Grunwald. *Carbon Footprints* 2022;1:5
DOI: 10.20517/cf.2022.03

Carbon Footprints

Review

Open Access

Check for updates

Artificial intelligence and soil carbon modeling demystified: power, potentials, and perils

Sabine Grunwald

Soil and Water Sciences Department, University of Florida, 2181 McCarty Hall, PO Box 110290, Gainesville, FL 32611, USA.

Correspondence to: Prof. Sabine Grunwald, Soil and Water Sciences Department, University of Florida, 2181 McCarty Hall, PO Box 110290, Gainesville, FL 32611, USA. E-mail: sabgru@ufl.edu

How to cite this article: Grunwald S. Artificial intelligence and soil carbon modeling demystified: power, potentials, and perils. *Carbon Footprints* 2022;1:5. https://dx.doi.org/10.20517/cf.2022.03

Received: 27 Feb 2022 **First Decision:** 2 Mar 2022 **Revised:** 16 Mar 2022 **Accepted:** 6 Apr 2022 **Published:** 18 Apr 2022

Academic Editor: P. K. Ramachandran Nair **Copy Editor:** Xi-Jun Chen **Production Editor:** Xi-Jun Chen

Abstract

The global soil carbon pool has been estimated to exceed the amount of carbon stored in the atmosphere and vegetation, though uncertainties to quantify below-ground carbon and soil carbon fluxes accurately still exist. Modeling soil carbon using artificial intelligence (AI) - machine learning (ML) and deep learning (DL) algorithms - has emerged as a powerful force in the carbon science community. These AI soil carbon models have shown improved performance to predict soil organic carbon (SOC) storage, soil respiration (R_s), and other properties of the global carbon cycle when compared to other modeling approaches. AI systems have advanced abilities to optimize fits between inputs (geospatial environmental covariates) and outputs (e.g., SOC or R_s) through advanced pattern recognition, learning algorithms, latent variables, hyperparameters, hyperplanes, weighting factors, or multiple stacked processing (e.g., convolution and pooling). These machine-oriented applications have shifted focus from knowledge discovery and understanding of ecosystem processes, carbon pools and cycling toward data-driven applications that compute digital soil carbon outputs. The purpose of this review paper is to explore the emergence, applications, and progress of AI-ML and AI-DL algorithms to model soil carbon storage and R_s at regional and global scales. A critical discussion of the power, potentials, and perils of AI soil carbon modeling is provided. The paradigm shift toward AI modeling raises questions how we study soil carbon dynamics and what conclusions we draw which impacts carbon science research and education, carbon management, carbon policies, carbon markets and economies, and soil health.

Keywords: Soil carbon, soil organic carbon, soil respiration, artificial intelligence, machine learning, deep learning, artificial neural networks

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

INTRODUCTION

The global soil carbon (C) pool has been estimated to exceed the amount of carbon stored in the atmosphere and vegetation[1-3], though uncertainties to quantify below-ground carbon accurately still exist. According to Scharlemann *et al.*[2] (2014), the total estimated global SOC stock mean is about 1500 Pg C with highest soil organic carbon (SOC) stored in boreal moist (~350 Pg C), cool temperate moist (~210 Pg C), and tropical moist ecosystems (~150 Pg C). However, the uncertainty of studies that assessed soil carbon stocks (0-1 m) is considerable ranging from 504 to 3000 Pg C among 27 different global assessments[2]. These uncertainties are due to the large variability in SOC, changing climate and environmental conditions that impact soil C dynamics, different methods and modeling approaches to estimate SOC, as well as data limitations and overuse of legacy data[4]. A substantial proportion of SOC has been measured only in the topsoil (< 30 cm) with sparser observations in the subsoils that have been estimated to store about half of the global soil carbon[5-7].

The spatial and temporal variability of soil carbon storage, soil carbon sequestration (SCseq), and carbon fluxes are critically important to address soil health, soil security, food security, regenerative agriculture, and climate-smart soil conservation management. Soil carbon provides an ecosystem service implicated in numerous soil functions, such as nutrient regulation and mitigation of greenhouse gas (GHG) emissions that are pivotal to emergent carbon economies and markets. The significance of soil carbon in global biogeochemical cycles is profound. To sustain multiple soil functions and preserve soil health and soil security several quantification methods, among them artificial intelligence (AI), have been utilized at escalating spatial scales. The purpose of this review paper is to explore the emergence, applications, and potential of AI - machine leaning (ML) and deep learning (DL) algorithms - to model soil carbon storage and soil respiration (R_s) at regional and global scales. A critical discussion of the power, potentials, and perils of AI soil carbon modeling is provided.

Soil carbon assessments and dynamics

Soils are considered net sinks for soil carbon with global net sequestration estimated at 1 Pg C yr^{-1}[8]. To enhance SOC sequestration agricultural practices, such as no-tillage, conservation tillage or reduced tillage, and land use conversions have been suggested to offset GHG emissions[9,10]. Estimates suggest that land use contributes about 25% of total global GHG emissions (mainly CO_2, CH_4 and N_2O) with 10%-14% directly from agricultural production, specifically via GHG emissions from soils and livestock management, and another 12%-17% from land cover change, including deforestation and conversion of grassland[11]. Specifically, emissions of N_2O and CH_4 from soils with high greenhouse warming potentials with 280-310 and 56-21 times that of CO_2 (20-100 years, respectively) are implicated in soil carbon gains and losses. Six *et al.*[12] (2004) found in a global meta-analysis that in no-tillage agricultural systems SCseq observations were positive +195, +213, +222 kg C ha^{-1} yr^{-1} in the topsoil in humid climate after 5, 10, and 20 years of measurements, respectively. However, initial SOC losses due to increased GHG emissions from soils were observed in the topsoil in temperate dry climate with SCseq observations of -306, -37, and +97 kg C ha^{-1} yr^{-1} after 5, 10, and 20 years. Sun *et al.*[13] (2020) in a global meta-analysis in no-tillage systems assessed that SCseq varied between -2.75 to +3.99 Mg C ha^{-1} yr^{-1} (0.35 ± 0.05 standard error) in the topsoil with climate dependent sequestration rates. However, besides climatic factors such as mean annual temperature and mean annual precipitation[13-16], other factors such as soil texture[17], crop frequency and legumes cover crops[18] can pose major influence on SCseq in no-tillage or conservation tillage systems. Agricultural-based GHG mitigation practices were estimated with wide ranges dependent on assumptions of C pricing [$US20 to US100 per Mg CO_2(eq)] up to a maximum technical potential: (1) biochar application: 1.0-1.8 Pg CO_2(eq) yr^{-1}; (2) grazing land management: 0.3-1.6 Pg CO_2(eq) yr^{-1}; (3) cropland management: 0.3-1.5 Pg CO_2(eq) yr^{-1}; (4) enhanced root phenotypes: about 1 Pg CO_2(eq) yr^{-1}; (5) restore degraded land: 0.1-0.7 Pg CO_2(eq) yr^{-1}; (6) restore Histosols: 0.3-1.3 Pg CO_2(eq) yr^{-1}; (7) rice management:

0.2-0.3 Pg CO_2(eq) yr^{-1}; (8) water management 0-0.07 Pg CO_2(eq) yr^{-1}; and (9) retirement of land (setaside): 0.01-0.05 Pg CO_2(eq) yr^{-1}[9,14].

Numerous science-informed initiatives and programs to enhance SCseq paint an optimistic carbon future. One prominent initiative, the "4 per Mille Soils for Food Security and Climate" initiative was launched at COP21 in 2015 aiming to increase global soil organic matter stocks by 4 per 1000 (or 0.4%) per year as a compensation for the global emissions of GHGs by anthropogenic source[19]. According to Minasny *et al.*[19] (2017), applying the 4 per mille in the top 1m of global agricultural soils, SOC sequestration was estimated between 2-3 Gt C yr^{-1}, which would effectively offset 20%-35% of global GHG emissions. Though White *et al.*[20] (2018) disputed that such global GHG offsets are gross overestimates and the 4 per mille rate of SCseq is not feasible. Other criticism raised in regards to the "4 per Mille Soils" initiative involve poor and inconsistent calculation of target and GHG emissions, the implausibility of upscaling results to global scale, and the fact that soil carbon storage is limited and non-permanent[21]. Poulton *et al.*[22] (2018) measured SOC increases at > 7‰ per year (0-23 cm depth) in 65% on long-term experimental plots at Rothamsted UK which approximated about 4 ‰ per year (0-40 cm depth). Though it was pointed out that practices favoring SOC sequestration are already implemented in many agro-ecosystems, farmers may not have the necessary resources (e.g., insufficient manure), and some practices may be uneconomic or limit crop yield which would be undesirable to achieve global food security. van Groenigen *et al.*[23] (2017) critiqued that available nitrogen and phosphorus is insufficient to achieve 4 per mille increase of soil carbon per year. Baveye *et al.*[24] (2018) cautioned that enhanced mineralization on addition of easily decomposable carbon (i.e., the priming effect) could potentially release even more CO_2 from soils, and amplified temperature increases and/or microbial activity may release large amounts of CO_2 from soils in the future. The question whether SOC storage can be increased by 0.4% (= 4 ‰) per year is a sensational hyperbole or realistic can only be answered through accurate global soil carbon assessments. The advancements in AI-soil carbon modeling offer opportunities to improve SOC stock, SCseq, and GHG emission assessments.

Artificial intelligence: machine learning and deep learning

Artificial intelligence emerged during WWII (1939-1945) when Alan Turing invented the bombe machine to crack the "Enigma" code used by Germans, which was the foundation for ML. In 1950 two undergraduate students (Marvin Minsky and Dean Edmonds) build the first neural network computer and in 1959 Donald Hebb conceptualized the Hebbian learning algorithms with many other algorithms to follow. The first adoption of the term "AI" occurred at the Dartmouth conference in computer science in 1956. But it was not until the 1990s onward when the increase of computational power enabled the blossoming of AI algorithms and integration of AI in science, technology, engineering, and mathematics research. Since the early 2000s the Big Data era brought forth AI-geoscience, AI-smart agriculture, and other AI-ML applications with more recent applications of AI-DL methods[25].

According to Russell and Norvig[25] (2020), AI is concerned with not just understanding but also building intelligent entities - machines that can compute how to act effectively and safely in a wide variety of novel situations. Machine learning refers to machines and systems that can learn from experience supplied by data and algorithms with a model training (or calibration phase) followed by a model validation phase with independent data. Machine learning is the science of getting computers to act without being explicitly programmed. In essence, machine-driven recognition of patterns and structures in data are revealed through "brute force fitting" between input and output data. Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction[26]. Deep learning is similar to ML because the former is still just another methodology of statistical learning that extracts features or attributes from raw data sets. But the advancement of DL algorithms is that they automatically extract features for classification with multiple layers of adjustable

computing elements (e.g., hidden nodes and hidden layers) with sophisticated learning algorithms that fit inputs and outputs[25]. Artificial neural networks (ANN) are inspired by the biology of the human brain, specifically the organic interconnections between neurons. The human brain analyzes information it receives and identifies it via neuron connections according to past information it has stored in memory. The brain does this by labeling and assigning information to various groups, and it does this in nanoseconds. Similarly, when a system receives an input, the DL algorithms train the artificial neurons to identify patterns and classify information to produce the desired output. But, unlike the human brain, ANNs operate via discrete layers, connections, and directions of data propagation[25,27].

AI MODELING OF SOIL CARBON STORAGE AND DYNAMICS

In the discipline of pedometrics, the adoption of AI algorithms in digital soil mapping emerged in the early 2000s. From 2015 onward advanced AI soil models were developed in the domains of proximal soil sensing and soil carbon modeling using large environmental data hypercubes[28]. A comprehensive review of AI-ML algorithms applied in digital soil mapping, including soil carbon modeling, was provided by Khaledian and Miller[29] (2020), a review of DL for digital soil mapping was provided by Padarian *et al.*[30] (2019), and a review of DL in agriculture was provided by Kamilaris and Prenafeta-Boldú[31] (2018). Recently, a comprehensive review of ML and remote sensing methods to estimate various soil indicators was presented by Diaz-Gonzalez *et al.*[32] (2022). In this section I present a brief overview of some of the most prominent AI methods that have been employed in soil carbon modeling which informs a critical discussion of the power, potentials, and perils of these methods.

Soil carbon AI models are build using hypercubes of environmental covariates as inputs [Figure 1 and 2A]. These environmental covariates represent the domains of soils (S), topography (T), ecology (E), parent material or lithology (P), atmosphere or climate (A), water or hydrology (W), biota with vegetation and organisms (B), and human activities/management (H)[4,33-35] similar to the conceptual framework of SCORPAN (McBratney *et al.*[36] 2003). The STEP-factors are relatively stable across the human lifetime, while the AWBH-factors are dynamic in space and across time. Each of these factors can be quantified through a set of variables. For example, the S factor can be characterized by soil data such as soil texture, pH, soil taxonomic class, cation exchange capacity, *etc.* derived from legacy soil maps or databases, proximal soil sensing (e.g., visible-near infrared spectroscopy, VNIR; mid-infrared spectroscopy), gamma ray sensing, and remotely sensed soil moisture data. The factor A may be populated by climatic data such as long-term average of mean annual precipitation, seasonal variation of minimum and maximum temperature, and long-term solar radiation, while B can be populated by satellite-derived land use/land cover maps, vegetation indices like the Normalized Difference Vegetation Index (NDVI) or Enhanced Vegetation Index (EVI) derived from satellite data, biodiversity, and habitat data. The factor H may be populated by variables from the social, cultural, economic, and political domains (e.g., greenhouse gas emission data, land management data such as tillage operations, and fertilization amount and type). The aim is to populate the STEP-AWBH factors with environmental geodata that influence the carbon cycle. Xiong *et al.*[37] (2014) exemplified the STEP-AWBH model and multiple AI-ML methods in Florida, United States, to develop prediction models for SOC stocks.

Commonly applied ai algorithms to model soil carbon

Classification and Regression Trees (CART) were introduced by Breiman[38] (1984) and have served as foundational approach onto which other ML have built on [Figure 2B]. According to Breiman[38] (1984), CART involves constructing a set of decision trees on the predictor variables. The trees are grown by repeatedly stratifying the dataset into successively smaller subsets (child node) with binary splits based on a single categorical or continuous predictor variable. The splitting procedure is applied until the best split is

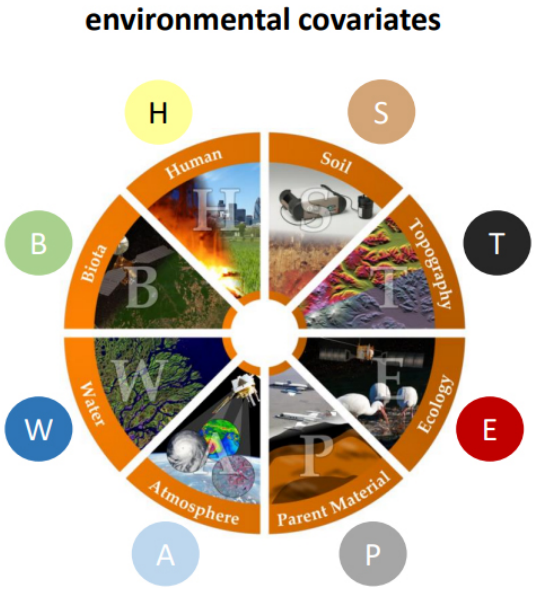

Figure 1. Environmental covariates from the domains of soils (S), topography (T), ecology (E), parent material (P), atmosphere (A), water (W), biota (B), and human (H). Each of these domains is represented by a wide variety of variables that facilitate AI-soil carbon modeling (image is courtesy of S. Grunwald).

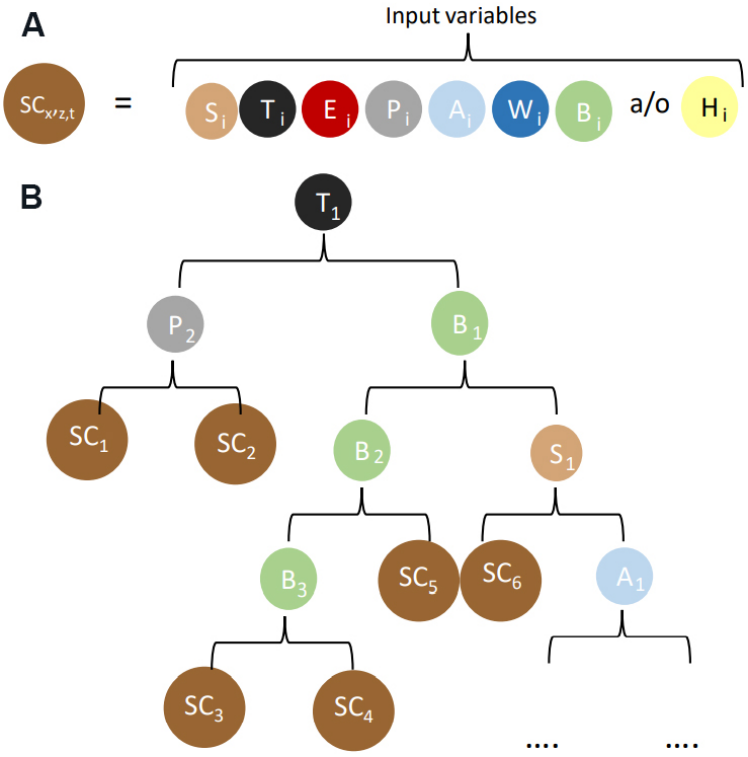

Figure 2. (A) Functional relations between environmental covariates (STEPAWBH factors with variables i = 1, 2, 3, ..., N). SC denotes a variable of the carbon cycle (target output), for example, soil organic carbon (SOC) stock, SOC density, total soil carbon, soil respiration (R_s), soil carbon sequestration (SCseq), soil carbon pools, soil carbon fractions, *etc.* x is spatial location (with xy coordinates; or latitude/longitude); z is soil depth with z = 1, 2, 3, ..., Z; and t is time with t1, t2, t3, ..., T. (B) AI model predicting soil carbon (SC) from environmental covariates. Simplified representation of a machine learning ensemble tree method (e.g., Classification and Regression Trees, CART, or Cubist) with tree branches and data splits.

chosen based on the one that maximizes the response into two homogenous groups (i.e., minimizing variability within each child node)[39].

A variant of CART is **Bagged Regression Trees (BaRT)** which is an ensemble decision tree method that involves the averaging of several individual trees to acquire a final prediction. Individual regression trees have shown somewhat erratic modeling results where small changes in input variables produces large differences in output trees[40,41]. This limitation of individual CART is overcome by bagging (i.e., bootstrap aggregation) in BaRT. Bagging is an ensemble learning method that is commonly used to reduce variance within noisy datasets. In bagging, a random sample of data in a training (or calibration) set is selected with replacement, which means that the individual data points can be chosen more than once. Thus, the procedure grows a regression tree from each bootstrap sample. To obtain the overall final prediction for a target variable the results of each individual tree are averaged[42].

Boosted Regression Trees (BoRT) belong to the Gradient Boosting Modelling family, which is one among many methods to predict the function F that maps the values of a set of predictor variables x = {x1,..,xp} into the values of the output variable y, by minimizing a specified loss function L. In BoRT the prediction is performed using boosting[43]. In general, boosting methods are applied to significantly improve the performance of a given estimation method, by generating instances of the method iteratively from a training data set and additively combining them in a forward "stagewise" procedure. BRT uses a specialized form (for regression trees) of the Stochastic Gradient Boosting[44]. The gradient boosting machine algorithm was described in detail by Friedman[44] (2001). The regression tree algorithm developed by Breiman[38] (1984) served as the foundation of BoRT, which has shown to boost accuracy compared with simple regression trees, mainly due to its stochastic gradient boosting procedure aiming at minimizing the risk of overfitting and improving its predictive power[45]. According to Hastie *et al.*[41] (2009), in BoRT trees are grown sequentially with each tree grown using the information from previously grown trees. The BoRT algorithm facilitates fitting the model to the data in an iterative process. At each iteration, individual regression trees, are fitted on a fraction (namely the bag fraction) of the dataset sampled without replacement. The main parameters for fitting BRT are the tree size and the learning rate.

Random Forest (RF) is a widely used ML method consisting of an ensemble of randomized classification and regression trees[38,46]. The RF algorithm grows different trees by randomly and repeatedly selecting predictor variables and training cases to develop a random population of trees. The algorithm grows an ensemble of regression trees based on binary recursive partitioning, where the predictor space at each tree node is partitioned based on binary splits on a subset of randomly selected predictors[47]. The output of RF is the average of individual tree predictions. It has been shown that the RF algorithm can be very efficient, especially when the number of descriptors is very large[48]. The RF model is capable of simultaneously handling categorical and continuous variables, as well as complex high-order variable relationships such as nonlinearity and interaction effects. Conditional quantiles can be inferred with **Quantile Regression Forests (QRF)**, a generalization of RF. Quantile regression forests is a non-parametric technique used to estimate the conditional quantiles of multidimensional predictor variables. The benefits of QRF is its ability to predict more accurate results for the conditional distribution of the response variable[49].

The **Support Vector Machines (SVM)** applies a projection of the input data into a high-dimensional feature space using a valid kernel function and then it uses a simple linear regression within this enhanced space[50]. This resulting linear regression function in the high-dimensional feature space corresponds to a non-linear regression in the original input space [Figure 3A]. In the new hyperspace, SVM aims to construct an

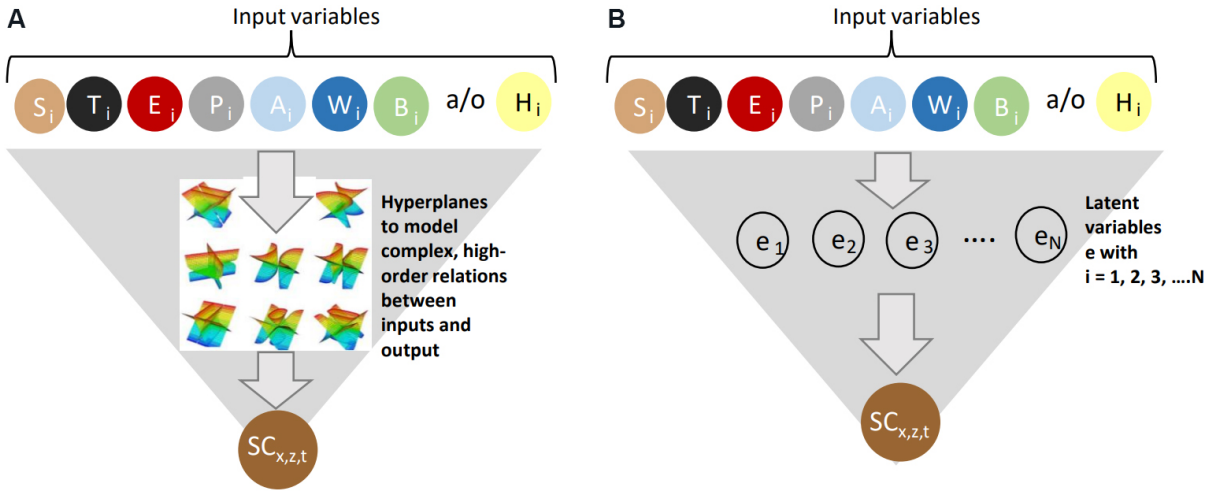

Figure 3. AI models predicting soil carbon (SC) from environmental covariates. Idealized model representation of (A) support vector machines (SVM), and (B) partial least squares regression (PLSR).

optimal hyperplane that separates classes and creates the widest margin between their data (i.e., classification), or that fits data and predicts (i.e., SVR) with minimal empirical risk and complexity of the modelling function[51,52]. **Support Vector Regression (SVR)** is a generalization of SVM and is used for nonlinear classification and regression[53]. The ε-SVR uses a loss function to define the borders (hyperplane) of the regression function. Hence the regression function lies between ± ε (maximum error). Therefore, the loss is equal to 0 if the difference between the predicted and measured values is less than ε[51].

The **Partial Least Square Regression (PLSR)** was developed by Wold[54] (1975) in econometrics and has since been widely used in many disciplines, including soil science and pedometrics. The PLSR algorithm relates the response variable (e.g., SOC) and a large number of highly collinear predictor variables (e.g., environmental covariates) through a multivariate model to identify successive orthogonal principal components (latent variables) that maximize the covariance between the response and predictor variables (Garthwaite[55], 1994). These latent factors are defined as linear combinations constructed between input variables (i.e., predictors) and response variables, such that the original multidimensionality is reduced to a lower number of orthogonal factors to detect the structure in the relationships between predictor variables and between these latent factors and the response variables [Figure 3B]. The extracted factors account for successively lower proportions of original variance[56,57]. According to Carrascal *et al.*[56] (2009), PLSR is especially suited to analyzing a large array of interrelated predictor variables (i.e., variables that are not truly independent). Soil carbon often covaries with other soil and environmental properties, and thus, PLSR is well suited to handle such multicollinearities.

The **Cubist (Cub)** algorithm is a decision tree model with piecewise linear models[58]. Cubist partitions the response data into subsets within which their characteristics are similar with respect to the predictors. A series of if-else conditions define rule-based partitions which are then arranged in a hierarchy. The simplest partition is based on only one predictor, though often multiple predictors are used to form a partition which are expressed in form of regression equations making models transparent for users [Figure 2B].

In general, an **ANN** is a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experiential knowledge and making it available for use [Figure 4A]. The benefits of ANN are (1) ability to learn and therefore generalize; (2) solve complex problems (e.g., complex

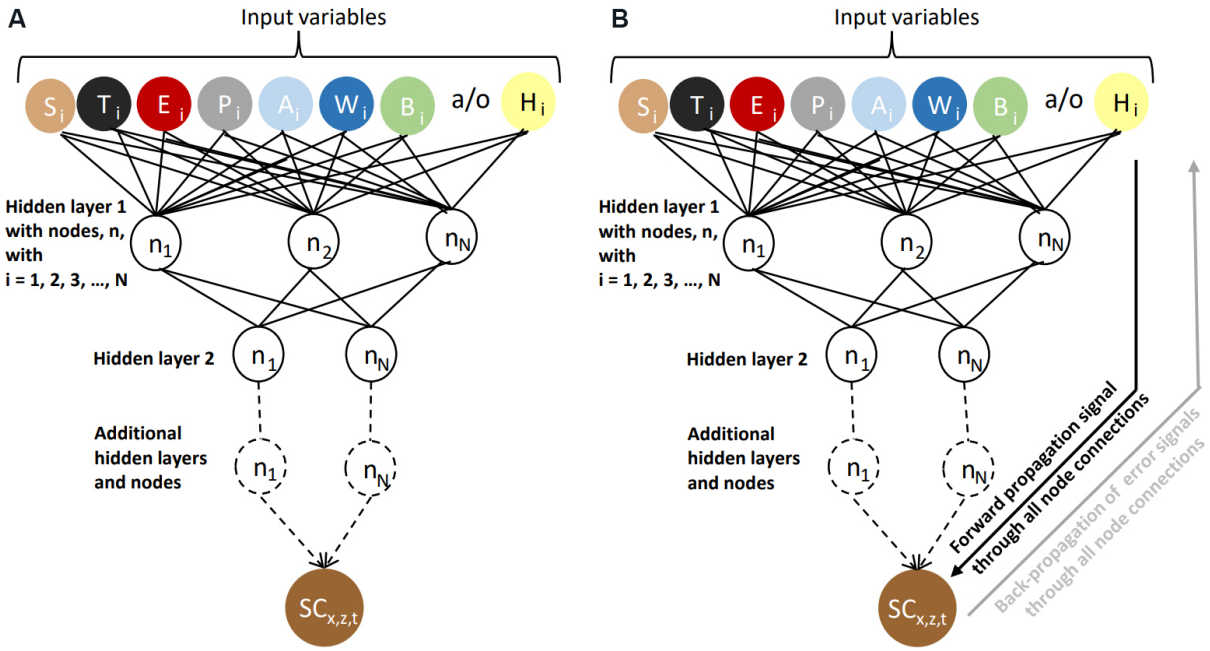

Figure 4. AI models predicting soil carbon (SC) from environmental covariates. Idealized model representation of (A) feedforward artificial neural network (fANN) and (B) backward propagation artificial neural network (bANN).

soil carbon-environmental relationships); (3) model linear, nonlinear, and high-order relations between inputs and outputs; and (4) provide input-output mapping (i.e., supervised learning)[59]. The **backpropagation ANN algorithm** is a multi-layer perceptron neural networks (i.e., a MLP neural nets) [Figure 4B]. The architecture of the MLP neural nets consists of input, one or multiple hidden, and output layers, each with a set of interconnected nodes (neurons) working in parallel to fit input data and output values through adjusting weights and cost function[60]. Hidden nodes represent abstract factors with no physical connection to ecosystems (i.e., the outside world). The purpose of hidden nodes is to transfer information from the input nodes to the output nodes. Backpropagation supervised learning is based on the error-correction learning rule. It consists of two passes through the different layers of the network: a forward pass and a backward pass. In the forward pass an activity pattern (input vector) is applied to the sensory nodes of the network and its effect propagates through the network layer by layer. Finally, a set of outputs is produced as the actual response of the network. During the forward pass the synaptic weights of the network are all fixed, while during the backward pass the synaptic weights are all adjusted in accordance with an error-correction rule. The actual response of the network is subtracted from a desired (target) response to produce an error signal. This error signal is then propagated backward through the network. During this backward pass the synaptic weights are adjusted to make the actual response of the network move closer to the desired response in a statistical sense[59]. Various backpropagation-based implementation methods including structure-fixed training and structure-adaptive training methods as well as sparse representation and dictionary learning methods were described in Wythoff[61] (1993). **Recurrent neural networks (RNN)** are algorithms for sequential data along a temporal sequence. These kind of algorithms remember its input due to an internal memory[59]. For example, RNN are suitable for soil carbon dynamics modeled over many years (e.g., SCseq modeling after conversion from conventional to no-tillage or modeling of SCseq and global climate change).

Convolutional Neural Networks (CNN) is a DL AI method that was described in detail by[27] for image, speech, and time series analysis. According to LeCun *et al.*[26] (2015), DL discovers intricate structures in complex and large datasets by using a backpropagation algorithm. A DL architecture is a multilayer stack of simple modules, all (or most) of which are subject to learning, and many of which compute non-linear input-output mappings. Deep neural networks exploit the property that many natural signals are compositional hierarchies, in which higher-level features are obtained by composing lower-level ones. Importantly, CNNs use convolutional layers to detect local conjunctions of features from previous layers with the pooling layer merging semantically similar features into one. CNNs are suited for SOC predictive modeling from environmental covariates because they allow convolution filtering (e.g., a 3 × 3 window filter) and pooling of multiple layers. Each unit of the feature map is linked to local patches in the feature maps of the previous layer through a set of weights; and the local weight sum is emulated through a non-linear transfer function. CNN allow use of data augmentation to represent soils within a region, which can reduce overfitting and also improve prediction accuracy. Another benefit of CNN is to predict different soil depths simultaneously in a model inherently taking into account the depth correlation of soil attributes. This allows to improve the prediction of SOC or other soil properties in deeper layers, which has been a common problem in other soil modeling studies with ML algorithms[30].

AI applications to model soil carbon storage and soil respiration

AI-based soil carbon stock and content modeling

Soil carbon models computed by AI methods allow explicit and rigorous evaluations computing various error metrics using cross-validation and/or validation with independent datasets. Another benefit of AI is the provision of spatially-explicit uncertainty assessment of soil carbon estimates [Table 1]. Commonly used evaluation metrics of soil carbon AI models include the coefficient of determination (R^2), root mean squared error (RMSE), mean absolute prediction error (MAE), residual prediction deviation (RPD), ratio of performance to inter-quartile range (RPIQ), and Lin's concordance correlation coefficient[62-64]. The RPIQ and RPD metrics take the variability of data into consideration though these metrics are often underreported in soil carbon studies. According to Bellon-Maurel *et al.*[62] (2010) a RPIQ < 1.00 is unreliable, 1.00 to < 1.60 is fair, 1.60 to < 2.00 is acceptable, and > 2.00 is excellent. For the RPD, a value of < 1.00 is not reliable, 1.00 to 1.40 is fair, 1.40 to < 2.00 is acceptable, and > 2.00 is considered excellent[65,66]. Several of the SOC AI models in Table 1 achieved excellent RPIQs, for example, Peng *et al.*[67] (2015) with RPIQ of 2.50 in Denmark, Ross *et al.*[68] (2019) with RPIQ of 2.10 in the southeastern U.S. The SOC models in Florida, U.S. achieved excellent RPDs up to 2.15[66] and acceptable RPD of 1.70[67], RPDs between 1.43 to 1.54 in Florida, U.S.[37] and between 1.32 and 1.88 in Florida, U.S.[69]. Regional soil carbon models derived from AI-ML and AI-DL methods showed a wide range of poor to excellent model fits with R^2 of 0.08[53] to 0.91[70], respectively [Table 1]. The RMSE results for soil carbon models shown in Table 1 need to be interpreted relative to the SOC observation range in the study region and the units of the specific soil carbon attribute. Some studies only predicted SOC concentrations, and not SOC stock, limiting interpretability in terms of soil health, soil functionality and the amount of carbon stored in soils. The model performance metrics suggests that in numerous of these studies there was substantial unexplained variability possibly linked to data limitations and/or sample densities.

Some sample sizes were small with only 220 soil samples in a study in Kenya[71], while other studies showed high numbers with 29,927 samples in East China[72], 70,803 in Australia[73], and about 150,000 soil profiles in a global study[74]. The environmental covariates (STEP-AWBH) incorporated in AI models varied widely with sometime ambiguous reporting, thus, interpretations which and how many covariates were incorporated in AI models is difficult. Though the H factor was rarely populated in most SOC models suggesting that land management, fertilization levels, GHG emissions, economic data, and other human and cultural dimensions are not given sufficient attention.

Table 1. Examples of strategically selected artificial intelligence (AI) methods to predict gridded soil organic carbon (SOC). Studies were selected to represent different geographic soilscapes, sample sizes, region sizes, and AI methods. Only the best performing models are reported from different studies

Target variable (soil depth, cm)	Units SOC	Location (approx. size, km^2)	Soil samples	Environ-mental covariates (number of variables, *n*)	SOC observations			AI method[1]	Eval	Independent validation[2]				Ref.
					Min.	Mean	Max.			R^2 or CCC	RMSE	RPD	RPIQ	
SOC stock (0-30)	Mg ha^{-1}	Eastern Mau Forest Reserve, Kenya, East Africa (650)	220	STE-AB (*n* = 19)	41.99	103.15	193.42	ANN RF SVM	Val.	0.61 0.53 0.64	15.46 17.57 14.88	- - -	- - -	Were *et al.*[71] (2015)
SOC content (0-30)	%	Skjer basin, Denmark (2500)	328	STEP-AWB	0.70	3.70	31.60	Cub (upland model C)	Val.	0.66	0.59	1.70	2.50	Peng *et al.*[67] (2015)
SOC stock (0-30; sampled at 5 increments)	Mg ha^{-1}	Eastern Australia (N/A)	564	STEP-AWB (*n* = 28)	5.08	24.80	88.23	BRT-All BRT-GA RF-All RF-GA	Val.	0.42 0.45 0.48 0.45	7.80 7.70 7.50 7.40	- - - -	- - - -	Wang *et al.*[77] (2018)
SOC stock (L1: 0-30, L2: 30-60, L3: 60-120, L4: 120-180, L5: 0-100)	kg m^{-2}	Santa Fe River Watershed, Florida, USA (3500)	554	STEP-AWB	- (L1) 1.84 (L5)	6.26 (L1) 11.79 (L5)	- (L1) 268.91 (L5)	RK/RT (L1) RK/RT (L2) RK/RT (L3) RK/RT (L4) RK/RT (L5)	Val.	- - - - -	3.69 6.31 9.31 3.01 18.48	0.65 0.97 0.21 1.04 0.38	- - - - -	Vasques *et al.*[7] (2010)
SOC stock (0-30)	kg m^{-2}	Argentina (30,000)	18,768 (5480 soil profiles)	STEP-AWB	-	-	-	QRF	Cross-val.	*0.63*	2.94	-	-	Heuvelink *et al.*[78] (2021)
SOC stock (0-20)	t C ha^{-1}	Zhejiang province, East China (102,646)	29,927	STEP-AWBH (*n* = 23)	1.18	49.74	213.55	BoRT RF	10-fold cross-val.	0.73 0.76	11.26 10.63			Deng *et al.*[72] (2018)
SOC content (0-20)	%	Florida, USA (150,000)	850	STEP-AWBH	0.13	2.68	38.57	PLSR PLSRmod RF SBIFmod	Val.	0.71 0.77 0.68 0.78	- - - -	1.85 2.08 1.782.15	0.45 0.51 0.44 0.53	Adi and Grunwald[66] (2019)[3]
SOC stock (0-20)	kg m^{-2}	Florida, USA (150,000)	1080	STEP-AWBH (*n* = 210; all relevant *n* = 43; minimum *n* = 4)	0.45	4.98	34.15	BaRT BoRT Cub RF	Val.	0.61 0.57 0.59 0.63	2.71 2.85 2.82 2.64	1.49 1.51 1.43 1.54	- - - -	Xiong *et al.*[37] (2014)
SOC stock (0-20)	kg m^{-2}	Florida, USA (150,000)	1014	STEP-AWBH (*n* = 327)	0.45	4.74	34.15	CaRT BaRT BoRT PLSR RF RK-RT SVM	Val.	0.57 0.70 0.68 0.64 0.72 0.63 0.66	3.42 2.48 2.56 2.82 2.39 2.99 2.62	1.32 1.81 1.75 1.59 1.88 1.51 1.71	0.94 1.30 1.26 1.14 1.35 1.08 1.23	Keskin *et al.*[69] (2019)[4]
SOC stock (L1: 0-20, L2: 20-100)	kg m^{-2}	South-eastern USA (350,000)	2564	STEP-AWB (*n* = 73)	1.10 (L1) 1.30 (L2)	3.70 (L1) 4.3 (L2)	12.60 (L1) 22.0 (L2)	RF (L1) RF (L2)	Val.	0.69 0.79	0.77 1.29	- -	2.10 1.96	Ross *et al.*[68] (2019)

SOC stock (0-20)	kg m^{-2}	France (640,679)	1,74	SE-AWB	0.25	-	26.0	BoRT (Cult model)	Val.	0.91	0.94	-	-	Martin *et al.*[70] (2011)
SOC stocks (L1: 0-5, L2: 5-15, L3: 15-30, L4: 30-60, L5: 60-100)	%	Chile (756,096)	1744	T-A	-	-	-	CNN (L1) CNN (L2) CNN (L3) CNN (L4) CNN (L5)	Val.	- - - - -	2.7 2.6 2.5 2.3 1.6	- - - - -	- - - - -	Padarian *et al.*[30] (2019)
SOC stocks (L1: 0-5, L2: 5-15, L3: 15-30, L4: 30-60, L5: 60-100, L6: 0-100)	log g/100 g	New South Wales, Australia (810,000)	5386	STEP-AWB (only results for whole models shown; local models also available)	-	-	-	Cub (L1) Cub (L2) Cub (L3) Cub (L4) Cub (L5) Cub (L6) SVR (L1) SVR (L2) SVR (L3) SVR (L4) SVR (L5) SVR (L6)	50-fold cross-val.	0.19 0.20 0.20 0.15 0.08 0.16 0.22 0.25 0.23 0.16 0.11 0.20	0.81 0.77 0.89 0.94 0.95 0.87 0.79 0.75 0.88 0.93 0.93 0.86	- - - - - - - - - - - -	- - - - - - - - - - - -	Somarathna *et al.*[53] (2016)
SOC content (L1: 0-5, L2: 5-15, L3: 15-30, L4: 30-60, L5: 60-100, L6: 100-200)	%	Australia (7.692 million)	70,803	STEP-AWB	0.001	1.73	36.32	Cub Cub-RK	Val.	-	-	-	-	Viscarra Rossel *et al.*[73] (2015)
SOC stock (0-20)	Mg ha^{-1}	USA (9.63 million)	3303	STEP-AWB	0.26	56.87	524.83	RF QRF	Val.	0.33 0.35	28.39 28.15	1.21 1.22	1.37 1.39	Cao *et al.*[75] (2019)
SOC stock (L1: 0-5, L2: 5-15, L3: 15-30, L4: 30-60, L5: 60-100)	g kg^{-1}	Canada (9.98 million)	39,366	STEP-AWB (*n* = 25 best model)	- Pred.: < 5 g kg^{-1}	-	- Pred.: > 300 g kg^{-1}	RF	5-fold cross-val.	0.72	79.8	-	-	Sothe *et al.*[76] (2022)
SOC bare topsoil	%	Europe (10.18 million)	7142	S-B, spectral data	0	1.68	43.84	BoRT	Val.	0.24	1.52	-	-	Safanelli *et al.*[117] (2020)
SOC stock	kg m^{-2}	Latin America	11,268	STEP-AWB	0	6.85	573.76	PLSR KK QRF SVM	Val. 5-fold cross-val.	Country specific r and RMSE are reported in form of graphs beyond the space of this table				Guevara *et al.*[79] (2018)
SOC content (L1: 0-5, L2: 5-15, L3: 15-30, L4: 30-60, L5: 60-100, L6: 100-200)	0/00 (g kg^{-1}) ----- SOC pred. on 250 m × 250 m global grid	Globe (510 million)	150,000 soil profiles	STEP-AWB (covariates incl. 158 remote-sensing derived properties)	-	-	-	Ensemble of 3 models (ANN, RF, BoRT)	10-fold cross-val.	0.64	32.8	-	-	Hengl *et al.*[74] (2017)
SOC content (L1: 0-5, L2: 5-15, L3: 15-30, L4: 30-60,	g kg^{-1} SOC pred. on 250 ×	Globe (510 million)	WoSIS *n* = 196,498 profiles And	STEP-AWB (400 covariates)				QRF	cross-val.		39.48			Poggio *et al.*[80] (2021)[5]

L5: 60-100, L6: 100-200)	250 (SoilGrids 2.0)	EU-LUCAS + Australia *n* = 240,000

[1]AI methods: ANN: Artificial neural network - multi layer perceptron; BaRT: bagged regression trees; BoRT: boosted regression trees; CaRT: classification and regression trees; CNN: convolutional neural networks; Cub: cubist; Cub-RK: regression kriging with cubist; GA: genetic algorithm (feature selection); KK: kernel-weighted nearest neighbors; PLSR: partial least square regression; PLSRmod: modified PLSR (with a two-step regression technique, 2Step-R that models categorical and continuous input data combining linear regressions (ridge regression - RR) and latent variable model (PLSR); RK/RT: regression kriging with regression trees; RF: random forest; SBIFmod: sparse Bayesian infinite factor regression modified with a two-step regression technique, 2Step-R that models categorical and continuous input data combining linear regressions (Bayesian linear regression) and latent variable model (sparse Bayesian infinite factor - SBIF); SVM: support vector machine; SVR: support vector regression; QRF: quantile random forest. [2]R^2: Coefficient of determination; CCC: Lin's concordance correlation coefficient[63]; RMSE: root mean square error; RPD: ratio of performance deviation[64]; RPIQ: ratio of performance to inter-quartile range[62]. [3]This study also modeled soil total (TC), recalcitrant (RC), moderately-available (MC), and hot-water extractable carbon (HC). [4]This study also modeled soil total carbon (TC), recalcitrant carbon (RC), and labile (hot-water extractable) carbon (HC). [5]This study also reported prediction interval coverage probability for soil organic carbon (SOC) for the six modeled soil layers.

No specific AI method stood out among reported soil carbon studies as superior. Random Forest was one prominent ML method applied in numerous SOC assessments[37,66,68,69,71,72,74-77] though the example studies assembled in Table 1 are not exhaustive. Deep learning algorithms are rarely used in soil carbon modeling (e.g., SOC assessment in Chile by Padarian *et al.*[30], 2019). More recently, the AI method QRF has gained interest to model SOC due to its ability to assess confidence intervals of estimates. For example, QRF was employed to model SOC in Argentina[78], Latin America[79], United States[75], and globally[80]. In studies that compared SOC models derived from multiple AI methods differences among AI methods were rather subtle[69].

Noteworthy, error metrics and uncertainty assessments were rarely provided by previous regionalized and global carbon assessments demonstrating the power of AI modeling. For example, the global potential of SOC sequestration through the adoption of conversation management practices and restorative land use was estimated at 0.9 ± 0.3 Pg C yr^{-1}, which was considered to offset one-fourth to one-third of the annual increase in atmospheric CO_2 estimated at 3.3 Pg C yr^{-1}[81]. Global SOC storage was assessed using the Harmonized World Soil Database by Köchy *et al.*[82] (2014) and a transfer function approach was used to map global soil carbon stock by Minasny *et al.*[19] (2017). The global soil carbon map (GSOCMap) on a 1 km × 1 km grid covering the topsoil (0-30 cm) by the FAO Global Soil Partnership is a joint effort of nations around the globe. The global carbon budget provided by Le Quéré *et al.*[83] (2015) used a budgeting approach to assess different carbon pools and fluxes whereby soil carbon was lumped into the category residual terrestrial carbon sink due to limited reliable data. Noteworthy, there was limited use of AI in these global SOC assessments. Recently, the achievable SOC sequestration in croplands and grasslands around the globe was estimated by Batjes[84] (2019) with two different approaches. The first one based on literature estimates of SOC gains by bioclimatic zones (M1) and the other assumed an annual C increase of 3 to 5 promille with respect to current SOC mass. According to M1, achievable gains ranged from 0.05-0.12 Pg C yr^{-1} to 0.14-0.37 Pg C yr^{-1}, with a technological potential of 0.32-0.86 Pg C yr^{-1}, while for M2 gains were 0.07-0.12 Pg C yr^{-1}, 0.21-0.35 Pg C yr^{-1}, and 0.60-1.01 Pg C yr^{-1} based on four different management scenarios. The provision of soil carbon values and/or SOC maps without explicit error and uncertainty analysis leaves major ambiguities due to lack in confidence in reported soil carbon values. AI and rigorous uncertainty assessment avoids such pitfalls.

AI-based soil respiration modeling

Soil respiration (R_s) provides one of the largest global fluxes of carbon dioxide (CO_2) to the atmosphere. Global R_s indicates the level of microbial activity and plays a major role in the global carbon cycle. It was conceptualized that rising global temperatures are expected to lead to substantial higher decomposition rates of soil carbon, and thus, CO_2 release from soils. However, despite its importance, the response of soil carbon to warming is still one of the great uncertainties in global carbon cycling[85,86]. Some studies found that R_s is mainly controlled by a range of biotic and abiotic factors, specifically temperature and other climatic factors[87-89], while other studies found that temperature is not the primary driver for the response of R_s to global warming. For example, Haaf *et al.*[85] (2021) found that global R_s is mainly controlled by interacting soil properties and secondarily by vegetation traits and plant growth conditions. Haaf *et al.*[85] (2021) pointed out that mechanistic controls of microbial soil R_s in response to global climate warming are well understood at the experimental laboratory and plot scale; however, soil properties are "hidden" from remote sensing and challenging to be mapped accurately at a spatial scale at which microbial soil properties and associated ecosystem processes vary in nature. In a global AI analysis, Huang *et al.*[90] (2020) found that land cover change, not climatic factors, played the most important role in regulating R_s changes specifically in temperate and boreal regions. AI modeling of site-specific R_s data coupled to gridded environmental datasets has afforded to discern the effects of climatic, biotic, edaphic, and other variables on R_s, heterotrophic respiration (R_h), and autotrophic respiration (R_a).

One of the first global soil respiration studies found that R_s increased by 0.1 Pg C yr^{-1} (1989 to 2008) with global R_s integrated over the Earth's surface amounting to 98 ± 12 Pg C implying a global R_s response to air temperature (Q_{10}) of 1.5[91]. Similar quantifications found that global R_s rates derived from flux measurements responded to the increase in air temperature at the rate of 3.3 Pg C yr^{-1} $°C^{-1}$, and Q_{10} of 1.4 for the period 1965 to 2012[92] and R_s of 94.3 ± 17.9 Pg C yr^{-1}[93]. These global R_s assessments used simple transfer regression functions approaches, while more current regional and global R_s assessments have incorporated AI. Recently, the AI algorithm RF was compared to ten different terrestrial ecosystem simulation models to compute global R_s with the former AI model outperforming all simulation models in a performance analysis using R_s measurements[89]. In this global study, the RF model showed excellent performance with R^2 of 0.89 for R_s and 0.86 for HR with 85.5 Pg C yr^{-1} for R_s and 50.3 Pg C yr^{-1} for R_h. The average global R_a (i.e., the difference between R_s and R_h) was 35.2 Pg C yr^{-1} for the RF model. In contrast, the estimated global R_s and R_h by the ten ecosystem models ranged from 61.4 to 91.7 Pg C yr^{-1} and 39.8 to 61.7 Pg C yr^{-1}, respectively, which indicates the wide variability in results derived from process-based simulation models. Findings suggest that mechanistic modeling of soil R_s metrics showed higher uncertainty than the AI model. Notably, the contribution of R_a to R_s highly varied among the ecosystem models (between 18% to 48%), which differed to the estimate computed by RF (41%)[89].

In another global study, Warner *et al.*[94] (2019) used plot-derived R_s measurements (n = 2657) and the AI-ML method QRF to make R_s predictions onto a 1 km × 1 km grid across the globe. Environmental predictor variables [mean annual temperature (MAT), mean annual precipitation (MAP), mean annual MODIS EVI, and mean precipitation from November through January] yielded a QRF prediction model with a global area-weighted mean annual R_s of 592.2 ± 368.9 g C m^{-2} yr^{-1} and a global sum of 87.9 Pg C yr^{-1}. The R^2, RMSE, and MAE were 0.63, 305.2 g C m^{-2} yr^{-1}, and 141.0 g C m^{-2} yr^{-1}, respectively. Recently, QRF was also used to model global R_s at a 1 km × 1 km spatial grid using large experimental datasets (small set n = 5173 and large set n = 10,366)[95]. In this study, the smaller dataset obtained a global R_s sum of 88.6 Pg C yr^{-1} (MAE = 29.9; Std. = 57.9 Pg C yr^{-1}), whereas the model with the larger R_s dataset yielded 96.5 Pg C yr^{-1} (MAE = 30.2; Std. = 73.4 Pg C yr^{-1}). The inclusion of new data from underrepresented regions (e.g., Asia, Africa, South America) to build the larger dataset resulted in overall higher model uncertainty. These are surprising findings

because commonly AI models tend to improve model performance when using larger datasets, though in some instances increasing the sample size may also increase data variability that may negatively affect model performance. The global R_h from the small dataset was 49.9-50.2 (mean 50.1) Pg C yr^{-1} and from the larger dataset it was 53.3-53.5 (mean 53.4) Pg C yr^{-1}. Other global R_s modeling involved the application of AI-DL methods (ANN) which computed a global average R_s of 93.3 ± 6.1 Pg C yr^{-1} from 1960 to 2012 and an increasing trend in average global annual R_s of 0.04 Pg C yr^{-1}. This global R_s model used climatic (MAP and MAP) and biome type as predictor variables resulting in an R^2 of 0.60.

The spatial and temporal variations in global R_s and their relationship with climate and land cover was assessed using a global dataset of R_s measurements (2000-2014), satellite data, and various AI algorithms (RF, SVR, and ANN) and a traditional method (multivariate nonlinear regression, MNLR). The selected models explained 62% to 84% of the interannual and intersite variabilities in annual R_s with an RMSE ranging from 107 to 413 g C m^{-2} yr^{-1}[90]. In the 10 different global biomes the MNLR model (R^2 between 0.20-0.55; RMSE between 140-519 g C m^{-2} yr^{-1}) was outperformed by all of the AI models estimating R_s. The model performance of the RF was best in 6 biomes (R^2 between 0.47-0.68; RMSE between 148-429 g C m^{-2} yr^{-1}), followed by SVM in 4 biomes (R^2 between 0.41-0.69; RMSE between 132-438 g C m^{-2} yr^{-1}). The ANN model estimating R_s showed moderate performance (R^2 between 0.35-0.62; RMSE between 158-446 g C m^{-2} yr^{-1}). Boreal, temperate, and tropical regions contributed 15%, 24%, and 61%, respectively, to the total mean annual global R_s. Land cover was the primary explanatory variable for global R_s. The areas with significant changes in short vegetation cover (i.e., all vegetation shorter than 5 m in height) showed more frequent changes in R_s than in areas with significant climate change.

A data-driven AI approach (RF) was also employed to assess the effects of climatic, edaphic and productivity on R_h with n = 455 at global scale[96]. In this study global R_h was 46.8 Pg C yr^{-2} (1985-2013) with a significant increasing trend of 0.03 Pg C yr^{-2}. In this study, water availability dominated R_h inter-annual variability. Water availability dominated in extra-tropical forest and semi-arid regions, while temperature strongly controlled R_h in tropical forests.

There are numerous factors that enabled the shift toward AI-R_s (and R_h and R_a) global modeling. First, the assembly of large databases that harmonized thousands of soil R_s plot-scale data enabling global AI modeling[88,97]. Although the presented R_s-AI studies were derived at global scale, the same AI approaches are also applicable to investigate R_s at regional scales. Site-specific R_s data coupled to geospatial environmental grids have allowed to go beyond descriptive assessment of global R_s change. AI models facilitated to upscale R_s onto a global grid with commonly used spatial resolutions of 1 km × 1 km. Global AI R_s models outperformed more traditional methods (multivariate regression), though there was still a substantial portion of unexplained variability in models. This points to data limitations rather than AI modeling limitations given the expansive cloud computing and supercomputer capabilities. One major data limitation to all global studies is the unbalanced distribution of soil R_s measurement sites around the globe which are concentrated in North America and Europe, but sparser in other regions. Jian *et al.*[98] (2018) cautioned that recent global R_s models showed a wide range from 68 to 98 Pg C yr^{-1}, which suggests considerable uncertainty impacting global carbon accounting. In Jian *et al.*[98]'s (2018) study a sensitivity analysis using RF was performed that varied timescales (daily, monthly, and annual) of R_s and climate data to predict global R_s which ranged from 66.62-100.72 Pg (1961-2014). Using monthly R_s data rather than annual data decreased global R_s by 7.43-9.46 Pg. In contrast, global R_s calculated from daily R_s data was only 1.83 Pg lower than the R_s from monthly data. Using mean annual precipitation and temperature data instead of monthly data caused +4.84 and -4.36 Pg C differences, respectively. These results suggest that temporal slicing of R_s and climatic data impact AI estimates of global R_s, and thus the global carbon budget.

THE POWER, POTENTIALS, AND PERILS OF AI-BASED SOIL CARBON MODELING

There is no doubt that AI modeling provides advanced capabilities to predict SOC stocks and R_s. Though these AI models are still data limited in explaining the spatial variability of soil carbon storage within landscapes. Specifically, SOC measurements used at large region and global scale are derived from legacy databases with a smaller amount of data that represent current field conditions. The temporal mismatch between up-to-date environmental covariates and legacy SOC data may be another limiting factor. Proximal soil sensing (VNIR and MIR spectroscopy) has been pivotal to counter these trends and estimate SOC and other soil properties rapidly, cost-effectively, densely, and accurately through the application of AI. For example, SOC was estimated by AI from proximal sensing data at global scale by Viscarra Rossel *et al.*[99] (2016), in Florida by Knox and Grunwald[100] (2018), in regions in India by Clingensmith *et al.*[101] (2019), in Brazil by Moura-Bueno *et al.*[102] (2021), and in China by Shi *et al.*[103] (2014). The incorporation of VNIR and MIR spectral data along with remote sensing data into AI models that upscale SOC storage to large regions is promising[67].

Soil respiration observations have been integrated into global open-access databases[87,97] to be shared and used by the scientific community. Regional R_s data can be spiked into these global databases which facilitates global research on AI-R_s. Soil respiration data represent carbon fluxes (i.e., temporal state of an ecosystem), while SOC storage infers on the spatially-explicit state of an ecosystem. AI models to predict SOC sequestration rates are still in its infancy due to data availability. One example, to model SCseq using CART in no-tillage systems compared to conventional tillage systems at global scale was provided by Sun *et al.*[16] (2020). The open-access approach of global R_s data repositories differs from SOC data. The latter are limited by the access to data with due to different purpose: (1) regional AI-SOC research projects with up-to-date data that represent field conditions; (2) some national SOC data that have restricted access while others are public (e.g., U.S.); and (3) global public SOC data - for example, the World Soil Information Service (WoSIS) database - which includes more legacy data than up-to-date SOC data. In summary, some limitations for AI soil carbon modeling are due to limited data sharing and currency of data. Though the contribution of community SOC data into larger open-access global databases rests on fair data sharing policies that acknowledges the labor and costs involved in field and laboratory operations. Investments to collect new soil samples analyzed for SOC (topsoil and subsoil), consistent SOC monitoring at benchmark sites around the globe, and boosting of R_s measurements would greatly benefit future AI soil carbon modeling.

Furthermore, ethical concerns entail the amplified focus and reliance on AI technologies and machine-generated model outputs that lack knowledge discovery and human interpretation of soil carbon dynamics across large and complex soil-ecosystems[33]. Wadoux *et al.*[104] (2020) presented results that compared RF models created with real SOC observations and one from pseudo ("false") variables for the same region. These AI models produced comparable results to predict SOC which raises major concerns about the possibility of AI generated "digital fake" versions of soil carbon storage. These concerns about AI models were echoed by Liao[105] (2020) who pointed out that AI methods are prone to erratic behavior of model outputs due to outliers or misclassified pixels and are sensitive to pseudo ("false") variables. As the collection of geospatial environmental datasets, specifically sensor-derived data, is steadily increasing the risk to incorporate spurious predictor variables into soil prediction models also increases[33]. Data-driven AI modeling of soil carbon dynamics is prone to identify relations between massive and diverse datasets of input variables (environmental covariates) and outputs (SOC stock, R_s, or others) that may statistically exist, but from a physical or biogeochemical knowledge perspective make less sense.

Protection against such pitfalls is the application feature selection processing methods pre-AI modeling. Commonly used pre-processing methods are Recursive Feature Elimination (RFE) analysis to select the best performing subset of covariates which was described by Guyon *et al.*[106] (2002). The RFE procedure starts with the maximum number of covariates and iteratively removes the weakest explanatory variable until a specified number of covariates is reached. Heuvelink *et al.*[78] (2021) used RFE to identify covariates to model SOC stocks in Argentina using the QRF AI algorithm and Poggio *et al.*[80] (2021) used RFE before running the QRF AI algorithm to model global SOC. The Boruta algorithm, often used in combination with RF, is another pre-processing method to strategically filter out the most important environmental covariates that relate most strongly to a target output[107]. Boruta was applied successfully to build parsimonious RF-AI SOC prediction models that substantially reduced large environmental covariate sets[37,69]. Xiong *et al.*[37] (2014) compared various pre-processing algorithms that discerned all-relevant variables (i.e., strong and weakly relevant variables selected with Boruta and RF), minimal-optimal variables (four optimization algorithms were tested: greedy forward, greedy backward, hill climbing, and simulated annealing), and irrelevant environmental covariates to model SOC stock using four different AI methods (BaRT, BoRT, RF, and Cubist) in Florida, USA. The initial environmental covariate set comprised 210 variables, while the best performing parsimonious model identified with the all-relevant and minimal-optimal feature selection comprised just four covariates to predict SOC stock[37]. This holistic AI environmental modeling framework was based on the consistent feature selections in ML approach developed earlier by Nilson *et al.*[108] (2007). The advantage of automated feature selection compared to expert-based selection of covariates is that human bias is reduced that may, even unintentionally, impact SOC modeling and upscaling of results to region or global scale.

Another powerful feature selection methods is sparse Least Absolute Shrinkage and Selection Operator (LASSO)[109]. For example, LASSO as well as Boruta feature selections were employed along with AI-ML and AI-DL algorithms to model SOC in two contrasting climatic regions[110]. Wang *et al.*[77] (2018) found that the genetic algorithm outperformed stepwise multivariate regression in strategically selecting predictor variables before applying RF and BoRT to model SOC stocks in rangeland in Eastern Australia. While the application of feature selections in AI-SOC modeling is prominent (see studies in Table 1), it seems relatively rare in global AI-R_s modeling studies. Interestingly, in global AI-R_s studies knowledge discovery is emphasized over technical information of AI modeling that often is only provided in small print in appendices and supplementary documents of publications. Instead, separate pre- or post-hoc analyses that complement AI modeling are commonly found in the global R_s literature[111].

The dichotomy between data-driven and knowledge-driven soil modeling was discussed in detail by Wadoux *et al.*[104] (2020). Authors cautioned about knowledge discovery purely from ML and pattern recognition processes. It was suggested that pedologically relevant environmental covariates should be selected through feature selection pre-processing. In similar vein, McBratney *et al.*[28] (2019) raised concerns around over-parameterization and the generation of nonsensical soil predictions from AI models. In my view, an important vision going forward with AI soil carbon modeling is to keep balance between (1) data-driven feature selection and ML and DL modeling; and (2) knowledge-driven approaches pre-AI modeling and post-interpretation. Scientific interpretation enhances legitimacy and confidence in AI-generated digital soil C output. Ideally, an integral scientific approach involves consultation of multiple other sources (environmental datasets, literature, expert-knowledge) and comparative analysis derived from other methods rooted in a modeling paradigm different from AI (e.g., process-based simulation modeling, geostatistics, hybrid stochastic-deterministic methods, Bayesian methods, structural equation modeling, participatory action research). Ensemble modeling to aggregate soil carbon output from various, ideally contrasting model paradigms among them AI, may lower the risks of spurious AI model output. Meta-

modeling, an approach rooted in integral ecology, was envisioned as a viable framework for integration of multiple models and data to address soil security issues, among them soil carbon sequestration[112]. An integrative vision for soil carbon assessments would avoid the inflated hype about AI in carbon sciences. Such integrative strategy lowers the risk to "blindly" belief machine-generated soil carbon assessments; even validation of AI-generated results cannot fully inoculate from potential spurious modeling results that may be replicated in training and validation modes. Honoring the diversity of modeling approaches that provide partial knowledge of soil carbon dynamics rather than idealizing AI as superior to all other methods will further enhance scientific understanding of complex soil-ecosystems and carbon dynamics. It may also help to form resilient liaisons and partnerships among AI specialists and soil and environmental scientists.

The shift from simple ML rooted in pattern recognition toward more complex DL models with multiple layers of nodes, processing strategies (e.g., convolution and pooling), and fitting strategies (e.g., latent factors or weights) makes these kinds of models (see Figures 2-5) more abstract losing more-and-more physical and pedological meaning. Theoretically, these fitting and learning strategies in ANN model variants if put to the extreme could achieve an ideal model fit (R^2 of 1), which has been approximated already in soil carbon modeling applications. For example, CNN and Cubist were used to model various soil properties, among them SOC, using VNIR and MIR spectral soil data using a large dataset (n = 14,594) from the U.S.[113]. In this study, the two-channel 1D CNN model was best performing with R^2 between 0.95 and 0.98 for six different soil properties, the R^2 was 0.98 for both SOC and soil total carbon, and the RPIQ was 2.27 (SOC) and 3.01 (soil total carbon). These results are "near-perfect" though one may wonder about the many hyperparameters and processing layers in the CNN for tuning to achieve such superb model performance [Figure 5]. The many hyperparameters, latent factors, and fitting weights in AI models make sense to the machine, but are less meaningful for interpretation by human users or carbon scientists to infer on carbon cycle processes, soil functions, or ecosystem services. What pedological or biophysical insights in regard to SOC or ecosystem processes were derived in Ng *et al.*[113] (2019) research study or similar AI-DL soil carbon models? The extraordinary capabilities of AI-DL algorithms to fit inputs and outputs have been hailed black-boxes and AI-ML algorithms gray boxes, respectively; in essence, AI models lack transparency[105]. Black-boxes or gray-boxes mean, for example, that SOC storage or the ecosystem process of SCseq are encoded in AI-ANN models in form of multiple nodes, layers, and weighing factors replacing human understanding and striving for meaning-making about the soil-environment into machine code.

From a philosophical perspective, the question is whether the physical environment or a simulated, virtual environment (digital worlds) is more real to us. Chalmers[114] (2022) in his book *Reality+* discerned between virtual simulated worlds and those we perceive as real and suggested that we can live meaningful lives in virtual reality. Applying Chalmer's vision to carbon science this suggests that we would be able to live meaningful lives in machine generated worlds in which a soil carbon map or model is as real and satisfying as a soil in nature. Interestingly, in such a virtual/AI soil carbon world human knowledge and understanding of soil-environmental relations, mechanisms, ecosystem processes, carbon fluxes and cycling, and global climate change become irrelevant.

Given the rapid expansion of AI into carbon science, as well as many other sciences, poses urgency to think about ethical implications implicated in AI[105,115]. Reality+ confronts us with the question what is "real and meaningful to us" - an observable soil in nature that we can touch, sense, and use (phenomenology), laboratory measurements of the soil carbon content (empiricism), proximal or remote sensors and AI providing inference on soil carbon, a digital image/map of soil and its carbon storage computed by AI (representation), or simulated worlds (simulacra) created with advanced AI and visualization techniques. These simulacra replace "environmental reality" with its representation according to Baudrillard[116] (1994).

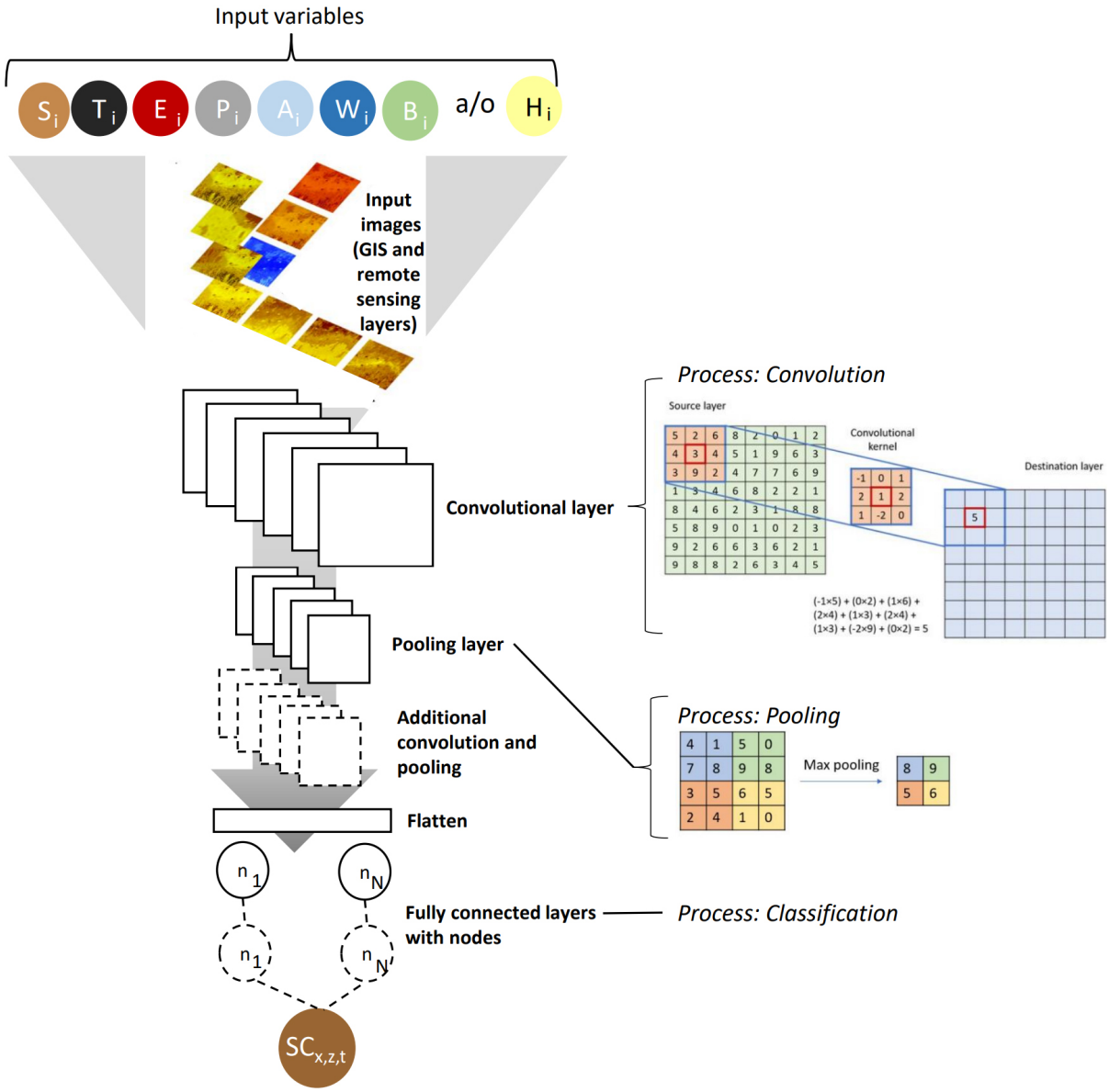

Figure 5. AI model predicting soil carbon (SC) from environmental covariates. Idealized model representation of a convolutional neural network (CNN).

He philosophized that society had become saturated with simulacra and that people live in a hyperreality in which meaninglessness prevails. While Baudrillard[116]'s (1994) vision was somewhat dystopian, Chalmer[114]'s (2022) Reality+ looks more optimistic. Whether the expansion of AI-DL and AI-ML soil carbon modeling is perceived as frustrating and frightening because only the machine knows leaving one confused, helpless, and meaningless or whether we get excited and enchanted by the beauty of machine-generated soil carbon data, maps, and models that we trust will have a profound impact on carbon science and how it is applied in carbon policies, carbon crediting, and carbon management. The risks involved are that AI-generated soil carbon hyperreality is prone to human manipulation (i.e., how the model is tuned and fitted) and misuse.

One key question is whether we are applying AI in data-driven or knowledge-driven ways to advance soil carbon science. The potentials and perils of AI in carbon science need to be carefully weighted to avoid pitfalls, and perhaps compute (surprising), spurious digital soil carbon predictions. The power of AI is the possibility of more accurate and precise soil carbon models that only machine algorithms can create. In the latter lies the profound potential of AI-ML and AI-DL to transform carbon science and modeling. Enhanced dialogue and awareness of AI model limitations may help to better understand soil carbon evolution and its ecosystem processes.

DECLARATIONS

Acknowledgments

I acknowledge the University of Florida which provided resources to work on this article.

Authors' contributions

The author contributed solely to the article.

Availability of data and materials

Not applicable.

Financial support and sponsorship

None.

Conflicts of interest

The author declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2022.

REFERENCES

1. Janzen H. Carbon cycling in earth systems—a soil science perspective. *Agric Ecosyst Environ* 2004;104:399-417. DOI
2. Scharlemann JP, Tanner EV, Hiederer R, Kapos V. Global soil carbon: understanding and managing the largest terrestrial carbon pool. *Carbon Manag* 2014;5:81-91. DOI
3. Shi Z, Allison SD, He Y, et al. The age distribution of global soil carbon inferred from radiocarbon measurements. *Nat Geosci* 2020;13:555-9. DOI
4. Grunwald S, Thompson JA, Boettinger JL. Digital soil mapping and modeling at continental scales: finding solutions for global issues. *Soil Sci Soc Am J* 2011;75:1201-13. DOI
5. James J, Devine W, Harrison R, Terry T. Deep soil carbon: quantification and modeling in subsurface layers. *Soil Sci Soc Am J* 2014;78:S1-S10. DOI
6. Koarashi J, Hockaday WC, Masiello CA, Trumbore SE. Dynamics of decadally cycling carbon in subsurface soils. *J Geophys Res* 2012;117:G03033. DOI
7. Vasques G, Grunwald S, Comerford N, Sickman J. Regional modelling of soil carbon at multiple depths within a subtropical watershed. *Geoderma* 2010;156:326-36. DOI
8. Rogelj J, den Elzen M, Höhne N, et al. Paris agreement climate proposals need a boost to keep warming well below 2 °C. *Nature* 2016;534:631-9. DOI PubMed
9. Paustian K, Lehmann J, Ogle S, Reay D, Robertson GP, Smith P. Climate-smart soils. *Nature* 2016;532:49-57. DOI PubMed
10. Rumpel C, Amiraslani F, Chenu C, et al. The 4p1000 initiative: opportunities, limitations and challenges for implementing soil

organic carbon sequestration as a sustainable development strategy. *Ambio* 2020;49:350-60. DOI PubMed PMC
11. Smith P, Clark H, Dong H. Agriculture, forestry and other land use (AFOLU). In: Krug T, Nabuurs GJ, editors. Climate change 2014: mitigation of climate change. Cambridge: Cambridge University Press; 2014. p. 813-922.
12. Six J, Ogle SM, Jay breidt F, Conant RT, Mosier AR, Paustian K. The potential to mitigate global warming with no-tillage management is only realized when practised in the long term. *Glob Chang Biol* 2004;10:155-60. DOI
13. Sun W, Canadell JG, Yu L, et al. Climate drives global soil carbon sequestration and crop yield changes under conservation agriculture. *Glob Chang Biol* 2020;26:3325-35. DOI PubMed
14. Lal R. Soil carbon stocks under present and future climate with specific reference to European ecoregions. *Nutr Cycl Agroecosyst* 2008;81:113-27. DOI
15. Smith P, Martino D, Cai Z, et al. Greenhouse gas mitigation in agriculture. *Philos Trans R Soc Lond B Biol Sci* 2008;363:789-813. DOI PubMed PMC
16. Stockmann U, Adams MA, Crawford JW, et al. The knowns, known unknowns and unknowns of sequestration of soil organic carbon. *Agric Ecosyst Environ* 2013;164:80-99. DOI
17. Ogle SM, Alsaker C, Baldock J, et al. Climate and soil characteristics determine where no-till management can store carbon in soils and mitigate greenhouse gas emissions. *Sci Rep* 2019;9:11665. DOI PubMed PMC
18. Nicoloso RS, Rice CW. Intensification of no-till agricultural systems: an opportunity for carbon sequestration. *Soil Sci Soc Am j* 2021;85:1395-409. DOI
19. Minasny B, Malone BP, Mcbratney AB, et al. Soil carbon 4 per mille. *Geoderma* 2017;292:59-86. DOI
20. White RE, Davidson B, Lam SK, Chen D. A critique of the paper 'soil carbon 4 per mille' by Minasny et al. (2017). *Geoderma* 2018;309:115-7. DOI
21. de Vries W. Soil carbon 4 per mille: a good initiative but let's manage not only the soil but also the expectations. *Geoderma* 2018;309:111-2. DOI
22. Poulton P, Johnston J, Macdonald A, White R, Powlson D. Major limitations to achieving "4 per 1000" increases in soil organic carbon stock in temperate regions: evidence from long-term experiments at Rothamsted Research, United Kingdom. *Glob Chang Biol* 2018;24:2563-84. DOI PubMed PMC
23. van Groenigen JM, van Kessel C, Hungate BA, Oenema O, Powlson DS, van Groenigen KJ. Sequestering soil organic carbon: a nitrogen dilemma. *Environ Sci Technol* 2017;51:4738-9. DOI PubMed
24. Baveye PC, Berthelin J, Tessier D, Lemaire G. The "4 per 1000" initiative: a credibility issue for the soil science community? *Geoderma* 2018;309:118-23. DOI
25. Russell S, Norvig P. Artificial intelligence: a modern approach. Available from: https://storage.googleapis.com/pub-tools-public-publication-data/pdf/27702.pdf [Last accessed on 12 Apr 2022].
26. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436-44. DOI PubMed
27. LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks. Cambridge: MIT Press; 1995.
28. Mcbratney A, de Gruijter J, Bryce A. Pedometrics timeline. *Geoderma* 2019;338:568-75. DOI
29. Khaledian Y, Miller BA. Selecting appropriate machine learning methods for digital soil mapping. *Appl Math Model* 2020;81:401-18. DOI
30. Padarian J, Minasny B, Mcbratney AB. Using deep learning for digital soil mapping. *Soil* 2019;5:79-89. DOI PubMed PMC
31. Kamilaris A, Prenafeta-boldú FX. Deep learning in agriculture: a survey. *Comput Electron Agric* 2018;147:70-90. DOI
32. Diaz-Gonzalez FA, Vuelvas J, Correa CA, Vallejo VE, Patino D. Machine learning and remote sensing techniques applied to estimate soil indicators - review. *Ecol Indic* 2022;135:108517. DOI
33. Grunwald S. Grand challenges in pedometrics-AI research. *Front Soil Sci* 2021;1:714323. DOI
34. Ma Y, Minasny B, Malone BP, Mcbratney AB. Pedology and digital soil mapping (DSM). *Eur J Soil Sci* 2018;70:216-35. DOI
35. Thompson JA, Roecker SM, Grunwald S, Owens PR. Digital soil mapping: Interactions with and applications for hydropedology. In: Lin HS, Editor. Hydropedology - synergistic integration of Pedology and Hydrology. Cambridge: Academic Press; 2012. p. 665-709. DOI
36. Mcbratney A, Mendonça Santos M, Minasny B. On digital soil mapping. *Geoderma* 2003;117:3-52. DOI
37. Xiong X, Grunwald S, Myers DB, Kim J, Harris WG, Comerford NB. Holistic environmental soil-landscape modeling of soil organic carbon. *Environ Model Softw* 2014;57:202-15. DOI
38. Breiman L. Classification and regression trees. London: Chapman & Hall; 1984.
39. Prasad AM, Iverson LR, Liaw A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* 2006;9:181-99. DOI
40. Breiman L. Bagging predictors. *Mach Learn* 1996:24;123-40. DOI
41. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York: Springer; 2009.
42. Efron B, Tibshirani R. An introduction to the bootstrap: monographs on statistics and applied probability. London: Chapman & Hall; 1993.
43. Freund Y, Schapire RE. Experiments with a new boosting algorithm. 13th International Conference on Machine Learning; 1996 Jan 22; Murray Hill. 1996. p. 148-56.
44. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001;29:1189-232. DOI

45. Lawrence R. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sens Environ* 2004;90:331-6. DOI
46. Breiman L. Random forests. *Mach Learn* 2001;45:5-32. DOI
47. Cutler A, Cutler DR, Stevens JR, Zhang C, Ma Y. Ensemble machine learning: methods and applications. Chicago: Media LLC; 2012.
48. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* 2003;43:1947-58. DOI PubMed
49. Meinshausen N, Ridgeway G. Quantile regression forests. *J Mach Learn Res* 2006;7:983-99.
50. Karatzoglou A, Meyer D, Hornik K. Support vector machines in R. *J Stat Soft* 2006;15:1-28. DOI
51. Smola AJ, Schölkopf B. A tutorial on support vector regression. *Stat Comput* 2004;14:199-222. DOI
52. Williams G. Data mining with rattle and R: the art of excavating data for knowledge discovery. New York: Springer; 2011.
53. Somarathna P, Malone B, Minasny B. Mapping soil organic carbon content over New South Wales, Australia using local regression kriging. *Geoderma Reg* 2016;7:38-48. DOI
54. Wold H. Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. *J Appl Probab* 1975;12:117-42. DOI
55. Garthwaite PH. An interpretation of partial least squares. *J Am Stat Assoc* 1994;89:122-7.
56. Carrascal LM, Galván I, Gordo O. Partial least squares regression as an alternative to current regression methods used in ecology. *Oikos* 2009;118:681-90. DOI
57. Tobias RD. An introduction to partial least squares regression. Proceedings of the Twentieth Annual SAS Users Group International Conference; Cary: SAS Institute Inc; 1995. p. 1250-7.
58. Quinlan JR. Programs for machine learning. Burlington: Morgan Kaufmannn; 1993.
59. Haykin S. Neural networks: a comprehensive foundation. Hoboken: Prentice Hall; 1999.
60. Hecht-Nielsen R. Theory of the Backpropagation Neural Network*. In: Wechsler H, editor. Neural Networks for Perception. Elsevier; 1992. p. 65-93. DOI
61. Wythoff BJ. Backpropagation neural networks: a tutorial. *Chemometr Intell Lab Syst* 1993;18:115-55. DOI
62. Bellon-Maurel V, Fernandez-Ahumada E, Palagos B, Roger J, Mcbratney A. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *Trends Analyt Chem* 2010;29:1073-81. DOI
63. McBratney AB, Minasny B, Stockmann U. Pedometrics. 1st ed. New York: Springer; 2018. DOI
64. Williams PC. Variables affecting near-infrared reflectance spectroscopic analysis. Near-infrared technology in the agriculture and food industries. Eagan: American Association of Cereal Chemists; 1987. p. 143-67.
65. Chang C, Laird DA, Mausbach MJ, Hurburgh CR. Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties. *Soil Sci Soc Am J* 2001;65:480-90. DOI
66. Adi SH, Grunwald S. Integrative environmental modeling of soil carbon fractions based on a new latent variable model approach. *Sci Total Environ* 2020;711:134566. DOI PubMed
67. Peng Y, Xiong X, Adhikari K, Knadel M, Grunwald S, Greve MH. Modeling soil organic carbon at regional scale by combining multi-spectral images with laboratory spectra. *PLoS One* 2015;10:e0142295. DOI PubMed PMC
68. Ross CW, Grunwald S, Vogel JG, et al. Accounting for two-billion tons of stabilized soil carbon. *Sci Total Environ* 2020;703:134615. DOI PubMed
69. Keskin H, Grunwald S, Harris WG. Digital mapping of soil carbon fractions with machine learning. *Geoderma* 2019;339:40-58. DOI
70. Martin MP, Wattenbach M, Smith P, et al. Spatial distribution of soil organic carbon stocks in France. *Biogeosciences* 2011;8:1053-65. DOI
71. Were K, Bui DT, Dick ØB, Singh BR. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecol Indic* 2015;52:394-403. DOI
72. Deng X, Chen X, Ma W, et al. Baseline map of organic carbon stock in farmland topsoil in East China. *Agric Ecosyst Environ* 2018;254:213-23. DOI
73. Viscarra Rossel RA, Chen C, Grundy MJ, Searle R, Clifford D, Campbell PH, et al. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. *Soil Res* 2015;53:845. DOI
74. Hengl T, Mendes de Jesus J, Heuvelink GB, et al. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One* 2017;12:e0169748. DOI PubMed PMC
75. Cao B, Domke GM, Russell MB, Walters BF. Spatial modeling of litter and soil carbon stocks on forest land in the conterminous United States. *Sci Total Environ* 2019;654:94-106. DOI PubMed
76. Sothe C, Gonsamo A, Arabian J, Snider J. Large scale mapping of soil organic carbon concentration with 3D machine learning and satellite observations. *Geoderma* 2022;405:115402. DOI
77. Wang B, Waters C, Orgill S, et al. Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. *Ecol Indic* 2018;88:425-38. DOI
78. Heuvelink GBM, Angelini ME, Poggio L, et al. Machine learning in space and time for modelling soil organic carbon change. *Eur J Soil Sci* 2021;72:1607-23. DOI
79. Guevara M, Olmedo GF, Stell E, et al. No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America. *Soil* 2018;4:173-93. DOI

80. Poggio L, de Sousa LM, Batjes NH, et al. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *Soil* 2021;7:217-40. DOI
81. Lal R. Soil carbon sequestration to mitigate climate change. *Geoderma* 2004;123:1-22. DOI PubMed
82. Köchy M, Don A, van der Molen MK, Freibauer A. Global distribution of soil organic carbon, based on the Harmonized World Soil Database-Part 2: Certainty of changes related to land-use and climate. *Soil Discussions* 2014;1:363-400. DOI
83. Le Quéré C, Moriarty R, Andrew RM, et al. Global carbon budget 2014. *Earth Syst Sci Data* 2015;7:47-85. DOI
84. Batjes NH. Technologically achievable soil organic carbon sequestration in world croplands and grasslands. *Land Degrad Dev* 2019;30:25-32. DOI
85. Haaf D, Six J, Doetterl S. Global patterns of geo-ecological controls on the response of soil respiration to warming. *Nat Clim Chang* 2021;11:623-7. DOI
86. Kirschbaum M. The temperature dependence of organic-matter decomposition—still a topic of debate. *Soil Biol Biochem* 2006;38:2510-8. DOI
87. Bond-Lamberty B, Thomson A. A global database of soil respiration data. *Biogeosciences* 2010;7:1915-26. DOI
88. Bond-Lamberty B. New techniques and data for understanding the global soil respiration flux. *Earth's Future* 2018;6:1176-80. DOI
89. Lu H, Li S, Ma M, et al. Comparing machine learning-derived global estimates of soil respiration and its components with those from terrestrial ecosystem models. *Environ Res Lett* 2021;16:054048. DOI
90. Huang N, Wang L, Song XP, et al. Spatial and temporal variations in global soil respiration and their relationships with climate and land cover. *Sci Adv* 2020;6:eabb8508. DOI PubMed PMC
91. Bond-Lamberty B, Thomson A. Temperature-associated increases in the global soil respiration record. *Nature* 2010;464:579-82. DOI PubMed
92. Hashimoto S, Carvalhais N, Ito A, Migliavacca M, Nishina K, Reichstein M. Global spatiotemporal distribution of soil respiration modeled using a global database. *Biogeosciences* 2015;12:4121-32. DOI
93. Xu M, Shang H. Contribution of soil respiration to the global carbon equation. *J Plant Physiol* 2016;203:16-28. DOI PubMed
94. Warner DL, Bond-lamberty B, Jian J, Stell E, Vargas R. Spatial predictions and associated uncertainty of annual soil respiration at the global scale. *Global Biogeochem Cycles* 2019;33:1733-45. DOI
95. Stell E, Warner D, Jian J, Bond-Lamberty B, Vargas R. Spatial biases of information influence global estimates of soil respiration: how can we improve global predictions? *Glob Chang Biol* 2021;27:3923-38. DOI PubMed
96. Yao Y, Ciais P, Viovy N, et al. A data-driven global soil heterotrophic respiration dataset and the drivers of its inter-annual variability. *Global Biogeochem Cycles* 2021:35. DOI
97. Jian J, Vargas R, Anderson-teixeira K, et al. A restructured and updated global soil respiration database (SRDB-V5). *Earth Syst Sci Data* 2021;13:255-67. DOI
98. Jian J, Steele MK, Thomas RQ, Day SD, Hodges SC. Constraining estimates of global soil respiration by quantifying sources of variability. *Glob Chang Biol* 2018;24:4143-59. DOI PubMed
99. Viscarra Rossel R, Behrens T, Ben-dor E, et al. A global spectral library to characterize the world's soil. *Earth Sci Rev* 2016;155:198-230. DOI
100. Knox NM, Grunwald S. Total soil carbon assessment: linking field, lab, and landscape through VNIR modelling. *Landscape Ecol* 2018;33:2137-52. DOI
101. Clingensmith CM, Grunwald S, Wani SP. Evaluation of calibration subsetting and new chemometric methods on the spectral prediction of key soil properties in a data-limited environment. *Eur J Soil Sci* 2019;70:107-26. DOI
102. Moura-Bueno JM, Dalmolin RSD, Horst-Heinen TZ, Grunwald S, ten Caten A. Environmental covariates improve the spectral predictions of organic carbon in subtropical soils in southern Brazil. *Geoderma* 2021;393:114981. DOI
103. Shi Z, Wang Q, Peng J, et al. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. *Sci China Earth Sci* 2014;57:1671-80. DOI
104. Wadoux AMJ, Samuel-Rosa A, Poggio L, Mulder VL. A note on knowledge discovery and machine learning in digital soil mapping. *Eur J Soil Sci* 2020;71:133-6. DOI
105. Liao SM. Ethics of artificial intelligence. Oxford: Oxford University Press; 2020.
106. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn* 2002;46:389-422. DOI
107. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Soft* 2010;36:1-13. DOI
108. Nilson R, Pena JM, Björkegren J, Tegnér J. Consistent feature selection for pattern recognition in polynomial time. *J Mach Learn Res* 2007;8:589-612. DOI
109. Xie Z, Xu Y. Sparse group LASSO based uncertain feature selection. *Int J Mach Learn & Cyber* 2014;5:201-10. DOI
110. Taghizadeh-Mehrjardi R, Schmidt K, Amirian-Chakan A, et al. Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by stacking machine learning models and rescanning covariate space. *Remote Sens* 2020;12:1095. DOI
111. Zhao Z, Peng C, Yang Q, et al. Model prediction of biome-specific global soil respiration from 1960 to 2012. *Earth's Future* 2017;5:715-29. DOI
112. Grunwald S, Mizuta K, Ceddia MB, et al. The meta soil model: an integrative multi-model framework for soil security. In: Field DJ, Morgan CLS, McBratney AB, editors. Global soil security. London: Springer Nature Publ.; 2017. p. 305-18. DOI
113. Ng W, Minasny B, Montazerolghaem M, et al. Convolutional neural network for simultaneous prediction of several soil properties

using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma* 2019;352:251-67. DOI
114. Chalmers DJ. Reality+: virtual worlds and the problem of philosophy. New York: W.W. Norton & Company; 2022.
115. Mitchell M. Artificial intelligence: a guide for thinking humans. New York: Farrar, Straus and Giroux; 2019.
116. Baudrillard J. Simulacra and simulation. Ann Arbor: University of Michigan Press; 1994.
117. Safanelli JL, Chabrillat S, Ben-Dor E, Demattê JAM. Multispectral models from bare soil composites for mapping topsoil properties over Europe. *Remote Sensing* 2020;12:1369. DOI