Version of Record: https://www.sciencedirect.com/science/article/pii/S0034425722000281
Manuscript_b7988b10cb10434362696cb778d9ad45

Using soil library hyperspectral reflectance and machine learning to predict soil organic carbon: assessing potential of airborne and spaceborne optical soil sensing

Sheng Wang[1,2*], Kaiyu Guan[1,2,3*], Chenhui Zhang[4], DoKyoung Lee[1,2], Andrew J. Margenot[1,2], Yufeng Ge[5], Jian Peng[3,4], Wang Zhou[1,2], Qu Zhou[1,2], Yizhi Huang[4]

1. Agroecosystem Sustainability Center, Institute for Sustainability, Energy, and Environment, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
2. Department of Natural Resources and Environmental Sciences, College of Agricultural, Consumer and Environmental Sciences, University of Illinois at Urbana Champaign, Urbana, IL 61801, USA
3. National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
4. Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
5. Department of Biological Systems Engineering, University of Nebraska-Lincoln, Lincoln, NE 68583, USA

*Corresponding author: shengwang12@gmail.com; kaiyug@illinois.edu

© 2022 published by Elsevier. This manuscript is made available under the Elsevier user license
https://www.elsevier.com/open-access/userlicense/1.0/

Abstract:

Soil organic carbon (SOC) is a key variable to determine soil functioning, ecosystem services, and global carbon cycles. Spectroscopy, particularly optical hyperspectral reflectance coupled with machine learning, can provide rapid, efficient, and cost-effective quantification of SOC. However, it remains largely unclear how to exploit soil hyperspectral reflectance to best predict SOC concentration, and the potential performance of airborne and satellite data for predicting surface SOC at large scales remains relatively underknown. This study utilized a continental-scale soil laboratory spectral library (37,540 full-pedon 350–2500nm reflectance spectra with SOC concentration of 0-780 $g \cdot kg^{-1}$ across the US) to thoroughly evaluate seven machine learning algorithms including Partial-Least Squares Regression (PLSR), Random Forest (RF), K-Nearest Neighbors (KNN), Ridge, Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) along with four preprocessed spectra, i.e. original, vector normalization, continuum removal, and first-order derivative, to quantify SOC concentration. Furthermore, by using the coupled soil-vegetation-atmosphere radiative transfer model, we simulated twelve airborne and spaceborne hyper/multi-spectral remote sensing data from surface bare soil laboratory spectra to evaluate their potential for estimating surface SOC concentration. Results show that LSTM achieved best predictive performance of quantifying SOC concentration for the whole data sets (R^2 = 0.96, RMSE = 30.81 $g \cdot kg^{-1}$), mineral soils (SOC $\leq$ 120 $g \cdot kg^{-1}$, R^2 = 0.71, RMSE = 10.60 $g \cdot kg^{-1}$), and organic soils (SOC > 120 $g \cdot kg^{-1}$, R^2 = 0.78, RMSE = 62.31 $g \cdot kg^{-1}$). Spectral data preprocessing, particularly the first-order derivative, improved the performance of PLSR, RF, Ridge, KNN, and ANN, but not LSTM or CNN. We found that the SOC models of mineral and organic soils should be distinguished given their distinct spectral signatures. Finally, we identified that the shortwave infrared is vital for airborne and spaceborne hyperspectral sensors to monitor surface SOC. This study highlights the high accuracy of LSTM with hyperspectral/multispectral data to mitigate a certain level of noise (soil moisture < 0.4 $m^3 \cdot m^{-3}$, green leaf area < 0.3 $m^2 \cdot m^{-2}$, plant residue < 0.4 $m^2 \cdot m^{-2}$) for quantifying surface SOC concentration. Forthcoming satellite hyperspectral missions like Surface Biology and Geology (SBG) have a high potential for future global soil carbon monitoring, while high-resolution satellite multispectral fusion data can be an alternative.

Keyword: Spectroscopy, Soil organic carbon, Hyperspectral reflectance, Radiative transfer modeling, Machine learning, Long Short-Term Memory, SBG

1. Introduction

Soil organic carbon (SOC) is a key variable that influences and integrates soil physical, chemical and biological processes (Bot and Benites, 2005). SOC concentration is an important indicator of soil quality and health (Wiesmeier et al., 2019). Soil with higher SOC concentration tends to have larger nutrient pools and greater water holding capacity to increase ecosystem productivity (Lal, 2016; Rice, 2004). Furthermore, soil serves an essential role in global carbon cycles, as soil is the largest carbon pool in the terrestrial ecosystems that can be either a source or sink for atmospheric CO_2 (Scharlemann et al., 2014; Schuur et al., 2015). Increasing SOC stocks is one important strategy to sequester atmospheric CO_2 to offset greenhouse gas emissions to mitigate anthropogenic-induced climate change (Sanderman et al., 2009). The SOC stock is computed by SOC concentration, bulk density, coarse mineral fragments, and profile depth considered. Accurate quantification of large-scale SOC concentration at regional to continental scale is challenged by the spatial and temporal variability of SOC, which in turn is highly influenced by soil parent materials, climate, topography, and biotic factors (Angelopoulou et al., 2019). Therefore, the first key step towards sustainable management of soil to improve its quality, functioning, and carbon sequestration potential is to develop accurate, efficient, robust, and cost-effective quantification of SOC concentration.

Conventional approaches such as Walkley-Black dichromate wet-chemistry methods (Walkley and Black, 1934) and dry combustion element analysis (Nelson and Sommers, 2015) can precisely quantify SOC concentration, but they are costly and time-consuming when dealing with large numbers of soil samples. Spectroscopic methods that utilize spectral information of soil to predict constituents can be a cost-effective alternative to rapidly and accurately quantify SOC concentration (Ben-Dor et al., 2009; Nocita et al., 2015). Numerous studies have demonstrated that soil spectra collected in the optical (350–2500 nm) or mid-

infrared (2.5–25 μm) ranges can be utilized to predict SOC concentration due to the absorption in infrared frequencies by polar covalent bonds, which for SOC include C-H, C=O, C-O, C=C and C-C (Margenot et al., 2016; Sanderman et al., 2021; Wijewardane et al., 2018). In particular, public soil spectra libraries -- such as the USDA Rapid Carbon Assessment (RaCA, Wills et al., 2014), LUCAS (Tóth et al., 2013), and ISRIC World Soil Reference Collection (Batjes, 2009) -- promote the applications of spectroscopy for SOC quantification. These laboratory soil spectral libraries have been combined with machine learning to predict SOC concentration and other soil properties (e.g., Wijewardane et al., 2016b), simulating airborne and spaceborne hyperspectral signals for sensor evaluation or design (e.g., Castaldi et al., 2016; Ward et al., 2019), and supporting soil-related sustainable agriculture development (Tziolas et al., 2020).

In recent years, machine learning-based techniques facilitate the use of hyperspectral reflectance data to generate accurate models for predicting soil properties (Padarian et al., 2020; Reichstein et al., 2019). A considerable number of studies have applied various machine learning, such as Partial Least Squares Regression (PLSR, Geladi and Kowalski, 1986), Random Forest (RF, Pal, 2005), K-Nearest Neighbors (KNN, Keller and Gray, 1985), Ridge (Hoerl and Kennard, 1970), and Artificial Neural Network (ANN, Hecht-Nielsen, 1989) with soil laboratory spectra data to quantify SOC concentration. For example, Wijewardane et al. (2016b) compared the performance of using PLSR and ANN for SOC concentration predictions and found that ANN (R^2 = 0.96, RMSE = 36.1 $g \cdot kg^{-1}$) outperformed PLSR in model accuracy. With the rapid development of machine learning techniques, deep learning such as convoluted neural networks (CNN, Fukushima et al., 1983), and Long Short-Term Memory (LSTM, Hochreiter and Schmidhuber, 1997) networks show greater potential to improve model predictive performance in remote sensing applications as scene classification (e.g., Bae et al., 2016) and time-series change detection (e.g., Xiao et al., 2019). Such deep learning methods need further assessment for SOC concentration predictions.

In addition to machine learning algorithms, spectral preprocessing transformations and data spiking can also lead to significant improvement of machine learning prediction (Dotto et al., 2018; Tziolas et al., 2020). For example, Dotto et al. (2018) evaluated scatter-correction and spectral-derivatives preprocessing on

PLSR and RF for SOC concentration predictions of 595 samples and found significant model improvement using continuum removal. However, the performance of spectral preprocessing with deep learning and large soil spectra libraries requires comprehensive analysis. Furthermore, several studies have reported that data spiking, which utilizes a small subset of local samples to augment the large soil library, can significantly improve predictions of SOC concentration (Guerrero et al., 2014; Lobsey et al., 2017; Tziolas et al., 2020). Improvements in SOC predictions with data spiking were mainly due to the fact that the localized model with subsets of specific soil categories, e.g. texture, horizons, and taxonomic orders, performed better than the generalized model that draws on the full soil spectral library (Lobsey et al., 2017; Wijewardane et al., 2016b). However, generalized spectral models for SOC predictions have the advantage of being scalable to all soil types. To develop generalized models, identifying key factors for model scalability is highly needed.

Optical sensing covering visible and near-infrared range, unlike mid-infrared spectroscopy, has the flexibility and scalability to be deployed on airborne and satellite platforms to quantify SOC at large scale. For example, Landsat and Sentinel multispectral imagery have been utilized to quantify regional surface SOC concentration with acceptable accuracy (e.g., Castaldi et al., 2019; Vaudour et al., 2021; Zhou et al., 2021). Furthermore, hyperspectral data from airborne and satellite remote sensing, e.g., AVIRIS-NG, HySpex, Hyperion, showed higher accuracy for surface SOC concentration estimation than multispectral data (Gomez et al., 2008; Hbirkou et al., 2012; Stevens et al., 2010). In particular, the recent and forthcoming satellite hyperspectral missions with improved signal-to-noise ratios, such as NASA Surface Biology and Geology (SBG Cawse-Nicholson et al., 2021), Italian Space Agency PRecursore IperSpettrale della Missione Applicative (PRISMA, Stefano et al., 2013), the Copernicus Hyperspectral Imaging Mission (CHIME, Nieke and Rast, 2018), Environmental Mapping and Analysis Program (EnMAP,Guanter et al., 2009), and German Aerospace Center Earth Sensing Imaging Spectrometer (DESIS, Krutz et al., 2018), all of which could provide unprecedented opportunities to monitor SOC on a global scale. To further evaluate the capability of these hyperspectral or multispectral optical remote sensing data for SOC concentration quantification, studies to utilize the laboratory spectra data to simulate signals detected by airborne or

spaceborne sensors are needed. For example, Castaldi et al. (2016) utilized LUCAS (713 samples) and PONMAC (163 samples) soil spectral library to simulate the current or forthcoming satellite (Hyperion, Landsat8, Sentinel2, EnMAP, PRISMA, and HyspIRI) signals to evaluate their capability for estimating SOC concentration with the PLSR approach. Ward et al. (2019) also explored using local PLSR approaches to predict SOC from LUCAS surface soil database and evaluated the capability of EnMAP for SOC predictions. These studies primarily focused on influences of atmospheric attenuation, sensor spectral responses and signal-to-noise ratios on surface SOC predictions. However, satellite or airborne based SOC predictions are often complicated with surface noises including variable surface soil moisture (Ge et al., 2014; Wijewardane et al., 2016a), green vegetation cover and plant residue (i.e. non-photosynthetic/senescent leaves) (Castaldi et al., 2019). Simulation studies of comprehensively evaluating the real-environment noises including soil moisture, green vegetation, plant residue, atmospheric attenuation, sensor spectral responses and signal-to-noise ratios on soil spectra are highly needed. In addition, considering the highly variable SOC concentration in terrestrial ecosystems and the rapid development of machine learning techniques, it is necessary to utilize a large soil database with advanced machine learning to evaluate the potential remote sensing data for SOC concentration predictions. Such evaluation with the simulated airborne and spaceborne data can provide insights into suitable remote sensing data and real-environment noises for SOC studies.

The overall objective of this study is to leverage and evaluate machine learning algorithms to make full use of hyperspectral reflectance to predict SOC concentration from measured laboratory hyperspectral reflectance and simulated air / space-borne optical data. The specific objectives of this study are: (1) to evaluate the accuracy of using PLSR, RF, KNN, Ridge, ANN, CNN, and LSTM to predict SOC concentration from soil laboratory spectra; (2) to assess the benefits of spectral preprocessing techniques (vector normalization, continuum removal, and first-order derivative) on SOC concentration predictions and identify key factors to influence spectral model generalization; and (3) to evaluate the potential of airborne and spaceborne optical data for surface SOC concentration predictions with simulated

hyperspectral or multispectral data using the surface bare soil spectra from the soil spectral library and the identified best machine learning model from objective 1 and 2. To mimic the real-environment noises on spectra, we applied the coupled soil-vegetation-atmosphere radiative transfer model to simulate noises including soil moisture, green vegetation, plant residue, atmospheric attenuation, and sensor signal-to-noise ratios on SOC concentration predictions.

2. Data

This study utilized a public soil spectral library from the USDA Rapid Assessment Carbon (RaCA) project, which was initiated by USDA Natural Resources Conservation Service in 2010 (Wijewardane et al., 2016b; Wills et al., 2014). This soil library has a total of 37,540 records covering all typical soil taxonomic classes, horizons, and textures across the US (Fig. 1). The 350–2500nm reflectance of air-dry soil samples was measured using ASD spectroradiometers (Malvern Panalytical, formally Analytical Spectral Devices, Boulder, Colorado, US) with soil particle sizes smaller than 2mm. The SOC concentrations of soil samples were determined through the USDA standard procedures, which used the dry combustion-based total carbon concentration to subtract inorganic carbon concentration from the soil calcium carbonate equivalent method (Wills et al., 2014). The SOC concentrations of all USDA RaCA samples span from 0 to 780 $g \cdot kg^{-1}$ and have a good representation of soil conditions across the US. The 80% of samples are mineral soil with around 20% of samples being organic soil with SOC concentration greater than 120 $g \cdot kg^{-1}$ (Fig. 1a). The detailed information for the sample size of each SOC concentration class, master horizon, texture, taxonomic order, land use, depth and US regions are shown in the supplementary Table S1. Notably, here we used a simple SOC concentration threshold (120 $g \cdot kg^{-1}$) to distinguish organic and mineral soils (Stolt and Bakken, 2014) other than with additional consideration of soil water saturation status and clay percentage (Soil Survey Staff, 2010), which have the threshold varying from 120 to 180 $g \cdot kg^{-1}$. Furthermore, soil reflectance spectra were stratified into the five categories of SOC concentration to illustrate the reflectance and SOC relationships (Fig. 1b). With the increase of SOC concentrations, the soil

reflectance spectra generally tend to have lower reflectance. However, the spectra of soil with SOC > 120 $g \cdot kg^{-1}$ showed higher reflectance in 1000–2500nm than those of soil with smaller SOC concentrations. For developing machine learning models, the soil laboratory reflectance and SOC data were used as feature inputs and labels, respectively. Details of soil pedon sampling locations can be found in Fig. 1c.

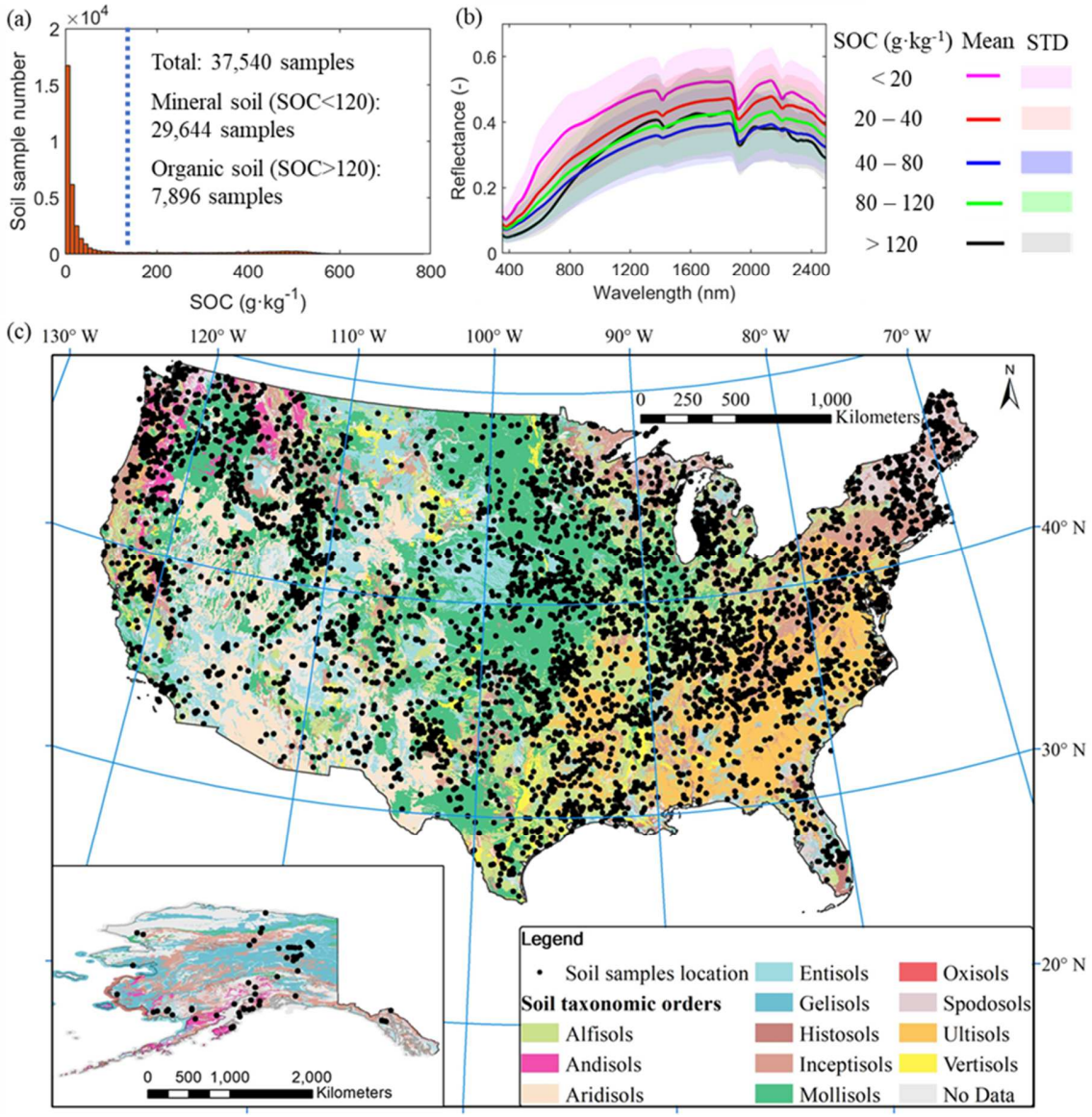


Fig. 1. Laboratory soil hyperspectral reflectance from the USDA Rapid Assessment Carbon soil spectral library. (a) Histogram distribution of measured SOC concentration from the dry combustion. The dashed blue line refers to the SOC concentration of 120 $g \cdot kg^{-1}$ to distinguish mineral and organic soil. (b) Measured

soil reflectance of five SOC concentration categories. The solid lines are mean reflectance, and the shaded areas refer to standard deviations of each SOC concentration category. (c)Soil pedon sampling locations for the USDA Rapid Assessment Carbon soil spectral library (base map: soil taxonomic orders from the US General Soil Map STATSGO2).

3. Methods

To identify the best machine learning algorithms and evaluate spectral preprocessing on machine learning models, we conducted vector normalization, continuum removal, and first-order derivative of soil spectra data (Fig. 2). With the original soil spectra data, we have four types of feature inputs for machine learning. In machine learning algorithms, we have selected PLSR, RF, Ridge, KNN, NN, CNN, and LSTM to identify the best combination of spectral preprocessing and machine learning algorithms for SOC concentration predictions. With the identified machine learning algorithms, we further analyzed the model uncertainties to assess the model's generalizability. Finally, we also tested the potential capability of current and forthcoming airborne and spaceborne multispectral and hyperspectral remote sensing data for surface SOC concentration quantification by using simulated spectra from the soil library. Specifically, we used the Earth bare surface database (Demattê et al., 2020) to identify RaCA surface bare soils potentially detected by remote sensing. By using RaCA surface bare soil spectra with the coupled soil-vegetation-atmosphere radiative transfer models, we thoroughly evaluated the influences of atmospheric attenuation, sensor spectral response, signal-to-noise ratios, surface soil moisture, green vegetation and plant residue on surface SOC concentration predictions.

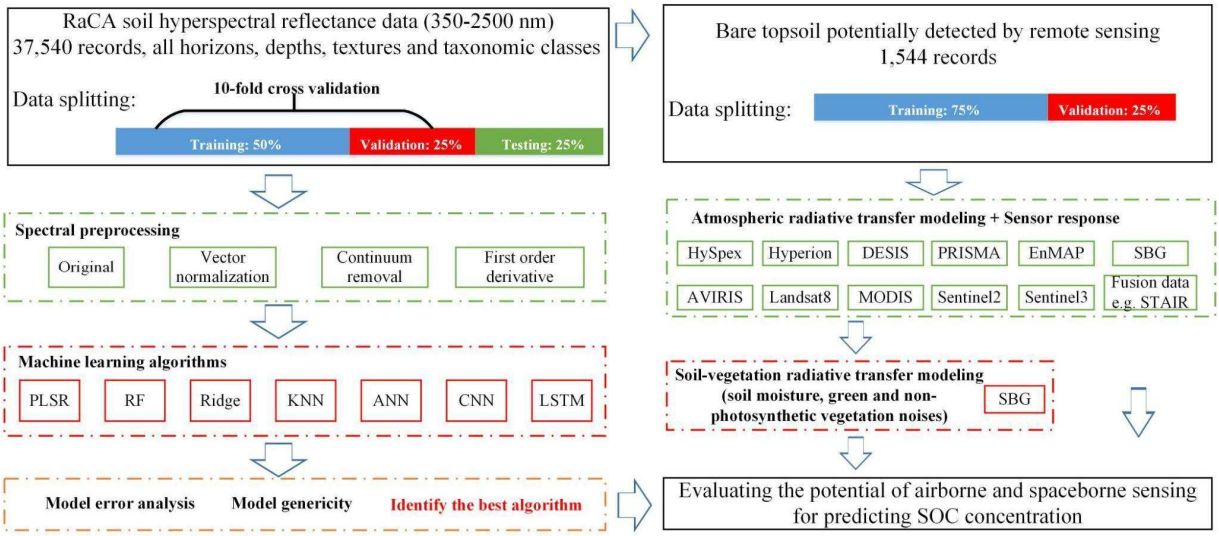


Fig. 2. Flowchart on predicting SOC concentration from laboratory soil spectra and simulated airborne and spaceborne remote sensing data. As a result of spectral preprocessing, the spectral inputs for machine learning include original soil reflectance, and processed spectra with vector normalization, continuum removal, and the first-order derivative. The machine learning algorithms include Partial Least Squares Regression (PLSR), Random Forest (RF), Ridge, K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM). After we identified the best model for SOC predictions, we analyzed model errors and generalizability. In the last step, we used the Earth bare surface database (Demattê et al., 2020) to identify bare soils from RaCA. Then we used the RaCA surface bare soil spectra with the soil-vegetation-atmosphere radiative transfer model to simulate airborne and satellite signals and evaluate the impact of surface soil moisture, green vegetation cover and plant residue cover on surface SOC predictions.

3.1 Spectral preprocessing techniques

To augment spectral signatures of soil samples, we used spectral preprocessing techniques, namely vector normalization, continuum removal, and first-order derivative (e.g. Ge et al., 2019). The vector normalization scaled each input spectral reflectance to a vector with a range length of 1. Continuum removal was to remove the convex shape of a spectrum to enhance reflectance spectral features. The first-order

213 derivative was to calculate the reflectance change with an incremental wavelength of 1nm. Besides these
214 three types of preprocessed spectra, this study also used the measured original reflectance spectra as
215 machine learning inputs. Four spectral feature inputs were combined with machine learning algorithms for
216 comparing SOC concentration predictions. This comprehensive comparison can evaluate the added values
217 of spectral preprocessing methods on each machine learning algorithm.

218 **3.2 Machine learning algorithms**

219 Seven commonly used supervised machine learning regression algorithms including PLSR, RF, Ridge
220 Regression, KNN, ANN, CNN, and LSTM were compared to assess their utility for SOC concentration
221 predictions. Among these methods, PLSR is a multivariate statistical regression method. RF is an ensemble
222 algorithm based on decision trees. Ridge regression is a variant of regularized linear regression. KNN and
223 ANN are simple nonlinear machine learning regression algorithms. Meanwhile, CNN and LSTM are deep
224 learning algorithms and can handle highly nonlinear regression tasks. We used Scikit-learn (Pedregosa et
225 al., 2011) and PyTorch (Paszke et al., 2019) packages to implement machine learning algorithms. We
226 briefly summarized each machine learning algorithm as below.

227 As a statistical approach widely used in chemometrics, Partial Least Square Regression (PLSR) is one of
228 the first methods used to predict soil properties including SOC from spectra (Reeves et al., 2002). This
229 algorithm first reduces data dimensionality of inputs through transforming highly collinear hyperspectral
230 reflectance to several orthogonal latent variables, which represent the major covariances of label data
231 variability (Geladi and Kowalski, 1986). In the second step, PLSR generates statistical linear regression
232 relationships between latent variables and labels. We optimized the number of PLSR latent variables by
233 minimization of the predicted residual error sum of squares (Wang et al., 2021).

234 Random Forest (RF) is a nonlinear ensemble decision tree-based algorithm, which deals with high-
235 dimensional input datasets through constructing and averaging a few randomized decision trees for
236 regression or classification. RF conducts regression between feature inputs and labels through decision

237 rules, which recursively grouped the input feature space into successively smaller tree branches (Belgiu and
238 Drăgu, 2016; Pal, 2005). RF also incorporates bootstrapping to overcome the overfitting issue of the
239 traditional tree regression. This study determined model parameters, e.g., the number of trees and the
240 maximum tree depth, through parameter screening to minimize out-of-bag prediction errors.

241 Ridge regression is a regularized least-squares linear regression technique but can handle high
242 multicollinearity datasets (Hoerl and Kennard, 1970). Ridge can usually generate parsimonious models for
243 nonparametric regression. Similar to PLSR, Ridge is well suitable for situations where observations are
244 fewer than predictor variables.

245 K-Nearest Neighbors (KNN) is one of the simplest supervised classification and regression algorithms
246 (Keller and Gray, 1985). Depending on their similarities, KNN categories the labels into groups, and then
247 adds the new unlabeled data to the most similar groups. The predicted outputs are based on the average of
248 k nearest groups. KNN performs well in solving problems depending on identifying similarity, but has a
249 drawback of slow processing large datasets with high dimensionality. In this study, KNN was employed to
250 identify groups with similar spectral relationships to enable SOC predictions from soil spectra.

251 Artificial Neural Networks (ANN) used in this study consists of one input layer, one output layer, and at
252 least one hidden layer. Each layer has a given number of nodes and each of them connects to the nodes in
253 the following layer. Nonlinearity is introduced through the activation function after each layer, and we used
254 ReLU (Glorot et al., 2011) as the activation function. In the forward pass, hyperspectral reflectance data
255 was fed into the network as batches, passed through each layer as matrix multiplications, and then applied
256 nonlinear activation functions. After the forward pass, the predicted SOC value was compared with the
257 ground truth through the loss function. In this study, Mean Squared Error (MSE) loss was used for all neural
258 network architectures. Then, in the backward pass, the gradients of the loss function with respect to the
259 network parameters were calculated through backpropagation (Kelley, 1960). The model parameters were
260 then updated with an optimizer at a certain learning rate. The ANN model's hyperparameters including the

number of nodes of each layer, the optimal number of hidden layers, the batch size of input spectra, the choice of the optimizer, and the learning rate were optimized through grid search to achieve the lowest RMSE. The ANN schematic network of this study had three hidden layers (Fig. 3a).

Convolutional Neural Networks (CNN) utilize convolution layers to extract the high-level features of input spectra before feeding them into the fully connected layers. In contrast to traditional hand-crafted convolution kernels, CNN's convolution kernels are learned through backpropagation. Through convolution layers, CNN can account for local connectivity in the input data, e.g., linkages across hyperspectral wavelengths, as opposed to ANN where all neurons in two consecutive layers are connected. After each convolution layer, max pooling is applied to the extracted feature to reduce the feature size and the size of the following convolution layers. As the soil reflectance has only one spectral dimension, this study applied a 1-dimensional CNN model (Riese and Keller, 2019) with hyperspectral data as inputs to predict SOC concentration as the model structure illustrated in Fig. 3b.

Long Short-Term Memory (LSTM) is one type of Recurrent Neural Networks (RNN) well-suited to model sequential data such as remote sensing data (Su et al., 2021; Pullanagari et al., 2021). Traditionally, recurrent neural networks could not model the long-term dependencies sequential data very well due to vanishing or exploding gradients in the backpropagation through time. However, LSTM utilizes memory cells to replace the recurrent hidden node to account for both short-term and long-term connections among neurons. As soil hyperspectral reflectance has sequential wavelengths, LSTM as Fig. 3c was used in this study to predict SOC concentration.

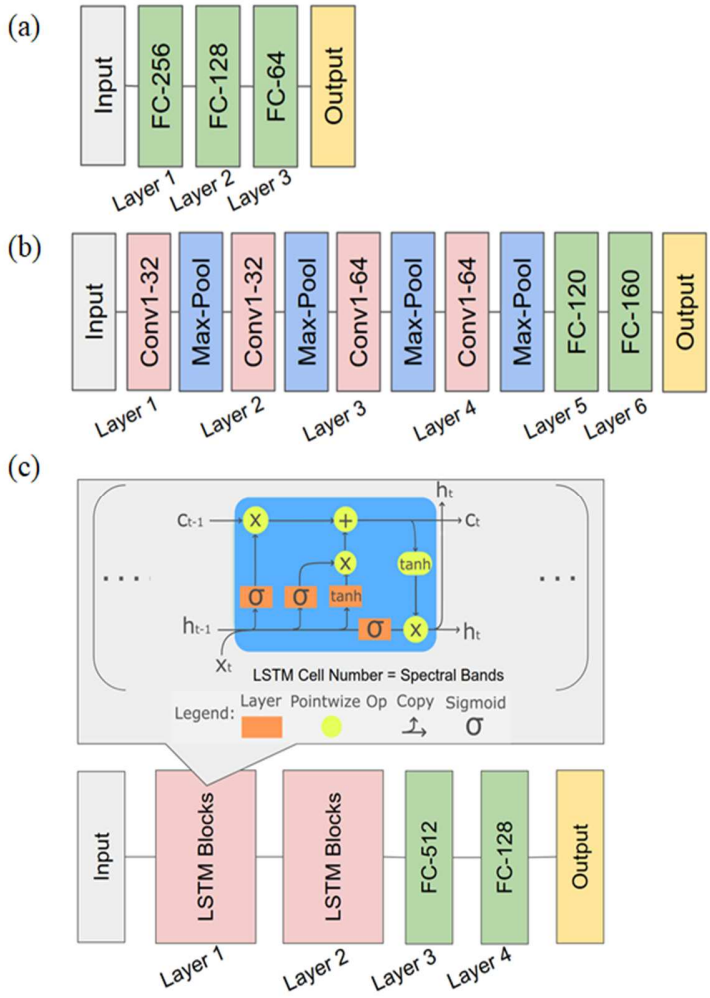


Fig. 3. Schematic representation of (a) Artificial Neural Networks (ANN), (b) Convolutional Neural Networks (CNN), and (c) Long Short-Term Memory (LSTM) used in this study.

For all the neural network architectures above, AdamW (Loshchilov and Hutter, 2017) was used as the optimizer to update the network parameters. Compared to the vanilla Stochastic Gradient Descent (SGD) method, Adam (Kingma and Ba, 2014) utilizes adaptive learning rate and momentum to better accommodate the highly non-convex loss landscape. Furthermore, AdamW decouples L2 regularization and weight decay due to their inequivalence in the context of adaptive gradient methods as opposed to the standard stochastic gradient descent, enabling better model generalization. In addition, a learning rate schedule was imposed to gradually reduce the learning rate when the change in evaluation metrics between

epochs starts to decrease. Before the training process, the weights in the network were initialized through Kaiming Initialization (He et al., 2015), a common weight initialization technique in ReLU networks to facilitate training and convergence.

3.3 Model uncertainty, generalizability and spectral importance assessment

After identifying the best machine learning algorithms for SOC concentration, we analyzed the relationships between simulation errors (relative absolute errors) and soil horizons, texture, taxonomic order, and SOC concentration levels to understand model generalizability. The imbalanced sample size for each soil class can influence model error distribution and induce machine learning training/testing biases (Leevy et al. 2018). Given limited sample sizes for a few soil classes in RaCA (see supplementary Table S1), we excluded the model uncertainty analysis for soil classes with sample sizes smaller than 250.

We further conducted spectral similarity analysis to understand spectral differences for soil classes. Specifically, we conducted principal component analysis (PCA, Wold et al., 1987) and t-distributed stochastic neighbor embedding (tSNE, Van Der Maaten and Hinton, 2008) to explore spectral differences in the high-dimensional hyperspectral data. PCA reduces the data dimensionality with maintaining the maximum variance information, whereas tSNE uses stochastic neighbor embedding to virtualize the high-dimensional datasets by minimizing the KL-Divergence of high-dimensional and low-dimensional data distribution (Van Der Maaten and Hinton, 2008). Furthermore, we applied a Gaussian Mixture Model-Expectation Maximization algorithm (Pernkopf and Bouchaffra, 2005) to unmix the RaCA soil spectral database into two clusters to evaluate SOC distributions by a histogram, and thus identify potential differences in the SOC concentration of these two clusters. Finally, we tested the machine learning performance of using the whole datasets or the stratified datasets (Cluster 1 and 2) for training and testing. Such analysis can provide insights into the pros and cons of applying generalized models and the factor influencing the spectral model generalization.

To assess the importance of each spectral wavelength for SOC concentration predictions, we analyzed the feature importance for the LSTM deep learning model as well as the commonly used PLSR and random forest models with raw spectra or first-order derivatives as inputs. We calculated the Variable Importance in Projection (VIP) scores as Wang et al. (2021) for PLSR to quantify feature importance. The feature importance of random forest was calculated from the mean and standard deviation values of the impurity decrease accumulation within each tree (Ishwaran, 2015). For deep learning model feature importance, we used permutation feature importance, which was determined by the decrease of a model evaluation metric (RMSE for this study) with a single feature input (each spectral wavelength) randomly shuffled (Altmann et al., 2010). Understanding spectral importance for SOC predictions can give insights into identifying key wavelengths and evaluating the suitability of remote sensing data for monitoring SOC concentration.

3.4 Potential of airborne and spaceborne data for surface SOC predictions

To assess the potential of major airborne and spaceborne hyper/multispectral missions for quantifying surface SOC concentration, we used RaCA soil spectra to simulate the airborne and spaceborne remote sensing signals. These simulated remote sensing missions included airborne hyperspectral data (HySpex, AVIRIS-NG), the past or current hyperspectral or multispectral satellite data (Hyperion, PRISMA, Landsat8, MODIS, Sentinel2, Sentinel3), satellite multispectral fusion data e.g. SaTellite dAta IntegRation (STAIR-2, Luo et al., 2020), International Space Station data (DESIS), and forthcoming space-borne hyperspectral data e.g. NASA SBG, CHIME, EnMAP. As SBG and CHIME sensor spectral configurations are close (Cawse-Nicholson et al., 2021), we only evaluated SBG utility here. Potentially, CHIME has two satellites and can provide higher temporal resolution data than other hyperspectral missions. Besides hyperspectral missions, Landsat and Sentinel2 are often used for monitoring SOC (Castaldi et al., 2016, 2019), this study also incorporated MODIS as it is often used for multi-source satellite fusion data (e.g. STAIR, Luo et al., 2018). Sentinel 3 and DESIS have rich spectral wavelengths in the visible-near infrared

range (400–1000 nm) not in the shortwave infrared range. Thus, incorporating these two missions can provide insights into the importance of short-wave infrared wavelengths for SOC prediction.

As optical remote sensing data can only detect the soil signals from a bare surface, we selected RaCA surface soils based on sampling depth information (supplementary Table S1). Additionally, we also used RaCA soil sampling locations with the Earth bare surface database (Demattê et al., 2020) to identify the bare soils for the satellite and airborne remote sensing signal simulations. We further excluded organic soils (SOC > 120 $g \cdot kg^{-1}$) as typical surface bare soils are mineral soils, particularly for croplands (Thaler et al., 2021).

In the remote sensing signal simulation, we followed previous studies (Verhoef et al., 2018; Castaldi et al. 2016) and used the MODerate resolution atmospheric TRANsmission model (MODTRAN, Berk et al., 2011) to simulate the atmospheric radiative transfer process to account for the atmospheric attenuation effects on airborne and satellite signals. Specifically, we conducted the atmospheric radiative transfer simulations through MODTRAN interrogation technique with a four-stream scheme to be coupled with bare soil reflectance (Verhoef and Bach, 2012; Verhoef et al., 2018). In the simulations, we also considered the sensor spectral response (supplementary Fig. S1) and signal-to-noise ratios (supplementary Table S2) to compare the performance for airborne and satellite missions. The simulated at-sensor radiance was resampled and convoluted to the airborne and spaceborne sensor data using the spectral response curves of each sensor. For sensor noises, we used the noise-equivalent delta radiance approach, which is simple but effective for optical sensors and insensitive to wavelengths (Verhoef et al., 2018). The sensor signal-to-noise ratios and calculated parameters for sensor models are shown in supplementary Table S2. We also excluded noisy signals in the atmospheric absorption windows and low signal-to-noise ratio wavelengths i.e. <400 nm, 1300–1400 nm, 1800–2000 nm and >2400 nm. Furthermore, we coupled MODTRAN with the soil-vegetation radiative transfer model Soil Canopy Observation, Photochemistry and Energy fluxes (SCOPE, van der Tol et al., 2009) to simulate surface noises on SOC concentration predictions. SCOPE is a soil-vegetation canopy radiative transfer model, which is based on leaf-scale PROSPECT-D model (Jacquemoud et al., 1990; Féret et al., 2017), canopy radiative transfer model 4SAIL (Verhoef, 1984), and

a general soil spectral vector Brightness-Shape-Moisture model BSM (Verhoef et al., 2018; Jiang and Fang, 2019). The PROSPECT-D model was used to simulate the green and senescent leaf (Table 1 shows the detailed parameters and their values). With the 4SAIL model to simulate different levels of leaf area index, we can get various fraction covers for green and senescent vegetation. We simulated 11 leaf area index scenarios from 0 to 1 $m^2 \cdot m^{-2}$ with an interval of 0.1 $m^2 \cdot m^{-2}$ and assumed leaf angle distribution close to horizontal. The soil spectral model BSM can simulate dry soil spectra, absorption of water film, water-soil Fresnel reflection, irregular water film thickness (Yang et al., 2020). In our simulations, we used the RaCA soil spectra as dry soil spectra with simulating 11 levels of volumetric soil moisture from 0.05 $m^3 \cdot m^{-3}$ (dry soils, the measured RaCA soil spectra) to 0.45 $m^3 \cdot m^{-3}$ (wet soils) with an interval of 0.05 $m^3 \cdot m^{-3}$. In the real environment, soil moisture can simultaneously vary with different green leaf areas or plant residue conditions. To explore the joint effects of vegetation cover and soil moisture, we used SCOPE to simulate SBG signals with 121 combinations of surface soil moisture and green vegetation, and 121 combinations of surface soil moisture and plant residue (Table 1).As such, we can evaluate the influence of surface conditions on the potential of remote sensing data for SOC concentration quantification.

Table 1. Key parameters and their values for MODTRAN and SCOPE in soil-vegetation-atmosphere radiative transfer modeling.

Model	Parameter	Value
MODTRAN	Visibility (km)	20
	Water vapor ($kg \cdot m^{-2}$)	10
	Sensor height (km)	Airborne 1; Spaceborne 700
PROSPECT-D	Chlorophyll content ($\mu g \cdot cm^{-2}$)	Green: 60; non-photosynthetic: 0
	Carotenoid content ($\mu g \cdot cm^{-2}$)	Green: 15; non-photosynthetic: 0
	Dry matter content ($g \cdot cm^{-2}$)	0.012
	Leaf water content ($g \cdot cm^{-2}$)	Green: 0.02; non-photosynthetic: 0.001

	Senescent material fraction	Green: 0; non-photosynthetic: 1
	leaf thickness parameter	1.4
4SAIL	Leaf area index ($m^2 \cdot m^{-2}$)	0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1
	Leaf angle distribution LIDFa	1
	Leaf angle distribution LIDFb	0
BSM	Volumetric soil moisture content ($m^3 \cdot m^{-3}$)	0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55

3.5 Model evaluation

In machine learning models, we split the whole dataset into three parts: training, validation and testing. The training and validation datasets consisted of 75% of the whole dataset, and we conducted 10-fold cross-validation for training and validation. Specifically, we split the training and validation datasets into 10 folds. For each time, we used 9-fold data for model training and 1-fold data for cross-validation. By iterating 10 times, we can obtain 10 models to get both model ensemble means and standard deviations. By also applying these 10 models to the testing dataset, we can obtain the mean and standard deviations for model testing. In model evaluation, we used statistics including coefficient of determination (R^2), mean absolute percentage errors (MAPE), root mean square errors (RMSE), relative RMSE (rRMSE), and the Ratio of Performance to InterQuartile distance (RPIQ). MAPE was calculated by using the ratio between mean absolute errors and mean values. rRMSE was the ratio between RMSE and data range. RPIQ was based on the ratio between the interquartile range (the third quartile minus the first quartile) and RMSE. The better model performance corresponds to the higher values of R^2 and RPIQ and the lower values of MAPE, RMSE, and rRMSE, or vice versa.

4. Results

4.1 Machine learning and preprocessing algorithms for SOC predictions

Through the comprehensive comparison of seven machine learning algorithms, we found that LSTM, CNN, and ANN outperformed other algorithms with high performance in predicting SOC concentration (Fig. 4). Particularly, LSTM had the best predictive performance to make the best use of hyperspectral reflectance to predict SOC concentration with R^2 around 0.96, MAPE around 19.17%, RPIQ around 1.24, RMSE around 30.81 $g \cdot kg^{-1}$, and rRMSE around 4.19% in model testing (see detailed statistics in Fig. 4 and supplementary Fig. S2). Both LSTM and CNN can take the sequential spectral connections into account and show high performance compared to ANN. For other machine learning, RF and KNN achieved good performance followed by Ridge and PLSR. The results agree with previous studies that ANN can obtain better predictive performance than multivariate regression methods such as PLSR (Wijewardane et al., 2016b).

For the feature preprocessing of hyperspectral reflectance, we did not see benefits for performance improvements of deep learning models such as LSTM and CNN. These two highly nonlinear deep learning models can fit spectra and SOC concentration without the aid of spectra preprocessing. Meanwhile, other machine learning models showed significant improvements when applying spectra preprocessing techniques. In particular, the highest model improvements occur using first-order derivatives (Fig. 4). Overall, the best model performance is achieved by LSTM and CNN. Particularly, LSTM can also be applied with multispectral data with limited wavelength inputs, while CNN operated convolution with high-dimensional feature data as hyperspectral reflectance.

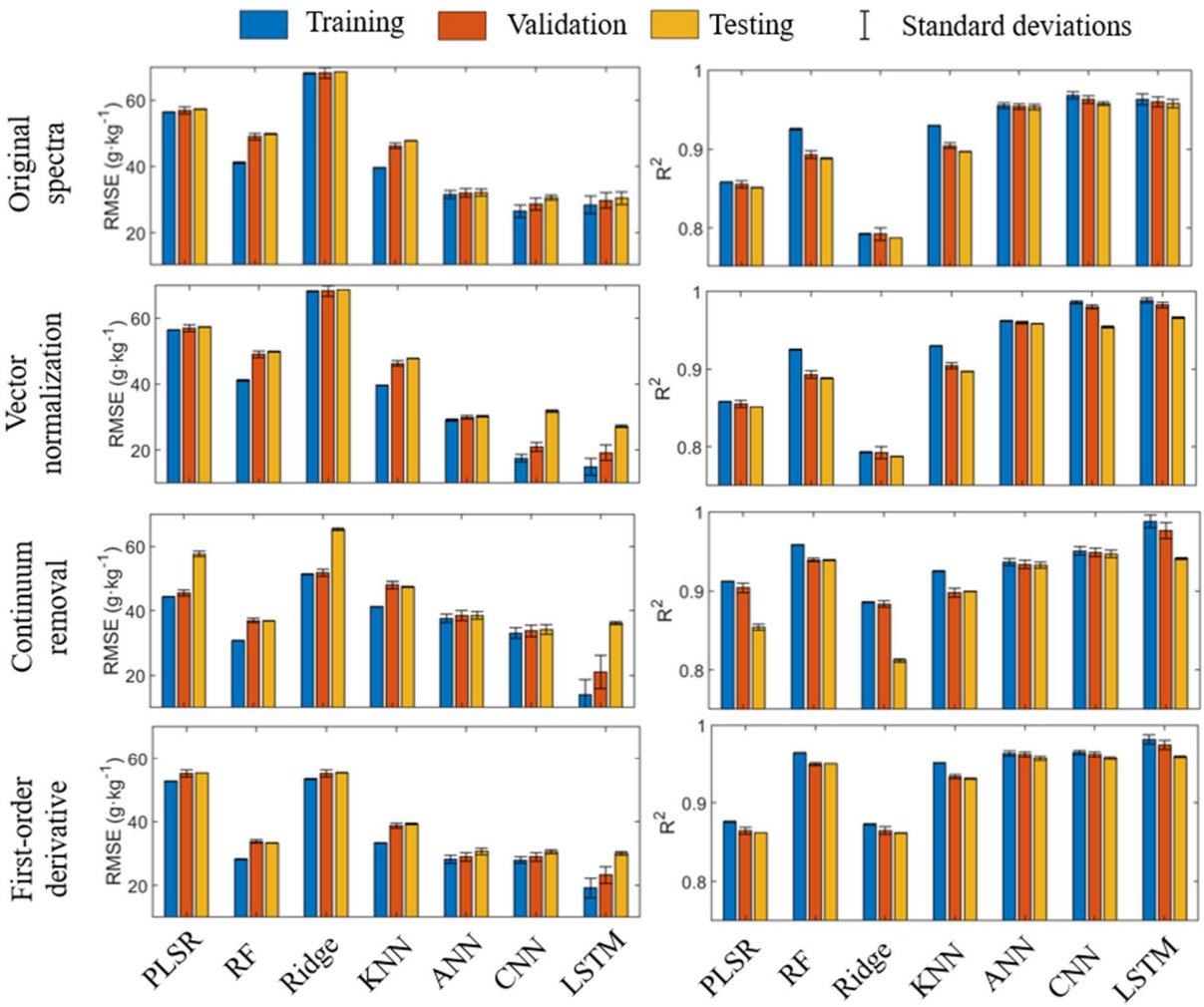
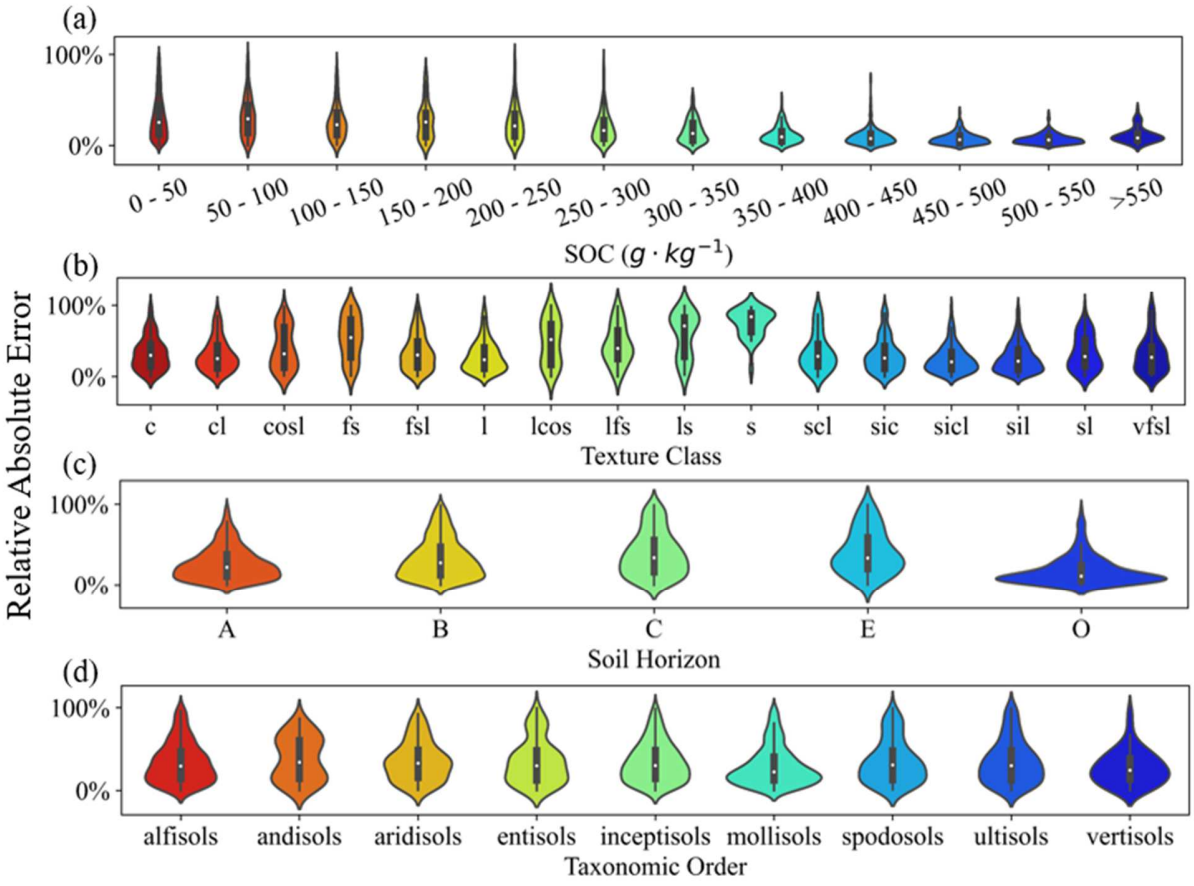


Fig. 4. Performance (RMSE and R^2) of predicting SOC concentration from soil spectra with machine learning algorithms (PLSR, RF, Ridge, KNN, ANN, CNN, and LSTM) and spectral preprocessing techniques (original reflectance spectra, vector normalization, continuum removal, and first-order derivative). The error bars represent the standard deviations of model simulation performance. RMSE is the root mean squared error. R^2 is the coefficient of determination.

4.2 Model uncertainty and generalizability

By stratifying the testing datasets of the best model LSTM simulations into 12 SOC concentrations (Fig. 5a), we found the strong dependency of the model error (relative absolute error) on SOC concentrations.

LSTM generally showed high errors in low SOC concentration parts (<250 g·kg^{-1}), while the model had relatively smaller errors in high SOC concentrations (>350 g·kg^{-1}). In the error analysis with soil horizons (Fig. 5b), there were high errors in C horizons (averaged SOC concentration in RaCA = 4.26 g·kg^{-1}) and low errors in O horizons (averaged SOC concentration = 353.23 g·kg^{-1}). Similarly, we also found that among soil textures (Fig. 5c), coarse sand (symbology cos, averaged SOC concentration = 4.59 g·kg^{-1}) and sand (symbology s, averaged SOC concentration = 6.10 g·kg^{-1}) soils have low SOC concentrations and show large errors. In Fig. 5d, aridisols and entisols, which generally have low SOC concentrations (averaged SOC concentration = 5.00 and 13.31 g·kg^{-1}, respectively), also showed relatively larger errors compared to other soil taxonomic orders. Such findings imply that the generalized spectra soil model for soils with low SOC (mineral soils) and high SOC (organic soils) have significantly different performances, although all soil spectra data (80% of samples were mineral soils) were used for model training.



435 **Fig. 5**. LSTM model uncertainty analysis with different soil classes. (a) the relative absolute error (y-axis)
436 with SOC concentrations (x-axis). (b) relative absolute error with soil horizons (x-axis, A: topsoil, B:
437 subsoil, C: parent material, E: eluviated, O: organic). (c) relative absolute error with soil textures (x-axis,
438 c: clay, cl: clay loam, cos: coarse sand, cosl: coarse sandy loam, fs: fine sand, fsl: fine sandy loam, l: loam,
439 lcos: loamy coarse sand, lfs: loamy fine sand, ls: loamy sand, s: sand, sc: sandy clay, scl: sandy clay loam,
440 si: silt, sic: silty clay, sicl: silty clay loam, sil: silt loam, sl: sandy loam, vfsl: very fine sandy loam). (d)
441 relative absolute error with soil taxonomic orders (x-axis). In the violin boxplots, the white dot on the black
442 bar refers to the median. The black bars inside the violin boxplots correspond to the 25th and 75th percentile.
443 The lower/upper black lines are the median minus/plus 1.5 times of the interquartile range (the 75th
444 percentile minus the 25th percentile). The violin shapes reflect the data distribution.
445
446 To understand the spectral characteristics of mineral and organic soils, we conducted tSNE (Fig. 6a) and
447 PCA (supplementary Fig. S3) to reduce the dimensionality of hyperspectral reflectance to two dimensions.
448 Both tSNE and PCA methods show that mineral and organic soils have distinct spectral features.
449 Particularly, tSNE is more obvious to show spectral differences of such high-dimensional data. Such
450 differences also agree with Fig. 2b, which shows the mean reflectance of organic soils is different from
451 spectra of other SOC categories. Furthermore, the histogram results (Fig. 6b) of two clusters identified by
452 the Gaussian Mixture Model-Expectation Maximization algorithm approximately correspond to mineral
453 soils ($SOC \leq 120\ g \cdot kg^{-1}$) and organic soils ($SOC > 120\ g \cdot kg^{-1}$). These findings on soil spectra characteristics
454 imply the significant reflectance spectra differences between mineral and organic soils, which influences
455 model generalizability.

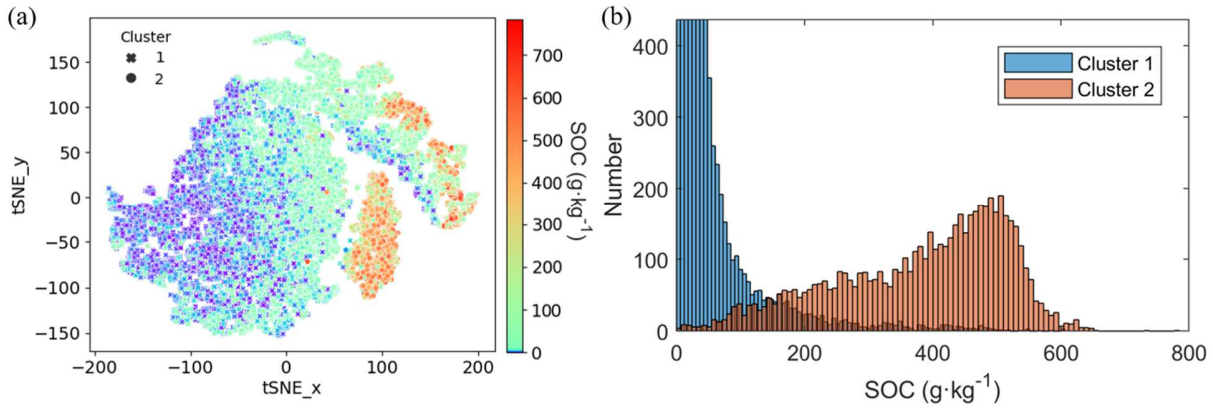


Fig. 6. Visualization of high-dimensional RaCA soil hyperspectral reflectance spectra into two clusters. (a) shows the visualized data with t-distributed stochastic neighbor embedding (tSNE) similarity by minimizing the divergence of the displayed low-dimensional and the original high-dimensional data distributions through stochastic neighbor embedding. Spectra of soils with high vs low SOC concentrations diverge. Results of PCA refer to supplementary Fig. S3. Based on the Gaussian Mixture Model-Expectation Maximization algorithm, we detected two clusters in the soil spectra data. (b) shows the SOC histogram distribution of soil spectra Cluster 1 and 2. Cluster 1 is mostly mineral soils (SOC ≤ 120 g·kg^{-1}), while the majority of Cluster 2 is organic soils (SOC > 120 g·kg^{-1}). Due to the high range of cluster 1, here we only scale the y-axis to the maximum value of 450. The entire distribution of cluster 1 could be found in supplementary Fig. S4.

The testing performance of the generalized LSTM model shows high predictive accuracy of estimating SOC concentration with R^2 around 0.96, MAPE around 19.17%, and RPIQ around 1.24 (Fig. 7a). However, when applying this LSTM model to only organic soils (Fig. 7b, SOC > 120 g·kg^{-1}) or mineral soils (Fig. 7c, SOC < 120 g·kg^{-1}) show contrasting differences. The model simulations for organic soils also had good performance with R^2 around 0.77, MAPE around 13.24%, RPIQ around 3.40 (Fig. 7b). Meanwhile, for mineral soils, the generalized model only achieved relatively low performance with R^2 around 0.57, MAPE around 51.84%, RPIQ around 0.97 (Fig. 7c). Such differences in model performance agree with the results of Fig.6 and demonstrate that the generalized model developed from the whole SOC datasets has large

errors for mineral soils. Furthermore, we found that using the only organic soil spectra (SOC > 120 g·kg^{-1}) to retrain a new LSTM model for organic soils (Fig. 7d) did not improve significantly compared to the generalized model (Fig. 7b). However, using the only mineral soil spectral datasets (SOC ≤ 120 g·kg^{-1}) to retrain the model can achieve much better results with R^2 around 0.71, MAPE around 36.56%, and RPIQ around 1.44 (Fig. 7e). Notably, such findings further confirm that the SOC models of mineral and organic soils should be distinguished. The differences between mineral and organic soils may result from that lower SOC concentration constraining SOC-related signals and high OC soils may have different quality (chemical compositions) of SOC. Cautions should be taken in applying SOC prediction models developed from datasets with both mineral and organic soils to mineral soils, which are the dominant soil types in most terrestrial ecosystems (Bot and Benites, 2005).

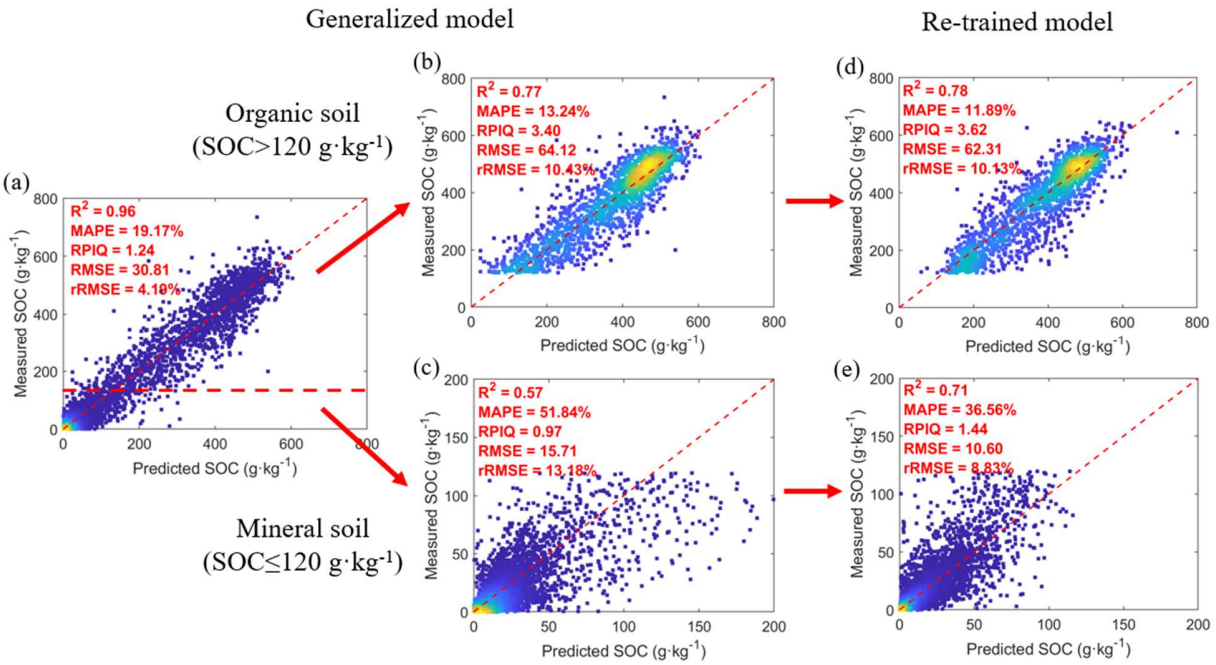


Fig. 7. Scatterplots of LSTM performance in the testing datasets. (a) LSTM trained with all RaCA soil spectra data. Model performance for all testing data. (b) The same model is as (a), but the model performance for testing data is for samples with SOC > 120 g·kg^{-1}. (c) The same model is as (a), but the model performance for testing data is for samples with SOC ≤ 120 g·kg^{-1}. (d) LSTM testing performance

with the model retrained with only organic soil data (SOC > 120 $g \cdot kg^{-1}$). (e) LSTM testing performance with the model retrained with only mineral soil data (SOC ≤ 120 $g \cdot kg^{-1}$). To illustrate the dataset splitting from (a) to (b) and (c), this figure added observations in the y-axis and predictions in the x-axis.

4.3 Simulated airborne and spaceborne data for surface SOC predictions

We analyzed the feature importance for the LSTM model developed from surface bare soils (Fig. 8b). The permutation feature importance identifies important wavelengths for SOC concentration predictions: 600–700nm, 1250–1350nm, and 2000–2400nm. These wavelengths are consistent with absorption features of organic components, such as lignin, cellulose, and starch, and agree with previous studies on the feature importance of using hyperspectral data for SOC concentration predictions (Bartholomeus et al., 2011, 2008). We also further compared the LSTM feature importance with PLSR and RF with original spectra and the first-order derivative (supplementary Fig. S5). The results show that LSTM and PLSR with the first-order derivative are very close and demonstrated that LSTM can make the flexible nonlinear transformation to exploit the spectral feature. Meanwhile, the feature importance of PLSR and RF with original spectra show the difference with that of LSTM and shows that PLSR and RF cannot make the full use of original spectra to predict SOC, which also led to the weaker model performance (Fig. 4). Compared to airborne and satellite sensor spectral wavelengths, Fig. 8 shows that missions with collecting spectral information on shortwave infrared, particularly on 2000–2400nm, have high potential for remote sensing-based surface SOC monitoring.

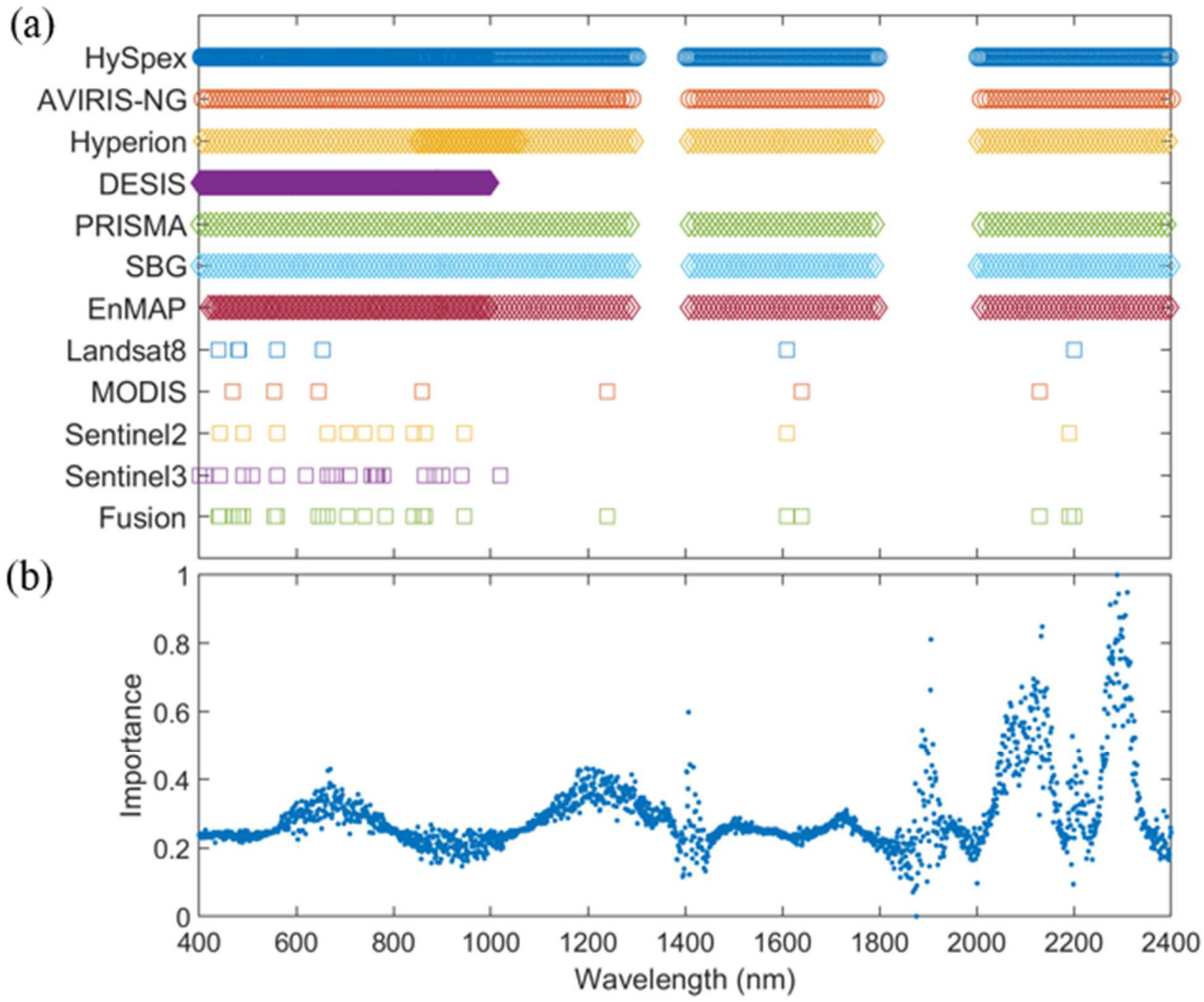
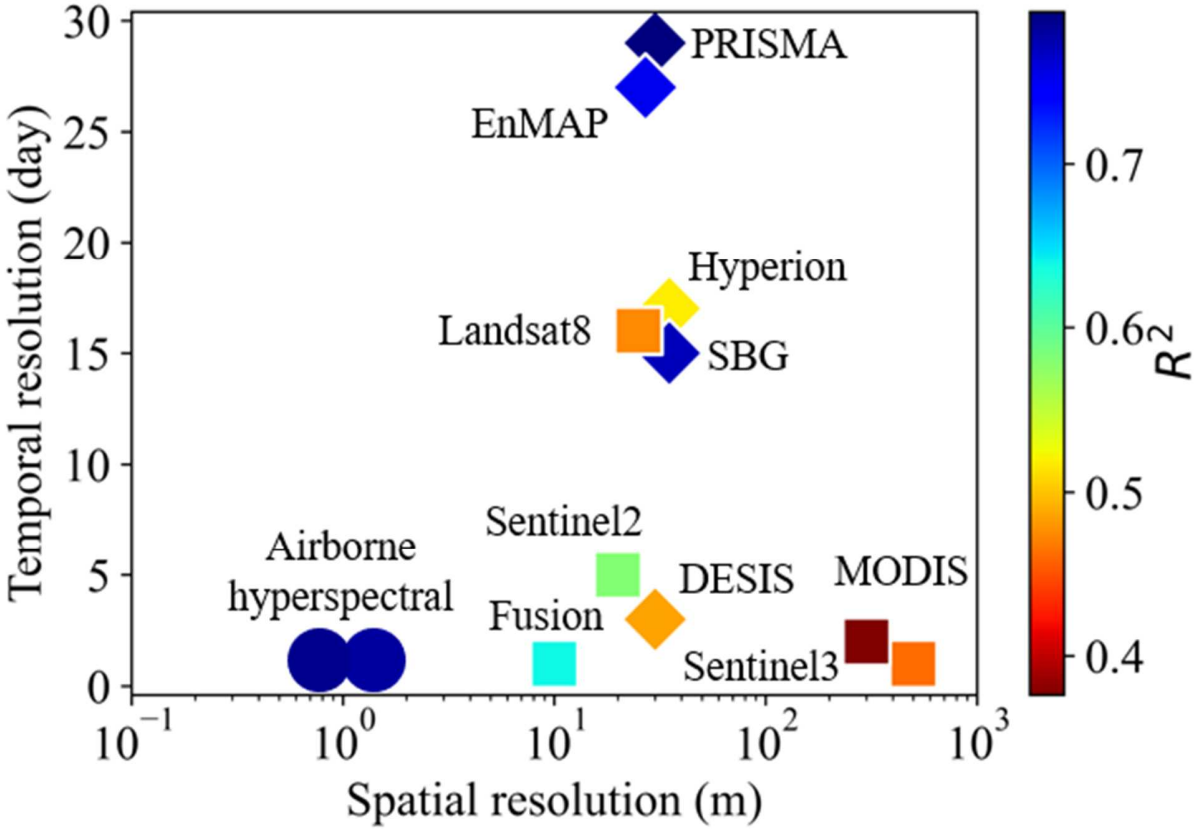


Fig. 8. Sensor wavelengths and feature importance. (a) Sensor spectral central wavelengths. (b) Importance of spectral wavelengths of the LSTM model in contributing to SOC concentration predictions. Note that (a) only shows the sensor central wavelengths and the used spectral response curves of airborne and spaceborne sensors are shown in the supplementary Fig S1.

We also used surface bare soil reflectance to simulate airborne and spaceborne data to further develop LSTM models to assess the potential for SOC concentration predictions. Fig. 9 shows the coefficient of determination R^2 of airborne and spaceborne remote sensing data for SOC concentration predictions. Other detailed statistics such as MAPE, RPIQ, RMSE, and rRMSE can be found in supplementary Table S3. Among the simulated airborne and spaceborne remote sensing data, airborne hyperspectral sensors, such as

HySpex and AVIRIS-NG, and spaceborne hyperspectral sensors including PRISMA, SBG, and EnMAP achieved the highest predictive performance ($R^2 \approx 0.75–0.79$, MAPE ≈ 43–58%, RPIQ ≈ 1.52–1.73, RMSE ≈ 14–16 $g \cdot kg^{-1}$) due to their high spectral resolutions. As the first hyperspectral satellite to monitor the globe, Hyperion hyperspectral data suffered from low signal-to-noise ratios (Gomez et al., 2008) and has relatively low performance. Nonetheless, the forthcoming next-generation satellite hyperspectral missions as SBG have much-improved signal-to-noise ratios. Results also show that PRISMA , SBG and EnMAP have similar and greater capability for global surface SOC monitoring. However, these satellite hyperspectral missions have a relatively low temporal frequency of 16-28 days. The fusion of multiple hyperspectral mission data including SBG, CHIME, PRISMA, and EnMAP can provide high temporal frequency for global soil monitoring (Cawse-Nicholson et al., 2021). As DESIS lacks shortwave infrared wavelengths, it did not achieve high performance to predict SOC concentration. Furthermore, the analysis of mainstream multispectral satellite data reveals that recent missions such as Sentinel2 can also achieve acceptable performance for SOC concentration predictions. We also found that using the satellite fusion data techniques such as STAIR 2.0 (Luo et al., 2020) to integrate spectral information from Landsat, Sentinel2, and MODIS can obtain high predictive results ($R^2 \approx 0.64$, MAPE ≈ 57.38%, RPIQ ≈ 1.32, RMSE ≈ 13.24 $g \cdot kg^{-1}$), which are even better than the spaceborne hyperspectral sensor DESIS. As satellite multispectral missions have consistent data quality and coverage, the satellite multispectral fusion data can provide rich and long historic records for quantifying surface SOC concentration in the past decades. In addition, the multispectral missions usually have consistently high signal-to-noise ratio data quality and are well capable of quantifying SOC. Given high spatial and temporal resolutions, long historic records, and cloud-free characteristics for satellite multispectral fusion data, multispectral fusion data can be an important alternative to satellite hyperspectral remote sensing data for global soil carbon monitoring.



543

544 **Fig. 9**. Coefficient of determination R^2 of SOC concentration predictions from the simulated airborne and
545 satellite spectra. Circles refer to the performance of airborne hyperspectral sensors (HySpex and AVIRIS-
546 NG). Diamonds show the performance of spaceborne hyperspectral sensors including Hyperion (350–
547 2500nm), DESIS (400–1000nm), PRISMA (400–2500nm), SBG(380–2500nm), EnMAP (420–1000nm).
548 Squares represent the performance of satellite multispectral sensors including Landsat8, MODIS, Sentinel2,
549 Sentinel3, and Fusion data, which can be generated by integrating Landsat8, MODIS and Sentinel2 as
550 STAIR (Luo et al., 2020). The x-axis is the spatial resolution. The y-axis is the potential maximum temporal
551 resolution (without considering the influences of cloud cover). Other statistics such as MAPE, RPIQ,
552 RMSE, and rRMSE can be found in supplementary Table S3.

553

554 The influences of the most common real-environment noises (soil moisture, green vegetation and plant
555 residue) on surface bare soil reflectance were simulated as shown in Fig. 10. With higher volumetric soil

moisture, soil reflectance tends to be lower and the spectral reflectance in the shortwave infrared (> 2000 nm) has more changes than the visible wavelengths (Fig. 10a). As green vegetation spectra are quite different from bare soil spectra, both visible and infrared parts of the spectrum changed significantly after adding more signals from green vegetation cover (Fig. 10b). Unlike green vegetation spectra, plant residue signals altered the shortwave infrared mostly and did not change visible parts significantly. The plant residue simulation results also agree with previous studies that the spectral signature of bare soil and crop residue are different in the shortwave infrared rather than visible wavelengths (Daughtry et al., 2008).

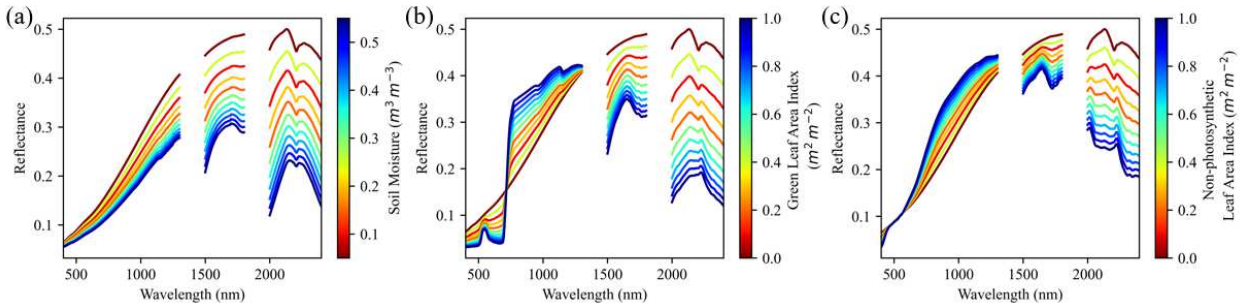


Fig. 10. Typical RaCA surface bare soil reflectance curve added with the simulated various levels of (a) soil moisture, (b) green vegetation leaf area index, and (c) plant residue leaf area index by using the coupled soil-vegetation-atmosphere radiative transfer models (SCOPE and MODTRAN).

By adding the simulated soil moisture, green vegetation leaf area index and plant residue signals to SBG hyperspectral reflectance, we further evaluated the model performance of retrained LSTM for SOC concentration predictions. Fig. 11 shows the testing performance (R^2) of the retrained LSTM with each level of noise. Other statistics of RMSE, MAPE and RPIQ can be found in supplementary Fig. S6–8. From Fig. 11, we found that within a certain level of surface noises (soil moisture < 0.4 $m^3 \cdot m^{-3}$, green leaf area < 0.3 $m^2 \cdot m^{-2}$ or plant residue < 0.4 $m^2 \cdot m^{-2}$), LSTM can mitigate noisy signals to achieve relatively similar and high performance for SOC prediction. Beyond this noise level, the model performance significantly degraded with the increased levels of surface noises. When soils turned to be extremely wet and covered by a high leaf area index, LSTM performance became the worst. By comparing Fig. 11a and b, we also

found the model performance with green leaf area index worse than that with the same level of plant residue, as green leaves alter both visible and infrared spectra (Fig. 10). To achieve acceptable model performance e.g. R^2 = 0.7, remote sensing spectra need to be collected with field volumetric soil moisture less than 0.4 $m^3 \cdot m^{-3}$, green leaf area index less than 0.3 $m^2 \cdot m^{-2}$ or plant residue less than 0.4 $m^2 \cdot m^{-2}$. This finding highlights the importance of identifying relatively pure soil pixels or using spectral unmixing techniques for using airborne or satellite remote sensing data to retrieve surface soil properties.

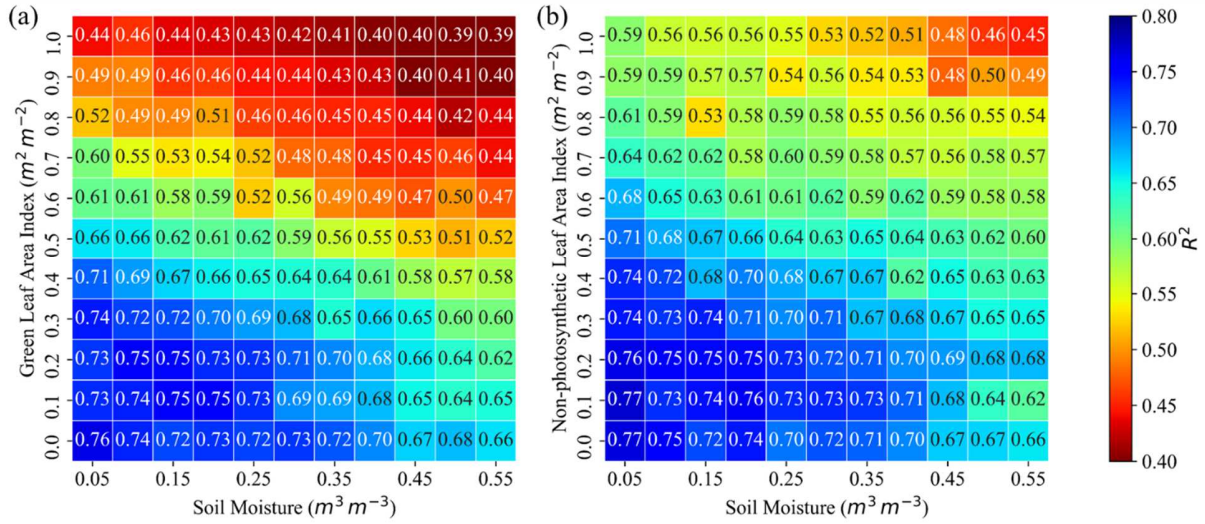


Fig. 11. Model performance (R^2) of LSTM to predict SOC concentration using the simulated SBG remote sensing data with various levels of (a) volumetric soil moisture and green vegetation leaf area index, (b) volumetric soil moisture and plant residue (non-photosynthetic leaf area index).

5. Discussion

This study utilized a large soil spectral library to conduct a comprehensive evaluation of machine learning algorithms and spectra preprocessing techniques for SOC concentration predictions. LSTM was identified as the most suitable algorithm for SOC concentration predictions. We also found that the spectral signatures of mineral and organic soils are significantly different, which should be considered when developing generalized SOC models. Forthcoming satellite hyperspectral missions such as SBG have the potential to accurately predict surface SOC concentration to facilitate global SOC monitoring. Identifying pure soil

pixels with limited noises of soil moisture, green vegetation and plant residue is important for using remote sensing to quantify surface SOC. Practical suggestions for machine learning and remote sensing data selection, implications for SOC model generalization, and uncertainties of this study are further discussed.

5.1 Selection of machine learning methods

Among seven state-of-the-art algorithms, this study identified LSTM as the most suitable to predict SOC concentration from spectra based on its high predictive performance (Fig. 4) and flexibility to be operated with both hyperspectral and multispectral data. CNN can also achieve high performance in predicting SOC concentration, but the kernel size limits the searching for spectral representation and the convolution often operates to find patterns from high-dimensional data. However, the drawbacks of such LSTM and CNN deep learning models are the requirements of large training datasets and computational power. As shown in Fig. 4, the simple machine learning algorithm, e.g. RF, combined with the first-order derivative spectral preprocessing to augment spectral signatures can also obtain high performance close to deep learning (Fig. 4 and supplementary Fig. S2). In practical applications, machine learning and feature augmentation methods should be selected according to data availability, computational sources, and requirements on model accuracy.

The emerging advanced machine learning techniques, for example, Transformer (Vaswani et al., 2017), can further interpret hyperspectral signatures to improve the accuracy of SOC concentration predictions. The Transformer model, utilizing the multi-head self-attention mechanism to effectively model long-term dependencies, has been proved to be powerful in sequential modeling, especially in natural language processing (Vaswani et al., 2017). Positional Encoding, an essential part of the Transformer architecture, can provide some sense of order information for better sequence interpretation in the absence of recurrent layers. Besides the innovations of model structure, deep learning techniques such as dropout layers (Pullanagari et al., 2021) and batch normalization (Ioffe and Szegedy, 2015) can be incorporated to improve model performances. Dropout layers can be used to effectively reduce overfitting and thus improve the

618 performance of testing data. Batch normalization or layer normalization can both speed up and stabilize the
619 training process while providing some regularization to the model. Furthermore, the latest developments in
620 deep imbalanced regression such as label distribution smoothing (LDS) and feature distribution smoothing
621 (FDS) (Yang et al., 2021) can be utilized to bridge the generalization gap brought by the long-tail
622 distribution of the dataset.

623 **5.2 SOC model generalization and improvement**

624 We found that the spectral difference between mineral and organic soils is a key limiting factor to develop
625 the generalized spectral model for SOC predictions. Since labels are divided into two major clusters, we
626 can train two separate models on low SOC concentration (mineral soils, Fig. 6e) and high SOC
627 concentration (organic soils, Fig. 6d) respectively. However, organic soils only account for a small
628 proportion of land area (Bot and Benites, 2005). The mineral soil LSTM model (Fig. 6e) could be suitable
629 as a generalized model for wide applications. Therefore, we selected soil spectra datasets with mineral soils
630 to simulate the airborne and satellite signals to evaluate their potential for SOC monitoring (Fig. 10).
631 Furthermore, another way to improve the SOC model generalization is to adjust cost functions to employ
632 evaluation metrics for imbalanced regression, in order to force the model to learn targeted observations of
633 interest in the training datasets. For example, to develop the SOC model specifically for mineral soils, cost
634 functions, such as the logarithmic or absolute difference instead of mean square errors, can be used to
635 increase sensitivity to low SOC concentrations. Additionally, we can assign a higher weight to multiply the
636 loss of low SOC concentration labels and a lower weight to high SOC concentration labels, to force the
637 model to target more on the low SOC concentration labels. Furthermore, as the distinct features of mineral
638 and organic soils, a binary classification model would be trained to predetermine the sample's prediction
639 range. If samples are likely to be mineral soils, we employ the mineral soil LSTM, otherwise the organic
640 soil LSTM. This setting may improve the model performance in both low SOC concentration and high SOC
641 concentration. Preprocessing techniques on label datasets, such as random undersampling (Galar et al.,

642 2013) and the Synthetic Minority Oversampling TEchnique (SMOTE, Chawla et al., 2002), could be an
643 alternative for generalized model development. Besides developing robust machine learning algorithms,
644 standard protocols to measure soil sample laboratory spectra and SOC concentration are also highly needed
645 to facilitate spectral model generalization (Dangal and Sanderman, 2020; Ge et al., 2011). In this study,
646 LSTM achieved an accuracy of $R^2 = 0.71$, RMSE = 10.60 $g \cdot kg^{-1}$ to predict SOC concentration for mineral
647 soils (SOC $\leq$ 120 $g \cdot kg^{-1}$) with optical hyperspectral reflectance. Such accuracy may not fit the need for
648 precise SOC quantification, for example, to detect land management effects on soil carbon change
649 (Sanderman et al., 2021). Other approaches, for example laboratory mid-infrared spectroscopy, could offer
650 an alternative for more precise quantification (Margenot et al., 2016; Wijewardane et al., 2018).

651 **5.3 Airborne and spaceborne remote sensing for monitoring SOC**

652 This study utilized the soil laboratory spectra data to simulate and evaluate the major airborne and
653 spaceborne remote sensing data for quantifying surface SOC concentration. In this simulation process, we
654 considered the effects of atmospheric attenuations, sensor spectral responses and signal-to-noise ratios,
655 surface soil moisture conditions, green vegetation cover, and plant residue cover. However, differences
656 could still exist between the simulated spectra based on laboratory spectra and the real-environment
657 collected airborne/spaceborne remote sensing signals. In our simulations, we added the same level of
658 surface noises to soil spectral datasets. Soil spectra collected from outdoor environments are more
659 complicated with mixing various levels of soil moisture (Ge et al., 2014), surface roughness (Chabrillat et
660 al., 2019), and crop residues (Stevens et al., 2010) than laboratory spectra of ground air-dry soil samples.
661 Therefore, airborne and spaceborne data need to be collected either at appropriate times or through frequent
662 time series to build bare soil reflectance composites. For example, when soils under agricultural
663 management are plowed and disced into < 5 cm aggregates with limited and more uniform crop residue
664 coverage, field soil spectra would be close to laboratory soil spectra measurements (Chabrillat et al., 2019).
665 Studies also demonstrated that remote sensing time series could be utilized to build reflectance composites

to derive bare soil reflectance for surface SOC predictions (Rogge et al., 2018; Demattê et al., 2018; Silvero et al., 2021; Vaudour et al., 2021). Particularly, recent studies show the promising performance of satellite remote sensing for quantifying surface SOC concentration, although the model performance for practical applications still needs improvement (Castaldi et al., 2019; Silvero et al., 2021; Zhou et al., 2021). Overall, using the real-environment spectra from airborne or satellite remote sensing for surface SOC estimation will be much more complex. Furthermore, beyond surface SOC, airborne and satellite remote sensing techniques need to be integrated with stratified field data collection (e.g. Potash et al., 2021) and agroecosystem carbon cycling modeling (e.g. Zhou et al., 2021) to provide a holistic view on the soil carbon stock changes. Nevertheless, this study leveraged a large soil database with diverse soil types and provide a comprehensive assessment of remote sensing data for quantifying surface SOC, which can give valuable insights into remote sensing mission designs and data selection (Castaldi et al., 2016).

Besides evaluating the potential of airborne and satellite data for SOC predictions, soil laboratory spectral models can facilitate large-scale SOC quantification. For example, regional-scale surface SOC mapping with airborne or satellite remote sensing requires a large amount of label data from targeted areas. The high financial and time costs of conventional dry combustion methods to quantify SOC concentrations, which constrain scaling of SOC monitoring, can be circumvented by spectra-based predictions (Nocita et al., 2015). With the strong ability to accurately quantify SOC concentration, the laboratory spectral models for SOC predictions can potentially serve as a reference approach to significantly reduce the cost and time for obtaining the ground truth of SOC data. Furthermore, the similarity between laboratory and field spectra enables the transfer learning and parameter fine-tuning of the laboratory spectral model to predict field surface SOC concentration from airborne or satellite remote sensing with limited field measurements (Liu et al., 2018; Tziolas et al., 2020).

6. Conclusion

This study utilized the USDA public soil spectral library with 37,540 records of soil 350–2500nm reflectance to assess the performance of the state-of-the-art machine learning and spectral preprocessing algorithms on SOC concentration predictions. These machine learning algorithms include Partial-Least Squares Regression (PLSR), Random Forest (RF), K-Nearest Neighbors (KNN), Ridge, Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM). LSTM achieved the best predictions for SOC concentration (R^2 = 0.96, RMSE = 30.81 $g \cdot kg^{-1}$, RPIQ =1.24) without any feature preprocessing techniques. CNN can obtain high predictive performance, but could not be operated with multispectral data with limited wavelengths in convolution. The first-order derivative spectral processing can significantly improve the model performance for PLSR, RF, KNN, and Ridge. Furthermore, we also found that the generalized SOC models developed from the whole datasets consisting of mineral and organic soils had more biases in mineral soils, which dominate terrestrial ecosystems. By using the only mineral soil spectra data, the mineral soil LSTM model performance has been significantly improved (R^2 = 0.71, RMSE = 10.60 $g \cdot kg^{-1}$, RPIQ =1.44). By using the coupled soil-vegetation-atmosphere radiative transfer modeling, we simulated the mainstream and forthcoming airborne and spaceborne hyper/multi-spectral data with consideration of atmospheric attenuations, sensor spectral responses and signal-to-noise ratios, surface soil moisture, green vegetation cover, and plant residue. Results demonstrate that with the same leaf area index, green vegetation has more negative impacts on SOC predictions than plant residue does. LSTM can mitigate noise signals (soil moisture < 0.4 $m^3 \cdot m^{-3}$, green leaf area < 0.3 $m^2 \cdot m^{-2}$ or plant residue < 0.4 $m^2 \cdot m^{-2}$) for surface SOC predictions. Satellite hyperspectral missions such as NASA SBG with both visible and shortwave infrared reflectance have great potential to enhance global SOC monitoring. The fusion data from multispectral satellite missions can also be used for global soil carbon monitoring and particularly tracing the historic surface soil carbon changes in the past decades.

Acknowledgment

This work was financed by the U.S. Department of Energy's Advanced Research Projects Agency-Energy (ARPA-E) SMARTFARM and SYMFONI projects. We would also like to thank the support from the seed fundings to S.W. and K.G. from Illinois Discovery Partners Institute (DPI), Institute for Sustainability, Energy, and Environment (iSEE), and College of Agricultural, Consumer and Environmental Sciences Future Interdisciplinary Research Explorations (ACES FIRE), Center for Digital Agriculture (CDA-NCSA), University of Illinois at Urbana-Champaign. This work was also partially funded by the USDA National Institute of Food and Agriculture (NIFA) Artificial Intelligence for Future Agricultural Resilience, Management, and Sustainability grant. Y.G. was supported by a USDA-NIFA grant (Award#: 2018-67007-28529). The authors thank the staff members at the Kellogg Soil Survey Lab of the USDA-NRCS for maintaining and querying the RaCA database. The authors would like to thank the editors and anonymous reviewers for the comments and suggestions to improve the quality of this paper.

Reference

Altmann, A., Toloşi, L., Sander, O., Lengauer, T., 2010. Permutation importance: A corrected feature importance measure. Bioinformatics 26, 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Angelopoulou, T., Tziolas, N., Balafoutis, A., Zalidis, G., Bochtis, D., 2019. Remote sensing techniques for soil organic carbon estimation: A review. Remote Sens. https://doi.org/10.3390/rs11060676

Bae, S.H., Choi, I., Kim, N.S., 2016. Acoustic Scene Classification Using Parallel Combination of LSTM and CNN. Proc. Detect. Classif. Acoust. Scenes Events 2016 Work. 11–15.

Bartholomeus, H., Kooistra, L., Stevens, A., van Leeuwen, M., van Wesemael, B., Ben-Dor, E., Tychon, B., 2011. Soil Organic Carbon mapping of partially vegetated agricultural fields with imaging spectroscopy. Int. J. Appl. Earth Obs. Geoinf. 13, 81–88. https://doi.org/10.1016/j.jag.2010.06.009

Bartholomeus, H.M., Schaepman, M.E., Kooistra, L., Stevens, A., Hoogmoed, W.B., Spaargaren, O.S.P., 2008. Spectral reflectance based indices for soil organic carbon quantification. Geoderma 145, 28–36. https://doi.org/10.1016/j.geoderma.2008.01.010

Batjes, N. H. 2009. Harmonized soil profile data for applications at global and continental scales: updates

738 to the WISE database. Soil Use and Management, 25(2), 124-127.
739 Belgiu, M., Drăgu, L., 2016. Random forest in remote sensing: A review of applications and future
740 directions. ISPRS J. Photogramm. Remote Sens. https://doi.org/10.1016/j.isprsjprs.2016.01.011
741 Ben-Dor, E., Chabrillat, S., Demattê, J.A.M., Taylor, G.R., Hill, J., Whiting, M.L., Sommer, S., 2009. Using
742 imaging spectroscopy to study soil properties. Remote Sens. Environ. 113, S38–S55.
743 Berk, A., Anderson, G.P., Acharya, P.K., Shettle, E.P., 2011. MODTRAN 5.2.1 User's Manual Spectral
744 Sciences, Inc., 4 Fourth Ave., Burlington, MA 01803-3304 Air Force Research Laboratory, Space
745 Vehicles Directorate, Air Force Materiel Command, Hanscom AFB, MA 01731-3010
746 Bot, A., Benites, J., 2005. The importance of soil organic matter: Key to drought-resistant soil and sustained
747 food production. Food & Agriculture Org.
748 Castaldi, F., Chabrillat, S., Don, A. and van Wesemael, B., 2019. Soil organic carbon mapping using
749 LUCAS topsoil database and Sentinel-2 data: An approach to reduce soil moisture and crop residue
750 effects. Remote Sensing, 11(18), p.2121.
751 Castaldi, F., Hueni, A., Chabrillat, S., Ward, K., Buttafuoco, G., Bomans, B., Vreys, K., Brell, M., van
752 Wesemael, B., 2019. Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction
753 in croplands. ISPRS J. Photogramm. Remote Sens. 147, 267–282.
754 https://doi.org/10.1016/j.isprsjprs.2018.11.026
755 Castaldi, F., Palombo, A., Santini, F., Pascucci, S., Pignatti, S., Casa, R., 2016. Evaluation of the potential
756 of the current and forthcoming multispectral and hyperspectral imagers to estimate soil texture and
757 organic carbon. Remote Sens. Environ. 179, 54–65. https://doi.org/10.1016/j.rse.2016.03.025
758 Cawse-Nicholson, K., Townsend, P.A., Schimel, D., Assiri, A.M., Blake, P.L., Buongiorno, M.F.,
759 Campbell, P., Carmon, N., Casey, K.A., Correa-Pabón, R.E., Dahlin, K.M., Dashti, H., Dennison,
760 P.E., Dierssen, H., Erickson, A., Fisher, J.B., Frouin, R., Gatebe, C.K., Gholizadeh, H., Gierach, M.,
761 Glenn, N.F., Goodman, J.A., Griffith, D.M., Guild, L., Hakkenberg, C.R., Hochberg, E.J., Holmes,
762 T.R.H., Hu, C., Hulley, G., Huemmrich, K.F., Kudela, R.M., Kokaly, R.F., Lee, C.M., Martin, R.,
763 Miller, C.E., Moses, W.J., Muller-Karger, F.E., Ortiz, J.D., Otis, D.B., Pahlevan, N., Painter, T.H.,

764 Pavlick, R., Poulter, B., Qi, Y., Realmuto, V.J., Roberts, D., Schaepman, M.E., Schneider, F.D.,
765 Schwandner, F.M., Serbin, S.P., Shiklomanov, A.N., Stavros, E.N., Thompson, D.R., Torres-Perez,
766 J.L., Turpie, K.R., Tzortziou, M., Ustin, S., Yu, Q., Yusup, Y., Zhang, Q., 2021. NASA's surface
767 biology and geology designated observable: A perspective on surface imaging algorithms. Remote
768 Sens. Environ. https://doi.org/10.1016/j.rse.2021.112349
769 Chabrillat, S., Ben-Dor, E., Cierniewski, J., Gomez, C., Schmid, T., van Wesemael, B., 2019. Imaging
770 Spectroscopy for Soil Mapping and Monitoring. Surv. Geophys. https://doi.org/10.1007/s10712-019-
771 09524-0
772 Chawla, N. V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-
773 sampling technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953
774 Dangal, S.R.S., Sanderman, J., 2020. Is standardization necessary for sharing of a large mid-infrared soil
775 spectral library? Sensors (Switzerland) 20, 1–17. https://doi.org/10.3390/s20236729
776 Daughtry, C.S.T. and Hunt Jr, E.R., 2008. Mitigating the effects of soil and residue water contents on
777 remotely sensed estimates of crop residue cover. Remote Sens. Environ., 112(4), pp.1647-1657.
778 Demattê, J.A.M., Fongaro, C.T., Rizzo, R. and Safanelli, J.L., 2018. Geospatial Soil Sensing System
779 (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images.
780 Remote Sensing of Environment, 212, pp.161-175.
781 Dotto, A.C., Dalmolin, R.S.D., ten Caten, A., Grunwald, S., 2018. A systematic study on the application of
782 scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic
783 carbon by Vis-NIR spectra. Geoderma 314, 262–274.
784 https://doi.org/10.1016/j.geoderma.2017.11.006
785 Fukushima, K., Miyake, S., Ito, T., 1983. Neocognitron: A Neural Network Model for a Mechanism of
786 Visual Pattern Recognition. IEEE Trans. Syst. Man Cybern. SMC-13, 826–834.
787 https://doi.org/10.1109/TSMC.1983.6313076
788 Galar, M., Fernández, A., Barrenechea, E., Herrera, F., 2013. EUSBoost: Enhancing ensembles for highly
789 imbalanced data-sets by evolutionary undersampling. Pattern Recognit. 46, 3460–3471.

790 https://doi.org/10.1016/j.patcog.2013.05.006
791 Ge, Y., Morgan, C.L.S., Ackerson, J.P., 2014. VisNIR spectra of dried ground soils predict properties of
792 soils scanned moist and intact. Geoderma 221–222, 61–69.
793 https://doi.org/10.1016/j.geoderma.2014.01.011
794 Ge, Y., Morgan, C.L.S., Grunwald, S., Brown, D.J., Sarkhot, D. V., 2011. Comparison of soil reflectance
795 spectra and calibration models obtained using multiple spectrometers. Geoderma 161, 202–211.
796 https://doi.org/10.1016/j.geoderma.2010.12.020
797 Ge, Y., Morgan, C.L. and Wijewardane, N.K., 2020. Visible and near-infrared reflectance spectroscopy
798 analysis of soils. Soil Science Society of America Journal, 84(5), pp.1495-1502.
799 Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Anal. Chim. Acta.
800 https://doi.org/10.1016/0003-2670(86)80028-9
801 Glorot, X., Bordes, A. Bengio, Y., 2011, June. Deep sparse rectifier neural networks. In Proceedings of the
802 fourteenth international conference on artificial intelligence and statistics (pp. 315-323). JMLR
803 Workshop and Conference Proceedings.
804 Gomez, C., Viscarra Rossel, R.A., McBratney, A.B., 2008. Soil organic carbon prediction by hyperspectral
805 remote sensing and field vis-NIR spectroscopy: An Australian case study. Geoderma 146, 403–411.
806 https://doi.org/10.1016/j.geoderma.2008.06.011
807 Guanter, L., Segl, K., Kaufmann, H., 2009. Simulation of optical remote-sensing scenes with application
808 to the EnMAP hyperspectral mission. IEEE Trans. Geosci. Remote Sens. 47, 2340–2351.
809 https://doi.org/10.1109/TGRS.2008.2011616
810 Guerrero, C., Stenberg, B., Wetterlind, J., Viscarra Rossel, R.A., Maestre, F.T., Mouazen, A.M., Zornoza,
811 R., Ruiz-Sinoga, J.D., Kuang, B., 2014. Assessment of soil organic carbon at local scale with spiked
812 NIR calibrations: Effects of selection and extra-weighting on the spiking subset. Eur. J. Soil Sci. 65,
813 248–263. https://doi.org/10.1111/ejss.12129
814 Hbirkou, C., Pätzold, S., Mahlein, A.K., Welp, G., 2012. Airborne hyperspectral imaging of spatial soil
815 organic carbon heterogeneity at the field-scale. Geoderma 175–176, 21–28.

https://doi.org/10.1016/j.geoderma.2012.01.017

He, K., Zhang, X., Ren, S. and Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (pp. 1026-1034).

Hecht-Nielsen, R., 1989. Theory of the backpropagation neural network. pp. 593–605. https://doi.org/10.1109/ijcnn.1989.118638

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Comput. 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoerl, A.E., Kennard, R.W., 1970. Ridge Regression: Applications to Nonorthogonal Problems. Technometrics 12, 69–82. https://doi.org/10.1080/00401706.1970.10488635

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: 32nd International Conference on Machine Learning, ICML 2015. pp. 448–456.

Ishwaran, H., 2015. The effect of splitting on random forests. Machine learning, 99(1), pp.75-118.

Kelley, H.J., 1960. Gradient theory of optimal flight paths. Ars Journal, 30(10), pp.947–954.

Keller, J.M., Gray, M.R., 1985. A Fuzzy K-Nearest Neighbor Algorithm. IEEE Trans. Syst. Man Cybern. SMC-15, 580–585. https://doi.org/10.1109/TSMC.1985.6313426

Kingma, D.P. Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Krutz, D., Venus, H., Eckardt, A., Walter, I., Sebastian, I., Reulke, R., Günther, B., Zender, B., Arloth, S., Williges, C., Lieder, M., Neidhardt, M., Grote, U., Schrandt, F., Wojtkowiak, A., 2018. DESIS - DLR Earth Sensing Imaging Spectrometer, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 356–368. https://doi.org/10.1007/978-3-319-92753-4_28

Lal, R., 2016. Soil health and carbon management. Food Energy Secur. https://doi.org/10.1002/fes3.96

Leevy, J.L., Khoshgoftaar, T.M., Bauder, R.A. and Seliya, N., 2018. A survey on addressing high-class imbalance in big data. Journal of Big Data, 5(1), pp.1-30.

Liu, L., Ji, M., Buchroithner, M., 2018. Transfer learning for soil spectroscopy based on convolutional

842 neural networks and its application in soil clay content mapping using hyperspectral imagery. Sensors
843 (Switzerland) 18. https://doi.org/10.3390/s18093169
844 Lobsey, C.R., Viscarra Rossel, R.A., Roudier, P., Hedley, C.B., 2017. rs-local data-mines information from
845 spectral libraries to improve local calibrations. Eur. J. Soil Sci. 68, 840–852.
846 https://doi.org/10.1111/ejss.12490
847 Loshchilov, I. Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.
848 Luo, Y., Guan, K., Peng, J., Wang, S., Huang, Y., 2020. STAIR 2.0: A generic and automatic algorithm to
849 fuse modis, landsat, and sentinel-2 to generate 10 m, daily, and cloud-/gap-free surface reflectance
850 product. Remote Sens. 12, 1–21. https://doi.org/10.3390/rs12193209
851 Margenot, A.J., Calderón, F.J., Goyne, K.W., Dmukome, F.N., Parikh, S.J., 2016. IR Spectroscopy, soil
852 analysis applications, in: Encyclopedia of Spectroscopy and Spectrometry. Elsevier, pp. 448–454.
853 Nelson, D.W., Sommers, L.E., 2015. Total Carbon, Organic Carbon, and Organic Matter. pp. 539–579.
854 https://doi.org/10.2134/agronmonogr9.2.2ed.c29
855 Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Dor, E. Ben,
856 Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M., Genot, V., Guerrero, C.,
857 Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla,
858 J.M., Shepherd, K.D., Stenberg, B., Towett, E.K., Vargas, R., Wetterlind, J., 2015. Soil Spectroscopy:
859 An Alternative to Wet Chemistry for Soil Monitoring. Adv. Agron. 132, 139–159.
860 https://doi.org/10.1016/bs.agron.2015.02.002
861 Padarian, J., Minasny, B., McBratney, A.B., 2020. Machine learning and soil sciences: A review aided by
862 machine learning tools. SOIL. https://doi.org/10.5194/soil-6-35-2020
863 Pal, M., 2005. Random forest classifier for remote sensing classification. Int. J. Remote Sens. 26, 217–222.
864 https://doi.org/10.1080/01431160412331269698
865 Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N.,
866 Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S.,
867 Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep

868 learning library, in: Advances in Neural Information Processing Systems.
869 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer,
870 P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M.,
871 Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
872 Pernkopf, F., Bouchaffra, D., 2005. Genetic-based EM algorithm for learning Gaussian mixture models.
873 IEEE Trans. Pattern Anal. Mach. Intell. 27, 1344–1348. https://doi.org/10.1109/TPAMI.2005.162
874 Potash, E., Guan, K., Margenot, A., Lee, D., DeLucia E., Wang, S., Jang, C. 2022. How to estimate soil
875 organic carbon stocks of agricultural fields? Perspectives using ex ante evaluation. Geoderma.
876 Pullanagari, R.R., Dehghan-Shoar, M., Yule, I.J., Bhatia, N., 2021. Field spectroscopy of canopy nitrogen
877 concentration in temperate grasslands using a convolutional neural network. Remote Sens. Environ.
878 https://doi.org/10.1016/j.rse.2021.112353
879 Reeves, J., McCarty, G., Mimmo, T., 2002. The potential of diffuse reflectance spectroscopy for the
880 determination of carbon inventories in soils. Environ. Pollut. 116. https://doi.org/10.1016/S0269-
881 7491(01)00259-7
882 Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep
883 learning and process understanding for data-driven Earth system science. Nature 566, 195–204.
884 https://doi.org/10.1038/s41586-019-0912-1
885 Rice, C.W., 2004. Carbon Cycle in Soils - Dynamics and Management, in: Encyclopedia of Soils in the
886 Environment. pp. 164–170. https://doi.org/10.1016/B0-12-348530-4/00183-1
887 Riese, F.M., Keller, S., 2019. SOIL TEXTURE CLASSIFICATION with 1D CONVOLUTIONAL
888 NEURAL NETWORKS BASED on HYPERSPECTRAL DATA, in: ISPRS Annals of the
889 Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 615–621.
890 https://doi.org/10.5194/isprs-annals-IV-2-W5-615-2019
891 Rogge, D., Bauer, A., Zeidler, J., Mueller, A., Esch, T. and Heiden, U., 2018. Building an exposed soil
892 composite processor (SCMaP) for mapping spatial and temporal characteristics of soils with Landsat
893 imagery (1984–2014). Remote Sensing of Environment, 205, pp.1-17.

894 Sanderman, J., Baldock, J.A., Dangal, S.R.S., Ludwig, S., Potter, S., Rivard, C., Savage, K., 2021. Soil
895 organic carbon fractions in the Great Plains of the United States: an application of mid-infrared
896 spectroscopy. Biogeochemistry. https://doi.org/10.1007/s10533-021-00755-1
897 Sanderman, J., Farquharson, R., Baldock, J., 2009. Soil carbon sequestration potential: a review for
898 Australian agriculture. Soil carbon sequestration potential a Rev. Aust. Agric.
899 Scharlemann, J.P.W., Tanner, E.V.J., Hiederer, R., Kapos, V., 2014. Global soil carbon: Understanding and
900 managing the largest terrestrial carbon pool. Carbon Manag. https://doi.org/10.4155/cmt.13.77
901 Schuur, E.A.G., McGuire, A.D., Schädel, C., Grosse, G., Harden, J.W., Hayes, D.J., Hugelius, G., Koven,
902 C.D., Kuhry, P., Lawrence, D.M., Natali, S.M., Olefeldt, D., Romanovsky, V.E., Schaefer, K.,
903 Turetsky, M.R., Treat, C.C., Vonk, J.E., 2015. Climate change and the permafrost carbon feedback.
904 Nature. https://doi.org/10.1038/nature14338
905 Silvero, N.E.Q., Demattê, J.A.M., Amorim, M.T.A., dos Santos, N.V., Rizzo, R., Safanelli, J.L., Poppiel,
906 R.R., de Sousa Mendes, W. and Bonfatti, B.R., 2021. Soil variability and quantification based on
907 Sentinel-2 and Landsat-8 bare soil images: A comparison. Remote Sensing of Environment, 252,
908 p.112117.
909 Staff, S.S., 2010. Keys to Soil Taxonomy. Soil Conserv. Serv. 12, 410.
910 Stefano, P., Angelo, P., Simone, P., Filomena, R., Federico, S., Tiziana, S., Umberto, A., Vincenzo, C.,
911 Acito, N., Marco, D., Stefania, M., Giovanni, C., Raffaele, C., Roberto, D.B., Giovanni, L., Cristina,
912 A., 2013. The PRISMA hyperspectral mission: Science activities and opportunities for agriculture and
913 land monitoring, in: International Geoscience and Remote Sensing Symposium (IGARSS). pp. 4558–
914 4561. https://doi.org/10.1109/IGARSS.2013.6723850
915 Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R., Hoffmann, L., van Wesemael, B., 2010.
916 Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy.
917 Geoderma 158, 32–45. https://doi.org/10.1016/j.geoderma.2009.11.032
918 Stolt, M.H., Bakken, J., 2014. Inconsistencies in Terminology and Definitions of Organic Soil Materials.
919 Soil Sci. Soc. Am. J. 78, 1332–1337. https://doi.org/10.2136/sssaj2014.02.0048n

Su, H., Zhang, T., Lin, M., Lu, W., Yan, X. H. 2021. Predicting subsurface thermohaline structure from remote sensing data based on long short-term memory neural networks. Remote Sens. Environ, 260, 112465.

Thaler, E.A., Larsen, I.J. and Yu, Q., 2021. The extent of soil loss across the US Corn Belt. Proceedings of the National Academy of Sciences, 118(8).

Tóth, G., Jones, A., Montanarella, L., 2013. The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. Environ. Monit. Assess. 185, 7409–7425. https://doi.org/10.1007/s10661-013-3109-3

Tziolas, N., Tsakiridis, N., Ogen, Y., Kalopesa, E., Ben-Dor, E., Theocharis, J., Zalidis, G., 2020. An integrated methodology using open soil spectral libraries and Earth Observation data for soil organic carbon estimations in support of soil-related SDGs. Remote Sens. Environ. 244. https://doi.org/10.1016/j.rse.2020.111793

Van Der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2625.

Van Tol, C., Verhoef, W., Timmermans, J., Verhoef, A. and Su, Z., 2009. An integrated model of soil-canopy spectral radiances, photosynthesis, fluorescence, temperature and energy balance. Biogeosciences, 6(12), pp.3109-3129.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in Neural Information Processing Systems. pp. 5999–6009.

Vaudour, E., Gomez, C., Lagacherie, P., Loiseau, T., Baghdadi, N., Urbina-salazar, D., Loubet, B., Arrouays, D., 2021. International Journal of Applied Earth Observations and Geoinformation Temporal mosaicking approaches of Sentinel-2 images for extending topsoil organic carbon content mapping in croplands. Int. J. Appl. Earth Obs. Geoinf. 96, 102277.

Verhoef, W. Bach, H., 2012. Simulation of Sentinel-3 images by four-stream surface–atmosphere radiative transfer modeling in the optical and thermal domains. Remote Sens. Environ., 120, pp.197-207.

Verhoef, W., Van Der Tol, C. and Middleton, E.M., 2018. Hyperspectral radiative transfer modeling to

explore the combined retrieval of biophysical parameters and canopy fluorescence from FLEX–Sentinel-3 tandem mission multi-sensor data. Remote Sens. Environ., 204, pp.942-963.

Walkley, A., Black, I.A., 1934. An examination of the Degtjareff method for determining organic carbon in soils: Effect of variation in digestion conditions and of inorganic soil constituents. Soil Sci. 63, 251–263.

Wang, S., Guan, K., Wang, Z., Ainsworth, E.A., Zheng, T., Townsend, P.A., Liu, N., Nafziger, E., Masters, M.D., Li, K. and Wu, G., 2021. Airborne hyperspectral imaging of nitrogen deficiency on crop traits and yield of maize by machine learning and radiative transfer modeling. Int. J. Appl. Earth Obs. Geoinf., 105, p.102617.

Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., von Lützow, M., Marin-Spiotta, E., van Wesemael, B., Rabot, E., Ließ, M., Garcia-Franco, N., Wollschläger, U., Vogel, H.J., Kögel-Knabner, I., 2019. Soil organic carbon storage as a key function of soils - A review of drivers and indicators at various scales. Geoderma. https://doi.org/10.1016/j.geoderma.2018.07.026

Wijewardane, N.K., Ge, Y., Morgan, C.L.S., 2016a. Moisture insensitive prediction of soil properties from VNIR reflectance spectra based on external parameter orthogonalization. Geoderma 267, 92–101. https://doi.org/10.1016/j.geoderma.2015.12.014

Wijewardane, N.K., Ge, Y., Wills, S., Libohova, Z., 2018. Predicting Physical and Chemical Properties of US Soils with a Mid-Infrared Reflectance Spectral Library. Soil Sci. Soc. Am. J. 82, 722–731. https://doi.org/10.2136/sssaj2017.10.0361

Wijewardane, N.K., Ge, Y., Wills, S., Loecke, T., 2016b. Prediction of Soil Carbon in the Conterminous United States: Visible and Near Infrared Reflectance Spectroscopy Analysis of the Rapid Carbon Assessment Project. Soil Sci. Soc. Am. J. 80, 973–982. https://doi.org/10.2136/sssaj2016.02.0052

Wills, S., Loecke, T., Sequeira, C., Teachman, G., Grunwald, S., West, L.T., 2014. Overview of the U.S. Rapid Carbon Assessment Project: Sampling Design, Initial Summary and Uncertainty Estimates, in: Soil Carbon. pp. 95–104. https://doi.org/10.1007/978-3-319-04084-4_10

Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemom. Intell. Lab. Syst. 2, 37–

52. https://doi.org/10.1016/0169-7439(87)80084-9

Xiao, C., Chen, N., Hu, C., Wang, K., Gong, J., Chen, Z., 2019. Short and mid-term sea surface temperature prediction using time-series satellite data and LSTM-AdaBoost combination approach. Remote Sens. Environ. 233. https://doi.org/10.1016/j.rse.2019.111358

Yang, Y., Zha, K., Chen, Y.C., Wang, H. Katabi, D., 2021. Delving into Deep Imbalanced Regression. arXiv preprint arXiv:2102.09554.

Zhou, T., Geng, Y., Ji, C., Xu, X., Wang, H., Pan, J., Bumberger, J., Haase, D., Lausch, A., 2021. Prediction of soil organic carbon and the C:N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images. Sci. Total Environ. 755. https://doi.org/10.1016/j.scitotenv.2020.142661

Zhou, W., Guan, K., Peng, B., Tang, J., Jin, Z., Jiang, C., Grant, R. and Mezbahuddin, S., 2021. Quantifying carbon budget, crop yields and their responses to environmental variability using the ecosys model for US Midwestern agroecosystems. Agricultural and Forest Meteorology, 307, p.108521.