our framework to address the increasing need of soil moisture datasets with higher temporal and spatial resolution at the global scale using similar or different satellite-derived soil moisture products.

5.5 Data Source and Scientific Replicability

The training soil moisture dataset used in this study is available (here: https://www.esa-soilmoisture-cci.org/) thanks to the ESA-CCI soil moisture initiative (Dorigo et al., 2017). The downscaled soil moisture predictions based on digital terrain analysis are provided (Guevara, et al., 2019, https://doi.org/10.4211/hs.b940b704429244a99f902ff7cb30a31f) in a set raster files with a *.tif extension for generic raster formats (e.g., 1 raster per year, folder: predicted_sm_global_15km) . This rasters (n=26, 1991-2016) can be imported to any GIS.

To ensure the replicability of this study we also provide a spatial data frame (in R native format *.rds) with the topographic terrain parameters (e.g., file: topographic_predictors_15km_grids.rds, also provided as *.tif raster in folder: prediction_factors_15km) used as prediction factors for the yearly means of the ESA-CCI soil moisture product. We provide the yearly means of the ISMN database that we used for evaluating the aforementioned soil moisture predictions. The ISMN yearly means are harmonized with the ESA-CCI soil moisture product and the downscaled soil moisture predictions based on terrain analysis, and provided in two separated files (e.g., files: harmonizedISMNvsESACCI.rds and harmonizedISMNvsPREDICTED.rds). These predictions are based on ML and digital terrain parameters and they are provided in a generic raster format (Guevara, et al., 2019, https://doi.org/10.4211/hs.b940b704429244a99f902ff7cb30a31f). Finally, we provide the processing R code used to develop our prediction framework (e.g., file: prediction_kknn_sm_terrain_global_15km_v0.R).

Acknowledgements

MG acknowledges support from a CONACYT doctoral fellowship (382790). RV and MT acknowledge support from the National Science Foundation grant CIF21 DIBBs: PD: Cyberinfrastructure Tools for Precision Agriculture in the 21st Century. We thank to Anita Z. Schwartz from the University of Delaware for her assistance preparing the global terrain dataset. We thank Ricardo Llamas from the University of Delaware for preparing Supplementary Figure S1.

Author contributions

MG, RV and MT conceptualized the project. MG performed analysis and wrote the first manuscript draft. This draft was revised, commented and edited by RV, MG and MT.

Supplement in:
https://www.earth-syst-sci-data-discuss.net/essd-2019-191/essd-2019-191-supplement.pdf

REFERENCES

Albergel, C., Dorigo, W., H, R. R., Balsamo, G., de Rosnay, P, MuñozSabater, J., Isaksen, L., de Jeu, R and Wagner, W.: Skill and Global Trend Analysis of Soil Moisture from Reanalyses and Microwave Remote Sensing, Journal of Hydrometeorology, 14(4), 1259–1277, 2013.

Alemohammad, S. H., Kolassa, J., Prigent, C., Aires, F. and Gentine, P.: Global downscaling of remotely sensed soil moisture using neural networks, Hydrology and Earth System Sciences, 22(10), 5341–5356, doi:10.5194/hess-22-5341-2018, 2018.

Al-Yaari, A., Wigneron, J.-P., Dorigo, W., Colliander, A., Pellarin, T., Hahn, S., Mialon, A., Richaume, P., Fernandez- Moran, R., Fan, L. and al, et: Assessment and inter-comparison of recently developed/reprocessed microwave satellite soil moisture products using ISMN ground-based measurements, Remote Sensing of Environment, 224, 289–303, doi:10.1016/j.rse.2019.02.008, 2019.

An, R., Zhang, L., Wang, Z., Quaye-Ballard, J. A., You, J., Shen, X., Gao, W., Huang, L., Zhao, Y. and Ke, Z.: Validation of the ESA CCI soil moisture product in China, International Journal of Applied Earth Observation and Geoinformation, 48, 28–36, doi:10.1016/j.jag.2015.09.009, 2016.

Becker, J. J., Sandwell, D. T., Smith, W. H. F., Braud, J., Binder, B., Depner, J., Fabre, D., Factor, J., Ingalls, S., Kim, S.-H., Ladner, R., Marks, K., Nelson, S., Pharaoh, A., Trimmer, R., Rosenberg, J. V., Wallace, G. and Weatherall, P.: Global Bathymetry and Elevation Data at 30 Arc Seconds Resolution: SRTM30_PLUS, Marine Geodesy, 32(4), 355–371, doi:10.1080/01490410903297766, 2009.

Berg, A. and Sheffield, J.: Climate Change and Drought: the Soil Moisture Perspective, Current Climate Change Reports, 4(2), 180–191, doi:10.1007/s40641-018-0095-0, 2018.

Bradford, J. B., Schlaepfer, D. R., Lauenroth, W. K., Yackulic, C. B., Duniway, M., Hall, S., Jia, G., Jamiyansharav, K., Munson, S. M., Wilson, S. D. and Tietjen, B.: Future soil moisture and temperature extremes imply expanding suitability for rainfed agriculture in temperate drylands, Scientific Reports, 7(1), doi:10.1038/s41598-017-13165-x, 2017.

Carslaw, D. C. and Ropkins, K.: openair — An R package for air quality data analysis, Environmental Modelling & Software, 27–28, 52–61, doi:10.1016/j.envsoft.2011.09.008, 2012.

Carslaw, D.C. The openair manual — open-source tools for analysing air pollution data. Manual for version 1.1-4, King's College London, 287pp., 2015.

Chang, J. C. and Hanna, S. R.: Air quality model performance evaluation, Meteorol Atmos Phys, 87(1), 167–196, doi:10.1007/s00703-003-0070-7, 2004.

Colliander, A., Fisher, J. B., Halverson, G. H., Merlin, O., Misra, S., Bindlish, R., Jackson, T. J. and Yueh, S. H.: Spatial Downscaling of SMAP Soil Moisture Using MODIS Land Surface Temperature and NDVI During SMAPVEX15, IEEE Geoscience and Remote Sensing Letters, 14, 2107–2111, doi:10.1109/LGRS.2017.2753203, 2017a.

Colliander, A., Jackson, T. J., Bindlish, R., Chan, S., Das, N., Kim, S. B., Cosh, M. H., Dunbar, R. S., Dang, L., Pashaian, L., Asanuma, J., Aida, K., Berg, A., Rowlandson, T., Bosch, D., Caldwell, T., Caylor, K., Goodrich, D., al Jassar, H., Lopez- Baeza, E., Martínez-Fernández, J., González-Zamora, A., Livingston, S., McNairn, H., Pacheco, A., Moghaddam, M., Montzka, C., Notarnicola, C., Niedrist, G., Pellarin, T., Prueger, J., Pulliainen, J., Rautiainen, K., Ramos, J., Seyfried, M., Starks, P., Su, Z., Zeng, Y., van der Velde, R., Thibeault, M., Dorigo, W., Vreugdenhil, M., Walker, J. P., Wu, X., Monerris, A., O'Neill, P. E., Entekhabi, D., Njoku, E. G. and Yueh, S.: Validation of SMAP surface soil moisture products with core validation sites, Remote Sensing of Environment, 191, 215–231, doi:10.1016/j.rse.2017.01.021, 2017b.

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, Geoscientific Model Development, 8(7), 1991–2007, doi:10.5194/gmd-8-1991-2015, 2015.

Crow, W. T., Berg, A. A., Cosh, M. H., Loew, A., Mohanty, B. P., Panciera, R., de Rosnay, P., Ryu, D. and Walker, J. P.: Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products, Reviews of Geophysics, 50(2), doi:10.1029/2011rg000372, 2012.

Dorigo, W., Oevelen, P. van, Wagner, W., Drusch, M., Mecklenburg, S., Robock, A. and Jackson, T.: A New International Network for in Situ Soil Moisture Data, Eos, Transactions American Geophysical Union, 92(17), 141–142, doi:10.1029/2011EO170001, 2011a.

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A. and al, et: ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions, Remote Sensing of Environment, 203, 185–215, doi:10.1016/j.rse.2017.07.001, 2017.

Dorigo, W. A., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Xaver, A., Gruber, A., Drusch, M., Mecklenburg, S., Oevelen, P. van, Robock, A. and Jackson, T.: The International Soil Moisture Network: a data hosting facility for global in situ soil moisture measurements, Hydrology and Earth System Sciences, 15(5), 1675–1698, doi:https://doi.org/10.5194/hess- 15-1675-2011, 2011b.

Dubayah, R. O. and Drake, J. B.: Lidar Remote Sensing for Forestry, Journal of Forestry, 98(6), 44–46, doi:10.1093/jof/98.6.44, 2000.

Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., Entin, J. K., Goodman, S. D., Jackson, T. J., Johnson, J., Kimball, J., Piepmeier, J. R., Koster, R. D., Martin, N., McDonald, K. C., Moghaddam, M., Moran, S., Reichle, R., Shi, J. C., Spencer, M. W., Thurman, S. W., Tsang, L. and Van Zyl, J.: The Soil Moisture Active Passive (SMAP) Mission, IEEE [online] Available from: https://dspace.mit.edu/handle/1721.1/60043 (Accessed 15 July 2019), 2010.

Fang, H., Beaudoing, H. K., Rodell, M., Teng, W. L., & Vollmer, B. E. Global Land Data Assimilation System (GLDAS) Products, Services and Application from NASA Hydrology Data and Information Services Center (HDISC). In ASPRS 2009 Annual Conference. Baltimore, Maryland. https://www.asprs.org/a/publications/proceedings/baltimore09/0020.pdf 2009.

Forkel, M., Carvalhais, N., Verbesselt, J., Mahecha, M. D., Neigh, C. S. R. and Reichstein, M.: Trend Change Detection in NDVI Time Series: Effects of Inter-Annual Variability and Methodology, Remote Sensing, 5(5), 2113–2144, doi:10.3390/rs5052113, 2013.

Forkel, M., Migliavacca, M., Thonicke, K., Reichstein, M., Schaphoff, S., Weber, U. and Carvalhais, N.: Codominant water control on global interannual variability and trends in land surface phenology and greenness, Global Change Biology, 21(9), 3414–3435, doi:10.1111/gcb.12950, 2015.

Garg, V., Nikam, B. R., Thakur, P. K., Aggarwal, S. P., Gupta, P. K. and Srivastav, S. K.: Human-induced land use land cover change and its impact on hydrology, HydroResearch, 1, 48–56, doi:10.1016/j.hydres.2019.06.001, 2019.

Green, J. K., Seneviratne, S. I., Berg, A. M., Findell, K. L., Hagemann, S., Lawrence, D. M. and Gentine, P.: Large influence of soil moisture on long-term terrestrial carbon uptake, Nature, 565(7740), 476–479, doi:10.1038/s41586-018-0848-x, 2019.

Greve, P. and Seneviratne, S. I.: Assessment of future changes in water availability and aridity, Geophysical Research Letters, 42(13), 5493–5499, doi:10.1002/2015gl064127, 2015.

Gruber, A., Dorigo, W. A., Crow, W. and Wagner, W.: Triple Collocation-Based Merging of Satellite Soil Moisture Retrievals, IEEE Transactions on Geoscience and Remote Sensing, 55(12), 6780–6792, doi:10.1109/tgrs.2017.2734070, 2017.

Gu, X., Zhang, Q., Li, J., Singh, V. P., Liu, J., Sun, P. and Cheng, C.: Attribution of Global Soil Moisture Drying to Human Activities: A Quantitative Viewpoint, Geophysical Research Letters, 46(5), 2573–2582, doi:10.1029/2018gl080768, 2019a.

Gu, X., Zhang, Q., Li, J., Singh, V. P., Liu, J., Sun, P., He, C. and Wu, J.: Intensification and Expansion of Soil Moisture Drying in Warm Season Over Eurasia Under Global Warming, Journal of Geophysical Research: Atmospheres, 124(7), 3765–3782, doi:10.1029/2018jd029776, 2019b.

Guevara, M. and Vargas, R.: Downscaling satellite soil moisture using geomorphometry and machine learning, edited by B. Poulter, PLOS ONE, 14(9), e0219639, doi:10.1371/journal.pone.0219639, 2019.

Guevara, M., R. Vargas, M. Taufer. Gap-Free Global Annual Soil Moisture: 15km Grids for 1991-2016, HydroShare, https://doi.org/10.4211/hs.b940b704429244a99f902ff7cb30a31f, 2019.

Hechenbichler K. and Schliep K.P. Weighted k-Nearest-Neighbor Techniques and Ordinal Classification, Discussion Paper 399, SFB 386, Ludwig-Maximilians University Munich, 2004.

Hengl, T.: Finding the right pixel size, Computers & Geosciences, 32(9), 1283–1298, doi:10.1016/j.cageo.2005.11.008, 2006.

Hijmans R. J. raster: Geographic Data Analysis and Modeling. R package version 2.9-23. https://CRAN.R- project.org/package=raster, 2019

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., Bonan, G., Cescatti, A., Chen, J., de Jeu,

R. and al, et: Recent decline in the global land evapotranspiration trend due to limited moisture supply, Nature, 467(7318), 951–954, doi:10.1038/nature09396, 2010.

Legates, D. R. and McCabe, G. J. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water resources research, 35(1):233–241, 1999.

Legates, D. R. and McCabe, G. J. A refined index of model performance: a rejoinder. International Journal of Climatology, 33(4):1053–1056, 2013.

Liu, Y. Y., Parinussa, R. M., Dorigo, W. A., De Jeu, R. A. M., Wagner, W., van Dijk, A. I. J. M., McCabe, M. F. and Evans,

J. P.: Developing an improved soil moisture dataset by blending passive and active microwave satellite-based retrievals, Hydrology and Earth System Sciences, 15(2), 425–436, doi:10.5194/hess-15-425-2011, 2011.

Liu, Y., Liu, Y. and Wang, W.: Inter-comparison of satellite-retrieved and Global Land Data Assimilation System-simulated soil moisture datasets for global drought analysis, Remote Sensing of Environment, 220, 1–18, doi:10.1016/j.rse.2018.10.026, 2019.

Lorenz, R. D., Pitman, A. J., Hirsch, A. L. and Jhan Srbinovsky: Intraseasonal versus Interannual Measures of Land– Atmosphere Coupling Strength in a Global Climate Model: GLACE-1 versus GLACE-CMIP5 Experiments in ACCESS1.3b, [online] Available from: https://www.semanticscholar.org/paper/Intraseasonal-versus-Interannual-Measures- of-in-a-Lorenz-Pitman/1327a707d832e98b6c011c2ba6dd1812d2e2c2d8 (Accessed 25 September 2019), 2015.

Mahecha, M. D., Reichstein, M., Carvalhais, N., Lasslop, G., Lange, H., Seneviratne, S. I., Vargas, R., Ammann, C., Arain,

M. A., Cescatti, A. and al, et: Global Convergence in the Temperature Sensitivity of Respiration at Ecosystem Level, Science, 329(5993), 838–840, doi:10.1126/science.1189587, 2010.

May, W., Rummukainen, M., Chéruy, F., Hagemann, S. and Meier, A.: Contributions of soil moisture interactions to future precipitation changes in the GLACE-CMIP5 experiment, Climate Dynamics, 49(5–6), 1681–1704, doi:10.1007/s00382-016-3408-9, 2016.

McColl, K. A., Alemohammad, S. H., Akbar, R., Konings, A. G., Yueh, S. and Entekhabi, D.: The global distribution and dynamics of surface soil moisture, Nature Geoscience, 10(2), 100–104, doi:10.1038/ngeo2868, 2017.

Minet, J., Bogaert, P., Vanclooster, M. and Lambot, S.: Validation of ground penetrating radar full-waveform inversion for field scale soil moisture mapping, Journal of Hydrology, 424–425, 112–123, doi:10.1016/j.jhydrol.2011.12.034, 2012.

Mohanty, B. P., Cosh, M. H., Lakshmi, V. and Montzka, C.: Soil Moisture Remote Sensing: State-of-the-Science, Vadose Zone Journal, 16(1), doi:10.2136/vzj2016.10.0105, 2017.

van der Molen, M. K., Dolman, A. J., Ciais, P., Eglin, T., Gobron, N., Law, B. E., Meir, P., Peters, W., Phillips, O. L., Reichstein, M. and al, et: Drought and ecosystem carbon cycling, Agricultural and Forest Meteorology, 151(7), 765–773, doi:10.1016/j.agrformet.2011.01.018, 2011.

Martens, B., de Jeu, R., Verhoest, N., Schuurmans, H., Kleijer, J. and Miralles, D.: Towards Estimating Land Evaporation at Field Scales Using GLEAM, Remote Sensing, 10(11), 1720, doi:10.3390/rs10111720, 2018.

Mason, D. C., Garcia-Pintado, J., Cloke, H. L. and Dance, S. L.: Evidence of a topographic signal in surface soil moisture derived from ENVISAT ASAR wide swath data, International Journal of Applied Earth Observation and Geoinformation, 45, 178–186, doi:10.1016/j.jag.2015.02.004, 2016.

Mishra, V., Tiwari, A. D., Aadhar, S., Shah, R., Xiao, M., Pai, D. S. and Lettenmaier, D.: Drought and Famine in India, 1870–2016, Geophysical Research Letters, 46(4), 2075–2083, doi:10.1029/2018gl081477, 2019.

Moeslund, J. E., Arge, L., Bøcher, P. K., Dalgaard, T., Odgaard, M. V., Nygaard, B. and Svenning, J.-C.: Topographically controlled soil moisture is the primary driver of local vegetation patterns across a lowland region, Ecosphere, 4(7), art91, doi:10.1890/es13-00134.1, 2013.

Murguia-Flores, F., Arndt, S., Ganesan, A. L., Murray-Tortarolo, G. and Hornibrook, E. R. C.: Soil Methanotrophy Model (MeMo v1.0): a process-based model to quantify global uptake of atmospheric methane by soil, Geoscientific Model Development, 11(6), 2009–2032, doi:https://doi.org/10.5194/gmd-11-2009-2018, 2018.

Nicolai-Shaw, N., Hirschi, M., Mittelbach, H. and Seneviratne, S. I.: Spatial representativeness of soil moisture using in situ, remote sensing, and land reanalysis data, Journal of Geophysical Research: Atmospheres, 120(19), 9955–9964, doi:10.1002/2015jd023305, 2015.

Peng, J., Loew, A., Merlin, O. and Verhoest, N. E. C.: A review of spatial downscaling of satellite remotely sensed soil moisture, Reviews of Geophysics, 55(2), 341–366, doi:10.1002/2016rg000543, 2017.

Piles, M., Ballabrera-Poy, J. and Muñoz-Sabater, J.: Dominant Features of Global Surface Soil Moisture Variability Observed by the SMOS Satellite, Remote Sensing, 11(1), 95, doi:10.3390/rs11010095, 2019.

Reich, P. B., Sendall, K. M., Stefanski, A., Rich, R. L., Hobbie, S. E. and Montgomery, R. A.: Effects of climate warming on photosynthesis in boreal tree species depend on soil moisture, Nature, 562(7726), 263–267, doi:10.1038/s41586-018-0582-4, 2018.

Samaniego, L., Thober, S., Kumar, R., Wanders, N., Rakovec, O., Pan, M., Zink, M., Sheffield, J., Wood, E. F. and Marx, A.: Anthropogenic warming exacerbates European soil moisture droughts, Nature Climate Change, 8(5), 421–426, doi:10.1038/s41558-018-0138-5, 2018.

Senanayake, I. P., Yeo, I.-Y., Tangdamrongsub, N., Willgoose, G. R., Hancock, G. R., Wells, T., Fang, B., Lakshmi, V. and Walker, J. P.: An in-situ data based model to downscale radiometric satellite soil moisture products in the Upper Hunter Region of NSW, Australia, Journal of Hydrology, 572, 820–838, doi:10.1016/j.jhydrol.2019.03.014, 2019.

Seneviratne, S. I., Wilhelm, M., Stanelle, T., Hurk, B., Hagemann, S., Berg, A., Cheruy, F., Higgins, M. E., Meier, A., Brovkin, V. and al, et: Impact of soil moisture-climate feedbacks on CMIP5 projections: First results from the GLACE- CMIP5 experiment, Geophysical Research Letters, 40(19), 5212–5217, doi:10.1002/grl.50956, 2013.

Singh, R. S., Reager, J. T., Miller, N. L. and Famiglietti, J. S.: Toward hyper-resolution land-surface modeling: The effects of fine-scale topography and soil texture on CLM4.0 simulations over the Southwestern U.S., Water Resources Research, 51(4), 2648–2667, doi:10.1002/2014WR015686, 2015.

Stocker, B. D., Zscheischler, J., Keenan, T. F., Prentice, I. C., Seneviratne, S. I. and Peñuelas, J.: Drought impacts on terrestrial primary production underestimated by satellite monitoring, Nature Geoscience, 12(4), 264–270, doi:10.1038/s41561-019-0318-6, 2019.

Tadono, T., Ishida, H., Oda, F., Naito, S., Minakawa, K. and Iwamoto, H.: Precise Global DEM Generation by ALOS PRISM, ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, II-4, 71–76, doi:10.5194/isprsannals-ii-4-71-2014, 2014.

Western, A. W., Grayson, R. B. and Blöschl, G.: Scaling of Soil Moisture: A Hydrologic Perspective, Annual Review of Earth and Planetary Sciences, 30(1), 149–180, doi:10.1146/annurev.earth.30.091201.140434, 2002.

Wilson, J. P.: Digital terrain modeling, Geomorphology, 137(1), 107–121, doi:10.1016/j.geomorph.2011.03.012, 2012.

Willmott, C. J., Robeson, S. M. and Matsuura, K.: A refined index of model performance, International Journal of Climatology, 32(13), 2088–2094, doi:10.1002/joc.2419, 2011.

Yee, M. S., Walker, J. P., Monerris, A., Rüdiger, C. and Jackson, T. J.: On the identification of representative in situ soil moisture monitoring stations for the validation of SMAP soil moisture products in Australia, Journal of Hydrology, 537, 367–381, doi:10.1016/j.jhydrol.2016.03.060, 2016.

Zhuo, L. and Han, D.: The Relevance of Soil Moisture by Remote Sensing and Hydrological Modelling, Procedia Engineering, 154, 1368–1375, doi:10.1016/j.proeng.2016.07.499, 2016.

Zhou, W., Hui, D. and Shen, W.: Effects of Soil Moisture on the Temperature Sensitivity of Soil Heterotrophic Respiration: A Laboratory Incubation Study, edited by S. Hu, PLoS ONE, 9(3), e92531, doi:10.1371/journal.pone.0092531, 2014.

CONCLUSIONS

This research described the development of a digital soil mapping strategy applied to the spatial variability of soil moisture and soil organic carbon from the country specific to the global scale. The main conclusion of this research is that soil moisture and soil organic carbon spatial trends can be continuously quantified with increased detail and increased accuracy across scales using different forms of machine learning (computer assisted statistical learning) and the synthesis of large amounts of soil related environmental information. As a fundamental component of soil science, this research shows how digital soil mapping provides increasing opportunities for quantifying the spatial variability of soil moisture and soil organic carbon across multiple spatial scales, land cover classes or periods of time.

The new knowledge on spatial variability of soil moisture and soil organic carbon could be used for identifying, with higher resolution and certainty, areas with soil moisture decline, land degraded areas or areas with the potential for soil carbon sequestration. With this research I provide new soil information and knowledge on soil moisture and soil organic carbon for obtaining improved estimates around the contribution of these variables to the regional and global carbon cycle. Specific conclusions related to the research questions at the end of the introductory section are explained in the following paragraphs.

Key conclusions:

There is no one single best method for statistical learning on digital soil mapping. Multiple models capture the spatial variability of soil values in different forms (data-driven, hypothesis driven) and they show different sources of errors. The combination of multiple prediction algorithms could be an efficient way to maximize the accuracy of digital soil organic carbon maps. As I show in **chapter 2**, to increase the accuracy of soil organic carbon predictions across Latin America is needed to solve discrepancies between country and global soil organic carbon estimates.

Terrain parameters derived from elevation data are key factors for efficient upscaling or downscaling soil organic carbon and soil moisture information from the country to the global scale. Terrain parameters, in combination with vegetation greenness, or climatic variables (e.g., water vapor pressure, precipitation or temperature) were top ranked predictors in the soil organic carbon modeling approach described in **chapter 3**. Using key prediction factors, I estimated a 'most probable' soil organic carbon stock between 46 and 47 Pg across Mexico and the conterminous United States for the first 30cm of soil depth during the period between 1991 and 2010 (which is conservative with previous global estimates). The soil organic carbon change between 1991-2001 and 2001-2010 (from 41 to 55 Pg) is challenging to explain or attribute due to the limited information across large areas for each period of time. The current lack of multi temporal soil covariates representing soil variability and depth relationships could be another source of uncertainty explaining the temporal variability of soil organic carbon at the continental scale of Mexico and the conterminous United States.

Soil organic carbon could also be predicted as a function of changes in soil moisture conditions as soil organic carbon is also indicator of soil hydrological functioning. Satellite soil moisture is able to represent with coarse grids the spatial and temporal variability of soil moisture conditions across the world. In **chapter 4**, the granularity and accuracy of satellite soil moisture was improved using digital terrain analysis and machine learning. Terrain parameters derived from elevation were also found to be strong predictors for soil moisture spatial trends and these results were confirmed from the continental scale of the conterminous United States (in **chapter 4**) to the global scale (in **chapter 5**). With potential implications in the capacity of soils to store soil organic carbon, in **chapter 5** I provide evidence of soil moisture decline across the global scale that is consistent between field soil moisture observations, satellite soil moisture records and modeled soil moisture data based on terrain parameters.

Research summary

From the country specific to the global scale, the overreaching goal of this research was to develop a digital soil mapping framework to increase the statistical accuracy of spatially continuous information on soil moisture and soil organic carbon across different scales of data availability (e.g., country-specific, regional, global). Comparing and testing multiple machine learning (e.g., tree-based, kernel-based, linear-based) and combined methods for predicting the spatial variability of soil organic carbon in a country specific basis, in Chapter 1, we found no best prediction algorithms. The selection of the best prediction was affected by a) the correlation of available data

for training models (within each country across Latin America) and its prediction factors, b) the spatial distribution (representativeness) of available data and c) the density of soil organic observations in relation to each country area.

Considering that no universal model is best for all data sets, selecting and combining multiple models (e.g., by the means of different ensemble learning approaches) could be one way to move forward towards increasing the accuracy of digital soil organic carbon maps. We highlight large discrepancies between country specific and global soil organic carbon estimates. We report a country-specific soil organic carbon stock of 77.8 ± 43.6 Pg (in the first 0-30cm of soil depth) across Latin America representing 'past-current' conditions (i.e., data from 1950's to date). This stock is 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) based on legacy datasets. These discrepancies between country and global soil organic carbon estimates could lead to inappropriate applications (e.g, on soil carbon assessments) and to increased bias in Earth system models predicting trends of land carbon uptake. The main conclusion from Chapter 1 is that the use of country-specific information and the use of different modeling approaches will enhance regional soil organic carbon mapping efforts and will provide insights to identify where and why different modeling approaches generate similar soil organic carbon estimates.

In Chapter 2 there is a transition from a country specific basis to a continental scale mapping soil organic carbon across two rich data countries: Mexico and United States. We focus on variable selection and explore multiple combinations of prediction

factors across the region using a recursive feature elimination of variables and a global search method (simulated annealing) to identify the main drivers of soil organic carbon spatial variability. We quantified soil organic carbon stocks and multiple forms of uncertainty for the period between 1991 and 2010 (46 ±12 Pg in the first 0-30cm of soil depth). Our results suggest that using key predictors (e.g., the first 5 top ranked predictors from our variable selection approach) yields a similar performance (~50% of explained variance) to a high-dimensional covariate space (e.g., n=150 environmental predictors). These results are consistent with a fully independent datasets supporting the reliability of our prediction framework. In chapter 2 I provide high spatial resolution (e.g., 250m pixels) soil organic carbon estimates that account for model uncertainty. As in chapter 1, we confirm that soil organic carbon stocks are lower (around 30%) compared with previous global soil organic carbon estimated stocks. In chapter 2 the results provide insights for interpreting the variance of multiple methods for estimating soil organic carbon stocks and multiple model-based uncertainty benchmarks for improving regional-to-global soil organic carbon monitoring efforts.

Soil organic carbon monitoring efforts can substantially benefit from remotely sensing data, for example, satellite soil moisture records are available in a daily basis from over forty years ago. However, there are limitations in remotely sensed soil moisture granularity and accuracy that could be improved by the means of digital soil mapping. In chapter 3, I describe a topography-modulated digital soil mapping framework applied to satellite soil moisture across the conterminous United States. The main goal was to characterize the spatial variability of soil moisture using satellite-

based information and topography derived terrain parameters as prediction factors. We focus on topography because regression analysis suggest that terrain parameters derived from elevation data are strong predictors for satellite soil moisture (r=0.8 in all model/years from 1991-2018). Modeled soil moisture data (satellite soil moisture + topography) increased the accuracy of the original satellite soil moisture product against field soil moisture observations by nearly 25%. We conclude that the machine learning fusion of topography and satellite soil moisture is useful to increase the spatial detail (granularity) and accuracy of soil moisture estimates. A trend detection test applied to the new generated soil moisture estimates reveal a significant soil moisture decline across 40% of conterminous United States (1991-2016), highlighting the need to develop effective soil water resources protection strategies. With chapter 3 I provide new soil moisture estimates that could be potentially good predictors for soil organic carbon (i.e.,soil organic carbon as a response of changing soil moisture conditions) and for obtaining new information about soil moisture and soil organic carbon responses to climate change.

Chapter 4 of this research is an extension of chapter 3 from the country specific to the global scale. Predicting soil moisture trends using a pattern recognition technique to couple satellite soil moisture estimates with topography information, we confirm the effectiveness of using topography information for predicting soil moisture patterns and improving the quality of satellite-based estimates against field soil observations and climate data. Our results suggest a soil moisture decline across the world that is consistent in the field soil moisture records, the original satellite estimates and the

modeled soil data. The main contribution of this chapter 4 is a novel approach for predicting soil moisture trends across the world that is independent of climate and vegetation data. A climate or vegetation independent soil moisture product is appealing for avoiding spurious correlations in further ecological applications analyzing the implications of soil moisture in vegetation and climate dynamics. In this paper we developed a variable importance approach based on permutation to show that topography has a similar prediction capacity on soil moisture compared with soil type and bioclimatic information, supporting the potential of terrain parameters for improving the spatial representation of soil moisture at the global scale. From the previous chapter to the global scale, my results suggest soil moisture decline across large areas of the world with potentially negative consequences in the regional to global carbon cycle. Increased soil moisture decline and aridity conditions may be threatening factors to the carbon storage capacity of soils in the upcoming years.

Future directions

In this research I described how different soil moisture and soil organic carbon mapping methods generate different results, as they capture the spatial variability of soil moisture and soil organic carbon following different assumptions (e.g., data or hypothesis driven). The results from this research could contribute with the development of soil carbon and soil water monitoring and information systems. Contributing with the development of soil information and monitoring systems, all methodologies cited or developed in this research as well as all soil spatial products (soil organic carbon and soil moisture maps, harmonized datasets) are public and

available through stable data repositories (e.g., data repositories supported by the National Science Foundation or Department of Energy) for its use by multiple soil information users.

This research is based on the synthesis of the exponentially increasing soil-related datasets and opens new possibilities to enhance our understanding of the current and future variability of soil moisture and soil organic carbon stocks. This research helped to fill some informational gaps around the spatial variability of soil moisture and soil organic carbon. However large uncertainties remain in our capacity to detect trends of soil moisture and soil carbon stocks of deeper soil layers along the complete soil profile. Many areas of the world lack of updated soil information for synthesis studies and the comparison and refinement of multiple models are still required to predict realistically spatial and temporal patterns of soil variability. This project provides a novel framework (including high quality data and novel methodologies) to identify emergent patterns of soil variability and generate environmentally relevant science that can be used for the formulation of public policy around soil and water conservation efforts. Increasing the spatial detail and accuracy of digital soil maps are the eternal challenges of digital soil mapping. With increased accuracy and increased detail of previous soil moisture and soil organic carbon estimates, the new knowledge represents a benchmark against which to assess the impact of climate and land cover changes on soil organic carbon stocks and soil moisture trends.

Appendix A

COPYRIGTH STATEMENTS OF PUBLISHED CHAPTERS

The Chapter 2 of this thesis (No silver bullet on digital soil mapping: country specific soil organic estimates across Latin America by Guevara et al., 2018) has been published in SOIL under the creative commons attribution license SOIL, 4, 173–193, 2018 DOI: https://doi.org/10.5194/soil-4-173-2018

The Chapter 3 of this thesis (Soil organic carbon across Mexico and the conterminous United States by Guevara et al., 2020) has been accepted by for its publication in Global Biogeochemical Cycles. The current state of this paper (towards Feb. 2020) is in production and can be used for this dissertation without allowing open access options DOI: https://doi.org/10.1029/2019GB006219

The Chapter 4 of this thesis (Downscaling satellite soil moisture using remote sensing and machine learning by Guevara and Vargas 2019) has been published in Plos One under the creative commons attribution license DOI: https://doi.org/10.1371/journal.pone.0219639

The Chapter 5 of this thesis (Gap-Free Global Annual Soil Moisture: 15km Grids for 1991–2016 by Guevara et al., 2019) is currently (Feb. 2020) under review in Earth System Science Data. This discussion paper is available under the creative commons attribution license DOI: https://doi.org/10.5194/essd-2019-191