Title

Soil Science-Informed Machine Learning

Authors

Minasny, Budiman

Bandai, Toshiyuki

Ghezzehei, Teamrat A

et al.

Publication Date

2024-12-01

DOI

10.1016/j.geoderma.2024.117094

Peer reviewed

Soil Science-Informed Machine Learning

Budiman Minasny[1], Toshiyuki Bandai[2], Teamrat A. Ghezzehei[3]; Yin-Chung Huang[1], Yuxin Ma[4], Alex. B. McBratney[1], Wartini Ng[1], Sarem Norouzi[5], Jose Padarian[1], Rudiyanto[6], Amin Sharififar[1], Quentin Styc[1], Marliana Widyastuti[1]

[1] School of Life and Environmental Sciences, The University of Sydney, NSW 2006, Australia.

[2] Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

[3] Life & Environmental Sciences Department, University of California, Merced, CA 95343, USA.

[4] New South Wales Department of Climate Change, Energy, the Environment and Water, Parramatta, NSW 2150, Australia.

[5] Department of Agroecology, Aarhus University, 8830 Tjele, Denmark.

[6] Faculty of Fisheries and Food Science, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia.

Abstract

Machine learning (ML) applications in soil science have significantly increased over the past two decades, reflecting a growing trend towards data-driven research addressing soil security. This extensive application has mainly focused on enhancing predictions of soil properties, particularly soil organic carbon, and improving the accuracy of digital soil mapping (DSM). Despite these advancements, the application of ML in soil science faces challenges related to data scarcity and the interpretability of ML models. There is a need for a shift towards Soil Science-Informed ML (SoilML) models that use the power of ML but also incorporate soil science knowledge in the training process to make predictions more reliable and generalisable. This paper proposes methodologies for embedding ML models with soil science knowledge to overcome current limitations. Incorporating soil science knowledge into ML models involves using observational priors to enhance training datasets, designing model structures which reflect soil science principles, and supervising model training with soil science-informed loss functions. The informed loss functions include observational constraints, coherency rules such as regularisation to avoid overfitting, and prior or soil-knowledge constraints that incorporate existing information about the parameters or outputs. By way of illustration, we present examples from four fields: digital soil mapping, soil spectroscopy, pedotransfer functions, and dynamic soil property models. We discuss the potential to integrate process-based models for improved prediction, the use of physics-informed neural networks, limitations, and the issue of overparametrisation. These approaches improve the relevance of ML predictions in soil science and enhance the models' ability to generalise across different scenarios while maintaining soil science principles, transparency and reliability.

1. Introduction

The 2024 Nobel Prize in Physics was awarded to researchers who utilised physics-based tools to develop methods that advance machine learning through artificial neural networks. Over the past two decades, the use of machine learning (ML) in soil research has surged. In 2023, an average of 8 papers per day were published on topics related to “machine learning” and “soil” (Scopus, June 2024). These advancements, highlight the growing importance of ML in various scientific fields, including soil research. (See Box 1 for the definition of ML.) Past reviews show that the application of ML in soil science spans various areas, including soil organic carbon (SOC), hydrology, contamination, remote sensing, erosion, ML methods and modelling, spectroscopy, and crops (Li et al., 2024; Padarian et al., 2020b). In particular, the application of ML is extensive in digital soil mapping (DSM) studies. The review by Wadoux et al. (2020) on ML in DSM indicated that most studies emphasise predicting soil properties (in particular SOC) and improving prediction accuracy. However, only a few studies account for existing soil knowledge in the modelling processes.

Undoubtedly, machine learning has revolutionised the processing of large soil databases, finding patterns which are difficult to uncover using traditional statistical models (Heung et al., 2016). Soil observational data, collected via field and laboratory techniques and numerous sensors, provide extensive datasets that conventional statistical models may not efficiently handle (Safanelli et al., 2021; Tziolas et al., 2020). ML models excel in discovering patterns within spatiotemporal soil data, which are often challenging for process-based models to address. In addition, ML facilitates the generation of detailed soil information, scaling from field-level observations to global insights (Helfenstein et al., 2024; Padarian et al., 2022b; Poggio et al., 2021; Rosin et al., 2023).

While ML can replicate observed patterns in training data, it often falls short of explaining observed phenomena, and the learned patterns are usually not generalisable. ML models require substantial volumes of data, yet soil data are limited and sparse. The efficacy of ML models is constrained by the quantity and quality of training data, hindering their ability to predict “unseen” phenomena. (Read Box 2 “Six Dangers of ML in Soil Science”.) There is an ongoing discussion on

incorporating soil knowledge in ML models and the interpretability of the calibrated ML models (Ma et al., 2019; Wadoux et al., 2020). There is growing interest in applying interpretable ML models to explain how models predict certain attributes, addressing the “black box” issue. Weindorf and Chakraborty (2024) argued for a balance between ML modeling with human insight and knowledge for contextualising findings, and ensuring the completeness, validity, and interpretation of AI-generated results. Increasingly, there is a call to incorporate fundamental domain knowledge and physical rules into ML models to enhance their reliability and accuracy by providing theoretical constraints and informative priors (von Rueden et al., 2023). Concurrently, there is a push to model soil biogeochemical processes using physical rules, foundational to numerous achievements in computational physics and chemistry (Tang et al., 2024). In physics, hydrology, and related fields, there is increasing interest in physics-informed machine learning, aiming to guide ML models towards solutions that are physically plausible (Karniadakis et al., 2021; Kashinath et al., 2021). Notably, physics-informed ML models have been investigated to model soil water movement (Bandai and Ghezzehei, 2022).

We propose Soil Science-Informed ML (SoilML), which integrates soil-specific knowledge—including pedology, physical, chemical, and biological, processes into ML models, expanding the scope beyond Physics-Informed ML (PIML). SoilML prioritises modelling soil systems, accounting for interactions such as water cycling, soil-water-plant dynamics, and biogeochemical transformations. This approach aims to enhance model interpretability, improve predictions in data-scarce environments, and ensure that outputs are consistent with real-world soil behavior.

The paper is structured as follows. Section 2 demonstrates ways for incorporating soil science knowledge or principles in ML models, moving beyond merely identifying important predictors. Subsequently in Section 3, we present specific examples from four key fields: pedotransfer functions, digital soil mapping, soil spectroscopy, and modelling soil properties in space and time. We discuss the capabilities and limitations of conventional approaches and explore the potential to integrate soil science knowledge under the SoilML framework, followed by implications (Kashinath et al., 2021). Section 4 provides a discussion of SoilML models to address the unique 3D structure and interactions in soil systems, the

issue of overfitting, enhancing model interpretability, reliability, and predictive accuracy for soil science applications.

Box 1. Definitions

Artificial Intelligence (AI) is the field of study focused on designing and developing intelligent machines capable of performing tasks which mimic human intelligence.

Machine Learning (ML) is a subset of AI that uses algorithms to perform specific tasks without explicit instructions. The models learn and make predictions based on patterns and inferences derived from data, focusing on prediction accuracy. In this context, we do not consider statistical models such as linear regression and partial least squares regression as ML as they rely on predefined functional assumptions.

Deep learning is a subset of ML that involves a type of algorithm called artificial neural networks. These neural networks are designed to recognise patterns in data by processing data through multiple layers of processing units.

Physical rules are fundamental principles that describe how physical systems behave and interact in the natural world based on scientific observations, experiments, theories, and mathematical models.

Mechanistic or process-based models are mathematical models that describe one or multiple processes based on the underlying mechanisms and interactions

among system components. Based on the principles of physics, chemistry, biology, and related sciences, combined with empirical relationships, these models aim to represent how different components of a system work together to produce observed behaviours.

118

119

120

121

Box 2. Six Dangers of Machine Learning in Soil Science

(1) Data science of soil materials, ML models without soil science context

ML modelling often follows a workflow of processes that apparently do not require in-depth soil knowledge. A simple search on Google Scholar for "machine learning in soil classification" yields numerous papers, primarily from the computer science field, that apply ML techniques to predict soil types based on properties or images of soil. These studies often treat soil merely as a material with inadequately informed labels. This approach can result in information that lacks practical relevance and does not contribute to a deeper understanding of soil science or soil.

(2) Unscrutinised machine-learned soil prediction models

Defining the objective of an ML modelling exercise is essential. If the goal is to achieve the highest accuracy for a specific problem, interpretability may not be a priority. Considering the complexity of natural phenomena and human limitations in understanding complex relationships, demanding complete transparency from ML models may not be feasible. Nonetheless, it is crucial to ensure that the model provides a valid generalisation of the phenomenon being studied. Overreliance on automated outputs without sufficient scrutiny could lead to misinterpretations. Many papers in soil science literature use ML modelling without attempting interpretability, raising questions about the utility of such work for advancing soil

science knowledge.

(3)Lack of transparency in proprietary soil models

The utility of ML spans beyond research into commercial domains, such as predicting soil properties using near-infrared spectroscopy or an online platform for predicting soil properties using remote sensing images. This is particularly relevant in agronomic, soil contamination and soil carbon accounting applications. While research settings often require transparency regarding methods and data, commercial entities tend to be secretive to maintain a competitive edge, raising concerns about the reliability of their model predictions. Consequently, soil prediction becomes proprietary. Addressing these concerns may require implementing methods like uncertainty assessment, independent validation tests, and reporting on the diversity of soil types used in training the models.

(4)Stagnation of theoretical advancement

The focus on applying ML to predict and model soil properties and processes could overshadow the need to develop new theoretical frameworks in soil science. Without ongoing theoretical advancements, soil science may become overly dependent on data analytics without fostering innovative ideas. Theory generates hypotheses and generally leads to efficient experimentation and data generation. It’s not clear at this stage that this is the case for ML-generated prediction models.

(5)Doing too much with too little

There is a tendency to produce regional or global maps of various interesting soil characteristics without acknowledging the limited data and the risk of extrapolating these models in areas where data is sparse or of poor quality, leading to unreliable or misleading results. Another tendency is solely relying on ML models to infer controls of soil properties prediction. This overextension can compromise the integrity of soil science research and its applications. Sensible guidelines are required for the data density required for such predictions; for example, global maps based on several hundred observations are probably questionable, whereas those based on tens of thousands of observations inspire more confidence.

(6)Decline in direct soil observations and human fieldwork

Overreliance on ML might lead to overconfidence and decreased fieldwork and the gathering of new observations of soil, which are important for understanding soil in its natural conditions and accurately interpreting data and models. This shift could reduce the practical understanding of soil conditions and processes, diminishing the empirical grounding of soil science. Since soil is dynamic and responds to human forcings, continued widespread real-world observation is essential. Often, modelling and prediction can be better improved by accruing new observations at key locations rather than through incremental improvements of new ML methods.

2. Incorporating soil science knowledge in ML models

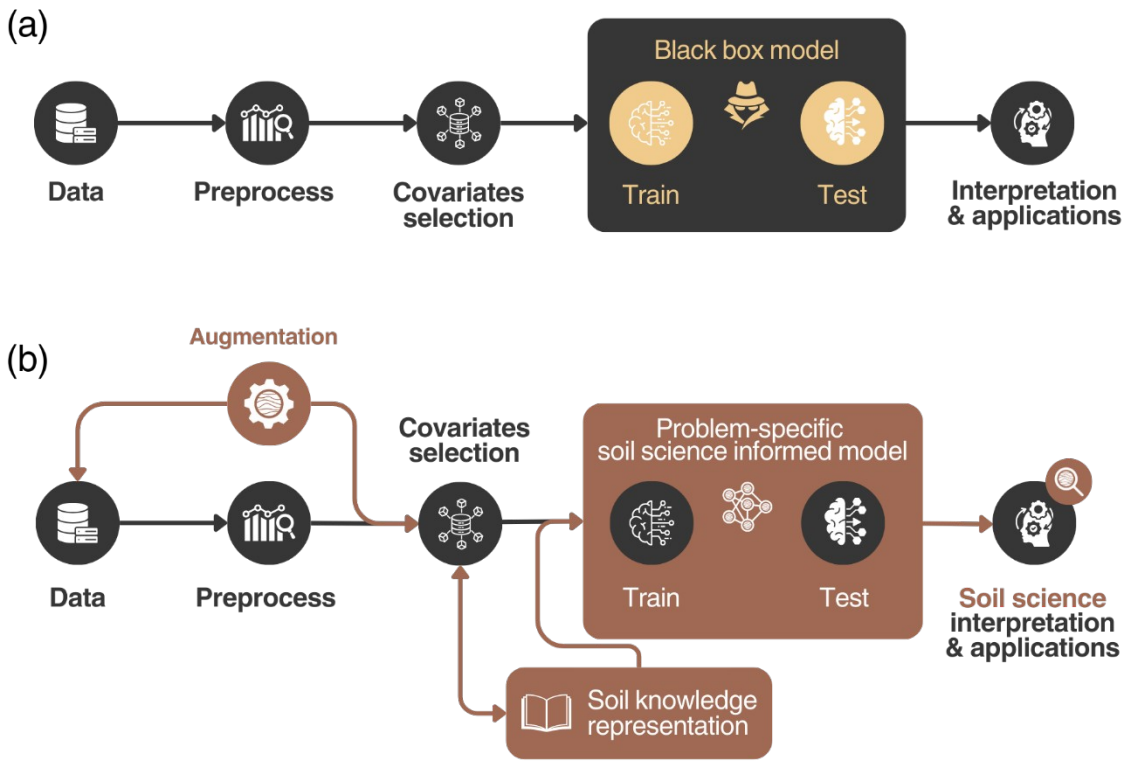


Figure 1. (a) The general workflow for ML applications in soil science. (b) the iterative and problem-specific nature of soil science-informed models. Conventional steps are in black circles, with steps involving soil science supervision in brown circles.

The steps for conducting an ML analysis on soil data typically follow the procedure depicted in Figure 1(a). The process starts with data collection and pre-processing to check for outliers, selection of covariates or predictors, followed by training of an ML model, which could include tuning the hyperparameters. The model is then tested on a proportion of the dataset which was not used in the training process. Finally, the model is interpreted using procedures such as variables of importance or Shapley values. (In ML, Shapley values quantify the relative contribution of each

predictor to a model’s output, for more details, see Padarian et al. (2020a) and Wadoux and Molnar (2022)). The entire procedure follows a standard ML workflow.

Most studies have utilised supervised learning, where ML algorithms are tasked with predicting labels based on a set of features. However, the concept of 'supervision' should extend beyond mere labels to encompass a broader prior knowledge framework. This knowledge is often embodied in functions or sets of rules that may depend on specific “labels”. Such supervision incorporates domain-specific insights that guide the learning process, enabling the algorithm to make more informed and contextually appropriate predictions. ML models should not only learn from observation data points but also integrate structured forms of knowledge effectively, thereby enhancing their predictive accuracy and relevance to soil science (see Figure 1(b), soil science supervision steps).

There is a lack of discussion on how to effectively incorporate soil science knowledge or physical rules in ML models. Here, we argue that ML models need to be iteratively designed and problem-specific, and they should be supervised to predict patterns conforming to soil phenomena. SoilML could deliver predictive models grounded in soil science, which not only achieve higher prediction accuracy but also enhance the models' ability to generalise predictions. Additionally, SoilML could improve transparency, thereby increasing the plausibility and reliability of these models (Kashinath et al., 2021; Wadoux et al., 2020).

2.1.1 Source of knowledge

The fields of Informed Machine Learning and Physics-Informed Machine Learning have emerged to address the empirical nature of current ML models by incorporating prior knowledge into the training process. von Rueden et al. (2023) discuss the source of knowledge, its representation, and how it is integrated into ML algorithms (Figure 2). The source of knowledge can be in three forms: specific scientific knowledge (in our case, soil science), general knowledge, and expert knowledge. Soil science knowledge could include laws or equations, principles, and rules. General knowledge is often intuitive and implicitly validated by human reasoning or empirical studies. Expert knowledge tends to be based on experience, for example, the relationship between certain soil properties and their covariates

(Lark et al., 2007) or mental models embedded in soil maps and their legends (Bui, 2004; Qi and Zhu, 2003).

These sources of knowledge can be represented as algebraic forms (e.g. Philip's infiltration equation) or differential equations (Richardson-Richards equation), simulation results, pseudo-observations, soil maps, rules, or probabilistic relationships (Hudson, 1992). In turn, these forms of knowledge can be integrated into the ML workflow through training data, model structure design, learning algorithms, and final evaluation. Current approaches to incorporating soil science knowledge in ML models involve several strategies. For example, in DSM, covariates may be selected to conform to soil-forming factors or *scorpan* covariates (See section 3.1). In soil moisture dynamics modelling, predictors could be selected based on components of a water-balance model. Another idea is incorporating pseudo-observations based on expert opinion on soil properties in areas which lack observations such as in high-elevation areas or extreme environments. Finally, interpretative tools such as Shapley value could interpret how predictors contribute to ML predictions, aligning predictions with existing knowledge (Padarian et al., 2020a; Wadoux and Molnar, 2022).

2.1.2 Incorporating soil science knowledge in ML models

Karniadakis et al. (2021) advocate three ways of incorporating soil science knowledge in ML models: observational priors, model structure design, and learning guidance (Figure 2).

(1)Observational priors

This approach involves augmenting training data to reflect underlying knowledge about the subject. Expert knowledge is mostly represented in DSM studies, including the addition of synthetic or pseudo-observations to the training data. DSM commonly relies on legacy soil data derived from laboratory measurements, which can be limited in spatial coverage. Field observations such as hand texture can provide a dense and complementary source of soil data, capturing variability across the landscape that laboratory data may miss. Eymard et al. (2024) demonstrated that integrating field observations of soil texture, even with potential biases, can

improve DSM predictions by identifying unique landscape features not represented in laboratory datasets, ultimately enhancing both model accuracy and the understanding of soil processes.

In addition, soil survey can be spatially biased due to preferential sampling patterns, and may have gaps in coverage due to inaccessible areas, such as steep terrain or remote regions. For example, Koch et al. (2019) used 13,000 boreholes to map the depth to the redox layer across Denmark using random forest regression kriging, but found that lowland areas were underrepresented. To address this, synthetic observations were added in these regions based on hydrogeological knowledge, improving lowland representation. Similarly, outputs from soil process-based models can be used to fill temporal gaps in observations, which will be discussed in Section 3.4. See also Box 3 on reducing overparametrisation in ML models via data augmentation.

(2)Model structure design

The architecture of ML models should be designed to ensure that their predictions are consistent with established soil science principles. This involves selecting appropriate model types, designing input and output layers and connections that can process and interpret soil data, and implementing mechanisms that incorporate domain-specific knowledge into the learning process. For example, ML structure needs to accommodate soil profile information. In the case of predicting soil at multiple depths within a profile, a multitasking ML model that predicts soil properties at multiple depths simultaneously would be preferable to creating independent soil depth prediction functions (Padarian et al., 2019b). In another example, conventional maps are updated or disaggregated using ML models with expert knowledge inputs, such as defining soil-landscape conditions in which a particular soil type could occur (Holmes et al., 2015; Lamichhane et al., 2021; Odgers et al., 2014; van Zijl et al., 2019).

(3)Learning guidance

The training of ML models can be directed using loss functions and constraints to ensure that the solutions align with soil science processes. Typically, ML models are

trained to minimise a loss or cost function; commonly, this involves adjusting the model parameters to minimise the mean squared error between the observed and predicted values.

Following Tang et al. (2024) we could define the loss function of an ML model as:

$$L = \text{observational constraints} + \text{coherency rules} + \text{prior constraints} \quad (1)$$

Observational constraints are usually defined as the mean squared error between observed and predicted values for continuous variables or classification error for categorical variables. For example, an ML model predicting the parameters of a soil water retention function would minimise the difference between measured and observed water retention at defined pressure heads.

Coherency rules, also known as regularisation or penalty function, aim to constrain the parameters to obey physical processes related to model parameters, thus avoiding overfitting. For instance, in water retention prediction, the relationship between water content and pressure head must adhere to the monotonicity of the curve (van Genuchten, 1980).

Prior or knowledge-based constraints, involve incorporating soil science and general knowledge or assumptions about the parameters or outputs, guiding the model towards more plausible solutions. For example, specifying ranges within which certain parameters must lie based on prior studies or expert knowledge, and imposing non-negativity constraints on parameters or outputs (e.g., ensuring that soil moisture content or soil thickness cannot be negative).

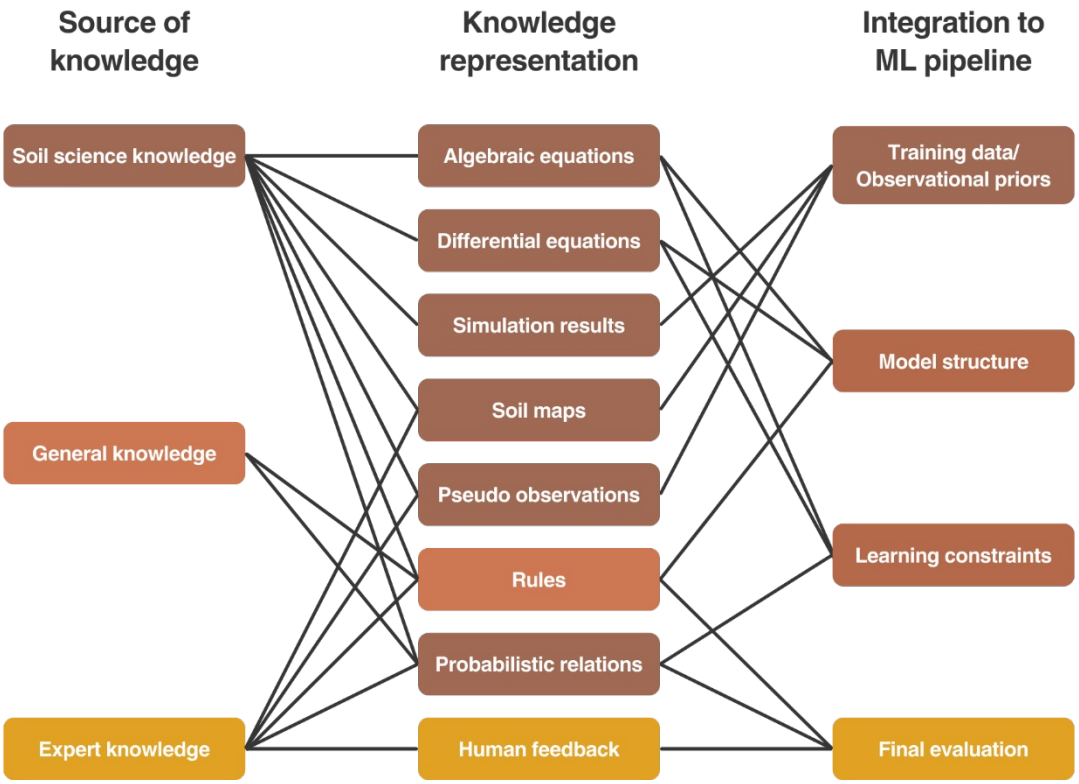
The three terms of the loss function can be weighted differentially, depending on the problem being solved.

The three approaches of knowledge incorporation outlined above are not standalone but could be combined to incorporate prior soil information and model constraints. Finally, soil scientists should evaluate the final outputs of the models in terms of the feasibility of the prediction or maps to evaluate against soil science knowledge or principles. For example, soil scientists could identify the congruency of soil-landscapes maps created by DSM or select digital maps of soil classes and

properties for implementing land suitability rules (Bui et al., 2020; Holmes et al., 2021).

Model structure design and learning guidance are typically applied together by modifying the ML input-output architecture and loss functions to be minimised. This approach requires a flexible ML framework that allows the structure and model loss function to be customised. ML models with a fixed structure, such as tree models, e.g. Cubist or random forest, may not be well-suited for such applications. Nevertheless, efforts could be made to modify the algorithms such as the spatial random forest model by Talebi et al. (2022).

A flexible ML framework that can accommodate these requirements includes neural networks with a generic input layer, one or several hidden layers and an output layer. The layers consist of multiple units connected via weights, allowing the model to learn a variety of functions. The structure of inputs and outputs can be modified to fit different dimensions of soil prediction, such as a 1-D, 2-D or 3-D. Additionally, convolutional layers could be added for filtering purposes, and custom objective or loss functions could be defined to align with specific goals. In the next section, we will explore examples of these models in greater detail.



283

284 Figure 2. Soil Science-Informed ML, pathways to supervise ML models with soil
285 science knowledge (adapted from von Rueden et al. (2023)).

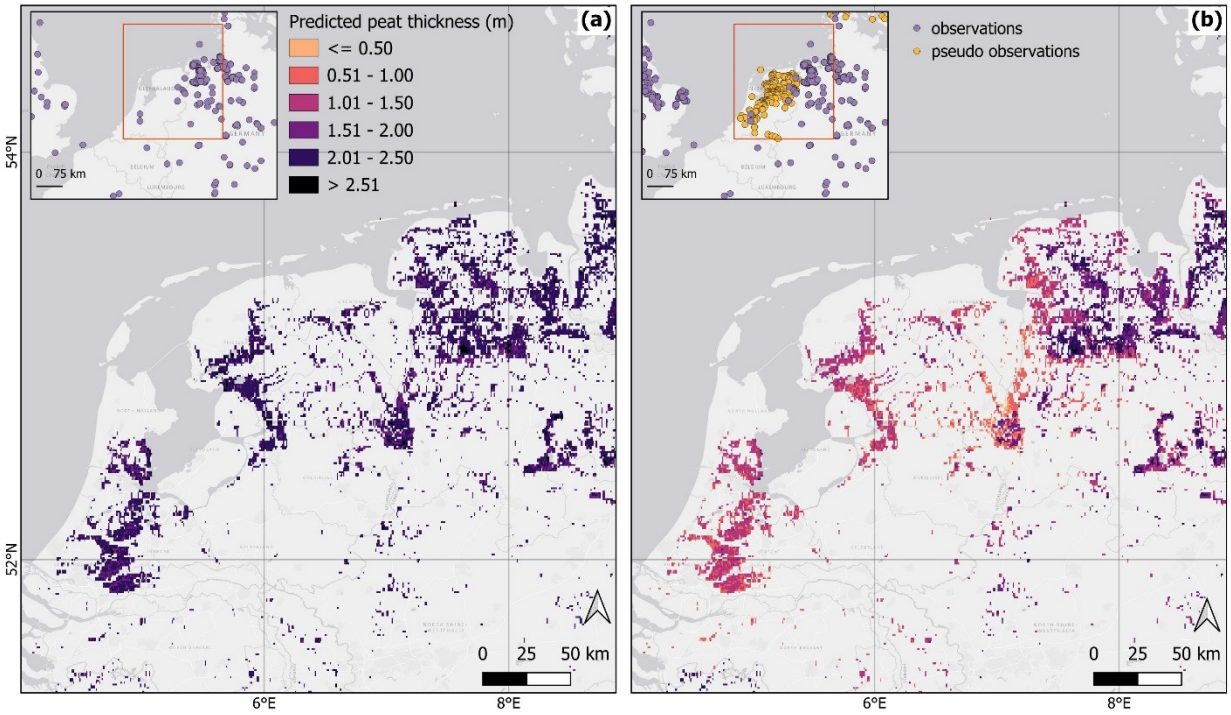


Figure 3. A comparison of peat thickness prediction in European peatlands (a) without and (b) with observational priors, the addition of data points from national peat thickness maps (adapted from Widyastuti et al. (2024)).

3. Applications of Soil Science-Informed Machine Learning (SoilML)

In this section, we introduce examples representing the application of SoilML through various forms of knowledge and their incorporation in ML models in several soil science domains, including digital soil mapping, soil spectroscopy, pedotransfer functions, and modelling dynamic soil properties. All these examples address soil security in terms of biomass production, carbon sequestration, and water cycling.

3.1 Digital Soil Mapping

Digital soil mapping (DSM) is a process of creating soil maps using spatial covariates that are combined with field observations, expressed as the "scorpan" model (McBratney et al., 2003)

$$S = f(s, c, o, r, p, a, n), \quad (2)$$

where S represents soil classes or attributes. This model provides empirical quantitative descriptions of relationships between soil and other spatially-referenced factors: soil (s), climate (c), organisms (o), topography (r), parent material (p), age (a), and spatial position (n).

In DSM, the procedures involved collecting geo-referenced soil observations that are intersected with environmental (scorpan) covariates. A spatial soil prediction function is built to relate observed soil properties of interest to these environmental covariables using ML models. The calibrated spatial soil prediction function can then predict and map soil properties across the area (Arrouays et al., 2020).

As discussed in Wadoux et al. (2020), examples of incorporating soil science knowledge in DSM procedures include experts selection of scorpan covariates which conform to the soil-forming processes of the region to be mapped. Pseudo observations could also be added in areas that lack field observations. In the global peat thickness mapping study by Widyastuti et al. (2024), many regions of the world lacked direct field observations. Incorporating pseudo-observations derived from national peat thickness maps can help guide the model. Figure 3 shows an example of peat thickness prediction using a random forest model. Initially, the random forest model was trained only with available field observations, which resulted in the peat thickness values being overpredicted by 2-3 m over the Netherlands and Germany. Incorporating 500 points from peat maps of Sweden, the Netherlands, and Denmark reduced the mean predicted thickness by over half (mean = 1.08 m), resulting in a more accurate and realistic map (Figure 3).

3.1.1 Case study: Contextual information for soil mapping using convolutional neural networks

Conventional approach: DSM models typically use point observations intersected with pixel-wise spatial covariates for calibration. Ideally, contextual information around the observations should be included as covariates. Studies include relative elevation around a point to provide contextual information. For example, Behrens et al. (2010) used differences in elevation from observation points to each of the surrounding neighbourhoods as predictors to capture the relative position of the

observation point on the landscape. Other approaches include calculating terrain attributes at various neighbourhood window sizes (Miller et al., 2015).

SoilML: Padarian et al. (2019b) demonstrated that a convolutional neural network (CNN) model using images of covariates (terrain and climate variables) can effectively explore spatial relationships between a point observation and its neighbouring pixels (Figure 4). The model also includes a 3-D stack of images as input, data augmentation to reduce overfitting, and simultaneous prediction of multiple depths.

Using a soil mapping example in Chile, the CNN model was trained to simultaneously predict SOC at multiple depths across the country. To increase data representation, data augmentation was employed to generate new samples by modifying the original data without changing its meaning. This included rotating a 3-D array by 90, 180, and 270 degrees. This step also acted as regularisation, reducing model variance and overfitting, and induced rotation invariance by ensuring the model responds similarly to rotated data, such as a soil profile next to a gully. The results showed that the CNN model reduced the error by 30% compared to conventional techniques that only used point information of covariates. For country-wide mapping at a 100 m resolution, a neighbourhood size of 3 to 9 pixels proved more effective than using a single point or larger neighbourhood sizes. Additionally, the CNN model produced less prediction uncertainty and more accurately predicted soil carbon at deeper layers.

Implications: The CNN framework is designed to accept images as input, capturing information about the observation and its spatial context. Its convolutional layers apply various filters, in the case of a DEM, it effectively mimics the calculation of terrain attributes across different window sizes (Taghizadeh-Mehrjardi et al., 2020). This contrasts with other ML models that require algorithm modifications to handle spatial data. For example, Talebi et al. (2022) developed a spatial random forest model that uses local spatial covariates, which were transformed into vectorised spatial patterns, as predictors. In addition, regularisation, or the addition of a penalty function to the loss function (Eq. 1), could constrain the model to follow certain soil-landscape rules, e.g. a penalty could be added to the loss function when soil thickness on the top of the hill is predicted to be larger than on the lower slope.

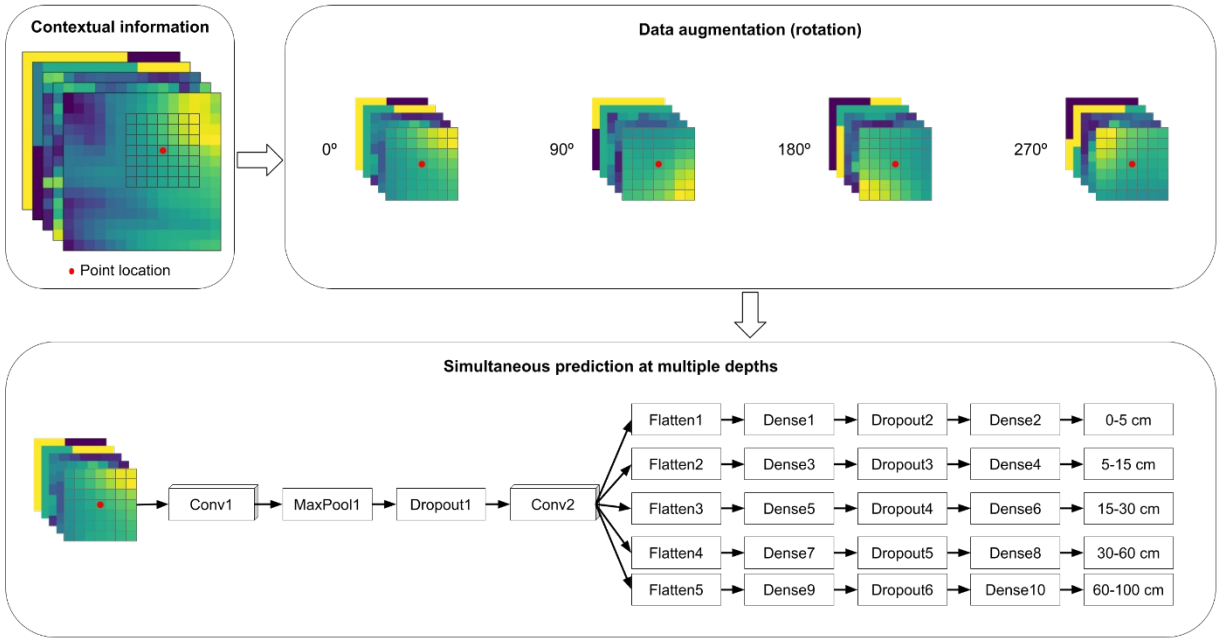


Figure 4. A deep learning framework for digital soil mapping incorporating contextual information and data augmentation for training a CNN model to predict soil properties at multiple depths.

3.1.2 Case study: 3-D soil mapping

Conventional approaches: The topic of mapping soil properties across space and depth has gained wide interest. However, soil profiles are usually observed via horizons, which vary in thickness and depth. In DSM, the variation of soil properties down a profile is often harmonised using the equal-area spline depth function approach. Soil observations at various depth intervals are first harmonised to pre-determined depth intervals. To create maps of soil at these defined depth intervals, models are trained to predict soil properties at several depth intervals simultaneously using either neural networks or other ML models capable of multivariate outputs.

Other studies propose that soil properties at any depth can be mapped using a model that incorporates depth along with spatial covariates as predictor variables, creating a ‘3D’ model. However, ML models consider depth as one of the covariates,

indifferent to spatial covariates. Due to the limited depth inputs, tree models such as random forest are sensitive to the training data and tend to predict soil properties at depths as stepped values (Ma et al., 2021).

SoilML: We propose designing a neural network to predict soil properties at regular depth intervals to alleviate such a problem. The model would take spatial covariates as inputs, and the training of the model disaggregates and predicts soil properties at all depths simultaneously.

For example, 59 soil cores, varying in depth between 85 and 130 cm were used in a study by (Fajardo et al., 2016). The cores had SOC measurements via visible-near-infrared (vis-NIR) and shortwave infrared (SWIR) spectroscopy (wavelengths between 350 and 2500 nm) at every 2 cm down to a depth of 1 m. To simulate horizon sampling, SOC observations were grouped by soil horizons (S). The spatial prediction model used the following covariates as inputs: terrain (elevation, wetness index, mid-slope position, altitude above channel network), remote sensing images (Vis-NIR and SWIR bands), and predicted SOC every 2 cm from the surface to 1 m. Simulating soil observations by layers, the training data of SOC observations were grouped by soil horizons. Thus, the loss function for the model is:

$$L=\sum_{i=1}^{n}\left(\sum_{j=1}^{m(i)}\left(S_{ij}-\widehat{s_{ij}}\right)^2\right) \qquad (3)$$

where n is the number of soil cores, m is the number of layer observations per core, and S is the observed SOC value per layer of observations. Note that the model predicts SOC at specific points and $\widehat{s_{ij}}$ refers to the aggregated or averaged predicted value corresponding to each observed layer. Figure 5 shows an example of the predicted SOC values across the profile.

Implications and prospects: Although numerous studies have incorporated depth as a covariate to generate 3D maps, it is important to be cautious about combining spatial covariates (covering geographical areas with grid spacing ranging from approximately 1 to 1000 m) with depth, which varies from about 0.01 to 2 m. ML models may struggle to distinguish the significant differences in scale and continuity between these types of measurements. Formal geostatistical approaches which predict in 3D by disaggregating the bulk depth measurements (Orton et al.,

2020), or using the Gaussian process regression (Wang et al., 2024), provide more robust solutions.

While soil properties at various depth intervals have been extensively mapped over the years, there remains a gap in mapping the distribution patterns of soil horizons and representing soil as a 3-D continuum. Soil is fundamentally a three-dimensional body composed of distinct horizons. Various studies have employed techniques such as electromagnetic induction and interpolation based on cone penetrometer resistance to map soil layers (Grunwald et al., 2001) and the thickness of soil horizons (Chaplot et al., 2010). For instance, Mendonça Santos et al. (2000) mapped the thickness of each of the 12 horizons in a Swiss floodplain in two dimensions, then stacked these results to represent a three-dimensional volume. Similarly, Gastaldi et al. (2012) combined logistic regression with ordinary regression to first model the occurrence of each horizon and subsequently their thickness. Advancements in ML, particularly neural networks, now offer the potential to model soil as a profile of horizons and predict each horizon's thickness and composition as observed.

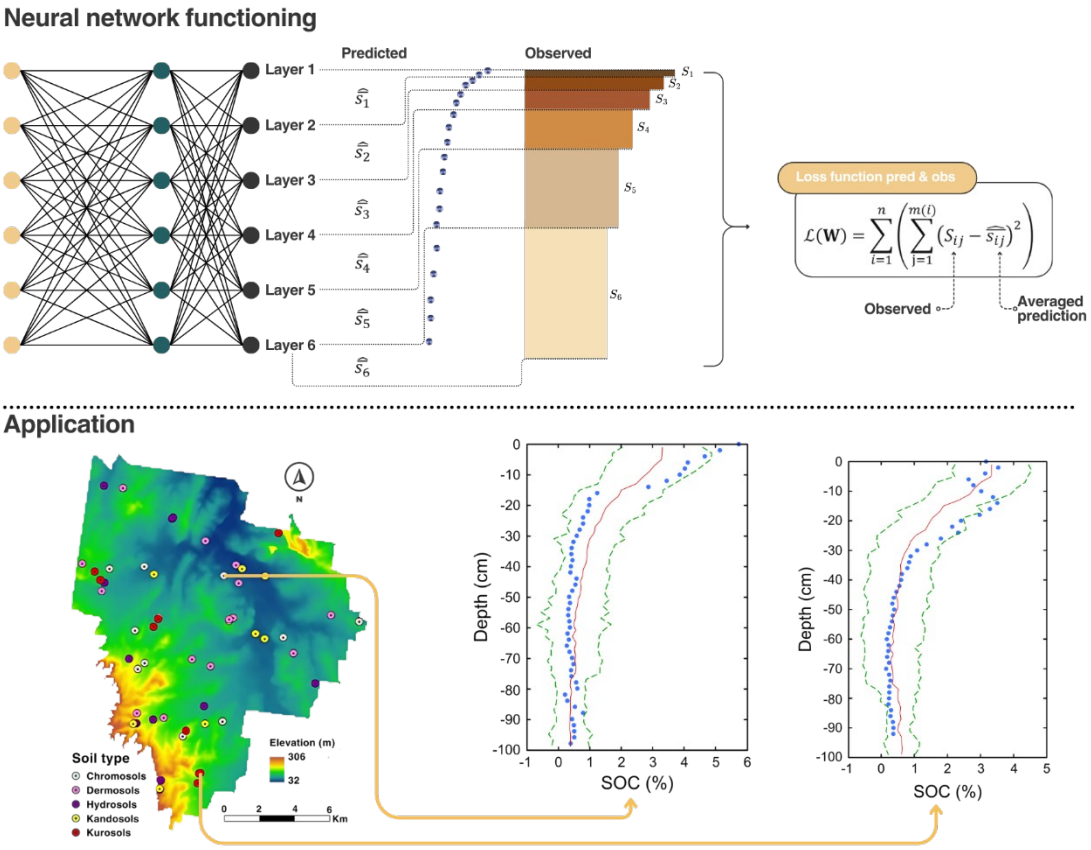


Figure 5. An example of a neural network model that predicts point soil observations along the soil profile depth using environmental covariates. The neural networks model was trained using soil profile data, with the loss function minimising the difference between the averaged predicted layer values and the observed soil layer values. The figure below shows the prediction of SOC across the Hunter Valley in NSW, Australia. The observed (blue dots), the predicted (red line), and the prediction interval (green dashed-line).

3.1.3 Case study: Soil class mapping incorporating taxonomic distance

Conventional approaches: Digital soil class mapping typically begins with the description of soil profiles and allocating the profiles to soil classes according to an established soil classification system. This process continues with correlating the observed soil classes with co-located covariates at each observation site. Most ML training in supervised classification involves minimising classification errors:

$$L = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} I\left(y_i \neq k\right) \quad (4)$$

where i = 1, 2, ... n is the number of observations, and k = 1, 2, ... c is the number of classes, $I\left(y_i \neq k\right)$ is an indicator when observed class y_i is not equal to class k. This error criterion assumes that the errors across all classes are of equal importance. However, this is not valid for soil classes and does not allow for situations where some errors are more important than others.

SoilML: Taxonomic distance between soil classes can be incorporated into a supervised classification routine. Minasny and McBratney (2007) calculated the taxonomic distance between soil classes based on a central concept, e.g. to define a modal soil profile for each soil class. The taxonomic distance matrix between soil classes can be represented as **D,** with $D_{j,k}$, represent the distance between class j and class k. The supervised classification loss function could be defined as the average misclassification cost:

$$L_a = \frac{1}{n}\sum_{i=1}^{n} D\left(C_i, \widehat{C}_i\right) \quad (5)$$

where L_a is the average taxonomic distance error, $D\left(C, \widehat{C}\right)$ is the taxonomic distance between observed class C and predicted class $\widehat{C}$. By using classification trees that minimise the taxonomic distance over misclassification error, the methodology is refined to model soil class relationships.

Implications: Defining taxonomic distance extends beyond predicting soil classes to encompass the development of soil classification units. While early numerical soil classification methods in the 1950s were constrained by limited data and technology, modern advances now allow taxonomic distance calculations to explore correlations between national systems and global frameworks like the World Reference Base or USDA Soil Taxonomy. For example, Michéli et al. (2016) used taxonomic distance to differentiate USDA soil great groups, demonstrating its utility in objectively refining classification criteria. Similarly, Hughes et al. (2017) showed that taxonomic distance calculations can aid in translating soil classes across various classification systems, enhancing global comparability and consistency.

Laborczi et al. (2019) compared topsoil (0-30 cm) texture classes in Hungary using two methods: directly compiled maps of clay, silt, and sand content for the 0–30 cm depth, and synthesised maps derived from the thickness-weighted average of the 0–5, 5–15, 15–30 cm layers. While the soil texture class maps produced by both methods are similar, taxonomic distances between the two maps reveal more pronounced discrepancies in certain regions. Significant differences are observed particularly in hilly and mountainous areas, which could pose challenges in erosion and sedimentation modelling and the prediction of flash floods. Additionally, inaccuracies in mapping salt-affected and hydromorphic soils could impact water management and irrigation planning. Nevertheless, DSM of soil classes still rarely considered the taxonomic distance in the ML training workflow.

3.2 Soil Spectroscopy

Soil can reflect, scatter, or emit electromagnetic radiation, resulting in a unique spectral signature. Soil responds uniquely to infrared radiation, making infrared spectrometers suitable for soil analysis because they can measure rapidly, cost-effectively, and non-destructively. An infrared spectrometer can predict multiple soil properties from a single-spectrum measurement. However, soil is a complex mixture of mineral and organic constituents, it is challenging to assign specific spectral features to particular physical, chemical, or biological components. Therefore, empirical multivariate calibration techniques are commonly employed to predict soil properties by relating spectra data to observed soil characteristics (Chen et al., 2023b; Hutengs et al., 2021; Vohland et al., 2022).

3.2.1 Case study: Physical model for soil spectra response to moisture and hydraulic properties

One major factor affecting soil reflectance is the presence of water (Lobell and Asner, 2002). Wet soil typically reflects less light than dry soil. This sensitivity of soil reflectance to moisture allows for the rapid estimation of soil water content through vis-NIR and SWIR reflectance measurements (Liu et al., 2002).

Conventional approaches: Various empirical models have been developed to relate soil reflectance to soil water content in the Vis-NIR-SWIR spectra (Babaeian et al.,

2019). These models include partial least squares regression (Bogrekci and S. Lee, 2006; Castaldi et al., 2015), principal component regression (Chang et al., 2001), and ML models (Hassan-Esfahani et al., 2015; Zaman et al., 2012). While these models are effective, they require extensive databases for calibration and their applicability is restricted to the specific soil conditions under which they were developed, as moisture response to NIR radiation depends on soil types and constituents (Babaeian et al., 2019).

SoilML: Radiative transfer models can effectively describe diffuse infrared radiation in soil. The Kubelka and Munk (KM) model (Kubelka and Munk, 1931) is a two-flux radiative transfer model that describes light transfer through a particulate medium, characterised by absorption (□) and scattering (□) coefficients. The model uses a set of differential equations to account for light travelling in two opposing directions and yields reflectance and transmittance as a function of *k*, *s,* and depth. The optical depth is assumed to be infinite for soil, and therefore the transmission becomes negligible.

Sadeghi et al. (2015) applied the KM model to explore the relationship between soil water content and reflectance. They proposed that the optical properties (*k* and *s*) of soil can be expressed by a linear volume averaging of the optical properties of its constituents, i.e., solid particles, water, and air. Based on this approach, they derived a physically based and linear equation that explicitly expresses SWIR to water content:

$$\frac{\theta}{\theta_s}=\frac{r-r_d}{r_s-r_d} \tag{6}$$

where θ is the volumetric water content ($m^3\ m^{-3}$), θ_s is the saturated water content ($m^3\ m^{-3}$), and r is the transformed reflectance. The parameters r_d, and r_s are the transformed reflectance of soil in dry and saturated states, respectively. Transformed reflectance (r) can be calculated from the measured reflectance (R) as follows:

$$r=\frac{(1-R)^2}{2R} \tag{7}$$

Norouzi et al. (2022) hypothesised that the two distinct forms of soil water, i.e., capillary and adsorbed water, impact soil reflectance differently (Figure 6). Building

on this hypothesis, they considered different optical properties for capillary and adsorbed water and derived a new model to describe the relationship between soil reflectance and water content:

$$r = r_d + \overbrace{c_a \theta_a^{p_a}}^{r_a} + \overbrace{c_c \theta_c^{p_c}}^{r_c} \tag{8}$$

The total transformed reflectance of wet soil (r) can be decomposed into three components: r_d, r_a, and r_c corresponding to the dry soil, adsorptive water, and capillary water, respectively. Parameters c_a, p_a, c_c, p_c are the optical properties related to adsorbed and capillary water. In this equation, θ_a and θ_c are the volumetric water contents of adsorbed and capillary water that can be derived from the soil retention curve. Norouzi et al. (2022) used the model by Lebeau and Konrad (2010) for soil water retention curve to partition the total water content (θ) into its components θ_a and θ_c:

$$\theta = \theta_c + \theta_a \tag{9}$$

where the capillary component is modelled based on Kosugi (1996):

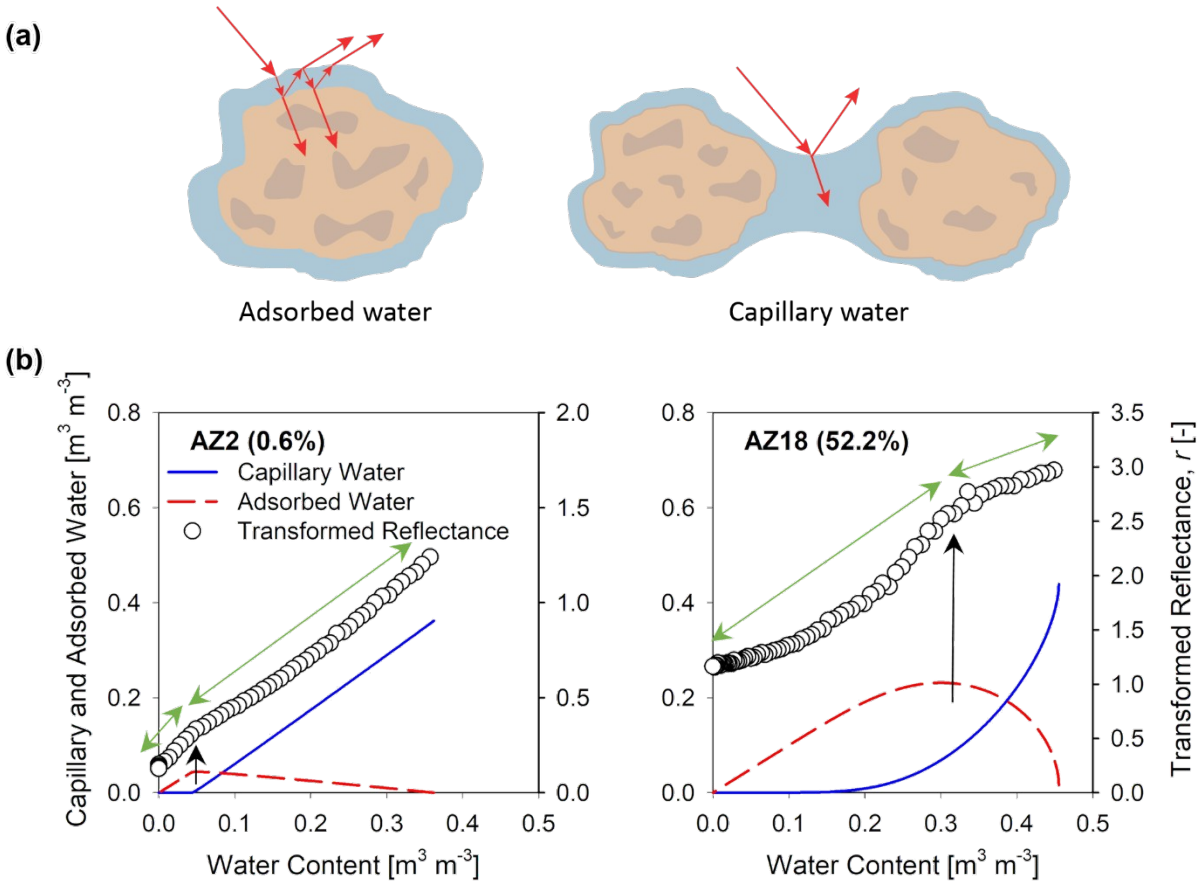
$$\theta_c = \frac{1}{2}\theta_s \, erfc\left[\frac{\ln(h/h_m)}{\sqrt{2}\,\sigma}\right] \tag{10}$$

where h is the pressure head, θ_s is the saturated volumetric water content, and *erfc* denotes the complementary error function; h_m, σ, and θ_o are fitting parameters.

The adsorptive component is represented using the Campbell and Shiozawa (1992) model for extremely low matric potentials, which linearly diminishes as the amount of capillary water increases:

$$\theta_a = \theta_o \left(1 - \frac{\ln|h|}{\ln|h_d|}\right)\left(1 - \frac{\theta_c}{\theta_s}\right) \tag{11}$$

where h_d is the pressure head at oven dryness and generally corresponds to a finite value of -10^7 cm (Campbell and Shiozawa, 1992). Equation (8), in combination with Equations (10) and (11), directly connects soil reflectance to the soil water retention curve.


(a)
Adsorbed water
Capillary water
(b)
AZ2 (0.6%)
Capillary Water
Adsorbed Water
Transformed Reflectance
Capillary and Adsorbed Water [m3 m-3]
Water Content [m3 m-3]
AZ18 (52.2%)
Transformed Reflectance, r [-]
Water Content [m3 m-3]

555

Figure 6. (a) An illustration of reflections and refractions of light beams (indicated by red arrows) as they interact with adsorbed and capillary water. (b) Effects of capillary and adsorbed water (left axis) on transformed reflectance measurements at 2210 nm (right axis) for two Arizona soils with 0.6% and 52.2% clay content. The black arrows indicate the points where the reflectance slope (shown by green arrows) changes sharply, marking the transition from capillary to adsorbed water regimes.

Norouzi et al. (2022) demonstrated that soil reflectance is influenced not only by the amount of water in the soil but also by the structure of water, specifically the capillary and adsorbed components. They showed that when the soil water retention curve is known, Equation (8) accurately describes the relationship between soil moisture and reflectance. As shown in Figure 6b, a noticeable change in reflectance at 2100 nm occurred at a specific point marked by a black arrow. The slope of the reflectance, marked by a green arrow, changes before and after this point, indicating that the shift corresponds to the maximum water content for adsorbed water, as seen when compared to the water components on the left axis. This transition signifies where capillary water recedes, and adsorbed water becomes the dominant component at the surface. This reflectance change is highly dependent on soil texture, occurring at lower water content for coarse-textured soils (e.g., at ~0.05 m^3 m^{-3} for AZ2) and at a higher water content for fine-textured soils (e.g., at ~0.3 m^3 m^{-3} for AZ18).

This also means that Equation (8) can be inverted to derive the soil water retention curve using NIR spectra of soil reflectance and moisture content measured during an evaporation experiment, optimising retention curve parameters and optical properties to match observed NIR reflectance. Validation with 21 soils of varying textures and mineralogy demonstrated accurate retrieval of the entire retention curve, from saturation to oven dryness (Norouzi et al., 2023).

Implications and prospect: Considering that the SWIR of a drying thin soil sample from saturation to air-dry can be measured within a few hours, the physics-based approach proposed by Norouzi et al. (2023) can be an efficient method for

measuring the soil water retention curve, which often takes several weeks. Although the above example is not an ML model, the water retention function could be defined as a neural network. This model can be further constrained by Richardson-Richards' equation to align with water content and time measurements collected during the evaporation experiment. Moreover, the radiation transfer model can integrate factors such as soil water content, particle size, and organic matter effect. It has the potential to predict soil texture and organic matter content by calibrating optical, absorption, and scattering coefficients (Wu et al., 2023). In combination with physics-informed ML, these models can improve both the predictability and interpretability of soil spectra models.

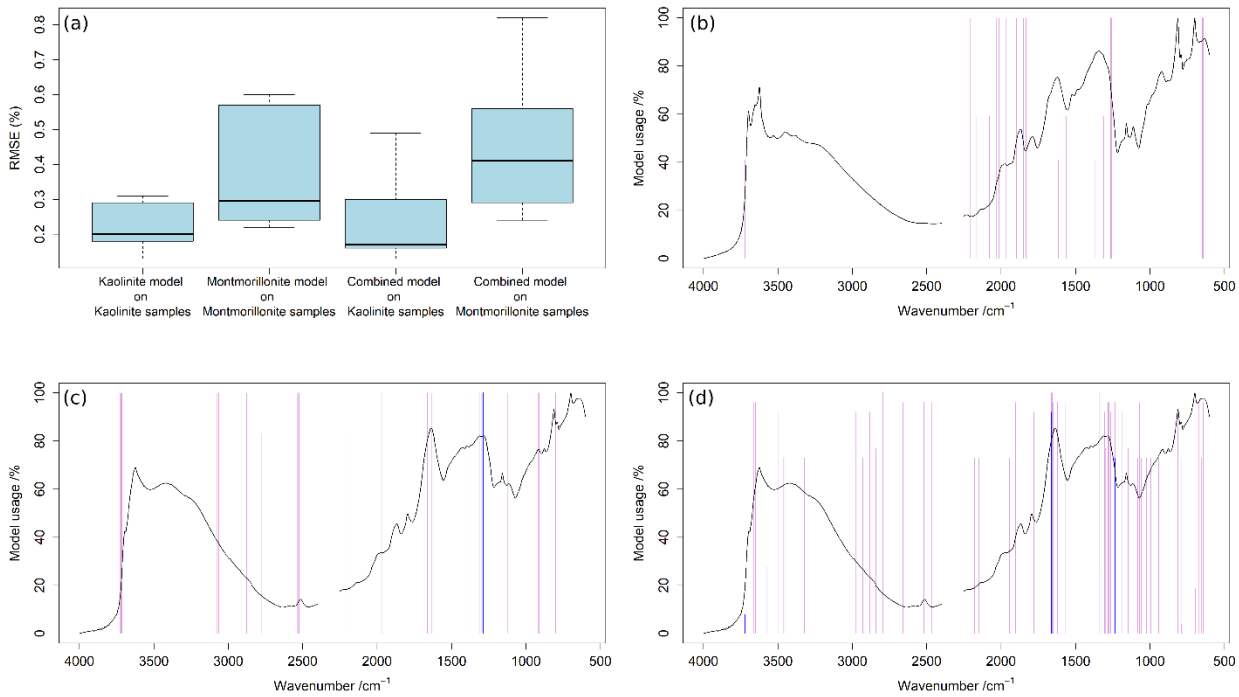
3.2.2 Case study: Building soil-based spectra functions

Conventional approaches: Soil spectra are typically pre-processed with smoothing or transformation to remove noise and serve as inputs of regression models or ML algorithms. The models are trained to minimise root mean square error (RMSE) and maximize the coefficient of determination (R^2). The working of the models can be explained by the importance or usage of variables in the model. For example, variable importance in projection (VIP) score is used in partial least square regression modelling to help identify which wavelength is mostly related to the soil property, and the model usage rate can be used to evaluate Cubist models (Chen et al., 2023a; Seidel et al., 2022). However, there is no information input from soil science knowledge when training the model, and the prediction result will merely depend on the relationship between the spectra features and the soil properties. In this way, soil science knowledge only serves the purpose of explaining outcomes, rather than being directly involved in model building.

SoilML: Prior soil information, e.g. morphological and mineralogical characteristics, can help divide the samples into homogeneous groups before modelling, and models will therefore be trained based on soils with shared properties. By comparing the effects of models trained on (1) all samples and (2) sample sets divided by prior information, this case study demonstrated the possibility of including soil knowledge in the modelling process.

In this case study, 370 Bt horizon soil samples with 0–5% carbon content were extracted from the Kellogg Soil Survey Laboratory (KSSL) dataset (Soil Survey Staff, 2014). X-ray diffraction analysis revealed kaolinite and montmorillonite as the dominant clay minerals. Modelling of total carbon content was conducted separately on 185 kaolinite-dominant samples and 185 montmorillonite-dominant samples with mid-infrared spectra and the Cubist regression tree model. Spectra were pre-processed with Savitzky-Golay smoothing, SNV transformation, and trimming off the CO_2 peak. Samples were randomly divided into a 70% calibration set and a 30% validation set for the training and testing of Cubist models, and this process was performed 10 times to get a distribution of results.

Results show that individual models based on dominant mineralogical components were more accurate than the total model (Figure 7a). The model created using all samples tended to have higher spread in the boxplot, which indicated less robustness than models from the pre-divided training set. The kaolinite model mainly used wavenumbers around 2000 cm^{-1} (Figure 7b), and the montmorillonite model more relied on multiple wavenumbers across the spectrum (Figure 7c). The combined model, on the other hand, utilised more conditions and variables than the individual models (Figure 7d), which might be due to the heterogeneity of soils dominated by different mineralogical characteristics. By including prior soil information in modelling, grouping the samples based on their mineralogical component improved the performance of models and enabled clearer differentiation of wavelengths used in the models.



640

641 Figure 7. (a) Boxplots of root mean square error (RMSE) results from ten repetitions
642 of Cubist modelling using various input data. (b)-(d) Variable importance of Cubist
643 models to predict total carbon content with mid-infrared spectra using: (b) kaolinite-
644 dominant samples for calibration, (c) montmorillonite-dominant samples for
645 calibration, and (c) combined kaolinite-dominant and montmorillonite-dominant
646 samples for calibration. Black lines are the mean spectra of calibration samples,
647 purple vertical lines are the variables used as predictors, while the blue vertical
648 lines are the conditions of Cubist models.

649

650 *Implications*: Analysing soil spectra alongside soil science knowledge enables the
651 identification of specific soil components, enhancing the effectiveness of statistical
652 methods and improving the understanding of soil properties and processes. By
653 grouping soils based on pedological information, such as soil order or soil horizon or
654 mineralogical component, researchers can refine models to achieve more accurate
655 and interpretable predictions. This approach encourages soil scientists to look
656 beyond mere prediction accuracy and develop a deeper understanding of the soil.
657 Incorporating soil knowledge can involve pre- or post-machine learning calibration,
658 such as inspecting spectra or grouping soils by mineralogy to guide the models.
659 After modelling, verifying that predictions align with soil science principles is crucial,

ensuring that ML applications do not overshadow the fundamental soil understanding (Ma et al., 2023).

3.3 Pedotransfer functions

Pedotransfer functions (PTFs) translate basic soil data into more complex, labour-intensive, and costly soil properties (Weber et al., 2024). They serve as predictive tools for estimating certain soil properties from easily measured or available data, thereby bridging the gap between available and required data. A prominent application of PTFs is in predicting the soil water retention curve, which describes the soil water content (θ), i.e., the volume of water per volume of soil under equilibrium at a given pressure head (h). Since measuring a soil water retention curve is time-consuming, PTFs offer a practical alternative by estimating it based on soil physical properties such as texture, bulk density, and SOC (Bagnall et al., 2022; Weber et al., 2024). PTFs are also used in various applications, including assessing irrigation, drainage, and evapotranspiration, enhancing DSM, and providing inputs for process-based simulation models to evaluate soil functions.

3.3.1 Case study: Prediction of water retention and hydraulic conductivity curves

Conventional approaches: The structure of a PTF typically involves using ML models to relate predictors (input data, such as soil texture and bulk density) to a predictand (output, such as water content at field capacity). In the context of predicting water retention curves using neural networks, there are three main model configurations (Figure 8):

Point PTFs: a point PTF predicts water contents (θ) at specific pressure heads (h) from basic soil properties such as sand, silt, clay, SOC, and bulk density. They require a training dataset that includes measurements of water retention at the specified pressure heads along with the basic soil properties.

Parametric PTFs: This configuration uses a hydraulic model capable of representing the data, focusing on predicting parameters of the hydraulic model. The output parameters are then used to form a continuous function describing the relationship between the dependent variable (θ) and the independent variable (h). This method

is favoured for its ability to provide a continuous prediction curve and is commonly employed in water retention modelling.

Direct Neural Network PTFs: In this setup, neural networks are directly applied to model water retention. The pressure head, along with basic soil properties, are used as inputs, allowing the model to learn a non-specific form of the soil water retention curve.

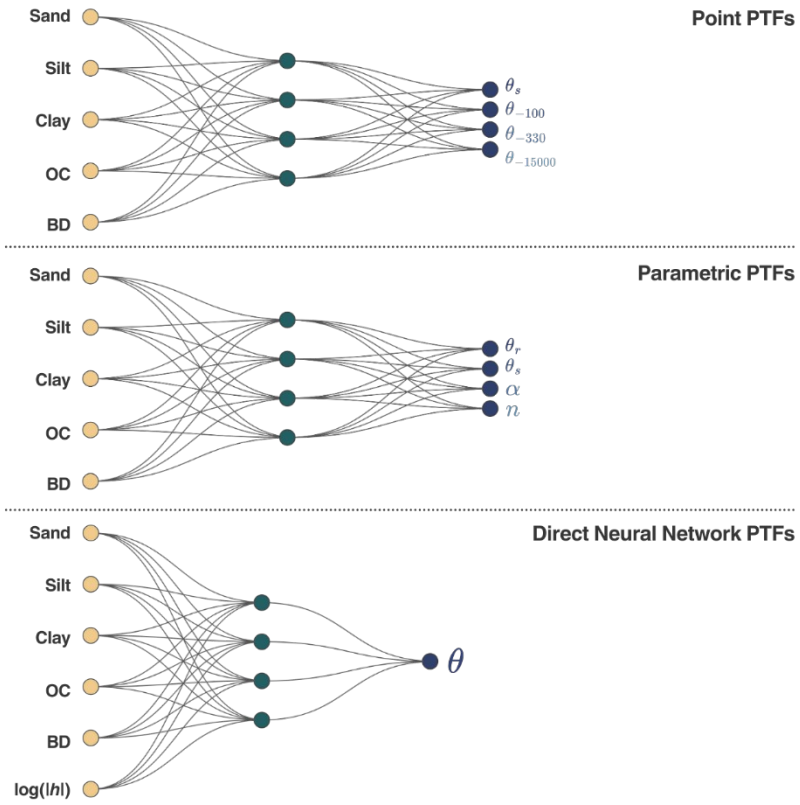


Figure 8. Three configurations of PTFs predicting soil retention: point, parametric, and direct neural networks (based on Haghverdi et al. (2012)).

In parametric PTFs, the van Genuchten equation (van Genuchten, 1980) is commonly used to model the water retention curve:

$$\theta(h)=\theta_r+(\theta_s-\theta_r)S_e(h)$$

$$S_e(h)=\left(\left[1+|\alpha h|^n\right]\right)^{-m} \quad (12)$$

where the water content (θ¿as a function of pressure head (h¿ is described by four parameters: θ_r, residual water content; θ_s, saturated water content; α,the inverse of air-entry pressure; and n, curve shape factor, with m defined as $m=1-1/n$. The van Genuchten model can be combined with the capillary theory model of Mualem (1976) to predict the unsaturated hydraulic conductivity curve, known as the Mualem-van Genuchten model.

Creating a parametric PTF first involves fitting the van Genuchten equation to observations to estimate the parameter vector ϕ = [$\theta_r, \theta_s, \alpha, n$¿. This is followed by forming relationships between basic soil properties (sand, clay, bulk density) and the parameters using neural networks (or other ML models) by minimising the following loss function:

$$L=\sum_{i=1}^{n}\left(\sum_{k=1}^{p}\left(\phi_{ik}-\widehat{\phi_{ik}}\right)^2\right), \quad (13)$$

where *n* is the number of observations, and *p* is the number of parameters to be estimated. The drawback of this approach is that the predicted parameters here do not necessarily bear a physical relationship.

SoilML: We can incorporate the van Genuchten function in training the ML model by requiring the estimated parameters to predict the observed water retention $[\theta(h)]$ rather than predicting each parameter of the van Genuchten equation independently. Minasny and McBratney (2002) used neural networks to predict the parameters of the van Genuchten function using soil properties (sand, clay, bulk density). The neural networks model predicted the van Genuchten parameters $[\hat{\theta}_r,\ \hat{\theta}_s,\ \hat{\alpha},\ \hat{n}]$but was trained to minimise the difference between the observed and predicted water content:

$$L_\theta=\sum\left(\theta(h)-\hat{\theta}\left(h\vee\hat{\theta}_r,\ \hat{\theta}_s,\ \hat{\alpha},\ \hat{n}\right)\right)^2 \quad (14)$$

In this case, the ML model is constrained to predict parameters that fit the water retention data (Figure 9). This led to more realistic prediction values and a more accurate estimation of the water retention relationship. Using soil water retention

data from Australia, Minasny and McBratney (2002) demonstrated that the PTFs trained using Eq. (14) predicted water retention much better compared to models that were trained using Eq. (13). In addition, the parameters of the van Genuchten model were better constrained according to theoretical expected values.

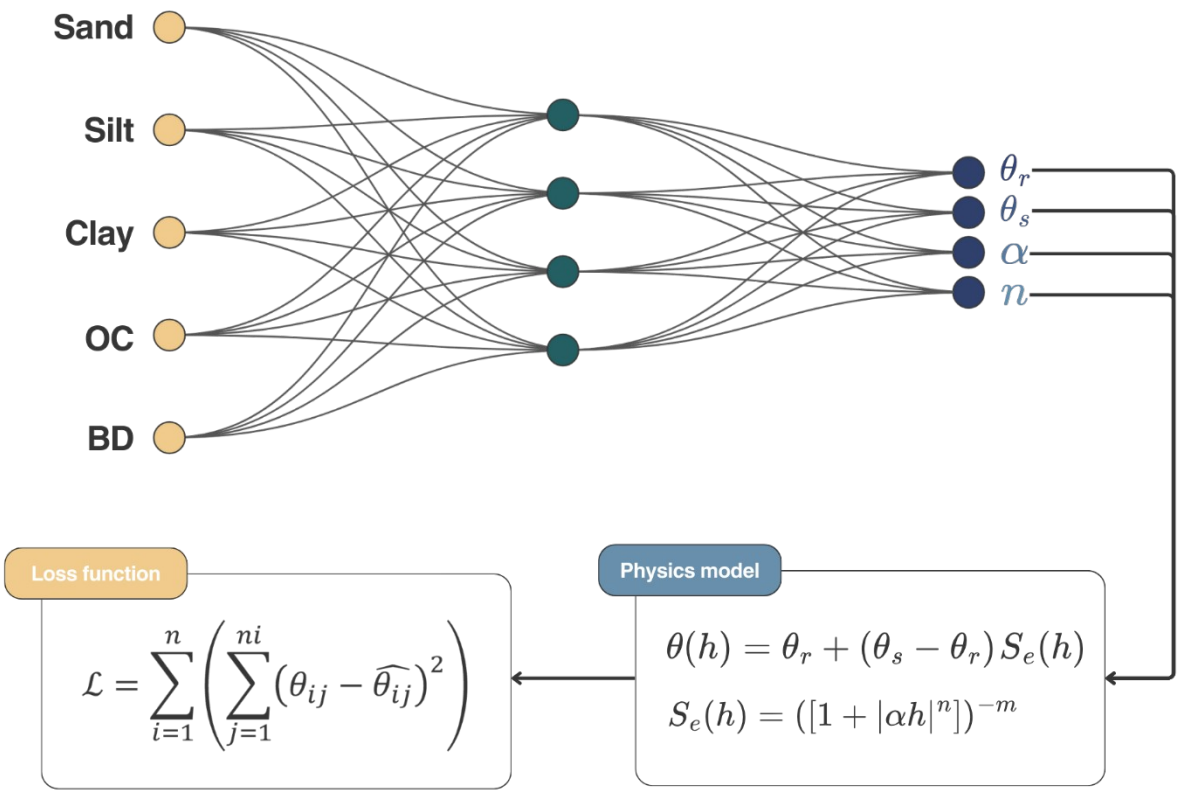


Figure 9. A physics-informed pedotransfer function for predicting a water retention curve function.

Recent research, such as that by Peters et al. (2024) and Weber et al. (2020), have highlighted the shortcomings of the Mualem-van Genuchten model (1976; 1980), particularly under dry conditions. With the residual water content θ_r as a fitted parameter, the model implies that the water content would never be lower than that value. It also focuses on hydraulic conductivity driven by capillarity and fails to

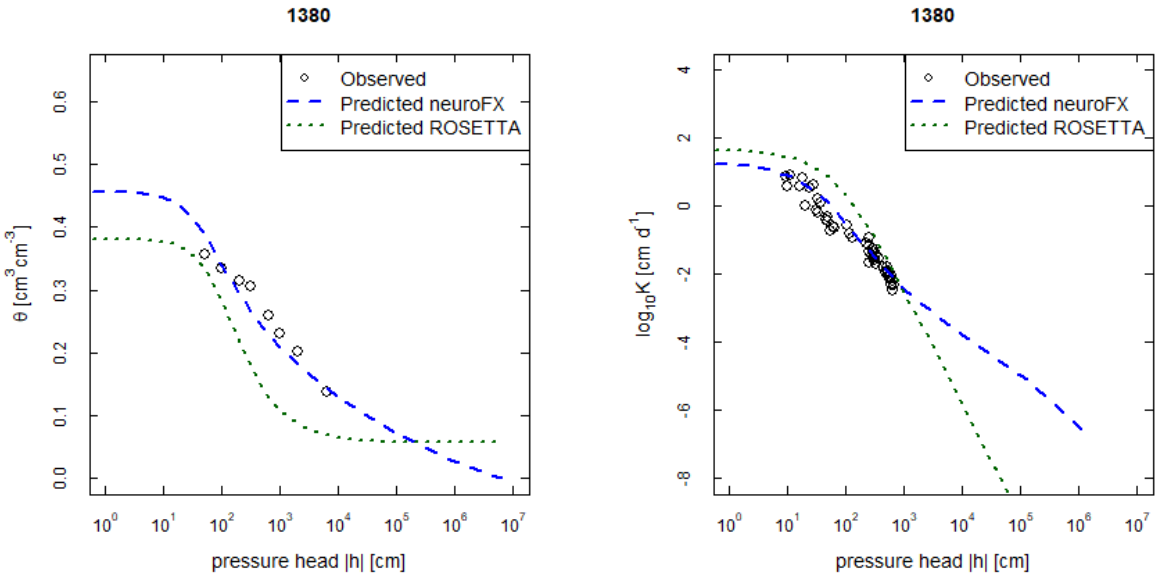
provide a reliable description of water retention and conductivity across the complete range of soil water content levels.

Rudiyanto et al. (2021) employed a comprehensive water retention and hydraulic conductivity model, referred to as the FXW model. It is based on the Fredlund and Xing (1994) water retention model and the hydraulic conductivity model of Wang et al. (2018). The FXW model can calculate the retention and hydraulic conductivity curves across the entire range of matric heads, from saturation to complete dryness. The water retention follows a series of functions that aim to scale the water content from saturation $\theta_s$to complete dryness ($\theta=0$¿at a defined pressure head h_0 at -6.3 x 10^6 cm, as shown in Eq. (A1) – (A6).

While more complicated, the number of parameters of the water retention curve [θ_s, α, n, m¿is the same as the van Genuchten model. Rudiyanto et al. (2021) developed a PTF called neuroFX that predicts parameters of the water retention curve using a neural network that takes sand, silt, clay, and bulk density as inputs. The loss function of the neural networks is defined in terms of measured versus predicted water content (Eq. A7).

Once the water retention parameters PTF was predicted [$\hat{\theta}_s$, $\hat{\alpha}$, $\hat{n}$, $\hat{m}$¿, the predicted parameters were used to calculate the effective saturation S_{ek}. The FXW parameters for hydraulic conductivity [log (K_s), L] were then estimated using another neural network function trained to minimise the difference in hydraulic conductivity values (Eq. A8).

These PTFs were shown to describe water retention and hydraulic conductivity more accurately than conventional PTFs. In sandy to loamy soils, conventional PTFs trained to predict the Mualem-van Genuchten parameters (ROSETTA) show an under-prediction of hydraulic conductivity in the dry range by several orders of magnitude (Zhang and Schaap, 2017). Moreover, ROSETTA produced non-zero water content at the dry end. The neuroFX PTF fits both water retention and hydraulic conductivity data well across the entire range of water contents (Figure 10).



774

775 Figure 10. An example of water retention and hydraulic conductivity curve of a
776 sandy loam predicted with PTF using neuroFX compared to the conventional Rosetta
777 model. The Rosetta model uses the Mualem-van Genuchten model, where the water
778 content does not reach zero as the soil is drying and the conductivity drops rapidly
779 at dry potentials.

780

781 In the direct neural networks PTFs, the neural networks are trained to model the
782 water retention function directly (Haghverdi et al., 2012). Since the neural network
783 learns the shape of the retention curve solely from measurements, the performance
784 of such PTFs is highly dependent on the quality, density, and distribution of the soil
785 water retention curve measurements within the training set (Haghverdi et al.,
786 2014). Norouzi et al. (2024) addressed this issue by imposing physical constraints
787 on the relationship between the pressure head in the input layer and water content
788 in the output layer. Specifically, four constraints were imposed: a monotonically
789 decreasing constraint between $\log(|h|)$ and water content (θ), enforcing linear
790 behaviour at the dry end of the retention curve, setting a specified range for the
791 pressure head at zero water content (h_0), and enforcing a constant water content
792 constraint above air-entry pressure. The loss function used for training the neural
793 network is given as:

$$J=\frac{\lambda_1}{N_{wet}}\sum_{i=1}^{N_{wet}}\left[\hat{\theta}^{(i)}-\theta^{(i)}\right]^2+\frac{\lambda_2}{N_{dry}}\sum_{i=1}^{N_{dry}}\left[\hat{\theta}^{(i)}-\theta^{(i)}\right]^2+\frac{\lambda_3}{S_1}\sum_{i=1}^{S_1}¿¿¿¿¿ \quad (15)$$

where pF is defined as the logarithm of the absolute value of the pressure head in centimetres. The first two terms focus on the mean squared error (MSE) between predicted and measured volumetric water contents, differentiated by wet-end ($pF \leq 4.2$) and dry-end ($pF > 4.2$) conditions. The parameters N_{wet} and N_{dry} are defined as the number of training examples from the wet-end and dry-end, respectively. The next four terms ensure the model adheres to physical laws. In particular, the third term enforces linearity at the dry end by setting the second derivatives in that region to zero for set 1 of the residual points (S_1). These residual points are specific combinations of data points generated within the input space (including sand, silt, clay, SOC, bulk density, and pF) that are used to enforce physical laws. The fourth and fifth terms bound the range of pressure head at zero water content using sets 2 and 3 of residual points, and the last term forces the water content to remain constant above the air-entry pressure using set 4 of the residual points. The monotonicity constraint is enforced by constructing inherently monotonic neural network architectures (Runje and Shankaranarayana, 2023). The Lambdas (λ) are weights that determine the relative contribution of each term in the loss function. The resulting neural network PTF is capable of predicting a non-specific form of the soil water retention curve from saturation to dryness and is differentiable with respect to the pressure head.

Implications and prospect: Overall, rather than predicting parameters of a soil function independently, incorporating the physical model in the training process can guide and constrain the ML models to predict physically-based values more accurately. Additional criteria could be added to the loss function to impose physical constraints. For example, the predicted soil water retention curve could be constrained to satisfy a realistic soil evaporation characteristic length calculated from the same water retention parameters. The characteristic length values must be in a realistic range (e.g. < 1 m) due to the limitation of capillary continuity of an evaporating soil surface (Or, 2020). This approach could also be used to map soil water retention curves. If we have observations of water retention data over an area, we could predict the water retention parameters from spatial covariates by

minimising the observed water retention data using Eq. (A7) or Eq. (15). Yang et al. (2015) provided an example of this using Bayesian hierarchical models.

3.4 Modelling soil properties in space and time

Modelling dynamic soil properties is crucial for understanding how soil changes over time and for improving land management practices. Static soil properties, which are assumed to be relatively constant over time, are mapped based on their spatial relationships with the landscape. In contrast, dynamic properties, such as soil water content, SOC, and nutrient availability, vary with time due to environmental and anthropogenic factors. Some properties change more rapidly than others (e.g. soil temperature versus soil pH), making it important to gauge the timescale of their prediction. Process-based models are effective in accounting for major soil processes within specific soil profiles or layers, but they require calibration to local conditions. The spatial application of these models can be challenging due to limited data for model initialisation and parameterisation and significant computational demands. On the other hand, ML models excel in spatial modelling but lack the capability to simulate processes.

To model the dynamic soil properties in space and time using the SoilML framework, several techniques can be used:

- Residual models: This approach involves using a ML model to predict the residuals of a physical-based model. It involves learning the errors in the physical-based model prediction as compared to observations and using this information to correct the predictions of the physical model (Willard et al., 2020). This residual modelling approach only learns what components are missing from the physical model and does not incorporate any informed knowledge.

- Meta or surrogate models: This approach involves using ML models to emulate physical-based models. This involves generating scenarios of various input soil and climate variables and running them through a simulation model to obtain simulation results that can be used as training data. An ML model is then trained to model the output as a function of these inputs (Perlman et al., 2014). The ML could identify the

sensitivity of the physical model and key variables influencing the model output, identifying under- or overrepresented inputs (Luo et al., 2019).

- Hybrid models (combination of ML and process-based models): Integrating ML models with process-based models can significantly improve the capacity to model soil properties across space and time. Soil data are often spatially well represented but temporally sparse. One approach involves using the outputs of the process-based model as additional training data (both spatially and temporally) for the ML model (Ma et al., 2019). Zhang et al. (2023) and Zhang et al. (2024) demonstrated that incorporating process-based model outputs as supplementary training data for ML models leads to higher prediction accuracy for soil carbon than using standalone ML models. Another approach is to use the output of process-based models over an area as a dynamic covariate, in combination with other static covariates, in an ML model. Xie et al. (2022) integrated the predictions of process-based soil carbon as a dynamic covariate into a DSM model and found improved prediction accuracy compared to both the standalone ML models and process-based models.

- Physics-Informed Neural Networks (PINN): Neural networks are used as function approximators by embedding physical laws in the learning process. The physical laws can be described using partial differential equations (PDE), allowing neural networks to model complex behaviours and dynamics accurately. PINN uses the backpropagation method of neural networks to calculate the partial derivative of the differential equations, and thus, the neural network solution adheres to the physical equations and observations during training (Bandai and Ghezzehei, 2021; Cai et al., 2021). Applications of PINN in soil studies include retrieval of soil moisture using GNSS reflectometry (Kilane, 2024) and soil water and heat flow (Wang et al., 2023b).

3.4.1 Case study: Modelling soil water flow using Physics-Informed Neural Network (PINN)

Conventional approaches: Soil moisture dynamics can be described by the Richardson-Richards' equation (Richards, 1931; Richardson, 1922), which is based on the conservation of mass and the Darcy-Buckingham law (Buckingham, 1907). The Richards equation incorporates water retention curves and hydraulic

conductivity functions to encode macroscopic soil hydraulic properties on the scale of interest. Commonly, parametric models (e.g., the Mualem- van Genuchten model) are used to represent the soil hydraulic functions. Their parameters are estimated via inverse modelling, where the parameters are adjusted by repeatedly solving Richards' equation to match the model output with the data. A widely used software, HYDRUS, has such an inverse modelling capability, where a finite element method solves Richards' equation (Šimůnek et al., 2016). A limitation of this approach is the inflexibility of the parametric models used to represent the soil hydraulic functions. For example, if the Mualem-van Genuchten model is used as the parametric model, inverse modelling would fail if the target soil's water retention curve exhibits a bimodal shape.

SoilML: Several studies have proposed a physics-informed neural network (PINN) approach for inverse modelling based on Richards' equation to improve the capability to analyse soil moisture data. In the original PINNs proposed by Raissi et al. (2019), fully-connected neural networks are used to represent the solution to partial differential equations as a function of the temporal and spatial coordinates. The neural networks are trained to minimise a loss function consisting of an observation constraint term and a PDE residual term (see Eq. 1). The PDE residual term can be computed by automatic differentiation (Baydin et al., 2018), which is implemented in the neural networks framework. Tartakovsky et al. (2020) employed PINNs to estimate the hydraulic conductivity function for a time-independent two-dimensional Richards' equation. Subsequently, Bandai and Ghezzehei (2021) developed PINNs for the time-dependent one-dimensional Richards' equation to estimate both water retention curves and hydraulic conductivity functions. In their framework (Figure 11), two monotonically constrained neural networks (Daniels and Velikova, 2010) are used to represent the soil hydraulic functions.

Through numerical experiments, they demonstrated that the PINNs framework has the potential to estimate soil hydraulic functions without initial and boundary conditions. Furthermore, several studies have improved upon their PINNs framework. To improve the stability of PINNs against sparse and noisy data, Depina et al. (2021) replaced the monotonic neural networks with the Mualem-van Genuchten model and estimated the model's parameters via a global optimisation algorithm. They validated their approach using both synthetic data and laboratory

infiltration experimental data. The model has been extended to model layered soils (Bandai and Ghezzehei, 2022). Recently, the PINNs approach has been extended to multi-physics problems in vadose zone hydrology, such as solute transport in unsaturated soils (Haruzi and Moreno, 2023) and coupled heat and water transport (Wang et al., 2023b).

While PINNs have shown some success using synthetic data and laboratory experimental data, it remains difficult to train PINNs and obtain consistent results with field-observed soil moisture data due to the limited amount and accuracy of the data. Additionally, training PINNs for long temporal domains is challenging because the original formulation of PINNs does not encode the temporal causality of dynamical systems. Although many methods have been proposed to alleviate those issues, as discussed in Wang et al. (2023a), they have not yet been applied to soil processes.

As an alternative approach, Bandai et al. (2024) proposed a hybrid method where the Richards' equation is solved using a traditional numerical method (i.e., finite volume method with Backward Euler method), and neural networks are embedded in the numerical model to represent soil hydraulic functions. This approach leverages the flexibility of neural networks to represent unknown functions in physics-based models (e.g., soil hydraulic functions) while ensuring the basic physics encoded in Richards' equation is maintained. This contrasts with the PINNs approaches, where Richards' equation is enforced in a soft manner as a loss term in the loss function, and therefore, the basic physics laws, such as the conservation of mass, Buckingham-Darcy law, and temporal causality, are not guaranteed. They demonstrated that their neural network approach better fit infiltration experimental data than using the Mualem- van Genuchten model.

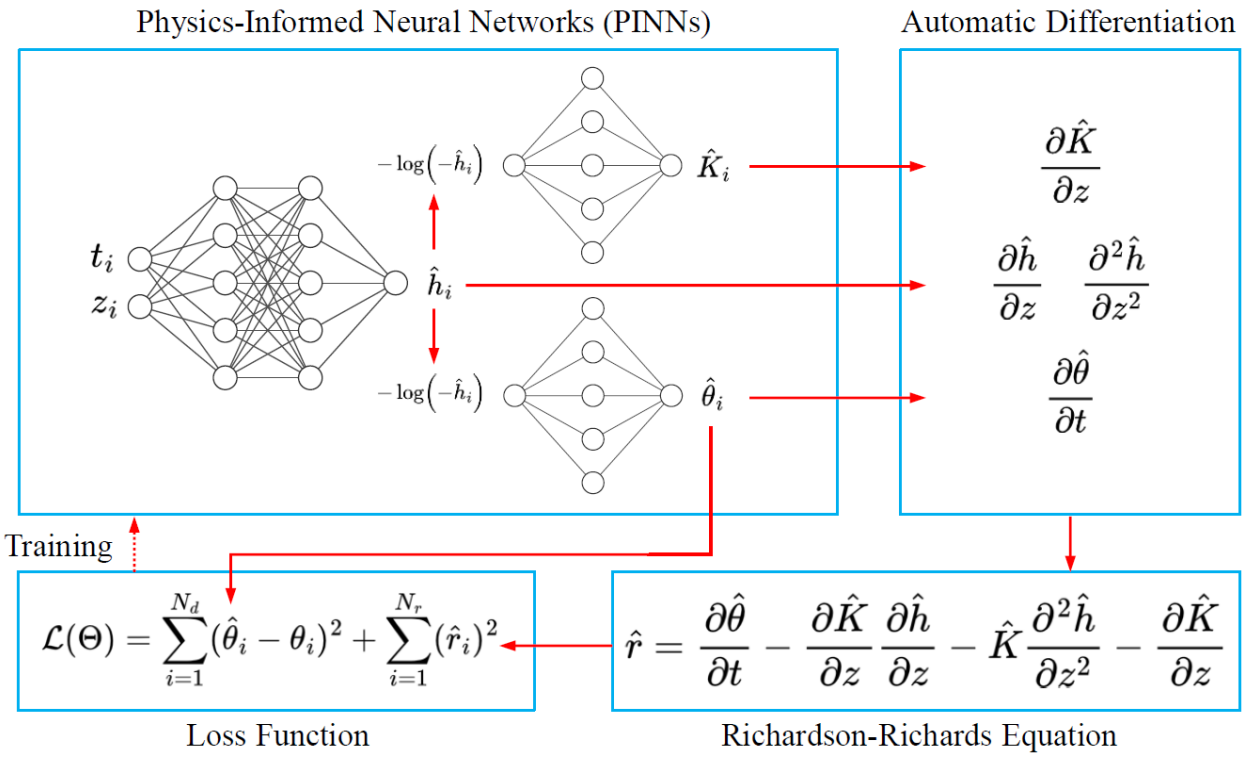


Figure 11. Physics-informed neural networks for the Richardson-Richards equation. The temporal and space coordinates (*t* and *z*) are fed into a fully connected neural network (a) to calculate the water potential h, which is then further converted into the hydraulic conductivity *K* and the volumetric water content θ by two monotonic neural networks ((b) and (c)), respectively. The three neural networks are trained simultaneously by minimising the loss function consisting of the data misfit term and the residual of the Richardson-Richards equation (Bandai and Ghezzehei, 2021).

3.4.2 Case study: Soil temperature modelling

Soil temperature is influenced by various soil properties, such as thermal conductivity and heat capacity, which are affected by factors like bulk density, moisture content, and organic matter (Jury and Horton, 2004). While ML models are often used to predict soil temperature based on historical data, these models can be limited by their reliance on specific observation periods and may not fully capture the underlying causes of temperature variations (Lembrechts et al., 2022). Although

ML algorithms can identify nonlinear relationships between soil temperature and air temperature along with other climate variables, they are generally incapable of showing the physical processes involved. Another drawback of uninformed ML models is their reliance on large amounts of data for reliable calibration and lack generalisability. Time series of soil temperature data are commonly recorded at meteorological stations, providing complete temporal coverage but sparse spatial coverage.

Soil temperature is governed by soil thermal properties, which vary with soil moisture level (assuming constant organic matter and bulk density over time) (Ochsner et al., 2001). The soil heat capacity and thermal conductivity determine the heat flow rate and, consequently, temperature change and fluctuations over time. Considering a static moisture level (i.e., at field capacity or wilting point), soil heat flow rate can vary due to spatial variations in soil bulk density, organic matter content, texture, and mineralogy, as these variables affect soil thermal diffusivity even under constant solar radiation and other environmental conditions.

Soil heat capacity and thermal conductivity can be used to calculate soil temperature in space and time using a physical rules-informed model such as the standard heat flow equation. The volumetric heat capacity of soil is defined as the amount of heat required to raise a unit volume of soil by one degree of temperature. As soil is a composite of air, water, and solid materials, soil heat capacity is described by the heat capacities of all the constituents, weighted by their volumetric fractions. Thus, volumetric soil heat capacity can be expressed as:

$$C_{soil} = X_a C_a + X_w C_w + \sum_{j=1}^{N} X_{sj} C_{sj} \qquad (16)$$

where X refers to the volume fraction, C is volumetric heat capacity, and the subscripts a, w, and sj refer to air, water, and solid constituent j (for N different solid materials in the soil). Soil thermal conductivity quantifies the rate at which heat energy is conducted through a unit area of soil under a unit temperature gradient in a direction perpendicular to the area. While soil thermal conductivity can be directly measured, it can also be estimated using PTFs (e.g. He et al., 2020; Wessolek et al., 2023; Zhang and Wang, 2017).

The amount of thermal energy that moves through an area of soil in a unit of time is known as soil heat flux or heat flux density. The ability of a soil to conduct heat determines how fast its temperature changes during the day or between seasons. The magnitude of this heat energy is a component of the soil surface energy balance, which varies with surface cover, moisture content, and solar irradiance. Heat energy is transported through soil by several mechanisms, including conduction, convection of heat by flowing liquid water and moving air, convection of latent heat, and radiation. However, the most important heat transfer in soil is by conduction, which refers to the heat transported by molecular collisions. The conductive heat flux for a pure substance in one dimension is described by Fourier's law:

$$J_{HC} = \lambda \frac{dT}{dz} \qquad (17)$$

where J_{HC} is the amount of thermal energy, λ is thermal conductivity, T is temperature, and z is soil depth. Combined with the continuity equation, we have the general heat transport equation that describes the change in temperature with time (Carslaw and Jaeger, 1959):

$$C \frac{\partial T}{\partial t} = \lambda \frac{\partial^2 T}{\partial z^2} \qquad (18)$$

Figure 12 demonstrates that using physics-based equations, we could calculate soil temperature from air temperature directly by considering soil thermal properties. Since heat transport governs soil temperature, a danger of blind use of ML in soil temperature modelling is the lack of physical rules in the prediction. For example, the boundary conditions (at the soil surface and different depths), temporal fluctuations of surface temperature, and heat flow variation to varying depths due to differences in soil thermal properties can significantly affect soil temperature estimation (Cichota et al., 2004; Gao et al., 2017; Ouzzane et al., 2014).

Implications and prospects: The nature of soil temperature dynamics and its spatio-temporal variations often prevent ML models from recognising the underlying phenomena and understanding why they vary at different scales. While efforts have been made to incorporate physical knowledge into ML to make it interpretable (e.g. Abimbola et al., 2021; Li et al., 2022), the specific physical rules driving soil

temperature within ML are still not well recognised. Xie et al. (2024) derived a PINN model for soil temperature prediction based on the heat transport equation (Eq. 18). Trained using historical soil temperature data, they were able to predict one-dimensional soil temperature at multiple soil depths accurately. ML models incorporating remote sensing products (such as land cover and land surface temperature) can help determine related soil and environmental properties that are not considered in the physical model. These properties include land cover, the rate of heat transport from the atmosphere, and other factors influencing soil temperature. Future work should focus on combining physical rules within ML algorithms to improve the accuracy and reliability of soil temperature predictions.

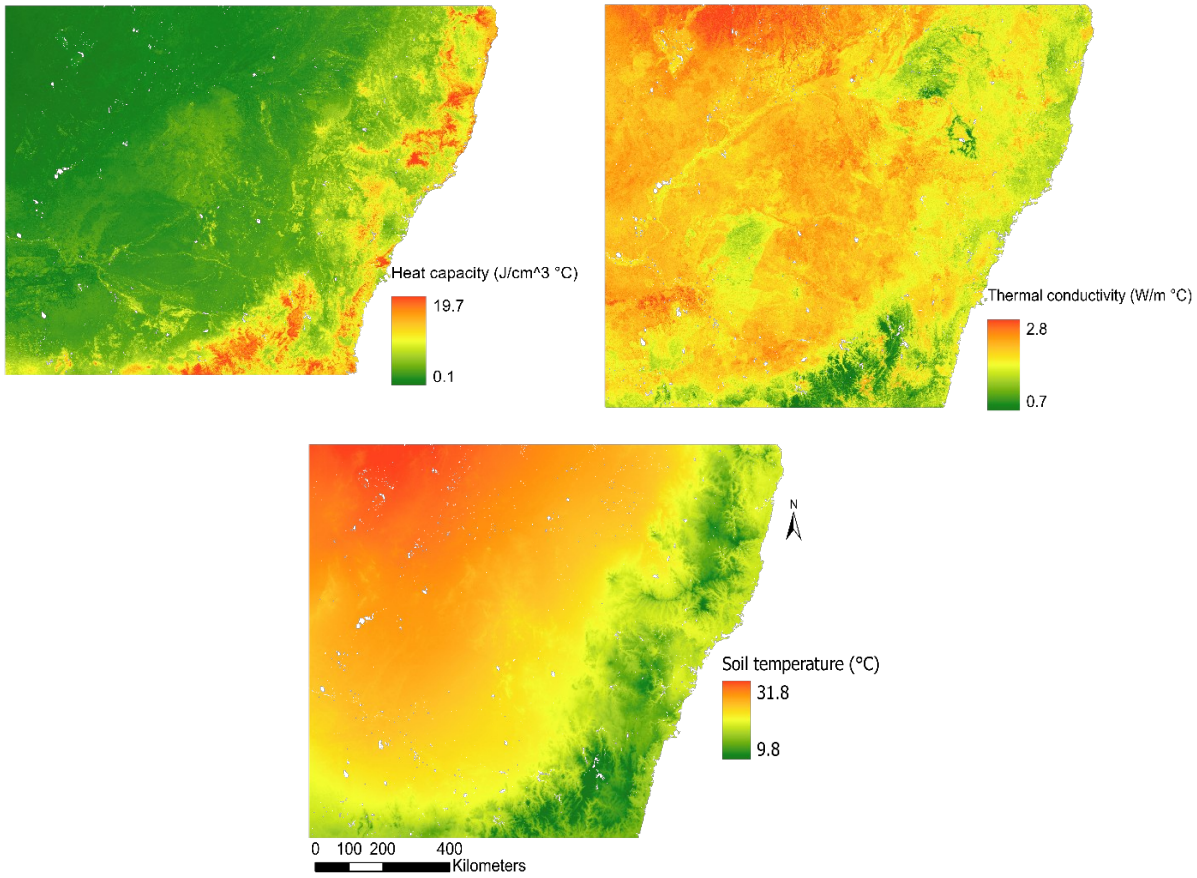


Figure 12. The figure shows an example of soil heat capacity and thermal conductivity for New South Wales state, Australia, at 5–15 cm soil depth layer with a moisture level of 60% of the field capacity. The lower map is the February 2022

average temperature map for the 5–15cm soil depth, obtained using a steady-state analytical method.

3.4.3 Prospective case study of modelling soil carbon dynamics

Conventional approaches: There is increasing interest in modelling SOC dynamics and SOC changes to infer carbon emission or sequestration from the atmosphere. Both physical rules-based and ML models are used to estimate SOC changes across regions. Physical rules-based models typically define several SOC pools, each characterised by a mean residence time derived from first-order kinetics. Conversely, ML models are often trained on sparse temporal soil data along with both static and dynamic covariates (Sun et al., 2021; Yang et al., 2022). Static covariates include soil characteristics, topography, and long-term climate patterns. Dynamic covariates often involve temporal data like land use or land cover and vegetation indices from remote sensing images. There are also cumulative temporal indices such as total rainfall since a specific period, and years since land cover changes to reflect temporal dynamics (Padarian et al., 2022b). These dynamic covariates track SOC dynamics from one state to another to infer SOC changes. However, the time scale of the changes in the remotely-sensed images is not aligned with the SOC dynamic changes, and fails to explain the underlying processes behind SOC changes. For example, a change in land cover from forest to cropping field will cause SOC to decline rapidly but this process could take several years to reach a steady state. Similarly, a change from cropping field to pasture could accumulate SOC slowly and take several years to achieve equilibrium. In addition, SOC changes can occur over shorter periods, such as crop rotation (Fang et al., 2018), or longer periods, such as decomposition. Furthermore, surface conditions detected by remote sensing may not adequately represent subsurface processes. Integrating these processes into models is crucial for a more comprehensive understanding of SOC changes.

SoilML: Physical rules-based SOC models often struggle to accurately resolve spatial information because decomposition constants may vary with soil types and topography. Conversely, ML models, lacking process-based insights, tend to produce abrupt changes in SOC when there is a shift in land use from one period to the next.

Physical rules-based approaches can be particularly useful in addressing the problem of limited observational data in soil carbon dynamics modelling. We can create an ML model that allows SOC dynamics to follow physical rules. Here, we provide a framework for integrating various soil properties and environmental factors, offering a way to enhance model reliability in predicting soil carbon changes.

SOC is observed in space and time $C_{x,t}$ and can be modelled as a mass balance of production and input (I), decomposition (k). The evolution of the organic carbon content (C) for a particular soil depth through space and time can therefore be expressed as (Andrén and Kätterer, 1997):

$$\frac{dC}{dt} = I - k_1 C_1 + h k_1 C_1 - k_2 C_2 \qquad (19)$$

Soil carbon change over time ($\frac{dC}{dt}$) can be modelled as consisting of a fast pool (C_1) and a slow pool (C_2). Carbon input I enters the fast pool with a rate constant of k_1, which in turn becomes humified and mineral-associated at a rate of h into the slow pool, which has a rate constant of k_2. The input I depends on the types of organic matter, above and below-ground, climate, soil type, depth, and management practices. The humification and decomposition constants vary in space and time, influenced by temperature, clay content, and moisture levels.

First, we build a neural network model which will predict parameters $\phi = [h, k_1, k_2]$ from soil characteristics and factors related to climate, topography, vegetation and human activities:

$$\phi = f\left(soil, topography, climate, vegetation\right) \qquad (20)$$

A second neural network can be constructed to predict soil C in space and time that conforms to the C dynamics equation. The neural networks would incorporate static inputs such as soil texture, topography, long-term mean rainfall, and temperature, along with dynamic inputs such as land use and vegetation indices, together with output from the first neural network:

$$C_{x,t} = f\left(soil, topography, climate, vegetation, I_{x,t}, k_{1x}, k_{2x}, h_x\right) \qquad (21)$$

where the input I is a function of:

$$I_{x,t} = f\left(soil, topography, climate, vegetation\right) \quad (22)$$

The network can be trained based on observed C concentration, with a loss function:

$$L = w_1 \sum \left(C_{x,t} - \widehat{C}_{x,t}\right)^2 + ¿ w_2 \sum \left(\frac{dC}{dt} + \frac{d\widehat{C}}{dt}\right)^2 ¿ \quad (23)$$

The first term accounts for the sparsely observed carbon in space and time, and the second term provides a constraint that the model will adhere to the dynamics defined in Equation (19), based on a series of long-term experimental data or simulated data. The terms w_1 and w_2 refer to the weights for the first and second terms.

4. Discussion, Assumptions and Limitations

We have demonstrated a variety of forms of soil science knowledge and how they can be incorporated into the ML training process. Here, we discuss aspects of ML that can improve soil science understanding.

Soil is a unique 3-D volume and ML models should be soil science-informed. It is important to design ML model applications that specifically accommodate the multidimensional nature of soil (Poggio and Gimona, 2014). Soil is not just a simple substrate but a three-dimensional body with unique properties varying by depth. Thus, when developing ML models, they need to be soil science-informed, incorporating architecture, variables, and data layers that reflect the unique characteristics of the soil.

Modify the ML models to suit our needs, not modify our data to suit ML needs. This means a shift in how we approach ML development. Traditionally, much of the focus in ML has been on adjusting, filtering, or transforming soil data to fit the requirements of existing algorithms and models. This approach could lead to loss of information or oversimplification of soil data. Instead, we should adapt ML models to work with soil data, such as modifying and regularising their loss functions. This soil-centric approach in model development ensures that the technology serves the specific needs of its applications, rather than forcing data into predefined, potentially limiting frameworks.

Overparametrisation and interpolation. Classical statistics promotes Occam's razor principle, which suggests selecting the hypothesis with the fewest assumptions among competing hypotheses. This translates to the preference for models with fewer parameters, as they are easier to interpret and less likely to overfit the data. However, ML models are usually overparametrised, having many more parameters compared to the size of the training data (Belkin, 2021). Some models (such as random forest and boosting methods) are designed to perfectly fit (interpolate) the training data, which is usually noisy. Interpolating noisy data using ML models, traditionally associated with detrimental overfitting, has been demonstrated to perform well on test data (Belkin, 2021).

Belkin et al. (2019) proposed the double descent phenomenon in ML where the error, when plotted against model complexity, shows a two-phase behaviour contrary to the traditional U-shaped bias-variance trade-off (Figure 13). Initially, as model complexity increases, the test error decreases up to an interpolation threshold where the model perfectly fits the training data, causing the test error to peak. Unexpectedly, if the complexity continues to increase beyond this point, the test error decreases again, leading to a second descent. This phenomenon has been observed across various ML models. However, the ability of ML modes to interpolate does not necessarily mean the model is more accurate and generalisable. Practices that focus on regularising the training rather than achieving a perfect fit are being advocated (See Box 3). Currently, there is still a lack of metrics to quantify an ML model’s complexity with respect to its ability to generalise (Dar et al., 2021). Soil science data are usually relatively small compared to disciplines such as image or language processing. There is still a lack of understanding of the interpolation effect of ML models trained on a relatively small dataset (e.g. less than 100 observations). Overparametrisation and the double descent phenomena do not necessarily lead to improved accuracy. Thus, an independent validation dataset is needed to evaluate the generalisability of the ML models.

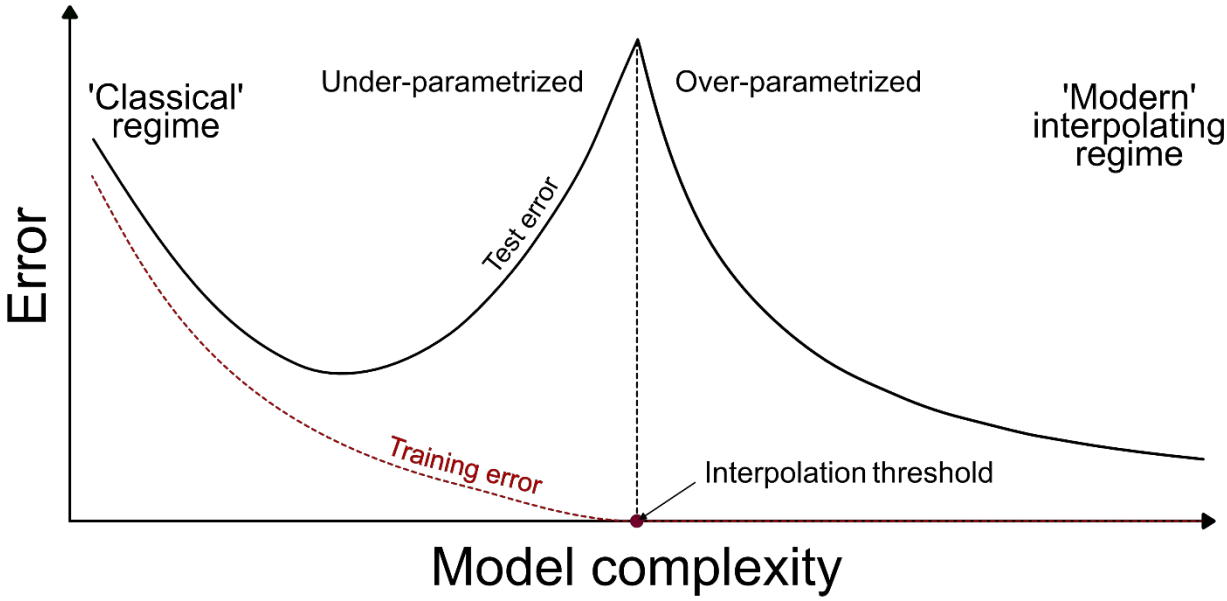


Figure 13. The double-descent error curve with “classical” and “modern interpolating” regimes, showing training error (red dashed line) and test error (solid line) as a function of model complexity. The left curve is the classical U-shaped risk curve arising from the bias–variance trade-off. The right curve, separated by the interpolation threshold, represents ML models with interpolation or zero training error. From Belkin et al. (2019).

Uncertainty analysis. This analysis helps acknowledge the limits of models. In soil science, these challenges are magnified due to the limited, sparse, and often heterogeneous nature of soil data (Libohova et al., 2019). In ML, uncertainties can stem from uncertain data and incomplete knowledge. Although some studies report prediction intervals and confidence levels, a comprehensive approach to uncertainty quantification remains a challenge, emphasising the need for better methods to evaluate and communicate the reliability of soil property predictions. Uncertainty quantification is essential for assessing prediction reliability, especially under unseen scenarios. Bayesian approaches are the standard method; however, the computational demands of modelling the full posterior distribution are very high. This challenge can be mitigated by using dropout techniques to approximate the posterior distribution. For example, Padarian et al. (2022a) used the Monte Carlo

dropout method for the prediction of SOC using vis-NIR-SWIR data, and demonstrated its capability to identify large uncertainty when new data presented is different from the data used during training.

Interpretability. Soil scientists are also interested in using ML models to have a better understanding of soil processes and formulate hypotheses (Padarian et al., 2020a; Wadoux and Molnar, 2022). The prevailing assumption in ML models is that, with sufficiently covariate capturing spatial dependence relationships, the spatial patterns of soils can be predicted, and the drivers of those spatial patterns can be identified (Bui, 2004; Bui et al., 2020). Interpretable ML models help clarify the significance of specific predictors in estimating soil properties, tackling the "black box" nature of many ML algorithms (Roscher et al., 2020). *Post hoc* analysis—including interpreting, visualising, and evaluating ML predictions—can determine how well a model aligns with established soil science knowledge. However, while ML primarily aims to minimise prediction error, soil scientists are more interested in uncovering underlying processes. Interpretability relies on human judgment, and just because a model highlights certain predictors as important doesn't necessarily mean they cause the observed effects or offer new insights. Techniques like Shapley values, while useful for generating hypotheses, could lead to biased conclusions if not carefully handled. Therefore, interpretability should not be conflated as a verification of a model’s generalisability. It also should not be viewed as a confirmation of a model’s accuracy but must be integrated with domain knowledge to validate and enhance predictions thoroughly.

Dynamic soil properties prediction in space and time. Soil data are often incomplete, noisy, and sparse in space and time. ML models, especially tree models, often struggle to interpolate these sparse data effectively. To overcome these limitations, it is beneficial to define the model's structure consistent with soil science principles, incorporate prior knowledge of soil in space (and time), and define a loss function that obeys physical principles. ML models are also often used to predict soil carbon fate under future climate scenarios, yet ML models usually perform poorly when used for extrapolation or predicting unseen or rare events.

Finally, SoilML should ensure *Reproducibility*, which involves a systematic approach to documenting, sharing, and verifying the processes and results of analysis. This

includes making the codes, and methodologies accessible to other researchers so that they can replicate the findings. Thorough evaluations are also crucial to identify any biases inherent in the model, data, or methodology. This might involve testing the model under various conditions, using diverse datasets to check for consistency, and scrutinising the assumptions underlying the model's predictions.

Box 3. Reducing overparametrisation in ML models

As ML is a data-hungry model, efforts are being made to reduce overparametrisation by either increasing the number of observations or simplifying the model architecture. Some of the approaches include (Dar et al., 2021):

- Data augmentation, generating additional synthetic data can increase the diversity of training data, which can help reduce overparametrisation. Data augmentation could include applying various covariates transformations (such as rotations) to existing data or based on prior information (see Figure 4).

- Transfer learning, involves using a pre-trained deep neural network (DNN) on a related problem to improve training efficiency and performance on a target problem with less data (Padarian et al., 2019a). Transfer learning is done by transferring and fine-tuning one or more layers from the source DNN. This approach can mitigate overparametrisation by leveraging the learned parameters from the source task.

- Pruning models, pruning highly parametrised ML models into less complex forms can improve the trade-off between generalisation performance and computational requirements. This approach is also useful for applications with limited storage space, computation time, and energy consumption.

5. Outlook

SoilML have been applied in four key areas: digital soil mapping, soil spectroscopy, pedotransfer functions, and dynamic soil property modeling. These applications

demonstrate how SoilML enhances model accuracy, improves interpretability, and preserves the principles of soil science.

ML approaches have successfully produced digital soil maps of continents and the world, but there is still a lack of modelling soil processes. Soil processes are commonly predicted statically using ML models without considering temporal processes. Currently most ML-derived outputs of soil maps are used as inputs of baseline data for assessing future conditions using process-based models. While physics-informed ML models are growing in environmental and earth modelling, their application in soil science is still minimal. With an increasing demand for quantifying soil functions, there is a need and potential to upscale our physical and chemical process models to a larger extent. SoilML has the potential to accelerate advancements in soil science by integrating soil-specific knowledge into the ML process. This can be achieved through the use of observational priors, tailored model structures, and informed loss functions that incorporate physical constraints and coherency rules.

There are still practical challenges of SoilML include high computational demands, the need for soil-specific priors, and difficulties in integrating multi-source data with varying spatial and temporal resolutions. Effective collaboration among the communities, including process-based modellers, pedometricians, remote sensing experts, and data scientists, is essential to advance the growth of SoilML for soil security assessment.

Acknowledgements

BM acknowledges funding from ARC Discovery Project Forecasting Soil Conditions, DP200102542.

Declaration of Generative AI in the writing process

In the preparation of this paper, ChatGPT 4o, was employed to assist with improving the clarity, grammar, and overall quality of the English language. Following its use, the authors thoroughly reviewed and edited the content to ensure accuracy and alignment with the intended meaning. The AI's involvement was limited to language

refinement, and no content generation, interpretation of data, or intellectual contributions were made by the AI. The responsibility for the scientific content, and conclusions of this paper lies entirely with the authors.

Appendix

The following are the equations of the water retention and hydraulic conductivity curves according to the FXW model.

The water retention curve according to Fredlund and Xing (1994):

$$\theta(h)=\theta_s S_e(h) \qquad \text{(A1)}$$

where θ_s is the saturated water content and $S_e(h)$ is the effective saturation, calculated as:

$$S_e(h)=C_f(h)\Gamma(h) \qquad \text{(A2)}$$

$$\text{With } C_f(h)=\left[1-\frac{\ln\left(1+\frac{h}{h_r}\right)}{\ln\left(1+\frac{h_0}{h_r}\right)}\right] \text{ and } \Gamma(h)=\left(\ln\left[\exp(1)+|\alpha h|^n\right]\right)^{-m} \qquad \text{(A3)}$$

The unsaturated hydraulic function, K(h), based on Wang et al. (2018) is defined as:

$$K(h)=K_s K_r(h) \qquad \text{(A4)}$$

where K_s is the saturated hydraulic conductivity and

$$K_r(h)=S_{ek}{}^{L}\gamma^2 \qquad \text{(A5)}$$

$$S_{ek}(h)=\frac{\Gamma(h)-\Gamma(h_0)}{1-\Gamma(h_0)} \text{ and } \gamma=\left[1-\left(1-\Gamma^{\frac{1}{m}}\right)^{1-\frac{1}{n}}\right]^2 \qquad \text{(A6)}$$

The neuroFX model minimises the following functions:

$$\text{Min: } L_\theta=\sum\left(\theta(h)-\hat{\theta}\left(h\vee\hat{\theta}_s,\,\hat{\alpha},\,\hat{n},\hat{m}\right)\right)^2 \qquad \text{(A7)}$$

$$\text{Min: } L_K=\sum\left(\log K(h)-\log\widehat{K}\left(h\vee\hat{\theta}_s,\,\hat{\alpha},\,\hat{n},\hat{m},\widehat{K_s},L\right)\right)^2 \qquad \text{(A8)}$$

References

Abimbola, O.P., Meyer, G.E., Mittelstet, A.R., Rudnick, D.R., Franz, T.E., 2021. Knowledge-guided machine learning for improving daily soil temperature prediction across the United States. Vadose Zone Journal 20(5), e20151.

Andrén, O., Kätterer, T., 1997. ICBM: The introductory carbon balance model for exploration of soil carbon balances. Ecological Applications 7(4), 1226-1236.

Babaeian, E., Sadeghi, M., Jones, S.B., Montzka, C., Vereecken, H., Tuller, M., 2019. Ground, proximal, and satellite remote sensing of soil moisture. Reviews of Geophysics 57(2), 530-616.

Bagnall, D.K., Morgan, C.L.S., Cope, M., Bean, G.M., Cappellazzi, S., Greub, K., Liptzin, D., Norris, C.L., Rieke, E., Tracy, P. et al., 2022. Carbon-sensitive pedotransfer functions for plant available water. Soil Science Society of America Journal 86(3), 612-629.

Bandai, T., Ghezzehei, T.A., 2021. Physics-informed neural networks with monotonicity constraints for Richardson-Richards equation: Estimation of constitutive relationships and soil water flux density from volumetric water content measurements. Water Resources Research 57(2), e2020WR027642.

Bandai, T., Ghezzehei, T.A., 2022. Forward and inverse modeling of water flow in unsaturated soils with discontinuous hydraulic conductivities using physics-informed neural networks with domain decomposition. Hydrol. Earth Syst. Sci. 26(16), 4469-4495.

Bandai, T., Ghezzehei, T.A., Jiang, P., Kidger, P., Chen, X., Steefel, C.I., 2024. Learning constitutive relations from soil moisture data via physically constrained neural networks. Water Resources Research 60(7), e2024WR037318.

Baydin, A.G., Pearlmutter, B.A., Radul, A.A., Siskind, J.M., 2018. Automatic differentiation in machine learning: a survey. J. Mach. Learn. Res. 18(1), 5595-5637.

Behrens, T., Schmidt, K., Zhu, A.X., Scholten, T., 2010. The ConMap approach for terrain-based digital soil mapping. European Journal of Soil Science 61(1), 133-143.

Belkin, M., 2021. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. Acta Numerica 30, 203 - 248.

Belkin, M., Hsu, D., Ma, S., Mandal, S., 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences 116(32), 15849-15854.

Bogrekci, I., S. Lee, W., 2006. Effects of soil moisture content on absorbance spectra of sandy soils in sensing phosphorus concentrations using UV-VIS-NIR spectroscopy. Transactions of the ASABE 49(4), 1175-1180.

Buckingham, E., 1907. Studies on the Movement of Soil Moisture. US Department of Agriculture, Bureau of Soils, Washington DC.

Bui, E.N., 2004. Soil survey as a knowledge system. Geoderma 120(1), 17-26.

Bui, E.N., Searle, R.D., Wilson, P.R., Philip, S.R., Thomas, M., Brough, D., Harms, B., Hill, J.V., Holmes, K., Smolinski, H.J. et al., 2020. Soil surveyor knowledge in digital soil mapping and assessment in Australia. Geoderma Regional 22, e00299.

Cai, S., Wang, Z., Wang, S., Perdikaris, P., Karniadakis, G.E., 2021. Physics-informed neural networks for heat transfer problems. Journal of Heat Transfer 143(6).

Campbell, G., Shiozawa, S., 1992. Prediction of hydraulic properties of soils using particle-size distribution and bulk density data, International workshop on

indirect methods for estimating the hydraulic properties of unsaturated soils. California: University of California, pp. 317-328.
Carslaw, H.S., Jaeger, J.C., 1959. Conduction of Heat in Solids. Second Ed. ed. Oxford University Press, Oxford.
Castaldi, F., Palombo, A., Pascucci, S., Pignatti, S., Santini, F., Casa, R., 2015. Reducing the influence of soil moisture on the estimation of clay from hyperspectral data: a case study using simulated PRISMA data. Remote Sens-Basel 7(11), 15561-15582.
Chang, C.-W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. Soil Science Society of America Journal 65(2), 480-490.
Chaplot, V., Lorentz, S., Podwojewski, P., Jewitt, G., 2010. Digital mapping of A-horizon thickness using the correlation between various soil properties and soil apparent electrical resistivity. Geoderma 157(3), 154-164.
Chen, S., Saby, N.P.A., Martin, M.P., Barthès, B.G., Gomez, C., Shi, Z., Arrouays, D., 2023a. Integrating additional spectroscopically inferred soil data improves the accuracy of digital soil mapping. Geoderma 433, 116467.
Chen, S., Xue, J., Shi, Z., 2023b. Spectral-guided ensemble modelling for soil spectroscopic prediction. Geoderma 437, 116594.
Cichota, R., Elias, E.A., de Jong van Lier, Q., 2004. Testing a finite-difference model for soil heat transfer by comparing numerical and analytical solutions. Environmental Modelling & Software 19(5), 495-506.
Daniels, H., Velikova, M., 2010. Monotone and partially monotone neural networks. IEEE Transactions on Neural Networks 21(6), 906-917.
Dar, Y., Muthukumar, V., Baraniuk, R.G., 2021. A farewell to the bias-variance tradeoff? An overview of the theory of overparameterized machine learning. arXiv preprint.
Depina, I., Jain, S., Mar Valsson, S., Gotovac, H., 2021. Application of physics-informed neural networks to inverse problems in unsaturated groundwater flow. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards 16(1), 21-36.
Eymard, A., Richer-de-Forges, A.C., Martelet, G., Tissoux, H., Bialkowski, A., Dalmasso, M., Chrétien, F., Belletier, D., Ledemé, G., Laloua, D. et al., 2024. Exploring the untapped potential of hand-feel soil texture data for enhancing digital soil mapping: Revealing hidden spatial patterns from field observations. Geoderma 441, 116769.
Fajardo, M., McBratney, A., Whelan, B., 2016. Fuzzy clustering of Vis–NIR spectra for the objective recognition of soil morphological horizons in soil profiles. Geoderma 263, 244-253.
Fang, J., Yu, G., Liu, L., Hu, S., Chapin, F.S., 2018. Climate change, human impacts, and carbon sequestration in China. Proceedings of the National Academy of Sciences 115(16), 4015-4020.
Fredlund, D.G., Xing, A., 1994. Equations for the soil-water characteristic curve. Canadian Geotechnical Journal 31(4), 521-532.
Gao, Z., Russell, E.S., Missik, J.E.C., Huang, M., Chen, X., Strickland, C.E., Clayton, R., Arntzen, E., Ma, Y., Liu, H., 2017. A novel approach to evaluate soil heat flux calculation: An analytical review of nine methods. Journal of Geophysical Research: Atmospheres 122(13), 6934-6949.

Gastaldi, G., Minasny, B., McBratney, A., 2012. Mapping the occurrence and thickness of soil horizons within soil profiles, Digital Soil Assessments and Beyond CRC Press, Balkema London, pp. 145-148.
Grunwald, S., Lowery, B., Rooney, D.J., McSweeney, K., 2001. Profile cone penetrometer data used to distinguish between soil materials. Soil and Tillage Research 62(1), 27-40.
Haghverdi, A., Cornelis, W.M., Ghahraman, B., 2012. A pseudo-continuous neural network approach for developing water retention pedotransfer functions with limited data. Journal of Hydrology 442-443, 46-54.
Haghverdi, A., Öztürk, H.S., Cornelis, W.M., 2014. Revisiting the pseudo continuous pedotransfer function concept: Impact of data quality and data mining method. Geoderma 226-227, 31-38.
Haruzi, P., Moreno, Z., 2023. Modeling water flow and solute transport in unsaturated soils using physics-informed neural networks trained with geoelectrical data. Water Resources Research 59(6), e2023WR034538.
Hassan-Esfahani, L., Torres-Rua, A., Jensen, A., McKee, M., 2015. Assessment of surface soil moisture using high-resolution multi-spectral imagery and artificial neural networks. Remote Sens-Basel 7(3), 2627-2646.
He, H., He, D., Jin, J., Smits, K.M., Dyck, M., Wu, Q., Si, B., Lv, J., 2020. Room for improvement: A review and evaluation of 24 soil thermal conductivity parameterization schemes commonly used in land-surface, hydrological, and soil-vegetation-atmosphere transfer models. Earth-Science Reviews 211, 103419.
Helfenstein, A., Mulder, V.L., Hack-ten Broeke, M.J.D., van Doorn, M., Teuling, K., Walvoort, D.J.J., Heuvelink, G.B.M., 2024. BIS-4D: mapping soil properties and their uncertainties at 25 m resolution in the Netherlands. Earth Syst. Sci. Data 16(6), 2941-2970.
Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62-77.
Holmes, K.W., Griffin, E.A., Odgers, N.P., 2015. Large-area spatial disaggregation of a mosaic of conventional soil maps: evaluation over Western Australia. Soil Research 53(8), 865-880.
Holmes, K.W., Griffin, E.A., van Gool, D., 2021. Digital soil mapping of coarse fragments in southwest Australia: Targeting simple features yields detailed maps. Geoderma 404, 115282.
Hudson, B.D., 1992. The soil survey as paradigm-based science. Soil Science Society of America Journal 56(3), 836-841.
Hughes, P., McBratney, A.B., Huang, J., Minasny, B., Micheli, E., Hempel, J., 2017. Comparisons between USDA Soil Taxonomy and the Australian Soil Classification System I: Data harmonization, calculation of taxonomic distance and inter-taxa variation. Geoderma 307, 198-209.
Hutengs, C., Eisenhauer, N., Schädler, M., Lochner, A., Seidel, M., Vohland, M., 2021. VNIR and MIR spectroscopy of PLFA-derived soil microbial properties and associated soil physicochemical characteristics in an experimental plant diversity gradient. Soil Biology and Biochemistry 160, 108319.
Jury, W.A., Horton, R., 2004. Soil physics. John Wiley & Sons.
Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L., 2021. Physics-informed machine learning. Nature Reviews Physics 3(6), 422-440.

Kashinath, K., Mustafa, M., Albert, A., Wu, J.-L., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A. et al., 2021. Physics-informed machine learning: case studies for weather and climate modelling. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 379(2194), 20200093.

Kilane, N.G., 2024. Application of physics-informed neural network approach in soil moisture retrieval using GNSS reflectometry, York University, Ontario, Canada.

Koch, J., Stisen, S., Refsgaard, J.C., Ernstsen, V., Jakobsen, P.R., Højberg, A.L., 2019. Modeling depth of the redox interface at high resolution at national scale using random forest and residual gaussian simulation. Water Resources Research 55(2), 1451-1469.

Kosugi, K.i., 1996. Lognormal distribution model for unsaturated soil hydraulic properties. Water Resources Research 32(9), 2697-2703.

Kubelka, P., Munk, F., 1931. Ein Beitrag zur Optik der Farbanstriche, 12. Zeitschrift für Technische Physik.

Laborczi, A., Szatmári, G., Kaposi, A.D., Pásztor, L., 2019. Comparison of soil texture maps synthetized from standard depth layers with directly compiled products. Geoderma 352, 360-372.

Lamichhane, S., Kumar, L., Adhikari, K., 2021. Updating the national soil map of Nepal through digital soil mapping. Geoderma 394, 115041.

Lark, R.M., Bishop, T.F.A., Webster, R., 2007. Using expert knowledge with control of false discovery rate to select regressors for prediction of soil properties. Geoderma 138(1), 65-78.

Lebeau, M., Konrad, J.-M., 2010. A new capillary and thin film flow model for predicting the hydraulic conductivity of unsaturated porous media. Water Resources Research 46(12).

Lembrechts, J.J., van den Hoogen, J., Aalto, J., Ashcroft, M.B., De Frenne, P., Kemppinen, J., Kopecký, M., Luoto, M., Maclean, I.M.D., Crowther, T.W. et al., 2022. Global maps of soil temperature. Global Change Biology 28(9), 3110-3144.

Li, Q., Zhu, Y., Shangguan, W., Wang, X., Li, L., Yu, F., 2022. An attention-aware LSTM model for soil moisture and soil temperature prediction. Geoderma 409, 115651.

Li, X., Nieber, J.L., Kumar, V., 2024. Machine learning applications in vadose zone hydrology: A review. Vadose Zone Journal 23(4), e20361.

Libohova, Z., Seybold, C., Adhikari, K., Wills, S., Beaudette, D., Peaslee, S., Lindbo, D., Owens, P.R., 2019. The anatomy of uncertainty for soil pH measurements and predictions: Implications for modellers and practitioners. European Journal of Soil Science 70(1), 185-199.

Liu, W., Baret, F., Xingfa, G., Qingxi, T., Lanfen, Z., Bing, Z., 2002. Relating soil surface moisture to reflectance. Remote Sens Environ 81(2), 238-246.

Lobell, D.B., Asner, G.P., 2002. Moisture Effects on Soil Reflectance. Soil Science Society of America Journal 66(3), 722-727.

Luo, Z., Eady, S., Sharma, B., Grant, T., Liu, D.L., Cowie, A., Farquharson, R., Simmons, A., Crawford, D., Searle, R. et al., 2019. Mapping future soil carbon change and its uncertainty in croplands using simple surrogates of a complex farming system model. Geoderma 337, 311-321.

Ma, Y., Minasny, B., Demattê, J.A.M., McBratney, A.B., 2023. Incorporating soil knowledge into machine-learning prediction of soil properties from soil spectra. European Journal of Soil Science 74(6), e13438.
Ma, Y., Minasny, B., Malone, B.P., Mcbratney, A.B., 2019. Pedology and digital soil mapping (DSM). European Journal of Soil Science 70(2), 216-235.
Ma, Y., Minasny, B., McBratney, A., Poggio, L., Fajardo, M., 2021. Predicting soil properties in 3D: Should depth be a covariate? Geoderma 383, 114794.
McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117(1), 3-52.
Mendonça Santos, M.L., Guenat, C., Bouzelboudjen, M., Golay, F., 2000. Three-dimensional GIS cartography applied to the study of the spatial variation of soil horizons in a Swiss floodplain. Geoderma 97(3), 351-366.
Michéli, E., Láng, V., Owens, P.R., McBratney, A., Hempel, J., 2016. Testing the pedometric evaluation of taxonomic units on soil taxonomy — A step in advancing towards a universal soil classification system. Geoderma 264, 340-349.
Miller, B.A., Koszinski, S., Wehrhan, M., Sommer, M., 2015. Impact of multi-scale predictor selection for modeling soil properties. Geoderma 239-240, 97-106.
Minasny, B., McBratney, A.B., 2002. The neuro-m nethod for fitting neural network parametric pedotransfer functions. Soil Science Society of America Journal 66(2), 352-361.
Minasny, B., McBratney, A.B., 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. Geoderma 142(3), 285-293.
Mualem, Y., 1976. A new model for predicting the hydraulic conductivity of unsaturated porous media. Water Resources Research 12(3), 513-522.
Norouzi, S., Pesch, C., Arthur, E., Norgaard, T., Møldrup, P., Greve, M.H., Beucher, A., Sadeghi, M., Zaresourmanabad, M., Tuller, M. et al., 2024. Physics-informed neural networks for estimating a continuous form of the soil water retention curve from basic soil properties. ESS Open Archive.
Norouzi, S., Sadeghi, M., Tuller, M., Ebrahimian, H., Liaghat, A., Jones, S.B., de Jonge, L.W., 2023. A novel laboratory method for the retrieval of the soil water retention curve from shortwave infrared reflectance. Journal of Hydrology 626, 130284.
Norouzi, S., Sadeghi, M., Tuller, M., Liaghat, A., Jones, S.B., Ebrahimian, H., 2022. A novel physical-empirical model linking shortwave infrared reflectance and soil water retention. Journal of Hydrology 614, 128653.
Ochsner, T.E., Horton, R., Ren, T., 2001. A new perspective on soil thermal properties. Soil science society of America Journal 65(6), 1641-1647.
Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214-215, 91-100.
Or, D., 2020. The tyranny of small scales—On representing soil processes in global land surface models. Water Resources Research 56(6).
Orton, T.G., Pringle, M.J., Bishop, T.F.A., Menzies, N.W., Dang, Y.P., 2020. Increment-averaged kriging for 3-D modelling and mapping soil properties: Combining machine learning and geostatistical methods. Geoderma 361, 114094.
Ouzzane, M., Eslami-Nejad, P., Aidoun, Z., Lamarche, L., 2014. Analysis of the convective heat exchange effect on the undisturbed ground temperature. Solar Energy 108, 340-347.

Padarian, J., McBratney, A.B., Minasny, B., 2020a. Game theory interpretation of digital soil mapping convolutional neural networks. SOIL 6(2), 389-397.
Padarian, J., Minasny, B., McBratney, A.B., 2019a. Transfer learning to localise a continental soil vis-NIR calibration model. Geoderma 340, 279-288.
Padarian, J., Minasny, B., McBratney, A.B., 2019b. Using deep learning for digital soil mapping. SOIL 5(1), 79-89.
Padarian, J., Minasny, B., McBratney, A.B., 2020b. Machine learning and soil sciences: a review aided by machine learning tools. SOIL 6(1), 35-52.
Padarian, J., Minasny, B., McBratney, A.B., 2022a. Assessing the uncertainty of deep learning soil spectral models using Monte Carlo dropout. Geoderma 425, 116063.
Padarian, J., Stockmann, U., Minasny, B., McBratney, A.B., 2022b. Monitoring changes in global soil organic carbon stocks from space. Remote Sens Environ 281, 113260.
Perlman, J., Hijmans, R.J., Horwath, W.R., 2014. A metamodelling approach to estimate global emissions from agricultural soils. Global Ecology and Biogeography 23(8), 912-924.
Peters, A., Durner, W., Iden, S., 2024. The PDI model system for parameterizing soil hydraulic properties. Vadose Zone Journal 23(4), e20338.
Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. SOIL 7(1), 217-240.
Poggio, L., Gimona, A., 2014. National scale 3D modelling of soil organic carbon stocks with uncertainty propagation — An example from Scotland. Geoderma 232-234, 284-299.
Qi, F., Zhu, A.X., 2003. Knowledge discovery from soil maps using inductive learning. International Journal of Geographical Information Science 17(8), 771-795.
Raissi, M., Perdikaris, P., Karniadakis, G.E., 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics 378, 686-707.
Richards, L.A., 1931. Capillary conduction of liquids through porous mediums. Journal of Applied Physics 1, 318-333.
Richardson, L.F., 1922. Weather prediction by numerical process. Cambridge University Press, Cambridge, United Kingdom.
Roscher, R., Bohn, B., Duarte, M.F., Garcke, J., 2020. Explainable machine learning for scientific insights and discoveries. IEEE Access 8, 42200-42216.
Rosin, N.A., Demattê, J.A.M., Poppiel, R.R., Silvero, N.E.Q., Rodriguez-Albarracin, H.S., Rosas, J.T.F., Greschuk, L.T., Bellinaso, H., Minasny, B., Gomez, C. et al., 2023. Mapping Brazilian soil mineralogy using proximal and remote sensing data. Geoderma 432, 116413.
Rudiyanto, Minasny, B., Chaney, N.W., Maggi, F., Goh Eng Giap, S., Shah, R.M., Fiantis, D., Setiawan, B.I., 2021. Pedotransfer functions for estimating soil hydraulic properties from saturation to dryness. Geoderma 403, 115194.
Runje, D., Shankaranarayana, S.M., 2023. Constrained monotonic neural networks, Proceedings of the 40th International Conference on Machine Learning. PMLR, Proceedings of Machine Learning Research, pp. 29338--29353.
Sadeghi, M., Jones, S.B., Philpot, W.D., 2015. A linear physically-based model for remote sensing of soil moisture using short wave infrared bands. Remote Sens Environ 164, 66-76.

Safanelli, J.L., Demattê, J.A.M., Chabrillat, S., Poppiel, R.R., Rizzo, R., Dotto, A.C., Silvero, N.E.Q., Mendes, W.d.S., Bonfatti, B.R., Ruiz, L.F.C. et al., 2021. Leveraging the application of Earth observation data for mapping cropland soils in Brazil. Geoderma 396, 115042.

Seidel, M., Vohland, M., Greenberg, I., Ludwig, B., Ortner, M., Thiele-Bruhn, S., Hutengs, C., 2022. Soil moisture effects on predictive VNIR and MIR modeling of soil organic carbon and clay content. Geoderma 427, 116103.

Šimůnek, J., van Genuchten, M.T., Šejna, M., 2016. Recent Developments and Applications of the HYDRUS Computer Software Packages. Vadose Zone Journal 15(7), vzj2016.2004.0033.

Sun, X.L., Minasny, B., Wang, H.L., Zhao, Y.G., Zhang, G.L., Wu, Y.J., 2021. Spatiotemporal modelling of soil organic matter changes in Jiangsu, China between 1980 and 2006 using INLA-SPDE. Geoderma 384, 114808.

Taghizadeh-Mehrjardi, R., Mahdianpari, M., Mohammadimanesh, F., Behrens, T., Toomanian, N., Scholten, T., Schmidt, K., 2020. Multi-task convolutional neural networks outperformed random forest for mapping soil particle size fractions in central Iran. Geoderma 376, 114552.

Talebi, H., Peeters, L.J.M., Otto, A., Tolosana-Delgado, R., 2022. A truly spatial random forests algorithm for geoscience data analysis and modelling. Mathematical Geosciences 54(1), 1-22.

Tang, J., Riley, W.J., Manzoni, S., Maggi, F., 2024. Feasibility of formulating ecosystem biogeochemical models from established physical rules. Journal of Geophysical Research: Biogeosciences 129(6), e2023JG007674.

Tartakovsky, A.M., Marrero, C.O., Perdikaris, P., Tartakovsky, G.D., Barajas-Solano, D., 2020. Physics-informed deep neural networks for learning parameters and constitutive relationships in subsurface flow problems. Water Resources Research 56(5), e2019WR026731.

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, 111793.

van Genuchten, M.T., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Science Society of America Journal 44(5), 892-898.

van Zijl, G., van Tol, J., Tinnefeld, M., Le Roux, P., 2019. A hillslope based digital soil mapping approach, for hydropedological assessments. Geoderma 354, 113888.

Vohland, M., Ludwig, B., Seidel, M., Hutengs, C., 2022. Quantification of soil organic carbon at regional scale: Benefits of fusing vis-NIR and MIR diffuse reflectance data are greater for in situ than for laboratory-based modelling approaches. Geoderma 405, 115426.

von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R. et al., 2023. Informed machine learning – A taxonomy and survey of integrating prior knowledge into learning systems. IEEE Transactions on Knowledge and Data Engineering 35(1), 614-633.

Wadoux, A.M.J.C., Minasny, B., McBratney, A.B., 2020. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews 210, 103359.

Wadoux, A.M.J.C., Molnar, C., 2022. Beyond prediction: methods for interpreting complex models of soil variation. Geoderma 422, 115953.
Wang, J., Filippi, P., Haan, S., Pozza, L., Whelan, B., Bishop, T.F.A., 2024. Gaussian process regression for three-dimensional soil mapping over multiple spatial supports. Geoderma 446, 116899.
Wang, S., Sankaran, S., Wang, H., Perdikaris, P., 2023a. An Expert's Guide to Training Physics-informed Neural Networks.
Wang, Y., Jin, M., Deng, Z., 2018. Alternative model for predicting soil hydraulic conductivity over the complete moisture range. Water Resources Research 54(9), 6860-6876.
Wang, Y., Shi, L., Hu, X., Song, W., Wang, L., 2023b. Multiphysics-Informed Neural Networks for Coupled Soil Hydrothermal Modeling. Water Resources Research 59(1), e2022WR031960.
Weber, T.K.D., Finkel, M., da Conceição Gonçalves, M., Vereecken, H., Diamantopoulos, E., 2020. Pedotransfer function for the Brunswick soil hydraulic property model and comparison to the van Genuchten-mualem model. Water Resources Research 56(9), e2019WR026820.
Weber, T.K.D., Weihermüller, L., Nemes, A., Bechtold, M., Degré, A., Diamantopoulos, E., Fatichi, S., Filipović, V., Gupta, S., Hohenbrink, T.L. et al., 2024. Hydro-pedotransfer functions: a roadmap for future development. Hydrol. Earth Syst. Sci. 28(14), 3391-3433.
Weindorf, D.C., Chakraborty, S., 2024. Balancing machine learning and artificial intelligence in soil science with human perspective and experience. Pedosphere 34(1), 9-12.
Wessolek, G., Bohne, K., Trinks, S., 2023. Validation of soil thermal conductivity models. International Journal of Thermophysics 44(2), 20.
Widyastuti, M.T., Minasny, B., Padarian, J., Maggi, F., Aitkenhead, M., Beucher, A., Connolly, J., Fiantis, D., Kidd, D., Ma, Y. et al., 2024. PEATGRIDS: Mapping thickness and carbon stock of global peatlands via digital soil mapping. Earth Syst. Sci. Data Discuss. 2024, 1-29.
Willard, J.D., Jia, X., Xu, S., Steinbach, M.S., Kumar, V., 2020. Integrating physics-based modeling with machine learning: A survey. ArXiv abs/2003.04919.
Wu, F., Tan, K., Wang, X., Ding, J., Liu, Z., 2023. A novel semi-empirical soil multi-factor radiative transfer model for soil organic matter estimation based on hyperspectral imagery. Geoderma 437, 116605.
Xie, E., Zhang, X., Lu, F., Peng, Y., Chen, J., Zhao, Y., 2022. Integration of a process-based model into the digital soil mapping improves the space-time soil organic carbon modelling in intensively human-impacted area. Geoderma 409, 115599.
Xie, X., Yan, H., Lu, Y., Zeng, L., 2024. Simulating field soil temperature variations with physics-informed neural networks. Soil and Tillage Research 244, 106236.
Yang, R.M., Liu, L.A., Zhang, X., He, R.X., Zhu, C.M., Zhang, Z.Q., Li, J.G., 2022. The effectiveness of digital soil mapping with temporal variables in modeling soil organic carbon changes. Geoderma 405, 115407.
Yang, W.-H., Clifford, D., Minasny, B., 2015. Mapping soil water retention curves via spatial Bayesian hierarchical models. Journal of Hydrology 524, 768-779.
Zaman, B., McKee, M., Neale, C.M.U., 2012. Fusion of remotely sensed data for soil moisture estimation using relevance vector and support vector machines. International Journal of Remote Sensing 33(20), 6516-6552.

Zhang, L., Heuvelink, G.B.M., Mulder, V.L., Chen, S., Deng, X., Yang, L., 2024. Using process-oriented model output to enhance machine learning-based soil organic carbon prediction in space and time. Science of The Total Environment 922, 170778.
Zhang, N., Wang, Z., 2017. Review of soil thermal conductivity and predictive models. International Journal of Thermal Sciences 117, 172-183.
Zhang, X., Xie, E., Chen, J., Peng, Y., Yan, G., Zhao, Y., 2023. Modelling the spatiotemporal dynamics of cropland soil organic carbon by integrating process-based models differing in structures with machine learning. Journal of Soils and Sediments 23(7), 2816-2831.
Zhang, Y., Schaap, M.G., 2017. Weighted recalibration of the Rosetta pedotransfer model with improved estimates of hydraulic parameter distributions and summary statistics (Rosetta3). Journal of Hydrology 547, 39-53.