



Empirical relationships between environmental factors and soil organic carbon produce comparable prediction accuracy to machine learning

Umakant Mishra¹  | Kyongmin Yeo² | Kabindra Adhikari³  | William J. Riley⁴ | Forrest M. Hoffman⁵ | Corey Hudson¹ | Sagar Gautam¹

¹Computational Biology & Biophysics, Sandia National Laboratories, Livermore, CA 94550, USA

²IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10562, USA

³USDA-ARS, Grassland, Soil and Water Research Laboratory, Temple, TX 76502, USA

⁴Earth & Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

⁵Climate Change Institute, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA

Correspondence

Umakant Mishra, Computational Biology & Biophysics, Sandia National Laboratories, Livermore, CA 94550, USA
Email: umishra@sandia.gov

Assigned to Associate Editor Kyungsoo Yoo.

Funding information

US Department of Energy Office of Science. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. Lawrence Berkeley National Laboratory (LBNL) is managed by the Regents of the University of California for the U.S. Department of Energy under Contract no. DE-AC02-05CH11231. Oak Ridge National Laboratory (ORNL) is managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract no. DE-AC05-00OR22725.

Abstract

Accurate representation of environmental controllers of soil organic carbon (SOC) stocks in Earth System Model (ESM) land models could reduce uncertainties in future carbon–climate feedback projections. Using empirical relationships between environmental factors and SOC stocks to evaluate land models can help modelers understand prediction biases beyond what can be achieved with the observed SOC stocks alone. In this study, we used 31 observed environmental factors, field SOC observations ($n = 6,213$) from the continental United States, and two machine learning approaches (random forest [RF] and generalized additive modeling [GAM]) to (a) select important environmental predictors of SOC stocks, (b) derive empirical relationships between environmental factors and SOC stocks, and (c) use the derived relationships to predict SOC stocks and compare the prediction accuracy of simpler model developed with the machine learning predictions. Out of the 31 environmental factors we investigated, 12 were identified as important predictors of SOC stocks by the RF approach. In contrast, the GAM approach identified six (of those 12) environmental factors as important controllers of SOC stocks: potential evapotranspiration, normalized difference vegetation index, soil drainage condition, precipitation, elevation, and net primary productivity. The GAM approach showed minimal SOC predictive importance of the remaining six environmental factors identified by the RF approach. Our derived empirical relations produced comparable

Abbreviations: GAM, generalized additive model; ML, machine learning; NDVI, normalized difference vegetation index; RF, random forest; SOC, soil organic carbon.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Soil Science Society of America Journal* published by Wiley Periodicals LLC on behalf of Soil Science Society of America.

prediction accuracy to the GAM and RF approach using only a subset of environmental factors. The empirical relationships we derived using the GAM approach can serve as important benchmarks to evaluate environmental control representations of SOC stocks in ESMs, which could reduce uncertainty in predicting future carbon–climate feedbacks.

1 | INTRODUCTION

Soils store a large and dynamic fraction of global terrestrial carbon (Sulman et al., 2020) and affect many ecosystem services (Adhikari & Hartemink, 2016; Lal, 2013). Soils can act as sources or sinks of atmospheric carbon dioxide, depending on land use, management interventions, and environmental conditions. Observation-based global soil organic carbon (SOC) stock estimates show large spatial heterogeneity (Batjes, 2016; Hengl et al., 2014). This observed spatial heterogeneity in SOC stocks is primarily controlled by the soil forming factors: climate, organisms, topography, parent material, and time (Jenny, 1941; McBratney et al., 2003). As a result, different combinations of these environmental factors have widely been used for spatial prediction of SOC stocks at different scales (Adhikari et al., 2020; Mishra et al., 2021; Vitharana et al., 2017). Despite their key role in determining the spatial heterogeneity of SOC stocks and regulating land–atmosphere exchanges of carbon, the control of these environmental factors on SOC stocks are not correctly characterized and represented in current land surface model process representations. As a result, land models poorly represent current SOC spatial heterogeneity (Carvalho et al., 2014; Todd-Brown et al., 2013), which contributes to large uncertainty in predicting future carbon–climate feedbacks (Arora et al., 2020; Friedlingstein et al., 2014). Therefore, to reduce uncertainty in future carbon–climate feedback projections, it is critical to accurately (i.e., consistent with observations) represent environmental controllers of SOC stocks in land surface models.

A variety of approaches have been applied to predict the spatial heterogeneity and infer environmental controllers of SOC stocks (Lamichhane et al., 2019; Minasny et al., 2013). Among different approaches applied for spatial predictions of SOC stocks, linear regression and ordinary kriging have been most widely used approaches (Minasny et al., 2013; Olaya-Abril et al., 2017; Zhang et al., 2017). Linear regressions quantify the strength and direction of relationships between environmental factors and SOC stocks and have been applied primarily due to their simplicity and ease of interpretation of the results obtained. Ordinary kriging uses the spatial autocorrelation among existing samples to predict the value of SOC stocks at an unsampled location.

However, several recent studies demonstrated use of nonlinear approaches to predict the spatial heterogeneity of SOC stocks. Among nonlinear methods, machine learning (ML) approaches are increasingly being applied to predict soil properties, including SOC stocks (Lamichhane et al., 2019; Padarian et al., 2020; Siewert, 2018). Heuvelink et al. (2021) used a quantile regression forest ML approach to predict the annual SOC stock of surface soils of Argentina between 1982 and 2017 and reported a larger temporal variation in comparison to the Intergovernmental Panel on Climate Change Tier 1 approach of predicting SOC change. Ottoy et al. (2017) compared four digital soil mapping approaches to predict SOC stocks at a regional scale and reported that boosted regression trees achieved highest prediction accuracy. Authors identified drainage condition, soil type, and vegetation type as important environmental predictors of SOC stocks. Vos et al. (2018) used various data mining approaches to identify and interpret main factors that controlled the cropland SOC stocks. Authors reported land use, land-use history, clay content, and electrical conductivity as main predictors of the topsoil SOC stocks, whereas bedrock material, relief, and electrical conductivity were main predictors of the subsoil carbon stocks. Bui et al. (2009) reported that SOC in Australian agricultural soils were related to vegetation, biomass, soil moisture, and temperature patterns. Authors reported that the structure in the multivariate relationships between environmental factors and soil properties were consistent with principles of pedogenesis and landscape ecology. In a review of digital soil mapping literature, Ma et al. (2019) documented that the pedological knowledge can be used in digital soil mapping and digital soil mapping can also lead to new knowledge discovery regarding the soil formation. However, Wadoux, Samuel-Rosa, et al. (2020) noted that the knowledge discovery based on ML needs to be treated with caution. Interestingly, authors demonstrated how pseudo-covariates not related to any soil-forming factors, and processes can also accurately predict SOC. Therefore, careful preselection and preprocessing of pedologically relevant environmental covariates and the posterior interpretation and evaluation of the recognized patterns can only provide meaningful insights.

More recently, ensembles of multiple approaches have been applied to improve the spatial prediction of SOC stocks (Riggers et al., 2019; Vařát et al., 2017). A recent study showed that the median prediction obtained from an ensemble

of ML approaches better predicts the spatial heterogeneity of SOC stocks in comparison to individual ML or hybrid approaches, such as regression kriging (Mishra et al., 2020). In a majority of previous studies, ML approaches were used to identify important environmental predictors and predict the spatial variation of SOC stocks. In a recent review of ML applications in soil science, Padarian et al. (2020) identified two primary research needs: (a) identification of parsimonious ML models and (b) interpretability of the applied ML models. Similarly, in another review, Wadoux, Minasny, and McBratney (2020) identified the need to incorporate pedological knowledge in ML algorithms to make these approaches more relevant to soil science. These authors identified plausibility, interpretability, and explainability as the greatest challenges in using ML approaches in soil science.

Earlier studies in soil science used a number of ML approaches primarily for digital soil mapping and to identify important predictors of SOC stocks. However, in these studies ML is usually used as a “black box” model, which does not report mathematical relationships between environmental factors and SOC stocks that can be used to predict the SOC stocks. Therefore, our overall objective here was to derive observationally based mathematical relationships describing environmental controls on SOC stocks. The specific objectives were (a) to use ML to select important environmental predictors of SOC stocks, (b) to derive empirical relationships between environmental factors and SOC stocks, and (c) to use the derived relationships to predict the SOC stocks and compare the prediction accuracy of simpler model developed using the derived relationships with ML predictions.

2 | MATERIALS AND METHODS

2.1 | SOC observations

We used field SOC measurements from the rapid carbon assessment project of the Natural Resources Conservation Service’s Soil Science Division of the USDA (Soil Survey Staff & Loecke, 2016). That assessment project was designed to produce a robust estimate of SOC stocks in different kinds of soils and land uses across the conterminous United States based on consistent and dedicated soil sampling. Over 6,200 sampling sites across the conterminous United States (Supplemental Figure S1) were established following a hierarchical sampling design consisting of major land resource areas as first-level strata, which were further stratified based on land use and land cover and soil types in a nested fashion. Soil samples at observation locations were collected from genetic horizons and were analyzed for SOC concentration and bulk density following the Soil Survey Laboratory Methods Manual (Burt, 2004; Grossman & Reinsch, 2002). However, this study considered SOC stock for only the top 30 cm of soil, cal-

Core Ideas

- We used machine learning to derive predictive relationships between environmental factors and SOC stocks.
- Soil drainage, evapotranspiration, and vegetation index were important controllers of SOC stocks.
- Derived relationships produced comparable prediction accuracy using only a subset of environmental factors.
- Derived relationships can be used to benchmark land model representations of SOC stocks.

culated after correcting it for coarse fragments (Equation 1). For soil samples with missing bulk density measurements, a pedotransfer function based on a RF approach was developed (Sequeira et al., 2014) and SOC stock was calculated as

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

where SOC_{stk} is the SOC stock (t C ha^{-1}), SOC is the SOC concentration ($\text{g C } 100 \text{ g}^{-1}$ soil), BD is the soil bulk density (g cm^{-3}), D is the soil layer thickness (cm), and CF is the volumetric fraction of the coarse fragments.

The average SOC stock of continental U.S. surface soil was 9.5 kg m^{-2} , ranging from 0.06 to 127 kg m^{-2} . The observed SOC stocks showed unimodal (kurtosis = 24.7) and positively skewed (coefficient of skewness = 4.3) distribution. After the SOC data was log-transformed, the skewness coefficient and mean dropped to -0.04 and 3.9 Mg ha^{-1} , respectively, with a standard deviation of 1.0 and a CV of 25.7%. About 70% of our SOC data was less than 100 kg m^{-2} , whereas 1.5% of SOC data was larger than $1,000 \text{ kg m}^{-2}$. If we train a ML model for this highly unbalanced dataset, the resulting model will be biased to fit the small fraction of the extremely large values. Therefore, the modeling tasks were conducted using log transformed SOC stocks.

2.2 | Environmental predictors of SOC stocks

We compiled 31 environmental variables from different sources and evaluated their usefulness as predictors of SOC in the study area (Supplemental Table S1). These variables were representative of major soil forming factors: climate, vegetation, topography, and parent material (Jenny, 1941; McBratney et al., 2003). Seven of the 31 variables were climatic variables, obtained from the parameter-elevation regressions on independent slopes model and global climate

and weather data: the 30-yr (1981–2010) annual average of minimum, mean, maximum, and dewpoint temperatures; precipitation as rainfall; rainfall during the wettest and driest quarter in a year; and potential evapotranspiration (Supplemental Table S1). Six of the 31 variables described vegetation characteristics: land use, land cover, potential vegetation cover, remote sensing data (median value of surface reflectance during the growing season), net primary production, and ecological regions. Ten variables related to topography were derived from the national digital elevation model at 30-m spatial resolution that was resampled to 100-m grid scale for this study: elevation, slope aspect, slope length factor, multiresolution valley bottom flatness index, melton ruggedness index, midslope position, wetness index, slope height, slope gradient, and valley depth. Five variables described parent material and soil climate: soil types, surface geology, natural drainage condition, hydrological unit, and soil temperature regime. For these 31 environmental variables, vector layers were rasterized when necessary, and all the raster layers and point SOC observations were projected to a common Universal Transverse Mercator projection system (NAD 1983). The NAD83 projection uses a geocentric datum and geographic coordinate system based on the 1980 Geodetic Reference System ellipsoid. The values of the environmental variables at sampling locations were then extracted and a matrix of SOC stock and 31 predictors (6,123 rows, 34 columns) was created for modeling. All the categorical variables were converted to integer variables before using in this analysis.

2.3 | Dimensionality reduction using RF

We used a random forest (RF) regression approach to identify important environmental predictors of SOC stocks. Random forest is based on a decision tree model and consists of an ensemble of randomized classification or regression trees with a bootstrap aggregation (Breiman, 2001). In RF, a training dataset is first randomly drawn with replacement from the original data set. Then, a decision tree is fitted to the training data set by randomly selecting a subset of the input variables at each branch split. Typically, only $p/3$ variables are used to decide a branch split for a regression tree, where p is the number of predictor variables. The process is repeated to build a large number of uncorrelated trees, hence the name “forest,” and the prediction is computed by averaging the predictions of each tree. Random forest is one of the most popular predictive models in ML due to its outstanding performance even with little parameter tuning (Hastie et al., 2001). The RF model was trained by using the “randomForest” package in R (Liaw & Wiener, 2002). The total number of regression trees (ntree) was set to 500, and $mtry = 10$ ($\cong 31/3$) variables were randomly selected to compute a split at each branch. The number

of minimum data points to stop growing a tree was set to $nodesize = 10$.

We used a “greedy” approach (Edmonds, 1971) to identify uncorrelated sets of environmental predictors of SOC stocks. In the “greedy” approach, the environmental predictors were first arranged according to the variable importance rank from the RF model. The Pearson’s correlation coefficients between the environmental predictors were calculated, and environmental predictors with absolute value of the correlation coefficients larger than a threshold (taken as 0.6), were removed from the dataset. After removing the correlated variables, there were 19 environmental predictors remaining in the dataset. Another RF model was trained with the remaining 19 environmental predictors.

The variable importance was computed by the random permutation method, where one of the environmental variables is randomly permuted between the out-of-bag samples and the change in the prediction accuracy [$R^2(1 - \text{residual sum of square}/\text{total sum of square})$ and RMSE] due to the random permutation provides a measure for the importance of the environmental variable (Hastie et al., 2001). The permutation-based importance is one of the most common approach to assess the relative importance between input variables in the RF approach. However, the variable importance rank provides only the qualitative importance of the environmental predictors. To quantitatively investigate the effects of environmental predictors on SOC stocks, we trained a set of RF models with different sets of environmental predictors and measured the changes in the prediction accuracy. The number of environmental predictors in RF models was varied from 1 to 19. The first RF model is trained with the environmental predictor with the highest importance and, in the successive models, the number of environmental predictors is gradually increased in the order of the variable importance.

2.4 | GAMs to derive empirical relationships between environmental predictors and SOC stocks

Random forest is a powerful ML technique due to its strength in computing nonlinear relations between input and output variables. However, RF is essentially a “black box” model, which does not provide detailed information about the relationships between the input and output variables. The function between predictor variables and response is particularly challenging to tease apart. This makes it difficult to use RF to find an empirical relationship between a particular environmental predictor and SOC, particularly when the data points are not uniformly distributed over the high-dimensional feature space. Therefore, we used a generalized additive model (GAM) to derive empirical relationships between the RF-identified environmental predictors and SOC stocks. In GAM,

the relationship in the data can be modeled as (Hastie & Tibshirani, 1990; Hastie et al., 2001):

$$Y = C + \sum_{i=1}^p f_i(X_i) \quad (2)$$

Here, Y is the target variable (e.g., observed SOC), X_i is an environmental variable, f_i is a smooth function, and C is a constant, which is usually a mean of Y . Generalized additive models can be thought as a generalization of multilinear regression, but without the linear assumptions. This is performed by replacing the linear β parameters of the form $Y = C + \sum_{i=1}^p \beta_i X_i$ with a smoothing function f , usually in the form of additive splines. This allows the influence of an individual predictor variable to be decoupled and compared with the target variable, without requiring linearity of relationship between the predictor variable and target variable.

The thin plate spline is used for the smoother, $f_i(X_i)$ (Wood, 2003). For a one-dimensional problem, the smoothing function is found by minimizing

$$\sum_{i=1}^N [Y^i - f(X^i)]^2 + \lambda \int \left[\frac{d^2 f(x)}{dx^2} \right]^2 dx \quad (3)$$

in which Y^i and X^i , respectively, denote the target and the input feature, N is the total number of data, and λ is a penalty parameter. The function that minimizes Equation 3 is given as

$$f(x) = \sum_{i=1}^N \delta_i \eta(|x - X^i|) + \sum_{j=1}^2 \alpha_j \phi_j(x) \quad (4)$$

Here, δ_i and α_j are unknown parameters, ϕ_j is the $(j - 1)$ -th order polynomial, and $\eta(r) = r^3$. Furthermore, δ is approximated by a reduced order basis as $\delta = \mathbf{U}_k \delta_k$, in which \mathbf{U}_k is a rank- k matrix. The rank of \mathbf{U}_k denotes the maximum degree of freedom of the thin plate spline. To prevent an overfitting, k is chosen to be four. The unknown parameters, \mathbf{U}_k , δ_k , and, α , are estimated from the data by solving a regularized optimization problem as shown in Wood (2003). The GAM model (Equation 2) is then computed by iteratively computing the one-dimensional thin plate splines for each environmental variable, using the backfitting algorithm (Hastie et al., 2001). For our analysis here, we used the “mgcv” package in R to train a GAM model (Wood, 2017) using a restricted maximum likelihood method, and a thin plate spline for the smooth functions (Wood, 2003).

2.5 | Evaluation of prediction accuracy

To evaluate the prediction accuracy of the RF, GAM, and analytical models, we calculated coefficient of determination

(R^2) and RMSE using a 10-fold cross validation approach. In this approach each model is refitted 10 times using 70% of the SOC observations, and the predictions obtained from the fitted models were compared with the remaining 30% of observations. For each model, R^2 and RMSE were calculated using the following equations:

$$R^2 = \left(1 - \frac{\text{SSE}}{\text{SST}} \right) 100\%$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n [\hat{\text{SOC}}(x_i) - \text{SOC}(x_i)]^2}$$

where SSE is the sum of squared errors at cross-validation points, SST is the total sum of squares, $\text{SOC}(x_i)$ is the measured SOC, $\hat{\text{SOC}}(x_i)$ is the model predicted SOC, and n is the number of validation observations. A coefficient of determination close to 1 indicates a perfect model (i.e., 100% of variation has been explained by the model). For optimal predictions, RMSE values should approach zero.

3 | RESULTS

3.1 | Dominant environmental predictors of continental U.S. surface SOC stocks

The importance of all the environmental predictors of SOC stocks in descending order, as estimated by RF, is provided in Figure 1a. The resulting variable importance shows that soil drainage has the dominant effect on continental U.S. surface SOC stocks, followed by normalized difference vegetation index (NDVI) and dry-season precipitation. We also found that many of the environmental predictors used in this study were correlated with each other (Figure 1b). Whereas RF offers a good predictive model, it lacks the capability to identify multicollinearity in the environmental predictors (Mishra et al., 2020). Hence, as explained in Section 2.4, we removed the correlated variables (resulting in 19 variables) and reapplied the RF approach with the reduced number of environmental predictors.

Variable importance ranking changed after correlated environmental predictors were removed (Figure 2a), as did the incremental changes in R^2 with respect to the number of the environmental predictors (Figure 2b). We found significant improvement in the RF performance (R^2 and RMSE) as the number of environmental predictors increased from 1 to 8 (with the predictors ordered by the RF-inferred importance; Figure 2b). However, after 12 environmental predictors, the improvement in model prediction accuracy was minimal. These results suggested that among all the environmental

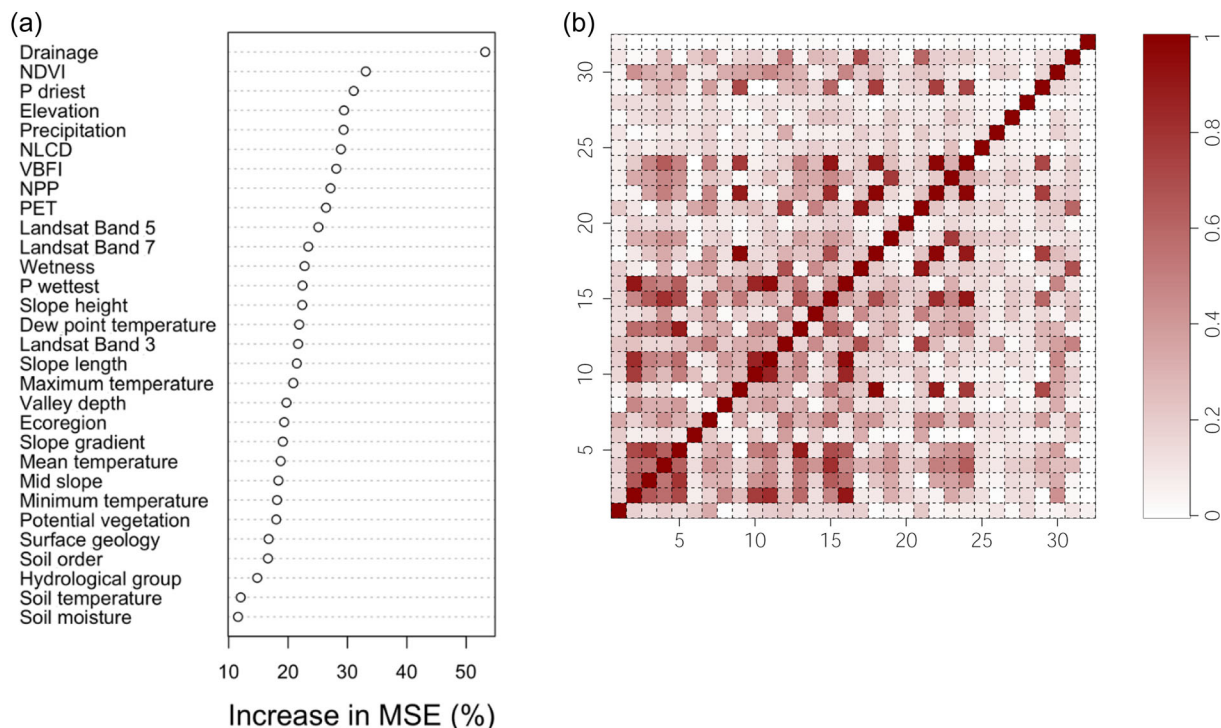


FIGURE 1 (a) Variable importance for the top 30 variables and (b) absolute values of the correlation coefficients between the variables. The index corresponds to the variable importance rank. MSE is mean squared error, NDVI is normalized difference vegetation index, P driest is precipitation in driest season, NLCD is national land cover database, VBFI is valley bottom flatness index, NPP is net primary productivity, PET is potential evapotranspiration, P wettest is precipitation of the wettest season

predictors we used, only 12 environmental predictors were the strongest predictors of SOC stocks.

3.2 | Nonlinear controls of environmental factors on SOC stocks

Using the 12 most important environmental predictors identified by RF as an input feature set, we trained the GAM to fit the log-transformed SOC stocks (Figure 3). The constant term in the GAM was $C = 3.98$. R^2 and RMSE were .52 and .69, respectively. The error metrics of GAM were slightly lower than for the RF ($R^2 = .56$, RMSE = 0.66). Whereas RF considers high-order nonlinear interactions between the environmental predictors, in GAM, SOC is modeled by a linear combination of nonlinear functions of each environmental predictor, not considering interactions between them, which may have resulted in a slightly lower prediction accuracy. Figure 3 shows the GAM-inferred relationships between environmental factors and log-transformed SOC stocks with respect to the 12 most important variables. Potential evapotranspiration, NDVI, and soil drainage condition are the three most important variables from RF (Figure 2a).

The empirical relationships between SOC stocks and environmental predictors were produced as splines by GAM. We next developed explicit analytical expressions by fitting the splines obtained from GAM. Figure 3 shows that the changes of SOC stocks with respect to many of the environmental variables ($n = 6$) are essentially negligible after considering the uncertainty. Hence, we identified only the following six important environmental variables: potential evapotranspiration, NDVI, soil drainage condition, precipitation of the wettest season, elevation, and net primary productivity;

- Potential evapotranspiration (PET):

$$\begin{cases} z = \frac{\text{PET}-641}{1,000} \\ Y_{\text{PET}} = \exp(0.44 - 1.24z - 1.51z^2 + 0.05z^3) - 0.6 \end{cases}$$

- Normalized difference vegetation index (NDVI):

$$Y_{\text{NDVI}} = \begin{cases} 0.078 + 1.87(\text{NDVI}16 - 0.4)^{1.62} & \text{if } \text{NDVI}16 > 0.4 \\ 0.078 - 4.36|\text{NDVI}16 - 0.4|^{2.44} & \text{otherwise} \end{cases}$$

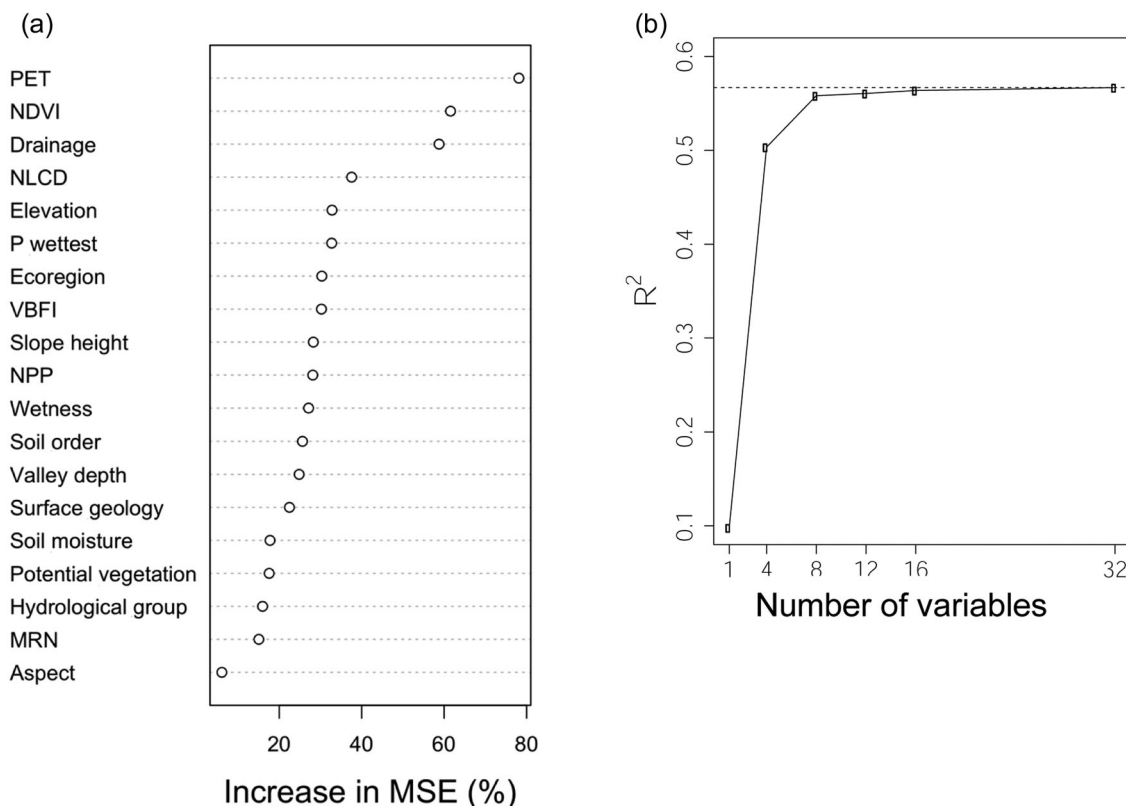


FIGURE 2 (a) Variable importance after removing correlated variables. (b) Changes in the model accuracy in terms of the number of the input variables of the random forest model. MSE is mean squared error, PET is potential evapotranspiration, NDVI is normalized difference vegetation index, NLCD is national land cover database, P wettest is precipitation of the wettest season, VBFI is valley bottom flatness index, NPP is net primary productivity, MRN is melton ruggedness number

- Soil drainage:

$$Y_{\text{Soil drainage}} = \begin{cases} -0.38 & \text{if Soil drainage} = 1 \\ -0.05 & \text{if Soil drainage} = 2 \\ 0.15 & \text{if Soil drainage} = 3 \\ 0.50 & \text{if Soil drainage} = 4 \\ 1.00 & \text{if Soil drainage} = 5 \end{cases}$$

- Elevation:

$$\begin{cases} z = \frac{\text{Elevation}}{1,000} \\ Y_{\text{Elevation}} = 0.17 - \exp[-1.34 - 0.75z(1 + 0.1z^2)] \end{cases}$$

- Precipitation:

$$\begin{cases} z = \frac{\text{Precipitation}}{250} \\ Y_{\text{Precipitation}} = 0.38 - \exp(-0.15 - 3.24z^{1.5}) \end{cases}$$

- Net primary productivity (NPP):

$$Y_{\text{NPP}} = 0.077 - 1.68 \times 10^{-5} \text{NPP}$$

The fitted curves accurately represented the splines from GAM (Figure 4). The log-transformed SOC stocks from the GAM approach were computed using the following equation:

$$\ln(\text{SOC}) = Y_{\text{PET}} + Y_{\text{NDVI}} + Y_{\text{Soil drainage}} + Y_{\text{Elevation}} + Y_{\text{Precipitation}} + Y_{\text{NPP}} + 3.98$$

Here, $\ln(\text{SOC})$ is log transformed SOC stocks, Y_{PET} is the empirical relation of PET with SOC stocks, Y_{NDVI} is the empirical relation of NDVI with SOC stocks, $Y_{\text{Soil drainage}}$ is the empirical relation of soil drainage with SOC stocks, $Y_{\text{Elevation}}$ is the empirical relation of elevation with SOC stocks, $Y_{\text{Precipitation}}$ is the empirical relation of precipitation with SOC stocks, and Y_{NPP} is the empirical relation of net primary productivity with SOC stocks. Among the environmental factors selected in GAM model, all environmental factors were continuous variables, except the soil drainage. As the soil drainage is a discrete variable, the equation of soil drainage is different than other reported equations.

Our results show that the analytical model we developed using only six environmental predictors (Figure 5) showed similar prediction accuracy as that obtained from the GAM

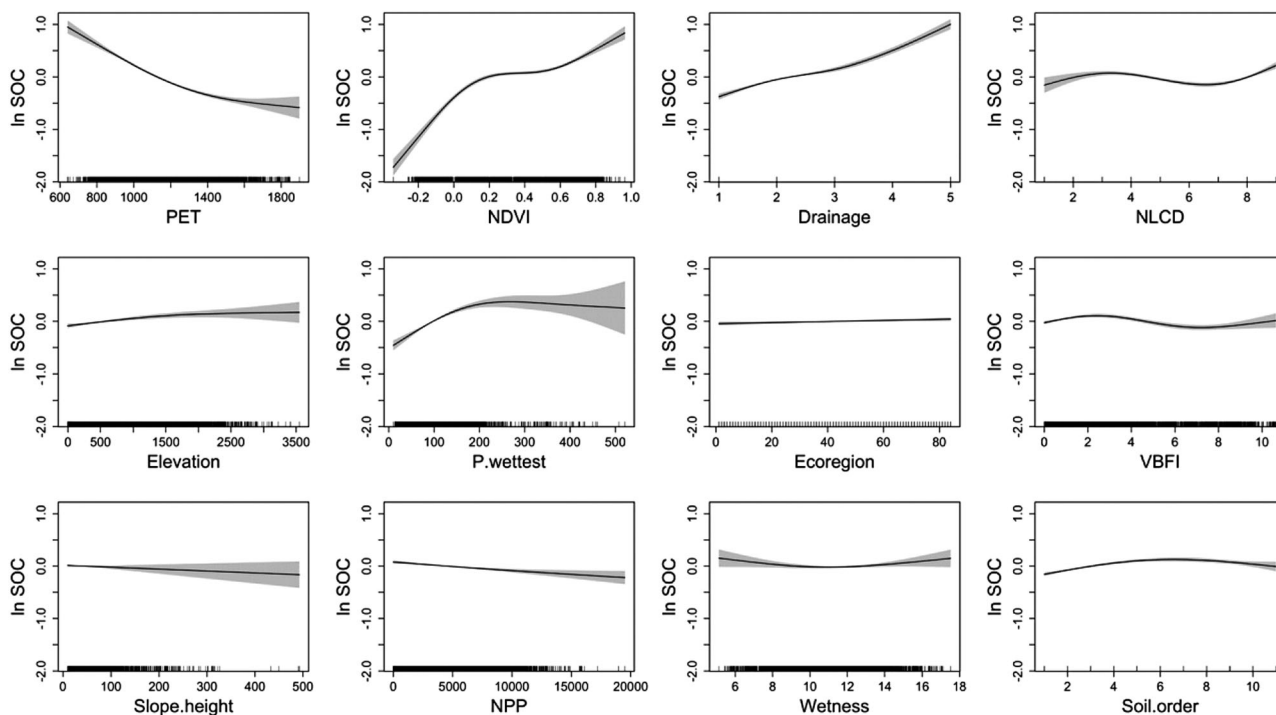


FIGURE 3 Variable-wise prediction of observed soil organic carbon ($\ln \text{SOC}$, $\ln \text{kg m}^{-2}$) by the generalized additive model. The shade around the solid line indicates 95% confidence interval. The minor ticks on the horizontal axis denote the values of data. PET is potential evapotranspiration, NDVI is normalized difference vegetation index, NLCD is national land cover database, P.wettest is precipitation of the wettest season, VBFI is valley bottom flatness index, and NPP is net primary productivity

with 12 variables. Using only the first three environmental predictors (potential evapotranspiration, NDVI, and soil drainage condition) together with the constant term (3.98), the analytical model achieved an R^2 of .48, indicating relatively marginal importance of the remaining three environmental factors (elevation, precipitation, and net primary productivity). Figures 5a and 5b show the comparison between the GAM model with all the 12 environmental variables and the analytical model with six environmental variables in predicting SOC stocks.

4 | DISCUSSION

Our study demonstrates use of ML to improve understanding of nonlinear controls of environmental factors on SOC stocks. We developed an approach to derive analytical expressions for observationally derived environmental controls on SOC stocks. In this approach, we first identified the dominant environmental predictors of continental U.S. surface SOC stocks using a RF approach. We then derived mathematical equations that captured environmental controls on SOC stocks using a GAM approach. The mathematical relations we derived produced comparable prediction accuracy consistent with the RF approach, using only a subset of environmental predictors

used in the RF approach. Our approach of deriving analytical relationships between environmental factors and SOC stocks can be used to evaluate ESM representations of environmental controls on SOC stocks. However, we note that our study quantified these relationships at a much finer resolution (100 m) than typically used in ESM land models for global simulations ($\sim 10\text{--}100 \text{ km}$). Therefore, a first step in evaluating ESM land model SOC predictions using these derived analytical relationships would be to run the models at fine resolution using appropriate forcing, initial conditions, and site characteristics. Such an analysis could point to deficiencies in the models' mechanistic representations so that evaluation at ESM resolutions could focus on spatial scaling methods.

Our analysis identified six environmental factors (potential evapotranspiration, vegetation index, soil drainage condition, precipitation, elevation, and net primary productivity) as dominant predictors of continental U.S. surface SOC stocks among the 31 environmental predictors we evaluated. Out of these 6 environmental factors, potential evapotranspiration, soil drainage condition, and NDVI were the most important environmental predictors of SOC stocks. Elevation and net primary productivity showed marginal importance in predicting continental US surface SOC stocks, although these are key environmental controls in current ESM land models.

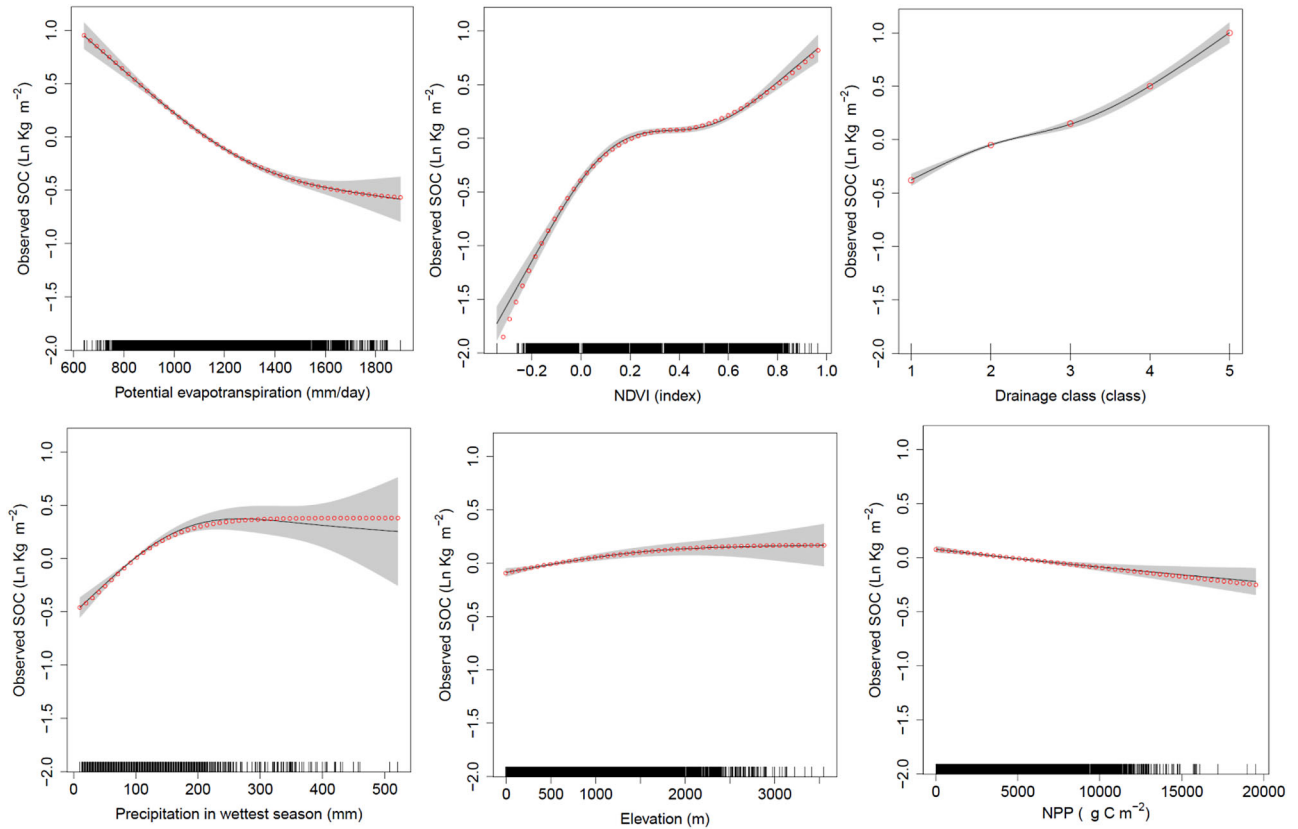


FIGURE 4 Curve fittings of the splines from the generalized additive model. the solid lines are the expectation values from the generalized additive model, and the circles are computed from the fitting curves. The shade around the solid line indicates 95% confidence interval

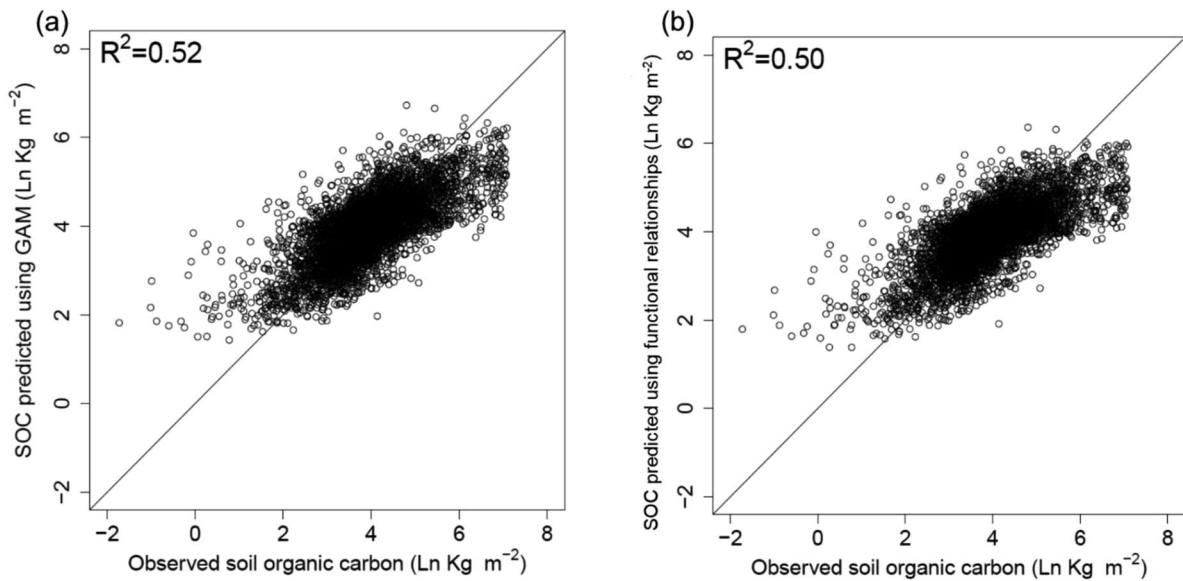


FIGURE 5 Comparison of the model predictions between (a) GAM (generalized additive model) with 12 variables, and (b) analytical model with six variables. SOC, soil organic carbon

Various earlier studies also used different combination of these environmental factors to predict SOC stocks at different scales and in different environmental conditions (Gonçalves et al., 2021; Lamichhane et al., 2019; Minasny et al., 2013; Mishra et al., 2020, 2021). Garten et al. (2009) reported that the control of soil moisture on bulk SOC and its fractions were greater than the controls of elevated carbon dioxide and temperature individually at a field scale. Consistent with this finding, the dominant controls of potential evapotranspiration, soil drainage condition, and precipitation demonstrate the control of soil moisture on SOC stocks across the continental U.S. Our results show that SOC stocks decreased exponentially with increases in potential evapotranspiration. In our dataset, higher potential evapotranspiration values are located in the southern United States (Supplemental Figure S1), which has higher air temperatures and solar radiation in comparison with other parts of the United States. Higher air temperatures and longer duration of solar radiation cause drier soil conditions, promoting SOC mineralization and lower total SOC stocks (Das et al., 2019; Hungate et al., 2002; Sherrod et al., 2005).

Our results show lower SOC stocks in excessively drained soils (Number 1) and higher SOC stocks in poorly drained soils (Number 5) across the continental United States. Excessively drained soils are generally coarse-textured soils with high saturated hydraulic conductivity. Similarly, poorly drained soils are often fine-textured soils with more of their pore space filled with water for longer periods of time. Our results are consistent with findings of earlier studies, which showed mean soil carbon concentration significantly differed across different soil drainage classes (Raymond et al., 2013; Wickland et al., 2010). Poorly and very poorly drained soils have lower soil respiration rates (Davis et al., 2010; Webster et al., 2008) compared with well-drained soils (Davidson et al., 1998; Savage & Davidson, 2001), resulting in higher SOC preservation. Some studies suggest precipitation has a strong positive correlation with SOC (Alvarez & Lavado, 1998; Burke et al., 1989; Evans et al., 2011), whereas other studies show precipitation has little to no influence on SOC (Doetterl et al., 2015; Percival et al., 2000). Our results show increased SOC stocks with increases in precipitation up to 200 mm yr⁻¹. Beyond 200 mm yr⁻¹, the effect of precipitation on SOC stocks was small. Considering precipitation as a proxy for soil moisture content, control of precipitation on SOC stocks is higher in drier areas of the continental United States than in areas with higher precipitation.

Our results indicate that with increased vegetation index, continental U.S. surface SOC stocks increased nonlinearly. We found large increases in SOC stocks as annual average NDVI values increased from -0.2 to 0.2, but the relationship between SOC and NDVI flattened at higher NDVI values (>0.2 to 1). This relationship could be due to nonlinear relationships between chlorophyll concentration of green biomass

and the calculated NDVI values (Yoder & Waring, 1994). Vegetation properties have been documented as strong predictors of SOC stocks (Guo et al., 2016; Jobbágy & Jackson, 2000; Li et al., 2010) and widely used in statistical and process-based models to predict SOC stocks (Gautam et al., 2020; Mishra et al., 2021).

Pedological knowledge about the study area should be used while selecting and using environmental variables in ML approaches (Wadoux, Samuel-Rosa, et al., 2020). A number of earlier studies have documented that ML selected environmental predictors were related with soil forming processes and factors in various environments. For example, Brungard et al. (2015) reported that the environmental covariates selected by ML provided information about the soil erosion and deposition processes in semiarid environments of the United States, which controlled soil type distribution across the study area landscapes. Hengl et al. (2017) used ML to select environmental covariates which represented factors of soil formation at global scale: climate, relief, living organisms, water dynamics, and parent material. Shi et al. (2018) reported that ML based feature selection methods provided a useful way to understand the relationships between soil properties and environmental variables and map soil properties accurately. Therefore, environmental factors selected by ML are not necessarily spurious but often result from processes that regulate the soil formation, and factors that determine the soil properties distribution across landscapes.

Our results are based on the investigation of the 31 environmental covariates that we used in this study. Though these environmental variables represent major soil forming factors, the empirical relationships between environmental factors and SOC stocks that we reported might change in presence of other environmental variables, which are currently not included in this study. Further, these environmental variables were collected from various sources with different spatial resolutions and accuracies. Our analysis has not included the effect of these uncertainties. Previous scaling studies of SOC stocks (Adhikari et al., 2020; Mishra & Riley, 2015) showed importance of different environmental factors on SOC stocks across different spatial scales. These findings suggest that the nonlinear mathematical algorithms we developed in this study may only be used to infer environmental controls on SOC stocks at the same spatial scale (100 m). Further, interactions between environmental factors in controlling the variability of SOC stocks need to be further investigated.

It is important to note the limitations of current ML approaches. First, the environmental covariates we used were a mixture of continuous and discrete variables. Whereas the discrete variables were converted to numeric variables by an integer encoding, such integer encoding may not correctly reflect the underlying structure of the environmental variable, which makes their importance ranked low in RF and eventually removed in the feature selection process. It

is an open question how to build an accurate ML model for such multimodal data. Second, when the data exhibits multicollinearity, it is difficult to select the true environmental factor that controls the dependent variable. Machine learning models may select any proxy that is highly correlated with the dependent variable. Hence, it is important to carefully review and include scientific understanding in the feature selection process. Lastly, in this study, we focused on identifying nonlinear structures between observed environmental controllers and SOC stocks. It is important to note that, however, the correlation does not imply causation. It is an area of active research to identify a causal structure from the observations.

5 | CONCLUSIONS

Appropriate representation of environmental controllers on SOC stocks in land models is required to project realistic rates of SOC change in response to land use and climate changes, and to understand feedbacks between the land and atmosphere. The nonlinear expressions we derived quantify controls of individual environmental factors on SOC stocks in the presence of other environmental factors. These observationally derived analytical expressions can be used for (a) benchmarking land model representations of environmental controls on SOC stocks, and (b) digital soil mapping. Our analysis showed potential evapotranspiration, NDVI, soil drainage condition, precipitation, elevation, and net primary productivity as important environmental controllers of continental U.S. surface SOC stocks. Out of these six environmental factors, potential evapotranspiration, NDVI, and soil drainage condition explained about 50% of the variability in observed SOC stocks (whereas the other three environmental variables explained another 6% of the variability). Our derived analytical expressions produced comparable prediction accuracy to the GAM and RF using only a subset of environmental factors. Future studies should investigate the functional forms of the relationships we derived to describe environmental controllers on SOC stocks, evaluate and compare with land model emergent relationships, and use these comparisons to improve mechanistic representations in land models.

ACKNOWLEDGMENTS

This study was supported jointly by the Laboratory Directed Research and Development program of Sandia National Laboratories and the Reducing Uncertainties in Biogeochemical Interactions through Synthesis and Computation Science Focus Area (RUBISCO SFA), which is sponsored by the Regional and Global Model Analysis (RGMA) activity of the Earth Environmental Systems Modeling (EESM) Program in the Earth and Environmental Systems Sciences Division (EESD) of the Office of Biological and Environmental

Research (BER) in the US Department of Energy Office of Science. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. Lawrence Berkeley National Laboratory (LBNL) is managed by the Regents of the University of California for the U.S. Department of Energy under Contract no. DE-AC02-05CH11231. Oak Ridge National Laboratory (ORNL) is managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract no. DE-AC05-00OR22725. Thanks to S. Wills for providing access to the SOC profile data. USDA is an equal opportunity provider and employer.

AUTHOR CONTRIBUTIONS

Umakant Mishra: Conceptualization; Formal analysis; Supervision; Writing – original draft. Kyongmin Yeo: Formal analysis; Writing – original draft; Writing – review & editing. Kabindra Adhikari: Data curation; Writing – review & editing. William J. Riley: Supervision; Writing – review & editing. Forrest M. Hoffman: Supervision; Writing – review & editing. Corey Hudson: Supervision; Writing – review & editing. Sagar Gautam: Writing – review & editing.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Umakant Mishra  <https://orcid.org/0000-0001-5123-2803>
Kabindra Adhikari  <https://orcid.org/0000-0002-1365-7015>

REFERENCES

- Adhikari, K., & Hartemink, A. E. (2016). Linking soils to ecosystem services: A global review. *Geoderma*, 262, 101–111. <https://doi.org/10.1016/j.geoderma.2015.08.009>
- Adhikari, K., Mishra, U., Owens, P., Libohova, Z., Wills, S., Riley, W. J., Hoffman, F. M., & Smith, D. (2020). Importance and strength of environmental controllers of soil organic carbon changes with scale. *Geoderma*, 375, 114472. <https://doi.org/10.1016/j.geoderma.2020.114472>
- Alvarez, R., & Lavado, R. S. (1998). Climate, organic matter and clay content relationships in the Pampa and Chaco soils, Argentina. *Geoderma*, 83(1–2), 127–141. [https://doi.org/10.1016/S0016-7061\(97\)00141-9](https://doi.org/10.1016/S0016-7061(97)00141-9)
- Arora, V. K., Katavouta, A., Williams, R. G., Jones, C. D., Brovkin, V., Friedlingstein, P., Schwinger, J., Bopp, L., Boucher, O., Cadule, P., Chamberlain, M. A., Christian, J. R., Delire, C., Fisher, R. A., Hajima, T., & Ziehn, T. (2020). Carbon–concentration and carbon–climate feedbacks in CMIP6 models and their comparison to CMIP5 models. *Biogeosciences*, 17(16), 4173–4222. <https://doi.org/10.5194/bg-17-4173-2020>

- Batjes, N. H. (2016). Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. *Geoderma*, 269, 61–68. <https://doi.org/10.1016/j.geoderma.2016.01.034>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., & Edwards, T. C. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*, 239–240, 68–83. <https://doi.org/10.1016/j.geoderma.2014.09.019>
- Bui, E. N., Henderson, B. L., & Viergever, K. (2009). Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia. *Global Biogeochemical Cycles*, 23, GB4033. <https://doi.org/10.1029/2009GB003506>
- Burke, I. C., Yonker, C. M., Parton, W. J., Cole, C. V., Flach, K., & Schimel, D. S. (1989). Texture, climate, and cultivation effects on soil organic matter content in US grassland soils. *Soil Science Society of America Journal*, 53(3), 800–805. <https://doi.org/10.2136/sssaj1989.03615995005300030029x>
- Burt, R. (2004). *Soil survey laboratory methods manual*. USDA-ARS.
- Carvalho, N., Forkel, M., Khomik, M., Bellarby, J., Jung, M., Migliavacca, M., Mu, M., Saatchi, S., Santoro, M., Thurner, M., Ahrens, B., Beer, C., Cescatti, A., Randerson, J. T., Reichstein, M., & Weber, U. (2014). Global covariation of carbon turnover times with climate in terrestrial ecosystems. *Nature*, 514(7521), 213–217. <https://doi.org/10.1038/nature13731>
- Das, S., Richards, B. K., Hanley, K. L., Kroumbi, L., Walter, M., Walter, M. T., Steenhuis, T. S., & Lehmann, J. (2019). Lower mineralizability of soil carbon with higher legacy soil moisture. *Soil Biology and Biochemistry*, 130, 94–104. <https://doi.org/10.1016/j.soilbio.2018.12.006>
- Davidson, E. A., Belk, E., & Boone, R. D. (1998). Soil water content and temperature as independent or confounded factors controlling soil respiration in a temperate mixed hardwood forest. *Global Change Biology*, 4(2), 217–227. <https://doi.org/10.1046/j.1365-2486.1998.00128.x>
- Davis, A. A., Compton, J. E., & Stolt, M. H. (2010). Soil respiration and ecosystem carbon stocks in New England forests with varying soil drainage. *Northeastern Naturalist*, 17(3), 437–454. <https://doi.org/10.1656/045.017.0306>
- Doetterl, S., Stevens, A., Six, J., Merckx, R., Van Oost, K., Pinto, M. C., Casanova-Katny, A., Muñoz, C., Boudin, M., Boeckx, P., & Venegas, E. Z. (2015). Soil carbon storage controlled by interactions between geochemistry and climate. *Nature Geoscience*, 8(10), 780–783. <https://doi.org/10.1038/ngeo2516>
- Edmonds, J. (1971). Matroids and the greedy algorithm. *Mathematical Programming*, 1(1), 127–136. <https://doi.org/10.1007/BF01584082>
- 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 Biogeochemical Cycles*, 25(3). <https://doi.org/10.1029/2010GB003945>
- Friedlingstein, P., Meinshausen, M., Arora, V. K., Jones, C. D., Anav, A., Liddicoat, S. K., & Knutti, R. (2014). Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks. *Journal of Climate*, 27(2), 511–526. <https://doi.org/10.1175/JCLI-D-12-00579.1>
- Garten, C. T., Classen, A. T., & Norby, R. J. (2009). Soil moisture surpasses elevated CO₂ and temperature as a control on soil carbon dynamics in a multi-factor climate change experiment. *Plant and Soil*, 319(1), 85–94. <https://doi.org/10.1007/s11104-008-9851-6>
- Gautam, S., Mishra, U., Scown, C. D., & Zhang, Y. (2020). Sorghum biomass production in the continental United States and its potential impacts on soil organic carbon and nitrous oxide emissions. *GCB Bioenergy*, 12(10), 878–890. <https://doi.org/10.1111/gcbb.12736>
- Gonçalves, D. R. P., Mishra, U., Wills, S., & Gautam, S. (2021). Regional environmental controllers influence continental scale soil carbon stocks and future carbon dynamics. *Scientific Reports*, 11(1), 1–10. <https://doi.org/10.1038/s41598-021-85992-y>
- Grossman, R., & Reinsch, T. (2002). Bulk density and linear extensibility. In Dane, J. H., & Topp, G. C. (Eds.), *Methods of soil analysis: Part 4. Physical methods* (pp. 201–228). SSSA. <https://doi.org/10.2136/sssabookser5.4.c9>
- Guo, X., Meng, M., Zhang, J., & Chen, H. Y. (2016). Vegetation change impacts on soil organic carbon chemical composition in subtropical forests. *Scientific Reports*, 6(1), 29607. <https://doi.org/10.1038/srep29607>
- Hastie, T., & Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, 46, 1005–1016. <https://doi.org/10.2307/2532444>
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning. Springer series in statistics*. Springer.
- Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotic, 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., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*, 12(2), e0169748. <https://doi.org/10.1371/journal.pone.0169748>
- Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J. G. B., Gonzalez, M. R., & Walsh, M. G. (2014). SoilGrids1km—global soil information based on automated mapping. *PLOS ONE*, 9(8), e105992. <https://doi.org/10.1371/journal.pone.0105992>
- Heuvelink, G. B., Angelini, M. E., Poggio, L., Bai, Z., Batjes, N. H., van den Bosch, R., Bossio, D., Estella, S., Lehmann, J., Sanderman, J., & Olmedo, G. F. (2021). Machine learning in space and time for modelling soil organic carbon change. *European Journal of Soil Science*, 72(4), 1607–1623. <https://doi.org/10.1111/ejss.12998>
- Hungate, B. A., Reichstein, M., Dijkstra, P., Johnson, D., Hymus, G., Tenhunen, J., Hinkle, C. R., & Drake, B. (2002). Evapotranspiration and soil water content in a scrub-oak woodland under carbon dioxide enrichment. *Global Change Biology*, 8(3), 289–298. <https://doi.org/10.1046/j.1365-2486.2002.00468.x>
- Jenny, H. (1941). *Factors of soil formation: A system of quantitative pedology*. McGraw-Hill Book Company.
- Jobbágy, E. G., & Jackson, R. B. (2000). The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecological Applications*, 10(2), 423–436. [https://doi.org/10.1890/1051-0761\(2000\)010%5b0423:TVDOSO%5d2.0.CO;2](https://doi.org/10.1890/1051-0761(2000)010%5b0423:TVDOSO%5d2.0.CO;2)
- Lal, R. (2013). Soils and ecosystem services. In *Ecosystem services and carbon sequestration in the biosphere* (pp. 11–38.). Springer.
- Lamichhane, S., Kumar, L., & Wilson, B. (2019). Digital soil mapping algorithms and covariates for soil organic carbon mapping and their

- implications: A review. *Geoderma*, 352, 395–413. <https://doi.org/10.1016/j.geoderma.2019.05.031>
- Li, P., Wang, Q., Endo, T., Zhao, X., & Kakubari, Y. (2010). Soil organic carbon stock is closely related to aboveground vegetation properties in cold-temperate mountainous forests. *Geoderma*, 154(3–4), 407–415. <https://doi.org/10.1016/j.geoderma.2009.11.023>
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18–22.
- Ma, Y., Minasny, B., Malone, B. P., & McBratney, A. B. (2019). Pedology and digital soil mapping (DSM). *European Journal of Soil Science*, 70, 216–235. <https://doi.org/10.1111/ejss.12790>
- McBratney, A. B., Santos, M. M., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, 117(1–2), 3–52. [https://doi.org/10.1016/S0016-7061\(03\)00223-4](https://doi.org/10.1016/S0016-7061(03)00223-4)
- Minasny, B., McBratney, A. B., Malone, B. P., & Wheeler, I. (2013). Digital mapping of soil carbon. *Advances in Agronomy*, 118, 1–47. <https://doi.org/10.1016/B978-0-12-405942-9.00001-3>
- Mishra, U., Gautam, S., Riley, W., & Hoffman, F. M. (2020). Ensemble machine learning approach improves predicted spatial variation of surface soil organic carbon stocks in data-limited northern circumpolar region. *Frontiers in Big Data*, 3, 40. <https://doi.org/10.3389/fdata.2020.528441>
- Mishra, U., Hugelius, G., Shelef, E., Yang, Y., Strauss, J., Lupachev, A., Harden, J. W., Jastrow, J. D., Ping, C.-L., Schuur, E. A. G., Matamala, R., Siewert, M., Nave, L. E., Koven, C. D., Fuchs, M., Palmtag, J., Kuhry, P., Treat, C. C., Zubrzycki, S., ... Riley, W. J. (2021). Spatial heterogeneity and environmental predictors of permafrost region soil organic carbon stocks. *Science Advances*, 7(9), eaaz5236. <https://doi.org/10.1126/sciadv.aaz5236>
- Mishra, U., & Riley, W. (2015). Scaling impacts on environmental controls and spatial heterogeneity of soil organic carbon stocks. *Biogeosciences*, 12(13), 3993–4004. <https://doi.org/10.5194/bg-12-3993-2015>
- Olaya-Abril, A., Parras-Alcántara, L., Lozano-García, B., & Obregón-Romero, R. (2017). Soil organic carbon distribution in Mediterranean areas under a climate change scenario via multiple linear regression analysis. *Science of the Total Environment*, 592, 134–143. <https://doi.org/10.1016/j.scitotenv.2017.03.021>
- Ottoy, S., De Vos, B., Sindayihubura, A., Hermy, M., & Van Orshoven, J. (2017). Assessing soil organic carbon stocks under current and potential forest cover using digital soil mapping and spatial generalisation. *Ecological Indicators*, 77, 139–150. <https://doi.org/10.1016/j.ecolind.2017.02.010>
- Padarian, J., Minasny, B., & McBratney, A. B. (2020). Machine learning and soil sciences: A review aided by machine learning tools. *soilless*, 6(1), 35–52. <https://doi.org/10.5194/soil-6-35-2020>
- Percival, H. J., Parfitt, R. L., & Scott, N. A. (2000). Factors controlling soil carbon levels in New Zealand grasslands is clay content important? *Soil Science Society of America Journal*, 64(5), 1623–1630. <https://doi.org/10.2136/sssaj2000.6451623x>
- Raymond, J. E., Fernandez, I. J., Ohno, T., & Simon, K. (2013). Soil drainage class influences on soil carbon in a New England forested watershed. *Soil Science Society of America Journal*, 77(1), 307–317. <https://doi.org/10.2136/sssaj2012.0129>
- Riggers, C., Poeplau, C., Don, A., Bammeringer, C., Höper, H., & Dechow, R. (2019). Multi-model ensemble improved the prediction of trends in soil organic carbon stocks in German croplands. *Geoderma*, 345, 17–30. <https://doi.org/10.1016/j.geoderma.2019.03.014>
- Savage, K., & Davidson, E. (2001). Interannual variation of soil respiration in two New England forests. *Global Biogeochemical Cycles*, 15(2), 337–350. <https://doi.org/10.1029/1999GB001248>
- Sequeira, C. H., Wills, S. A., Seybold, C. A., & West, L. T. (2014). Predicting soil bulk density for incomplete databases. *Geoderma*, 213, 64–73. <https://doi.org/10.1016/j.geoderma.2013.07.013>
- Sherrod, L. A., Peterson, G. A., Westfall, D. G., & Ahuja, L. R. (2005). Soil organic carbon pools after 12 years in no-till dryland agroecosystems. *Soil Science Society of America Journal*, 69, 1600–1608. <https://doi.org/10.2136/sssaj2003.0266>
- Shi, J., Yang, L., Zhu, A. -X., Rin, C., Liang, P., Zeng, C., & Pei, T. (2018). Machine-learning variables at different scales vs. knowledge-based variables for mapping multiple soil properties. *Soil Science Society of America Journal*, 82, 645–656. <https://doi.org/10.2136/sssaj2017.11.0392>
- Siewert, M. B. (2018). High-resolution digital mapping of soil organic carbon in permafrost terrain using machine learning: A case study in a sub-Arctic peatland environment. *Biogeosciences*, 15(6), 1663–1682. <https://doi.org/10.5194/bg-15-1663-2018>
- Soil Survey Staff & Loecke, T. (2016). *Rapid carbon assessment: Methodology, sampling, and summary*. USDA-NRCS.
- Sulman, B. N., Harden, J., He, Y., Treat, C., Koven, C., Mishra, U., O'Donnell, J. A., & Nave, L. E. (2020). Land use and land cover affect the depth distribution of soil carbon: Insights from a large database of soil profiles. *Frontiers in Environmental Science*, 146. <https://doi.org/10.3389/fenvs.2020.00146>
- Todd-Brown, K., Randerson, J., Post, W., Hoffman, F., Tarnocai, C., Schuur, E., & Allison, S. (2013). Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations. *Biogeosciences*, 10(3), 1717–1736. <https://doi.org/10.5194/bg-10-1717-2013>
- Vašát, R., Kodešová, R., & Borůvka, L. (2017). Ensemble predictive model for more accurate soil organic carbon spectroscopic estimation. *Computers & Geosciences*, 104, 75–83.
- Vitharana, U. W. A., Mishra, U., Jastrow, J. D., Matamala, R., & Fan, Z. (2017). Observational needs for estimating Alaskan soil carbon stocks under current and future climate. *Journal of Geophysical Research: Biogeosciences*, 122(2), 415–429. <https://doi.org/10.1002/2016JG003421>
- Vos, C., Jaconi, A., Jacobs, A., & Don, A. (2018). Hot regions of labile and stable soil organic carbon in Germany: Spatial variability and driving factors. *soilless*, 4(2), 153–167. <https://doi.org/10.5194/soil-4-153-2018>
- Wadoux, A. M. J. C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews*, 210, 103359. <https://doi.org/10.1016/j.earscirev.2020.103359>
- Wadoux, A. M. J. C., Samuel-Rosa, A., Poggio, L., & Mulder, V. L. (2020). A note on knowledge discovery and machine learning in digital soil mapping. *European Journal of Soil Science*, 71, 133–136. <https://doi.org/10.1111/ejss.12909>
- Webster, K., Creed, I., Bourbonniere, R., & Beall, F. (2008). Controls on the heterogeneity of soil respiration in a tolerant hardwood forest. *Journal of Geophysical Research: Biogeosciences*, 113(G3). <https://doi.org/10.1029/2008JG000706>
- Wickland, K. P., Neff, J. C., & Harden, J. W. (2010). The role of soil drainage class in carbon dioxide exchange and decomposition in

- boreal black spruce (*Picea mariana*) forest stands. *Canadian Journal of Forest Research*, 40(11), 2123–2134. <https://doi.org/10.1139/X10-163>
- Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1), 95–114. <https://doi.org/10.1111/1467-9868.00374>
- Wood, S. N. (2017). *Generalized additive models: An introduction with R*. CRC Press.
- Yoder, B. J., & Waring, R. H. (1994). The normalized difference vegetation index of small Douglas-fir canopies with varying chlorophyll concentrations. *Remote Sensing of Environment*, 49(1), 81–91. [https://doi.org/10.1016/0034-4257\(94\)90061-2](https://doi.org/10.1016/0034-4257(94)90061-2)
- Zhang, G.-l., Feng, L., & Song, X.-d. (2017). Recent progress and future prospect of digital soil mapping: A review. *Journal of Integrative Agriculture*, 16(12), 2871–2885. [https://doi.org/10.1016/S2095-3119\(17\)61762-3](https://doi.org/10.1016/S2095-3119(17)61762-3)

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Mishra, U., Yeo, K., Adhikari, K., Riley, W. J., Hoffman, F. M., Hudson, C., & Gautam, S. (2022). Empirical relationships between environmental factors and soil organic carbon produce comparable prediction accuracy to machine learning. *Soil Science Society of America Journal*, 86, 1611–1624. <https://doi.org/10.1002/saj2.20453>