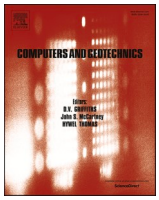
Computers and Geotechnics 177 (2025) 106823

Contents lists available at ScienceDirect

Computers and Geotechnics

journal homepage: www.elsevier.com/locate/compgeo



Indirect models for SWCC parameters: reducing prediction uncertainty with machine learning

Xuzhen He [a,*], Guoqing Cai [b,c], Daichao Sheng [a]

[a] School of Civil and Environmental Engineering, University of Technology Sydney, NSW 2007, Australia
[b] Key Laboratory of Urban Underground Engineering of Ministry of Education, Beijing 100044, China
[c] School of Civil Engineering, Beijing Jiaotong University, Beijing 100044, China

ARTICLE INFO

Keywords:
Probabilistic indirect model
Machine learning
Soil–water characteristic curve

ABSTRACT

The soil–water characteristic curve (SWCC) is crucial for modelling the transport of water and hazardous materials in the vadose zone. However, measuring SWCC is often cumbersome and time-consuming. This paper introduces indirect models that predict SWCC parameters in probabilistic distributions using easily measurable quantities such as particle-size distributions and porosity. This paper starts with building a joint normal model and the derived conditional probability from it serves as a predictive model. However, this model had extremely high prediction uncertainty. To reduce such uncertainty, various machine-learning techniques were explored, including introducing the dependence of variation scale on predictors, using artificial neural networks (ANN) to model nonlinear dependence, incorporating additional predictive features, and generating a larger dataset. The final machine-learning model successfully reduces prediction variability and has been rigorously tested on a separate set of samples to prevent overfitting.

1. Introduction

Constitutive models are essential for predicting and designing geotechnical structures (Dafalias and Taiebat, 2016; He, et al., 2020). However, even with a "perfect" model, the effectiveness and accuracy of numerical predictions depend largely on having reliable input parameters for these models. For predicting water flow and the movement of hazardous materials through the vadose zone, the soil–water characteristic curve (SWCC) and its parameters are crucial (Fredlund and Rahardjo, 1993; Zhou et al., 2012; Cai et al., 2020). Unfortunately, these parameters are difficult to measure directly, and flow permeability can vary by several orders of magnitude across the full range of saturation. Numerous laboratory and field methods exist to measure unsaturated soil hydraulic parameters, but these methods are often cumbersome and time-consuming (Chen et al., 2024).

In practice, it is easier to conduct simpler tests and estimate these SWCC parameters indirectly using empirical models. By treating pores as "idealised" cylindrical pores, Laplace's law (Fredlund and Rahardjo, 1993) can link pressure heads to pore sizes. The pore sizes of soils are related to particle-size distributions (PSD), packing state (i.e., fabric, primarily porosity), and organic matter content. Consequently, it is sensible to build indirect models and estimate SWCC parameters from measurements of these attributes (Sakaki et al., 2014; Zhai et al., 2020b; Zhang et al., 2022b; Es-haghi et al., 2023; Satyanaga et al., 2024).

The most straightforward indirect models are deterministic empirical equations. For instance, Sakaki et al. (2014) demonstrated the relationship between the air entry value and characteristic particle sizes such as d_{30} and d_{50}, which represent the particle sizes at which the mass cumulative percentages (MCP) are 30 % and 50 %, respectively. However, since predictors like particle-size distribution and porosity do not encompass all the necessary information to fully determine the water retention capability of soils, it is not expected to have a very accurate prediction using these models. Instead, the prediction should carry a high degree of uncertainty. Consequently, deterministic indirect models are generally not very useful in practice.

Indirect models should aim to provide a probabilistic description of the SWCC given predictors/inputs. Zhang et al. (2022a) developed probabilistic indirect models using a large dataset, which they used as priors for estimating SWCC parameters in conjunction with experimental data and Bayesian updating. Their probabilistic indirect model was essentially a linear regression. The objective of the present paper is to construct probabilistic indirect models and attempt to reduce prediction uncertainty using machine-learning techniques (He et al., 2021; Zhang et al., 2021; He, Xu et al., 2022; Zhang et al., 2022a). With this indirect model that reduces uncertainty in predicted SWCC parameters,

* Corresponding author.
E-mail addresses: xuzhen.he@uts.edu.au (X. He), guoqing.cai@bjtu.edu.cn (G. Cai), daichao.sheng@uts.edu.au (D. Sheng).

https://doi.org/10.1016/j.compgeo.2024.106823

Available online 19 October 2024

Nomenclature

List of symbols

f_{sg}^{-1}	Inverse of the sigmoid function
$f_{sp}^{-1}(\cdots)$	Inverse of the softplus function
f_{VG}	The van Genuchten equation
$\mathrm{MCP}_{2\mu m}$	Mass cumulative percentages at the particle size of 2 μm
n	SWCC parameter for the van Genuchten equation
$p(\cdots)$	Probability density functions
$p(\cdots\|\cdots)$	Conditional probability
α	SWCC parameter for the van Genuchten equation
γ^*, β_*^*, λ^*, ξ_*^*, ϕ	Indirect model parameters
θ_r	Residual water content, SWCC parameter for the van Genuchten equation
θ_s	Saturated water content, SWCC parameter for the van Genuchten equation

Acronym

ANN	Artificial neural network
HMC	Hamiltonian Monte Carlo
MCP	Mass cumulative percentages
NLL	Negative log likelihood
PDF	Probability density functions
PSD	Particle-size distributions
SWCC	Soil-water characteristic curve
USDA	United States Department of Agriculture
UNSODA	UNsaturated SOil hydraulic DAtabase
VB	Variational Bayes

unsaturated soil models are expected to be more widely adopted in practical engineering, addressing the current challenges posed by the unavailability of SWCC parameters in practice.

There are several models available in the literature for fitting SWCC., the mostly used is probably the van Genuchten equation (van Genuchten, 1980):

$$\theta = f_{VG}(s;\theta_s,\theta_r,\alpha,n) = \theta_r + \frac{\theta_s-\theta_r}{\left(1+\left(\frac{s}{\alpha}\right)^n\right)^{1-\frac{1}{n}}} \tag{1}$$

Here, θ is the volumetric water content, s is the suction, whose unit is cm in the present paper, measuring the rise of water columns. This choice of this unit is to be consistent with the dataset used. θ_s is the saturated water content, and θ_r is the residual water content. α is a parameter that is related to the air entry value, and its unit is the same with suction. n ($n > 1$) is a parameter measuring the pore-size distribution. The objective is therefore to provide a probabilistic indirect model for the SWCC parameters (θ_s, θ_r, α and n). In the language of probability theory, this involves determining the conditional probability of these parameters given predictors like PSD and porosity, denoted as $p(\theta_s, \theta_r, \alpha, n|\mathrm{PSD}, \mathrm{Porosity})$. This study focuses on indirect models for parameters of the van Genuchten Equation: But this framework is generic, and it could also be applied to other SWCC models.

The models are built on the UNSODA dataset (Nemes et al., 2015), which is a collection of data for 790 soil samples, including features such as dry density, particle density, porosity, saturated water content, organic matter content, PSD, and SWCC data measured in laboratory or field. The data underwent rigorous quality control measures, including cross-checks against other datasets and internal consistency checks. Outliers and erroneous data points were identified and either corrected or excluded from the final dataset. These procedures (Nemes et al., 2015) ensure that UNSODA is a reliable resource, contributing to its popularity and extensive usage in the literature. Since the UNSODA database contains more data for the drying branch of SWCC compared to the wetting branch, we thus concentrate on the parameters for the drying curve in this study (Zhai et al., 2020a).

The structure of this paper is organised as follows: Section 2 explains the estimation of SWCC parameters for all soil samples. Section 3 discusses the method used to process PSD data and how to select consistent features to characterise the PSDs. Section 4 addresses how to choose an appropriate value to measure porosity from multiple sources. The building of models begins with a joint normal model with reduced predictors, which is presented in Section 5. Sections 6 and 7 discuss various techniques to extend the model, including introducing dependence of variation scale on predictors, using artificial neural network (ANN) to model the nonlinear dependence, incorporating more features as predictors, and generating a larger training-validation set. Finally, Section 8 provides a summary and some useful conclusion.

2. SWCC parameters by Bayesian inference

To construct the indirect models, our initial step involves estimating the SWCC parameters for each soil sample. Typically, the traditional procedure involves determining a single set of parameters that best fit the measurement data for each soil sample. However, due to the scarcity and noise often present in measurement data, coupled with the imperfect fit of some SWCCs to the van Genuchten model, this inverse problem of parameter estimation becomes ill-posed – numerous parameter combinations may yield similarly minimal errors. Hence, the estimation of SWCC parameters is approached within a Bayesian framework.

For a given soil sample, with its SWCC parameters $(\theta_s,\theta_r,\alpha,n)$ and a measurement point with suction as s_i, we can compute the predicted water content as $f_{VG}(s_i;\theta_s,\theta_r,\alpha,n)$. We postulate that the measured water content θ_i follows a normal distribution with the predicted value as its mean, denoted in probability theory as $\theta_i \sim \mathcal{N}(f_{VG}(s_i;\theta_s,\theta_r,\alpha,n),\sigma)$. Here, σ represents the standard deviation, encompassing the measurement error of suction and/or water content, and accounts for the error scale associated with utilising the possible imperfect van Genuchten model to fit the measurement data. With this assumption, the likelihood of each measurement point (s_i,θ_i) given all parameters can be calculated and denoted as $p(s_i,\theta_i|\theta_s,\theta_r,\alpha,n,\sigma)$. It's important to note that the standard deviation is now in the parameter list to be estimated.

Suppose there are N measurement points for an SWCC, by assuming independence between them, the total likelihood is $\prod_{i=1}^{N} p(s_i,\theta_i|\theta_s,\theta_r,\alpha,n,\sigma)$. Incorporating a prior, the parameters can then be estimated within a Bayesian framework as $p(\theta_s,\theta_r,\alpha,n,\sigma|s_{i=1:N},\theta_{i=1:N}) \propto \prod_{i=1}^{N} p(s_i,\theta_i|\theta_s,\theta_r,\alpha,n,\sigma)p(\theta_s,\theta_r,\alpha,n,\sigma)$. Because we do not have other sources of information regarding the parameters, we use non-informative priors, i.e., $p(\theta_s,\theta_r,\alpha,n,\sigma) = \text{constant}$.

Two methods are used to approximate the posteriors $p(\theta_s, \theta_r, \alpha, n, \sigma|s_{i=1:N},\theta_{i=1:N})$ – Hamiltonian Monte Carlo (HMC) and Variational Bayes (VB). HMC is an algorithm to draw random samples according to a target probability distribution, particularly useful when direct sampling proves challenging. Its convergent rate is often faster than traditional Markov chain Monte Carlo methods such as the Metropolis–Hastings algorithm. HMC performs optimally for unconstrained random variables that closely resemble normal distributions. However, the parameters to be estimated are bound by certain conditions (e.g., $1 > \theta_s > \theta_r \geq 0$, $\alpha > 0$, $n > 1$ and $\sigma > 0$). Consequently, transformations with bijectors become necessary. The transformed variables are denoted $X_1 = f_{sg}^{-1}(\theta_s)$, $X_2 = f_{sg}^{-1}(\theta_r/\theta_s)$, $X_3 = ln(\alpha)$, $X_4 = ln(n-1)$ and $X_5 = ln(\sigma)$. Here, f_{sg}^{-1} denotes the inverse of the sigmoid function and ln the natural logarithm. It's straightforward to confirm that these transformed

variables are all unconstrained, and there exists a bijection between them and the original variables.

For each soil sample, we run 10 independent HMC chains to check the convergence. The "burn-in" phase consists of 20,000 iterations to allow the chains to reach stationary. During the first 70 % of this phase, the step size is adjusted by a simple algorithm (Andrieu and Thoms, 2008) to reach an optimal 65 % acceptance rate. After the "burn-in" phase, an additional 1,000 steps are run, resulting in a total of 10,000 samples. In Fig. 1, histograms represent the distributions of HMC-estimated parameters (θ_s, θ_r and α), plotted alongside SWCC measurements. It can be visually confirmed that the estimates are reasonable.

Variational Bayes is a method to find an analytical approximation to a target probability distribution, achieved by selecting a family of simpler surrogate distributions and finding the set of distribution pa-

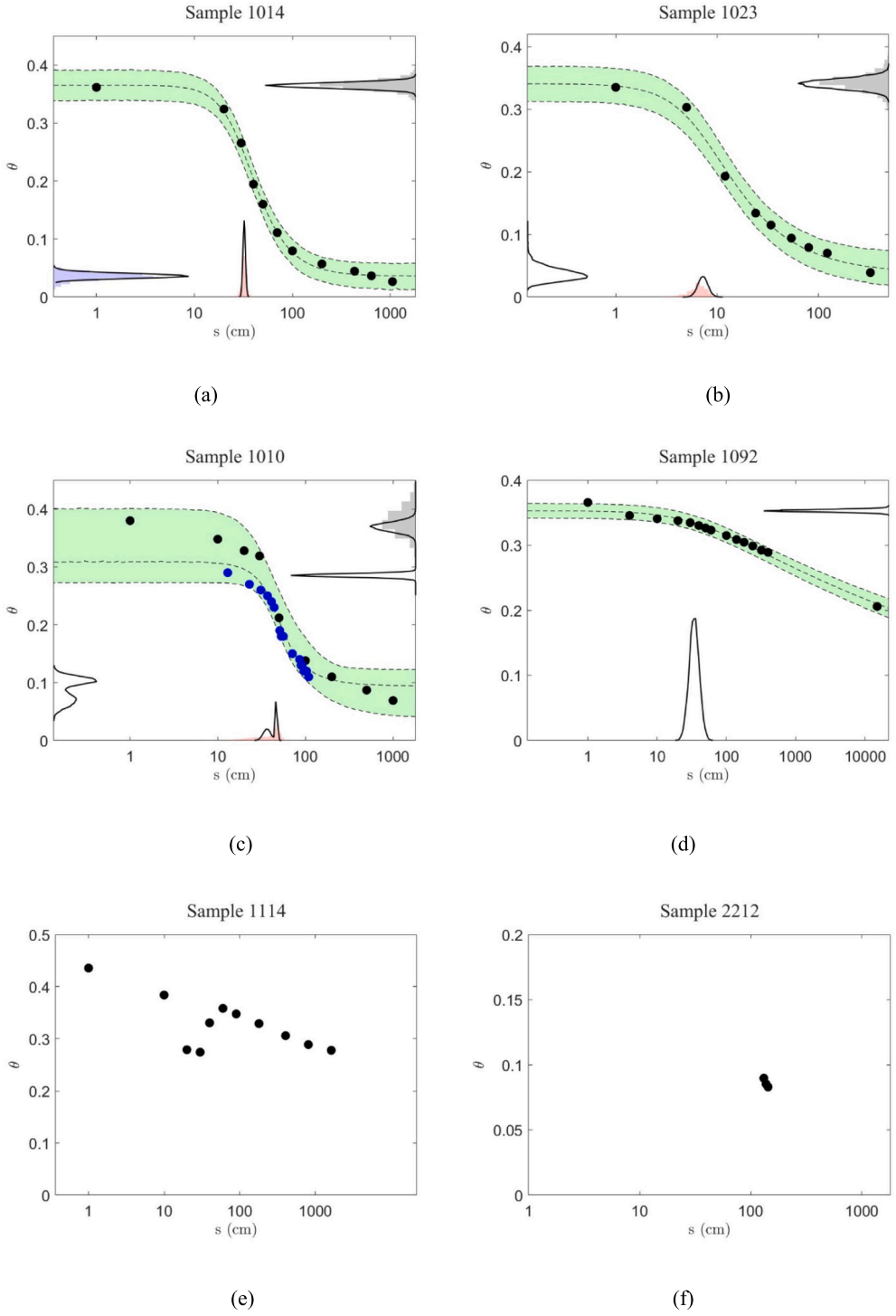


Fig. 1. Bayesian estimation of SWCC parameters (black dots: SWCC measured in laboratory; blue dots: SWCC measured in field; histograms: distributions of θ_s – grey, θ_r – blue, and α – red inferred by Hamiltonian Monte Carlo; Solid black lines: distributions of parameters inferred by Variational Bayes; Filled green area with dash lines: 95% interval for SWCC predictions) (a) a sample for which both HMC and VB give convergent estimations; (b) a sample for which HMC cannot give convergent estimations for some parameters; (c) a sample with two sets of SWCC measurements; (d) A sample for which the estimated θ_r is invalid; (e) a sample with erroneous data; (f) a sample with insufficient data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

rameters that minimise the difference between the surrogate and the target distributions. For the SWCC parameters, we choose the multivariate normal distribution as surrogate for the transformed variables $X_1 \sim X_5$, and determine the distribution parameters (means and covariance matrix) that minimise the Kullback–Leibler divergence. The distributions of VB-estimated parameters are presented as solid lines in Fig. 1.

Fig. 1a displays a soil sample (1014) where both HMC and VB methods converge and showing similar results. HMC results should be considered the true distributions, while VB results are approximations. However, HMC does not always converge for all parameters or for all samples – checked by $\widehat{R} < 1.1$ as suggested by Gelman et al. (2014). Taking Sample 1023 as an example as in Fig. 1b, HMC provides convergent estimates for θ_s and α, but not for θ_r. In contrast, VB shows easier convergence, indicated by continuously reduced and then plateaued Kullback–Leibler divergence. The VB-estimated θ_r of Sample 1023 is shown as solid line Fig. 1b, which is a reasonable estimation. The non-convergence of HMC often arises from either (1) flawed model assumptions or (2) insufficient or erroneous data. For parameters where VB converges but HMC does not, the non-convergence in HMC is likely due to incorrect assumptions about the normal distribution of the transformed parameters or the possibility that the SWCC cannot be accurately modelled by the van Genuchten equation. With sufficient data, VB can always provide an estimate because it seeks the distribution closest to the true distribution; even if the two distributions (true and assumed) differ, it can still produce a convergent estimate.

For some soil samples, both laboratory and field measurements are available. We conduct Bayesian inference separately to these two sets of measurements and assume that the final soil SWCC parameters are a mixture of the parameters estimated from each set. Fig. 1c illustrates this with Sample 1010, where field measurements suggest a lower θ_s compared to laboratory measurements. The VB-estimated parameters, as mixtures, exhibit bimodal patterns to reflect this.

Of the total 790 soil samples in UNSODA, VB provides convergent estimations for most samples (708). Some samples do not converge either due to (1) data errors as illustrated in Fig. 1e or (2) insufficient data as illustrated in Fig. 1f. The problematic samples are listed in Table 1. Fig. 2 shows the estimated 95 % intervals for θ_r, α and n for all samples. θ_s is omitted because it is not a target in the indirect models analysed in this study. There are very few samples (fewer than 3, circled in Fig. 2) where HMC gives convergent estimations, but VB does not. For samples with convergence from both HMC and VB, the results are largely consistent.

An important observation is that for some samples, such as Sample 1092 shown in Fig. 1d, the VB-estimated θ_r is very close to zero with extremely low variability (black intervals in Fig. 2, variability is so small that they visually appear as black dots). This occurs because, although there is enough data for a convergent estimation, the measured data are primarily at lower suction levels, providing limited or no information about the residual water content. Thus, these estimated θ_r is considered invalid. Consequently, only 378 out of 790 samples have valid θ_r estimations, while 708 out of 790 have valid estimations for α and n.

Table 1
Samples whose SWCC estimations are not convergent with Variational Bayes.

	laboratory measurements	field measurements
Data errors	1114, 1460, 4284	1132, 3163, 3195
Insufficient data	1320, 1461, 1462 2212, 2214, 2215, 2180–2181, 2213, 2216–2217, 2253, 2374, 2690 3050, 3175, 3225, 3330 4191–4204, 4211–4212, 4221–4224, 4550–4551, 4580–4583, 4720	1122–1123, 1134, 1330 2470–2472 3032–3033, 3102, 3113 3132, 3151, 3153, 3224–3225, 3240, 3242–3243, 3251–3253, 3260–3262, 3264, 3270, 3283, 3340–3341

3. Particle-size distribution

Measuring particle-size distributions is a standard practice in geotechnical projects due to its straightforward nature. These distributions bear a strong relationship with the SWCC, rendering them valuable for indirect modelling purposes. However, a soil's PSD forms a continuous curve, we instead need discrete characteristic quantities for predictive purpose. Three options are available: (1) using mass cumulative percentage (MCP) values at specific particle sizes; (2) using particle sizes at specific MCP values; (3) fitting the PSD curve with an equation and using the fitting parameters. Option 3 is disregarded due to the versatile shapes of PSDs from natural soils, making it challenging to find a single equation that adequately fits all PSDs. Options 1 and 2 are comparable, but Option 1 holds an advantage. PSDs are typically measured using standard sieves, the raw data already present MCP values at corresponding sieve sizes, thus minimising the necessity for fitting or interpolation.

One problem for the PSDs of UNSODA samples is that they are not all compatible, i.e., measured across varying sieve sizes. Predominantly, MCP readings are available as 610 values at 2 μm, 535 values at 50 μm, and 601 values at 2000 μm. These specific particle sizes hold significance as they aid in distinguishing clay, silt, sand, and gravel particles. Denoted as $MCP_{2\mu m}$, $MCP_{50\mu m}$ and $MCP_{2000\mu m}$, respectively, these readings serve as inputs for soil texture classification following the USDA (United States Department of Agriculture) system. To maximise data availability without resorting to interpolation, readings close to these key sizes are also used, which is named as horizontal shift strategy in this paper.For instance, if a sample lacks a reading at 2 μm, any reading falling within the range of 1 μm to 3 μm is considered as $MCP_{2\mu m}$. This principle extends to $MCP_{50\mu m}$ for readings between 32μm and 63μm, and to $MCP_{2000\mu m}$ for readings between 1000μm and 3350μm. Following this procedure, the MCP values are augmented to include 655 values at 2 μm, 679 values at 50 μm, and 672 values at 2000 μm.

To better characterise the PSDs, additional MCP values across various particle sizes are required. Additionally, having more MCP values provides more information, which generally aids machine learning in identifying better models. We aim to incorporate three additional readings between 2 μm and 50 μm (e.g., 5 μm, 10 μm, 20 μm), and three more readings between 50 μm and 2000 μm (e.g., 100 μm, 250 μm, 500 μm). However, the original dataset lacks sufficient values, containing only 85, 94, 238, 157, 280, and 302 data points for them. Hence, to expand the dataset, we resort to interpolation and the horizontal shift strategy.

Nemes et al. (1999) conducted a comprehensive investigation into the interpolation of particle-size distributions. Their findings underscored that fitting with spline functions significantly enhances accuracy with an adequate number of measurements. Conversely, fitting with the Gompertz equation exhibits less sensitivity to the availability of measurement. This is corroborated in our study, as depicted in Fig. 3. For the gap-graded Sample 1021 in Fig. 3a, the Gompertz equation ($f(x) = e^{-be^{-c(x-d)}}$ utilised in this study) fails to yield satisfactory results, whereas a monotonic cubic spline with dense measurement points appears reasonable. Conversely, for a sample (1081) with limited measurements, the monotonic cubic spline proves less effective, while the Gompertz equation offers some viable insights.

To interpolate MCP values at the desired particle sizes, we adhere to the following procedures with priority from highest to lowest: (1) Firstly, we examine if any neighbouring measurement falls within a 5 % error margin (log scale) and horizontally shift if available. (2) If there are more than three measurements, and the particle size to be interpolated falls between two measurements, the fitted value using a monotonic cubic spline is used. (3) In all other scenarios, interpolation is performed using the Gompertz equation. Following this protocol, 711 samples now possess MCP values at these particle sizes. The

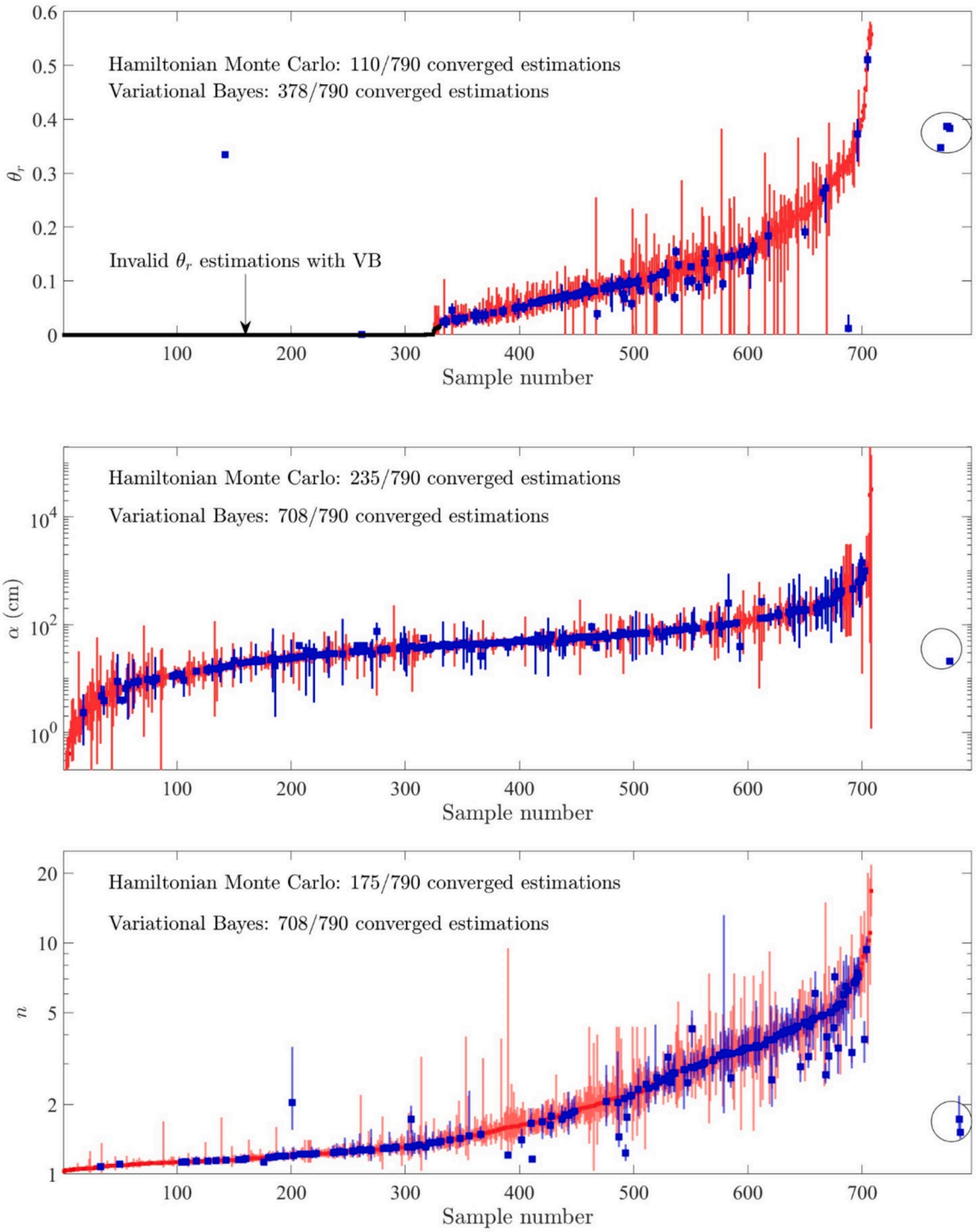


Fig. 2. Estimated SWCC parameters for all samples (red: 95% interval inferred with Variational Bayes; blue: 95% interval inferred with Hamiltonian Monte Carlo; samples plotted with ascending VB-inferred median; circled samples: convergent with HMC but not convergent with VB). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

distributions of MCP values at different particle sizes for all samples are illustrated in Fig. 4.

4. Porosity

The SWCC of soils depends on the packing of soil particles, of which porosity being the most important factor. Porosity, defined as the volume of voids per bulk volume, can also be determined from dry density, i.e., Porosity = 1 – (dry density/particle density), by assuming uniform particle materials with a known particle density. Moreover, by definition, the saturated volumetric water content is equal to the porosity. Thus, for the UNSODA samples, a measure of porosity can be derived from four distinct sources: (1) measured porosity (370 data points); (2) calculated from the provided dry density and particle density (395 data points); (3) experimental water content of water-saturated samples (305 data points); and (4) inferred θ_s from SWCCs as outlined in Section 2.

In Fig. 5, data are presented when two types of porosity measurement are available for the same samples. The inferred θ_s constitutes a

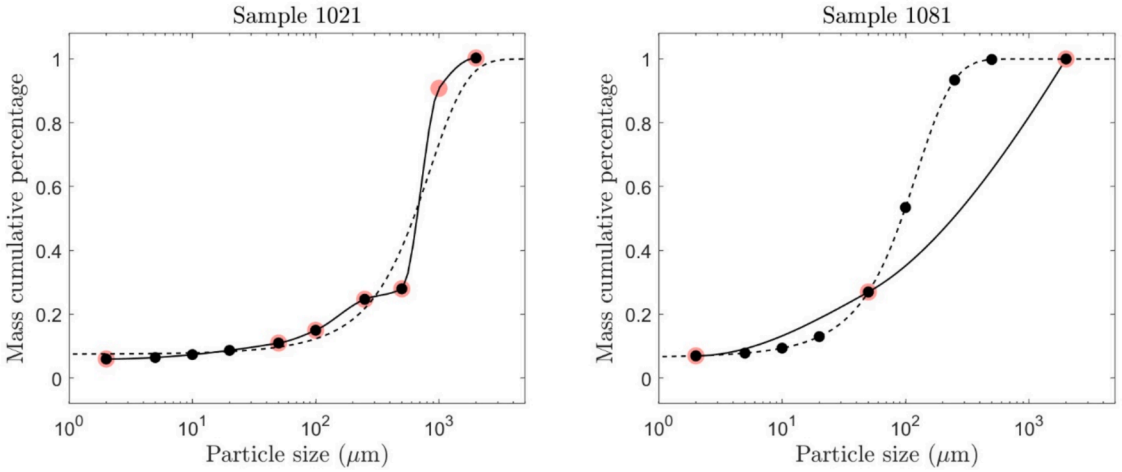


Fig. 3. Interpolating the required mass cumulative percentage (MCP) values (red dots: experimental data; solid lines: monotonic cubic spline; dash lines: fitting with Gompertz function; black dots: Interpolated MCP values at particle size of 2, 5, 10, 20, 50, 100, 250, 500, and 2000 μ m). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

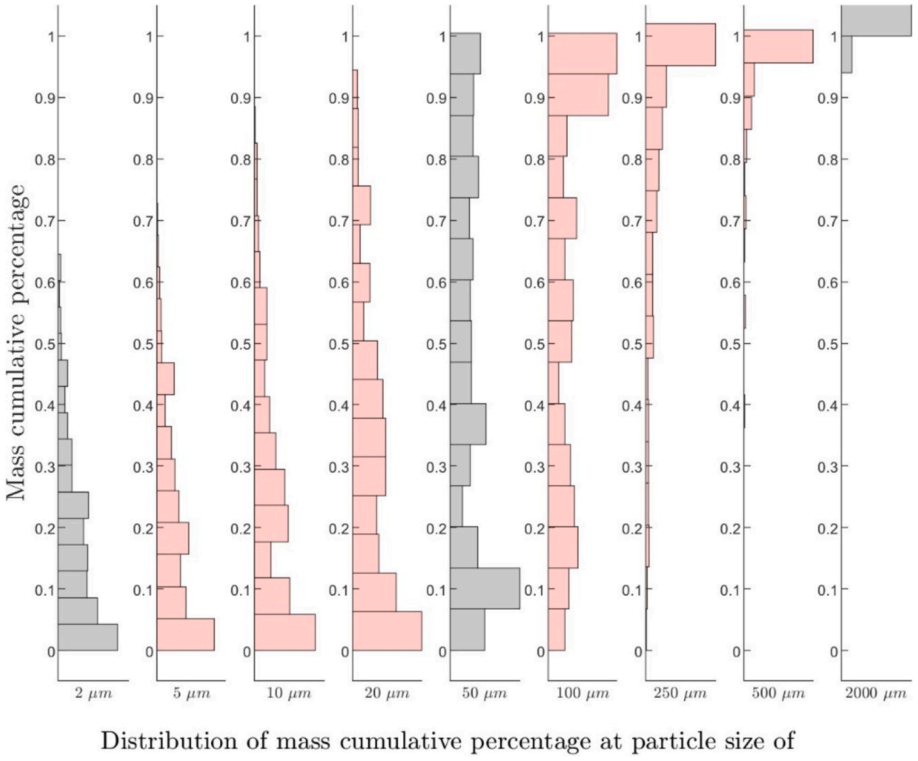


Fig. 4. Distributions of mass cumulative percentage (MCP) values for all samples.

distribution and is depicted as 95 % intervals. Most data points align closely with or fall upon the 1:1 line, while some discrepancies arise due to measurement errors. In the indirect models, a single porosity is used as predictors. To maximise porosity measurements, we choose the porosity value for samples in the following order: (1) provided θ_s, (2) calculated from dry density and particle density, (3) measured porosity, and (4) median of inferred θ_s. The measured θ_s provided by UNSODA is most preferred because it directly reflects the water retention capacity of soils. The porosity calculated from dry density and particle density is the next best option, as it involves minimal measurement error and is highly reliable. The provided porosity value is less preferred than the previous two, as some values are simply derived from dry density, and some are clearly erroneous. The inferred θ_s is the least preferred due to the higher uncertainty involved in the inference process. Following this procedure, 777 samples now possess a porosity measurement.

5. Joint normal model

Porosity is easy to measure and is an important predictor for indirect models. Given that the saturated water content, θ_s, is inherently equivalent to porosity, it is thus not targeted for prediction. As illustrated in Fig. 4, $MCP_{2000\mu m}$ is 1 for nearly all UNSODA samples, offering limited or no informative value and hence is disregarded as a predictor. Consequently, our initial effort involves constructing a basic model incorporating only $MCP_{2\mu m}$, $MCP_{50\mu m}$, and porosity as inputs features to predict the distribution of SWCC parameters, denoted as a conditional

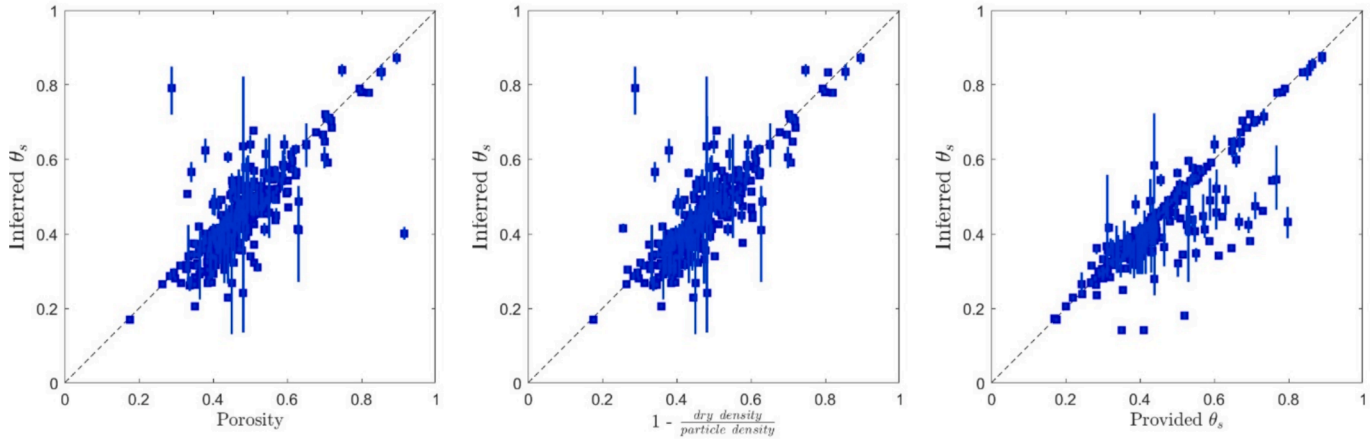


Fig. 5. Relationship between measurement of porosity from different sources.

probability $p(\theta_r, \alpha, n|\mathrm{MCP}_{2\mu m}, \mathrm{MCP}_{50\mu m}, \mathrm{Porosity})$. Conditional probability can be derived from the joint probability of all features, including predictors and targets. Thus, we commence by establishing a joint normal model for all features.

To ensure an unbiased assessment of subsequent models, the 790 UNSODA samples are partitioned into a test set comprising 158 samples and a training-validation set containing 632 samples. Given that the estimated SWCC parameters are as distributions, we perform $N_{d/s} = 10$ iterations of sampling for each soil sample in the training-validation set to generate a dataset of 6320 instances. It is important to note that some rows of data may have missing values for certain features. These gaps have minimal impact on constructing the joint normal model because its parameters are estimated from marginal distributions and pairwise correlations using a bootstrapping method. However, when building machine-learning models, any row with missing values in the input or output features will be simply excluded.

Soil classification has long been used by engineers to aid in understanding soil and estimating parameters. With $\mathrm{MCP}_{2\mu m}$ and $\mathrm{MCP}_{50\mu m}$, we can readily classify the soils according to the USDA system, Fig. 6 illustrates the distribution of SWCC parameters across different types of soils, revealing that sands typically exhibit lower residual water content θ_r and higher n values compared to clays. Regarding α, the distribution of sands appears narrower in contrast to clays. These findings align with empirical knowledge, validating the effectiveness of the preceding data processing procedure and instilling our confidence in constructing indirect models utilising these features.

The distributions of the features are illustrated in Fig. 7, revealing a substantial deviation from normality. To address this, we adopt a similar procedure outlined by Ching et al. (2014), by firstly identifying transformed variables with a normal marginal distribution, followed by constructing a joint normal distribution for these transformed variables. Considering that $\mathrm{MCP}_{2\mu m}$, $\mathrm{MCP}_{50\mu m}$, porosity and θ_r all lie within the range of (0,1), it is a common technique in statistical modelling to bring them onto the real number line with an inverse sigmoid transformation. Then, normality is check for the transformed variables. Similarly, α and n, with lower bounds, can be transformed onto the real number line using the logarithmic function or the softplus function. Denoting the transformed variables with a superscript 't', we have

$$\begin{aligned}
\mathrm{MCP}^t_{2\mu m} &= f_{sg}^{-1}\left(\mathrm{MCP}_{2\mu m}\right)\\
\mathrm{MCP}^t_{50\mu m} &= f_{sg}^{-1}\left(\mathrm{MCP}_{50\mu m}\right)\\
\mathrm{Porosity}^t &= f_{sg}^{-1}(\mathrm{Porosity})\\
\theta_r^t &= f_{sg}^{-1}(\theta_r)\\
\alpha^t &= ln(\alpha)\\
n^t &= ln(n-1)
\end{aligned} \tag{2}$$

In the diagonal of Fig. 8, histograms depict the marginal distributions of transformed variables (using logarithmic function for α and n), all visually resembling normal distributions. Additionally, we conducted Anderson–Darling test, Jarque–Bera test, and quantile–quantile plots, all confirming the normality of marginal distributions. However, normality of marginal distributions does not guarantee a joint normal distribution. The Henze-Zirkler test for multivariate normality resulted in rejection, with a p-value of zero (<0.05). Despite this, we proceed to construct a joint normal model and evaluate its performance.

The upper triangular portion of Fig. 8 displays correlations among the transformed variables. Positive correlations are observed among

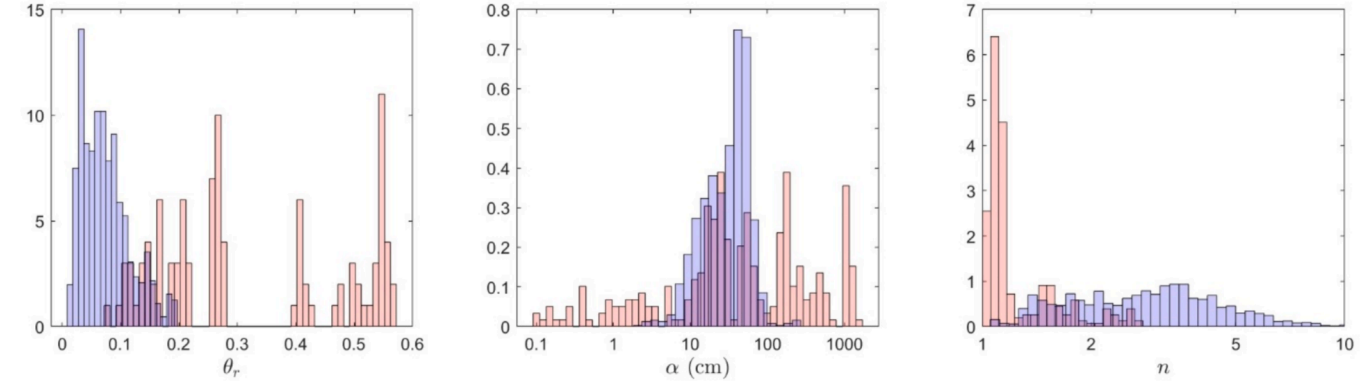


Fig. 6. Distribution of SWCC parameters for different types of soils (red: clay, sandy clay, or silty clay; blue: sand or loamy sand). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

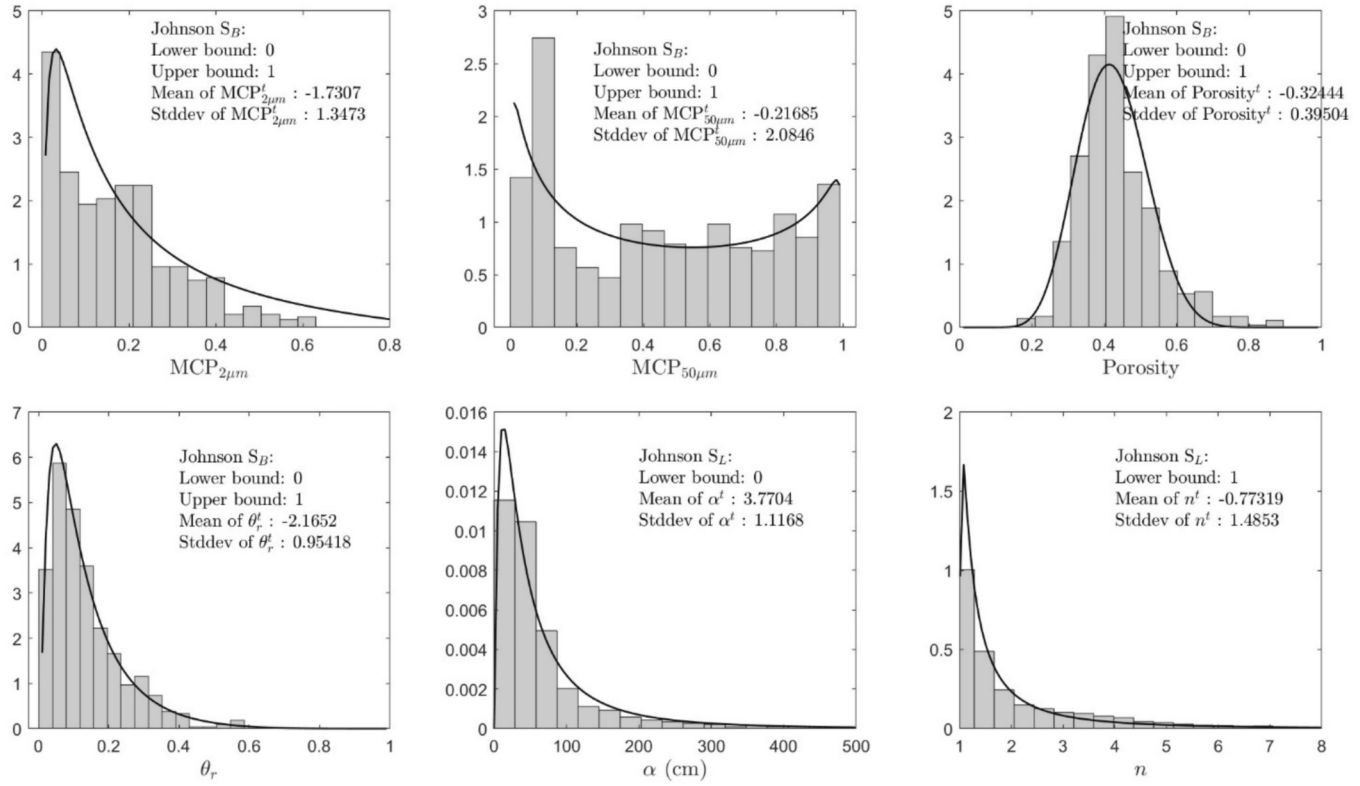


Fig. 7. Distribution of predictors (MCP value at particle size of 2 μm and 50 μm and porosity) and targets (θ_r, α and n) for all samples.

$MCP_{2\mu m}$, $MCP_{50\mu m}$ and porosity, which is consistent with our knowledge – larger $MCP_{50\mu m}$ values often correspond to larger $MCP_{2\mu m}$ values, which are generally from fine-grained soils that tend to have higher porosity. The residual water content shows a positive correlation with these predictors because fine-grained soils typically exhibit larger θ_r values, as similarly demonstrated in Fig. 6. Consistent with findings in Fig. 6, n displays a negative correlation with these predictors, but α shows no strong correlation with any of these features.

If the inverse sigmoid transformation of a variable is normally distributed, it follows a Johnson S_B distribution, which is fully determined by four parameters. Similarly, if a variable follows a Johnson S_L distribution with three parameters, its logarithmic transformation is normally distributed. The solid lines in Fig. 7 represent the probability density functions (PDFs) of the Johnson S_B *or* S_L distributions, demonstrating a good fit for the original variables.

To estimate the Pearson correlations among transformed variables, we utilised the bootstrapping method suggested by Ching et al. (2014). With the estimated means μ_i standard deviations σ_i and the Pearson correlation matrix ρ_{ij} (i or j = θ_r^t, α^t, n^t, $MCP^t_{2\mu m}$, $MCP^t_{50\mu m}$ and Porosity^t), the joint normal model for these transformed variables is fully determined.With such joint probability, we can make predictions using conditional probability. For a multivariate normal distribution, the conditional probability is still multivariate normal. In our case, the transformed SWCC parameters follow a multivariate normal distribution as

$$\begin{bmatrix} \theta_r^t \\ \alpha^t \\ n^t \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\mu} = \begin{bmatrix} \gamma^{\vartheta} \\ \gamma^{\alpha} \\ \gamma^{n} \end{bmatrix} + \begin{bmatrix} \beta_2^{\vartheta} & \beta_{50}^{\vartheta} & \beta_p^{\vartheta} \\ \beta_2^{\alpha} & \beta_{50}^{\alpha} & \beta_p^{\alpha} \\ \beta_2^{n} & \beta_{50}^{n} & \beta_p^{n} \end{bmatrix} \begin{bmatrix} MCP^t_{2\mu m} \\ MCP^t_{50\mu m} \\ Porosity^t \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^{\vartheta} & \sigma^{\vartheta\alpha} & \sigma^{\vartheta n} \\ \sigma^{\vartheta\alpha} & \sigma^{\alpha} & \sigma^{\alpha n} \\ \sigma^{\vartheta n} & \sigma^{\alpha n} & \sigma^{n} \end{bmatrix} \quad (3)$$

where the means are linear functions of the predictors, and the scale of variation is constant. γ^*, β_*^* and σ^{**} are models constants. They can be simply calculated from μ_i, σ_i and ρ_{ij} and are listed in Table 2.

The prediction performance is illustrated in the first sub-figures of Figs. 9–11. The blue vertical lines represent the 95 % intervals of the VB estimations, while the red lines indicate the model's predictions. For all three SWCC parameters, the prediction uncertainty is substantial. For instance, for many samples, the predicted θ_r has a 95 % probability in (0, >0.3), providing minimal useful information. Similarly, for almost all samples, the predicted α has a 95 % probability in range between ~ 4 cm and ~ 400 cm, which is extremely broad. Thus, the aim of this study is to reduce such prediction uncertainty.

Another naïve model involves ignoring the predictors and using the distribution of the targets estimated from all data to make predictions. Based on the dataset, the 95 % intervals for θ_r, α, and n are (0.018, 0.399), (4.59 cm, 319.1 cm), and (1.025, 8.05), respectively. These intervals are marked as horizontal dashed lines in Figs. 9–11. Visually, this naïve model appears to perform comparably to the joint model.

To quantify the performance of models, we need a metric. For probabilistic predictions, a natural one is likelihood. For example, when predicting α, the prediction is a conditional probability $p(\alpha|X)$, which is a function of both α and the predictors X. By inserting the measured values $p(\alpha = \alpha_i^{true}|X = x_i)$, we obtain the likelihood value. In this study, for a given soil sample i, we do not have the true value α_i^{true}, but rather a

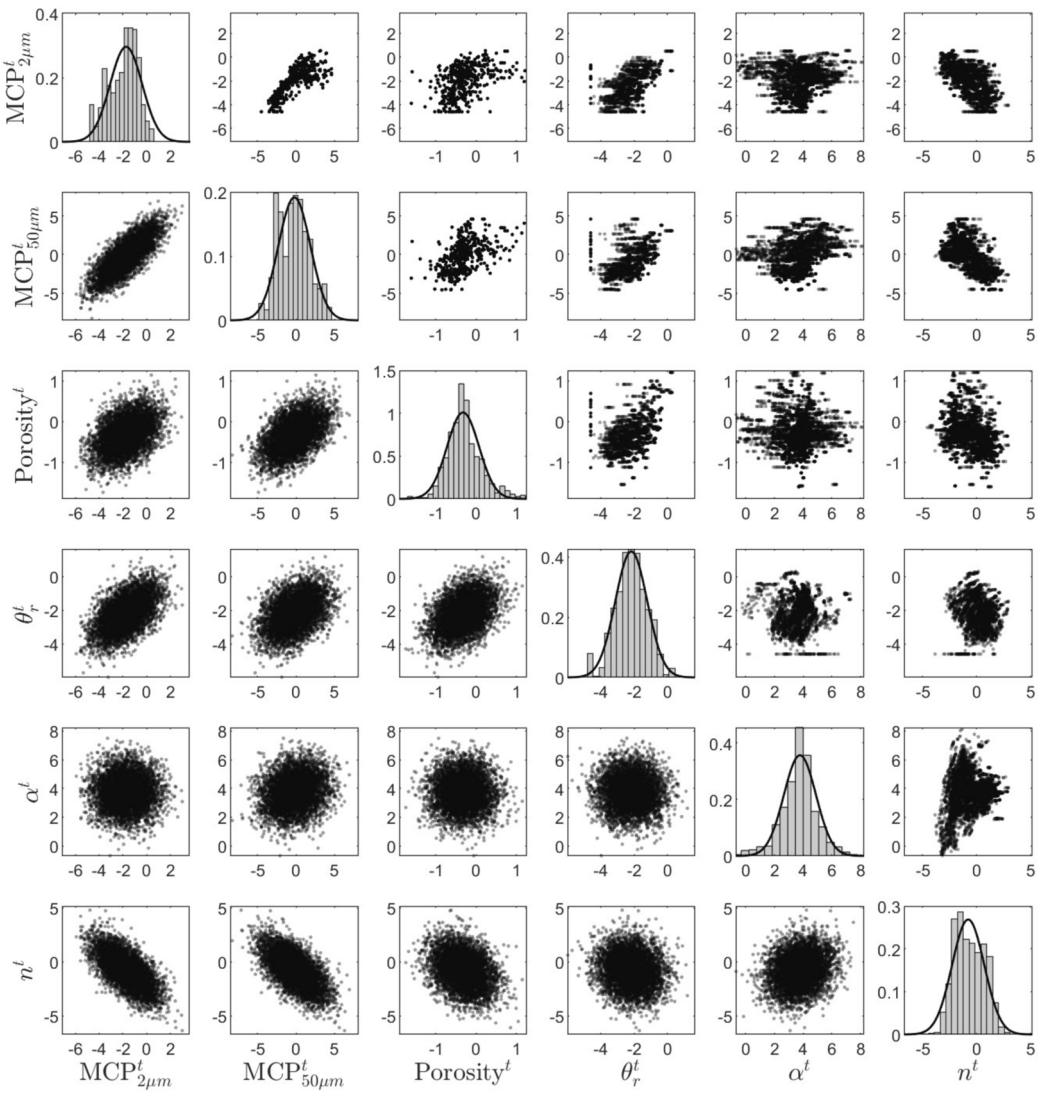


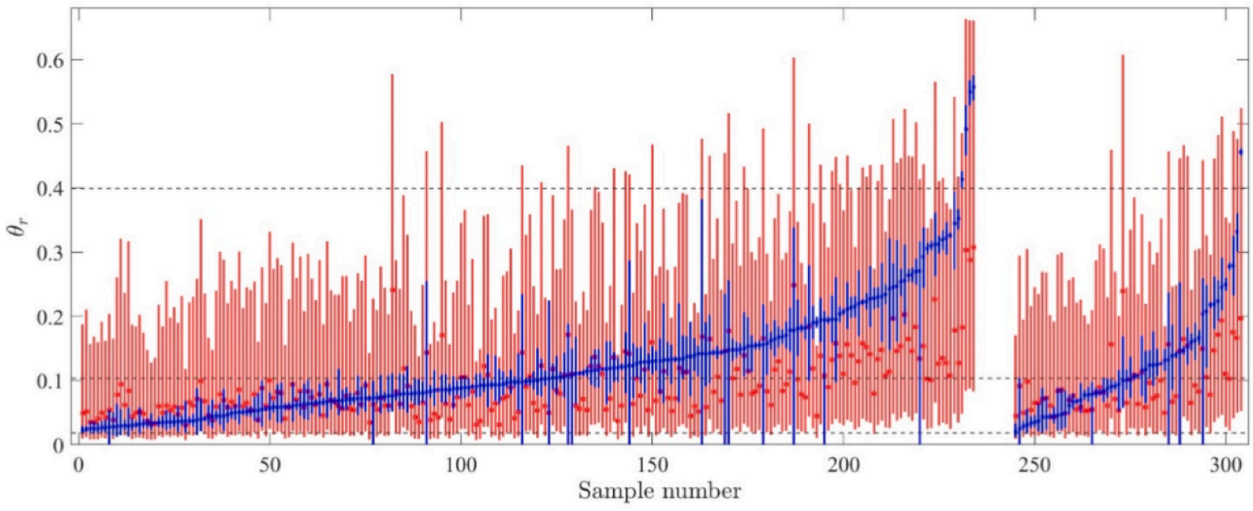
Fig. 8. Marginal distribution of transformed variables and corelations among them (diagonal: distributions with histogram from data and solid lines from fitted normal; upper triangular: corelations from data; lower triangular: corelations from model simulations).

Table 2
Parameters of linear models.

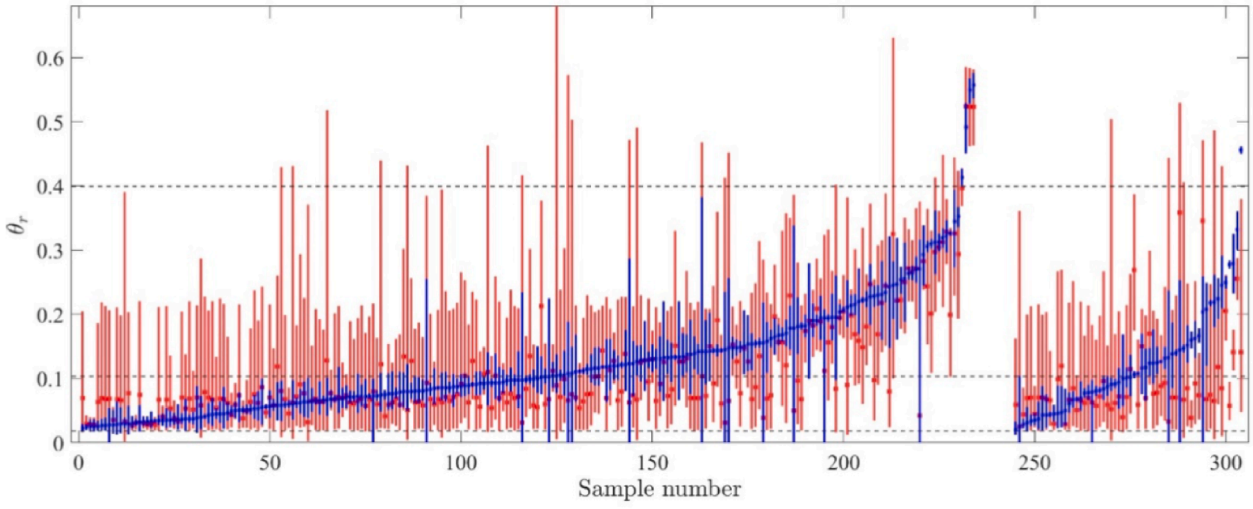
	γ^*	β^*_*			σ^{**}		
Joint normal model	−1.334	0.375	−0.054	0.593	0.601	−0.0047	0.314
	3.170	−0.314	0.298	−0.376	−0.0047	1.114	0.441
	−1.637	−0.482	−0.207	0.043	0.314	0.441	1.175
Linear-mean constant-scale fixed-parameter	−1.279	0.439	−0.039	0.194	0.777		
	2.728	−0.388	0.386	−0.463		1.333	
	−1.816	−0.486	−0.163	−0.038			0.944
Linear-mean constant-scale uncertain-parameter	−1.320	0.368	−0.073	0.110	0.755		
	−1.141	0.453	−0.010	0.289	0.799		
	2.638	−0.487	0.369	−0.503		1.306	
	2.854	−0.383	0.434	−0.287		1.360	
	−1.866	−0.541	−0.181	−0.119			0.925
	−1.712	−0.468	−0.134	0.038			0.964

distribution, $p_i^{true}(\alpha)$. We can then calculate a mean likelihood for each soil sample as $L_i = \int p(\alpha|X = x_i)p_i^{true}(\alpha)d\alpha$, where the integration is evaluated by 10,000 samples of $p_i^{true}(\alpha)$.

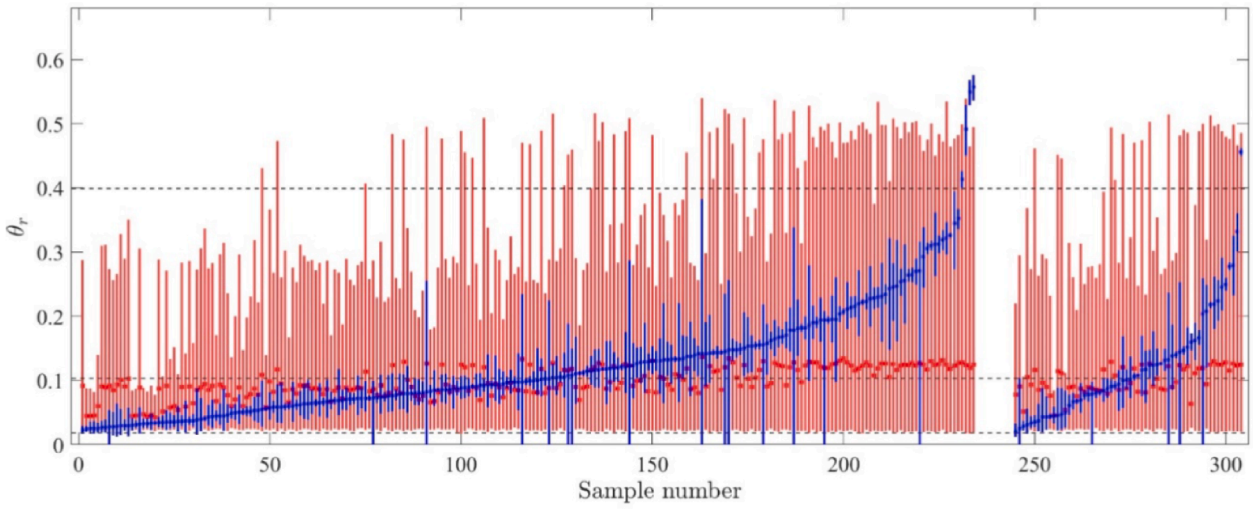
For a set of N soil samples, the mean negative log likelihood is given by $\mathrm{NLL} = -\frac{1}{N}\sum_{k=1}^{N}\ln(L_k)$. NLL measures how well the evidence (e.g., data) supports the model, with smaller NLL indicating better models. Using the joint model to predict θ_r, the NLL for the training-validation set and testing set is 1.041 and 1.047, respectively. For the naïve model, the NLL is 1.284 and 1.269, respectively. This indicates that the joint model performs better according to this metric. Similarly, the joint model also performs better in predicting n, but is only comparable in predicting α as shown in Table 3. Additionally, the comparable NLL



(a) Prediction by the joint normal model

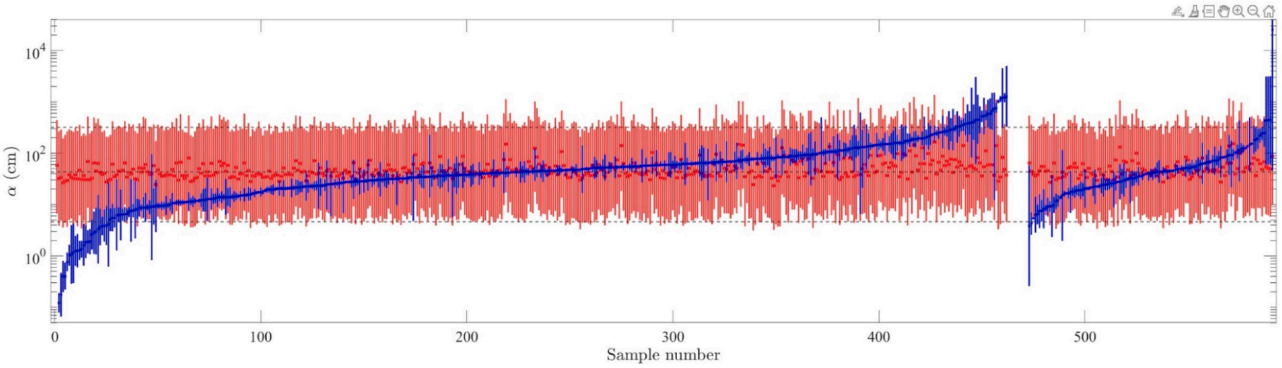


(b) Prediction by an overfitting model

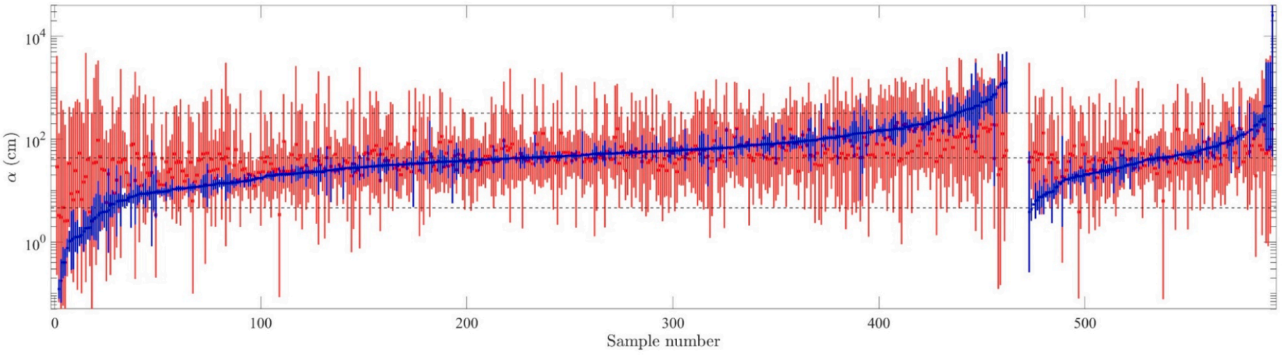


(c) Prediction by the final machine-learning model

Fig. 9. Prediction of residual water content by various models (blue: 95% interval of VB estimations; red: 95% interval of model predictions; dash lines: predicted 95% interval of the naïve model; left part: training-validation set; right part: testing set; samples ordered in ascending VB-estimated median. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) Prediction by the joint normal model



(b) Prediction by the final machine-learning model

Fig. 10. Prediction of $\boldsymbol{\alpha}$ by various models (blue: 95% interval of VB estimations; red: 95% interval of model predictions; dash lines: predicted 95% interval of the naïve model; left part: training-validation set; right part: testing set; samples ordered in ascending VB-estimated median. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

values between the training-validation set and testing set suggest that neither the joint model nor the naïve model overfits.

6. Linear models

The prediction from the joint model results in a joint distribution for all three targets. We can also make prediction for each individual target by ignoring their corelations, which is expressed as a model as:

$$\begin{aligned} &y^* \sim \mathcal{N}(\mu^*, \sigma^*) \qquad\qquad y^* \text{is either } \theta_r^t,\ \alpha^t \text{ or } n^t \\ &\mu^* = \gamma^* + \beta_2^* \mathrm{MCP}_{2\mu m}^t + \beta_{50}^* \mathrm{MCP}_{50\mu m}^t + \beta_p^* \mathrm{Porosity}^t \\ &\sigma^* = \text{constant} \end{aligned} \tag{4}$$

This is a model for which the prediction mean is linearly related to all predictors, and the scale of variation σ^* is constant. Because of this constant scale, minimising the NLL for a given dataset is equivalent to reducing the mean square error between the prediction mean μ^* and the true target values. Consequently, finding the coefficients (γ^* and β_*^*) is the same as performing ordinary linear regression. Once we have the estimated γ^* and β_*^* from linear regression, the scale σ^* can be estimated from the difference between the predicted means and the target values. The determined parameters for this linear-mean constant-scale model are in Table 2 and are very close to those of the joint model. Its performance is also like the joint model, as shown in Table 3.

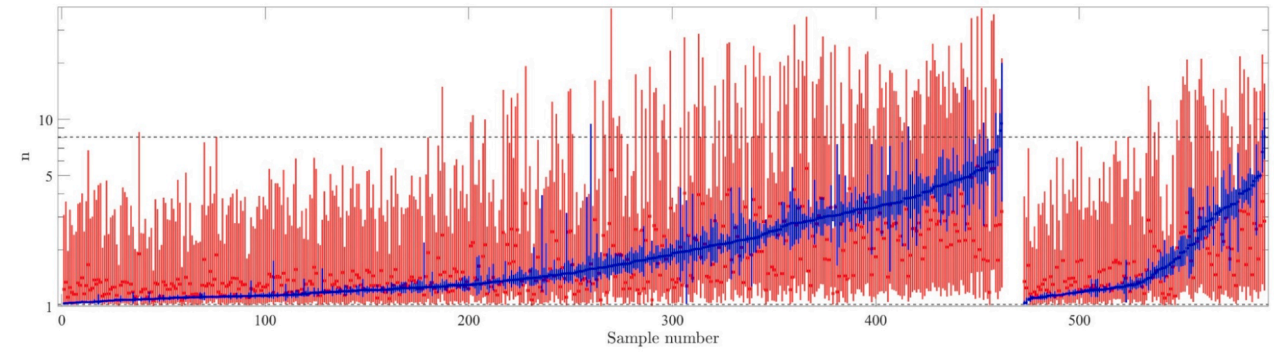
Writing the predictors as $X = \left(\mathrm{MCP}_{2\mu m}^t, \mathrm{MCP}_{50\mu m}^t, \mathrm{Porosity}^t\right)$ and the parameters as $\phi^* = (\gamma^*, \beta_2^*, \beta_{50}^*, \beta_p^*, \sigma^*)$, the model of Eq. 4 defines the likelihood of y^* given predictors X and parameters ϕ^*, i.e. $p(y^*|\phi^*, X)$. Therefore, we can adopt a Bayesian framework to estimate the posterior of parameters through $p(\vartheta^*|X,\ y^*) \propto p(y^*|\phi^*,\ X)p(\phi^*)$. Using non-informative priors ($p(\phi^*) = \text{constant}$), the posterior of parameters is estimated with MCMC and is presented in Table 2 as 95 % intervals. Even though these parameters are determined as distributions, their variability is very low and remain very close to those from the joint model or linear regression. This type of model, which accounts for parameter uncertainty, is also called a model with epistemic uncertainty in the literature. For each sample of the model parameters ϕ^*, we can insert them into the model and evaluate the NLL for the data, thus making the NLL a distribution as well. For this linear-mean constant-scale uncertain-parameter model, the NLLs are shown in Table 3 as 95 % intervals. Like the joint model and linear regression, this model does not improve performance.

So far, the prediction of α has been poor (Table 3), performing no better than the naïve model that predicts the same 95 % range (4.59 cm to 319.1 cm) for all samples. This is contradictory to our empirical knowledge that different soils have different air entry values and α is related to the air entry value. The primary issue is that we assumed a constant scale of variation for all predictions. As shown in Fig. 6, the median α values for sands and clays are comparable, but the variation for clays is much higher, which cannot be captured by a model with a constant scale of variation.

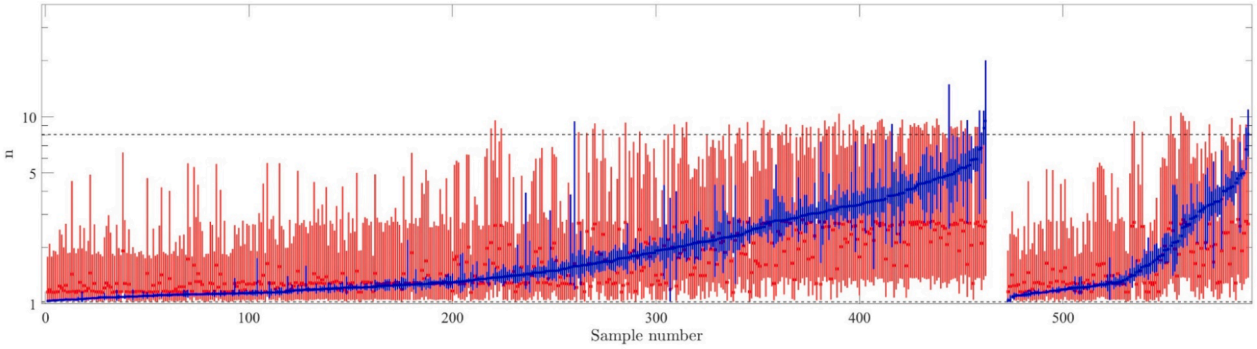
To address this, we could extend the model to have a linear dependence of the scale on predictors as:

$$\begin{aligned} &y^* \sim \mathcal{N}(\mu^*, \sigma^*) \qquad\qquad y^* \text{is either } \theta_r^t,\ \alpha^t \text{ or } n^t \\ &\mu^* = \gamma^* + \beta_2^* \mathrm{MCP}_{2\mu m}^t + \beta_{50}^* \mathrm{MCP}_{50\mu m}^t + \beta_p^* \mathrm{Porosity}^t \\ &f_{sp}^{-1}(\sigma^*) = \lambda^* + \xi_2^* \mathrm{MCP}_{2\mu m}^t + \xi_{50}^* \mathrm{MCP}_{50\mu m}^t + \xi_p^* \mathrm{Porosity}^t \end{aligned} \tag{5}$$

where f_{sp}^{-1} is the inverse of the softplus function. The point estimates of



(a) Prediction by the joint normal model



(b) Prediction by the final machine-learning model

Fig. 11. Prediction of n by various models (blue: 95% interval of VB estimations; red: 95% interval of model predictions; dash lines: predicted 95% interval of the naïve model; left part: training-validation set; right part: testing set; samples ordered in ascending VB-estimated median. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Negative log likelihood of models.

	θ_r		α		n	
No. of samples in brackets –>	Training-validation (234)	Testing (60)	Training-validation (462)	Testing (119)	Training-validation (462)	Testing (119)
Naïve model	1.283	1.269	1.754	1.396	1.685	1.647
Joint model	1.041	1.047	1.705	1.361	1.334	1.326
Linear-mean constant-scale fixed-parameter	1.032	1.028	1.672	1.453	1.328	1.333
Linear-mean constant-scale uncertain-parameter	0.991	0.999	1.667	1.446	1.325	1.326
	1.021	1.029	1.673	1.470	1.331	1.337
Linear-mean linear-scale fixed-parameter	0.965	1.003	1.568	1.374	1.294	1.320
Linear-mean linear-scale uncertain-parameter	0.958	0.996	1.561	1.361	1.291	1.314
	0.989	1.022	1.571	1.385	1.301	1.322
Nonlinear fixed-parameter	X is 4 among 1, 2, 3, 4, 5, 10, 50, and 100.		X is 100 among 5, 10, 50, 100, 500 and 1000.		X is 10 among 2,5,10, 50, 100, and 500.	
	0.981	1.096	1.321	1.325	1.204	1.266
Nonlinear uncertain-parameter	X is 10 among 5,10, 50, 100, and 500.		X is 5 among 5,10, 50, 100, and 500.		X is 10 among 5,10, 50, 100, and 500.	
	0.964	0.997	1.560	1.351	1.293	1.310
	1.033	1.060	1.576	1.402	1.317	1.330
Nonlinear fixed-parameter using more predictors	X is 2 among 1, 2, 3, 5,10, and 50.		X is 30 among 5, 20, 30, 40, 50, and 100.		X is 5 among 2, 3, 4, 5,10, 50, and 100.	
	0.971	1.063	1.155	1.181	1.167	1.203
Nonlinear fixed-parameter by generating more training data	X is 2 among 1, 2, 3, 5,10, and 50.		X is 30 among 5, 20, 30, 40, 50, and 100.		X is 5 among 2, 3, 4, 5,10, 50, and 100.	
	0.980	1.078	1.159	1.238	1.172	1.238

***X=Optimal number of hidden neurons.

the model parameters (γ^*, β_*^*, λ^*, and ξ_*^*) is obtained by minimising the NLL. For this model, we could not conduct ordinary linear regression anymore, but have to resort to optimisation – the Adam stochastic gradient descent algorithm in this study. The learning rate determines the size of corrective steps in optimisation. A high learning rate shortens the training time but may result into to a local minimum instead of the desired global minimum, whereas a lower learning rate results in a longer training process. After several experiments (with learning rate as 0.0001, 0.001, 0.01, 0.1 or 1), a learning rate of 0.01 was found to be suitable for this study. In training, 80 % of the data from the training-

validation samples are used for training and the rest 20 % for validation. Once the parameters are determined, we check the model performance on the training-validation set and the testing set, yielding NLL of 1.568 and 1.374, respectively. These scores are lower than those of previous models, indicating that incorporating a linear dependence of scale on predictors improves performance and reduces prediction uncertainty. Applying the same model to n also improves prediction, but no improvement is observed for θ_r, as shown in Table 3 (the row linear-mean linear-scale fixed-parameter).

Since the model outputs a distribution, we can similarly estimate the model parameters using a Bayesian framework. All following uncertain-parameter models have many model parameters, so we use VB to estimate the parameters to achieve efficiency. We use independent normal distributions as the surrogates for them. For complex models, using non-informative priors may lead to non-convergence. Therefore, we use diffusive priors in all following inference: γ^*, β_*^*, λ^*, and ξ_*^* are all independent and follow a 'spike-and-slab' distribution, which is a mixture of $\mathcal{N}(0, 10)$ and $\mathcal{N}(0, 100)$. The estimated parameters are close to that of the fixed-parameter model with very low level of variation. The distributions of NLL are also like that of the fixed-parameter model, listed as linear-mean linear-scale uncertain-parameter model in Table 3.

7. Nonlinear models

Until now, the dependence on predictors has been confined to linear relationships. To capture more complex interactions, we test several nonlinear models, specifically artificial neural networks (ANNs), expressed as follows:

$$y^* \sim \mathcal{N}(\mu^*, \sigma^*) \quad y^* \text{is either } \theta_r^t,\ \alpha^t \text{ or } n^t$$

$$\left(\mu^*, f_{sp}^{-1}\left(\sigma^*\right)\right) = f_{ANN}\left(\mathrm{MCP}_{2\mu m}^t, \mathrm{MCP}_{50\mu m}^t, \mathrm{Porosity}^t, \cdots; \phi\right) \tag{5}$$

The outputs of the ANNs are the mean μ^* and the transformed scale $f_{sp}^{-1}(\sigma^*)$. These outputs are then used to create a distribution as the final prediction. Given the limited number of features (<10), a fully connected neural network with a single hidden layer is sufficient (He et al., 2021; He et al., 2022). The sigmoid activation function is used for its smoothness. We denote all the trainable parameters in the ANN as ϕ. A point estimate of these parameters can be obtained by minimising the LNN through optimisation, resulting in fixed-parameter models.

Firstly, fixed-parameter models are built to predict the residual water content θ_r. The dashed lines in Fig. 12a show how the number of neurons in the hidden layer affects model performance on the training-validation set (blue) and the testing set (red). More neurons mean more powerful models with the ability to capture more complex relationships. As the number increases, the NLL on the training-validation set decreases continuously. However, the NLL on the testing set also increases. When the number is larger than 2, the NLL on the testing set becomes larger than that on the training-validation set, indicating overfitting, where the model corresponds too closely to the training data and fails to generalise to new data. Fig. 9b illustrates the predictions of an overfitting model for all samples. The model performs exceptionally well on the training-validation set, displaying reduced prediction uncertainty compared to both the joint model (Fig. 9a) and the final machine-learning model (Fig. 9c). However, this overconfidence results in poor performance on the testing set. Therefore, 2 neurons in the hidden layer are optimal.

The solid lines in Fig. 12a represent models with dropout (rate = 0.3), a technique often used in ANNs to prevent overfitting (He et al., 2021). The model does not show clear signs of overfitting with 10 neurons, but overfitting still occurs with a larger number of neurons. Compared to models without dropout, dropout does help prevent overfitting to some extent and is therefore used in all subsequent models. The leftmost point represents the performance of the linear model and even the optimal nonlinear models perform only comparably, indicating that using a nonlinear ANN does not improve the prediction of θ_r.

The performance of nonlinear fixed-parameter models for α are shown in Fig. 12b, suggesting that an optimal number of neurons in the hidden layer is 100. Similarly, the leftmost point represents the linear model, so using nonlinear ANN improves the performance in predicting α. Similar improvement is also observed for n (Table 3).

We can similarly estimate the trainable parameters ϕ within a Bayesian framework, resulting in nonlinear uncertain-parameter models. Fig. 12c illustrates the impact of the number of neurons on model performance, using the prediction of θ_r as an example. The vertical lines represent the 95 % interval. As the number of neurons increases, the NLL median remains relatively stable, with only a slight increase in variability. Beyond a critical point (10–50 neurons), model performance declines sharply, accompanied by increasing variability. The leftmost point still indicates the linear model. Thus, using a nonlinear model with epistemic uncertainty does not offer an improvement over a linear model. This is observed across all targets (Table 3) and in all cases, even when more predictors are included later. Therefore, nonlinear uncertain-parameter models are not considered further in this paper.

The previous models utilised only three features as predictors and characterised the PSD with just two features. As mentioned in Section 3, we were able to interpolate the MCP values at various particle sizes for most samples. By incorporating all these PSD-characterising features along with porosity to construct nonlinear fixed-parameter models, we achieved improved predictions: 1.063, 1.181, and 1.203 for θ_r, α, and n, respectively on the testing set as shown in Table 3 and Fig. 13. This is the best model till now. This improvement is anticipated, as more features provide more information about each sample, thereby reducing uncertainty.

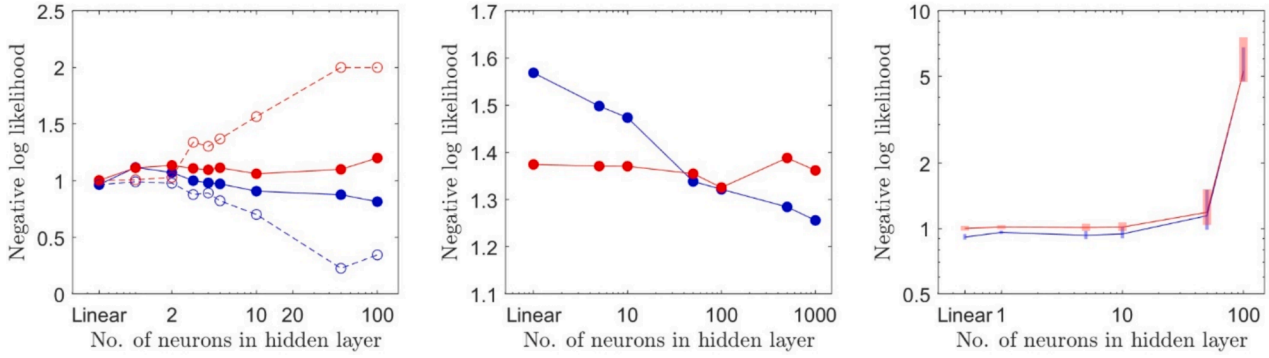


Fig. 12. Relationship between model performance and complexity for nonlinear models (blue = training-validation; red = testing) (a) Fixed-parameter model for $\boldsymbol{\theta_r}$; solid lines = with dropout; dash lines = without dropout; (b) Fixed-parameter model for $\boldsymbol{\alpha}$ with dropout; (c) Uncertain-parameter model for $\boldsymbol{\theta_r}$; vertical lines represent 95 % intervals. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

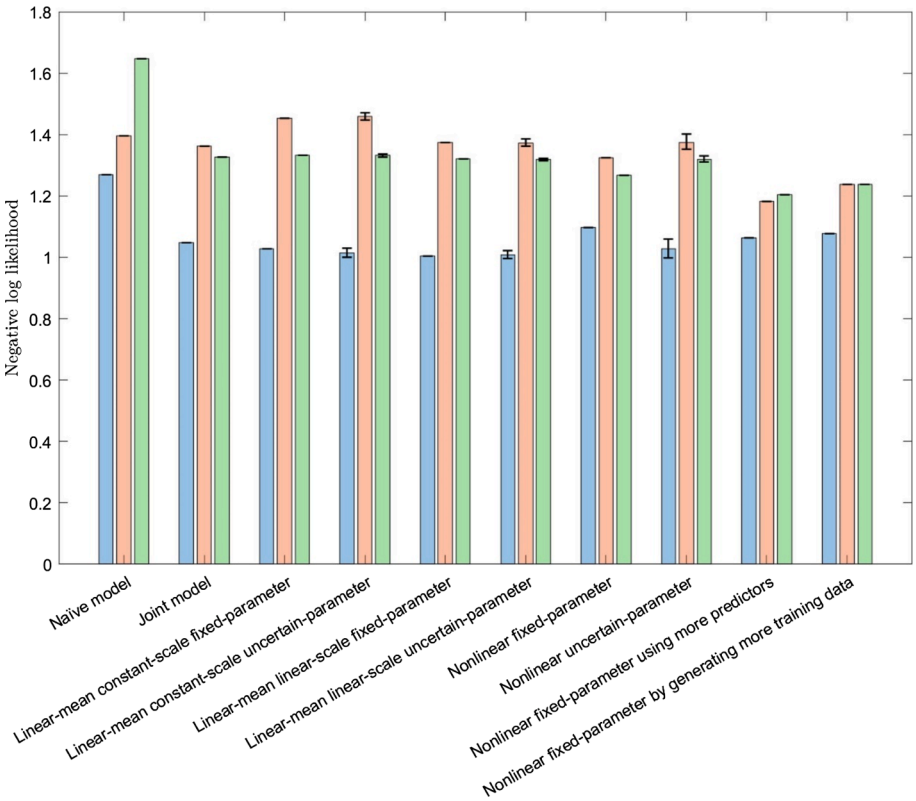


Fig. 13. Performance of all models on the testing set (blue = θ_r; red = α; green = n). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Another common technique for enhancing the performance of machine-learning models is collecting more data. The more diverse and high-quality data we have, the better our models can learn and generalise. In our previous models, the 6,320-row dataset was generated by performing $N_{d/s} = 10$ iterations of sampling for each soil sample in the training-validation set. We experimented with performing $N_{d/s} = 100$ iterations of sampling, resulting in a larger dataset of 63,200 rows. Using the same nonlinear fixed-parameter model and full set of predictors, the model's performance was similar to that of the model trained on the smaller dataset. This indicates that the smaller dataset likely already covered the full range of input/output interactions the model is expected to handle, and additional data may not be beneficial unless it includes new and interesting cases.

8. Conclusion

The SWCC is crucial for modelling the transport of water and hazardous materials in the vadose zone. However, measuring SWCC in the laboratory or field is often cumbersome and time-consuming. This paper introduces a framework to develop indirect models that predict SWCC parameters in probabilistic distributions using easily measurable quantities such as particle-size distributions and porosity. The primary focus is on reducing prediction uncertainty by employing multiple machine-learning techniques. The models are built on the UNSODA dataset, which is a collection of data for 790 soil samples.

This paper started with building convectional statistical joint models with only three predictors. The conditional probability of this joint model constitutes a predictive model. However, the prediction uncertainty is extremely large, not significantly better than a naïve model that ignores all the predictors and predict the same distribution for all samples. It was found that that this conditional probability from the joint model is equivalent to a linear regression. Subsequently, various techniques in machine learning were explored to improve prediction.

First, introducing the dependence of variation scale on predictors enhanced performance, as a constant scale could not account for observations that different types of soils had similar medians for α but varied scales of variation. Replacing linear dependence with ANN also improved model performance by capturing complex interactions between inputs and outputs. Additionally, incorporating more features to characterise the PSD of soils could provide more information, which helped better determine the possible distribution of SWCC parameters. Generating a larger dataset from training-validation samples yielded little gain, likely because the smaller dataset already covered the full range of input/output interactions.

The best model was a nonlinear fixed-parameter model trained on the complete set of predictors. This model significantly reduced prediction variability. Specifically, for predicting α and n, the NLL was reduced to 1.181 and 1.203 from 1.361 and 1.326 in the joint normal models, thereby achieving our initial goal. The final machine-learning model for predicting residual water content performed comparably to the joint model (1.063 vs. 1.047), indicating that a linear regression is sufficient for this parameter.

CRediT authorship contribution statement

Xuzhen He: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Guoqing Cai:** Writing – review & editing, Funding acquisition. **Daichao Sheng:** Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

the work reported in this paper.

Data availability

Data will be made available on request.

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

Andrieu, C., Thoms, J., 2008. A tutorial on adaptive MCMC. Statistics and Computing 18 (4), 343–373. https://doi.org/10.1007/s11222-008-9110-y.

Cai, G., et al., 2020. Soil water retention behavior and microstructure evolution of lateritic soil in the suction range of 0–286.7 MPa. Acta Geotechnica 15 (12), 3327–3341. https://doi.org/10.1007/s11440-020-01011-w.

Chen, K., et al., 2024. Influences of ink-bottle effect evolution on water retention hysteresis of unsaturated soils: An experimental investigation. Engineering Geology 330, 107409. https://doi.org/10.1016/j.enggeo.2024.107409.

Ching, J., Phoon, K.-K., Chen, C.-H., 2014. Modeling piezocone cone penetration (CPTU) parameters of clays as a multivariate normal distribution. Canadian Geotechnical Journal 51 (1), 77–91. https://doi.org/10.1139/cgj-2012-0259.

Dafalias, Y.F., Taiebat, M., 2016. SANISAND-Z: zero elastic range sand plasticity model. Géotechnique 66 (12), 999–1013. https://doi.org/10.1680/jgeot.15.P.271.

Es-haghi, M.S., Rezania, M., Bagheri, M., 2023. Machine learning-based estimation of soil's true air-entry value from GSD curves. Gondwana Research 123, 280–292. https://doi.org/10.1016/j.gr.2022.06.012.

Fredlund, D.G., Rahardjo, H., 1993. Soil Mechanics for Unsaturated Soils. Wiley. https://doi.org/10.1002/9780470172759.

Gelman, A. *et al.* (2014) *Bayesian Data Analysis, Third Edition (Texts in Statistical Science)*, *Book*. Doi: 10.1007/s13398-014-0173-7.2.

He, X., et al., 2020. Work–energy analysis of granular assemblies validates and calibrates a constitutive model. Granular Matter 22 (1), 28. https://doi.org/10.1007/s10035-019-0990-7.

He, X., et al., 2021. Deep learning for efficient stochastic analysis with spatial variability. Acta Geotechnica 1. https://doi.org/10.1007/s11440-021-01335-1.

He, X., Xu, H., Sheng, D., 2022. Ready-to-use deep-learning surrogate models for problems with spatially variable inputs and outputs. *Acta Geotechnica* [Preprint]. https://doi.org/10.1007/s11440-022-01706-2.

Nemes, A. *et al.* (1999) 'Evaluation of different procedures to interpolate particle-size distributions to achieve compatibility within soil databases', *Geoderma*, 90(3–4), pp. 187–202. Doi: 10.1016/S0016-7061(99)00014-2.

Nemes, A. *et al.* (2015) *UNSODA 2.0: Unsaturated Soil Hydraulic Database. Database and program for indirect methods of estimating unsaturated hydraulic properties*.

Sakaki, T., Komatsu, M., Takahashi, M., 2014. Rules-of-Thumb for Predicting Air-Entry Value of Disturbed Sands from Particle Size. Soil Science Society of America Journal 78 (2), 454–464. https://doi.org/10.2136/sssaj2013.06.0237n.

Satyanaga, A., et al., 2024. Modelling Particle-Size Distribution and Estimation of Soil–water Characteristic Curve utilizing Modified Lognormal Distribution function. Geotechnical and Geological Engineering 42 (3), 1639–1657. https://doi.org/10.1007/s10706-023-02638-8.

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. https://doi.org/10.2136/sssaj1980.03615995004400050002x.

Zhai, Q., et al., 2020a. Estimation of the wetting scanning curves for sandy soils. Engineering Geology 272. https://doi.org/10.1016/j.enggeo.2020.105635.

Zhai, Q., et al., 2020b. Framework to estimate the soil-water characteristic curve for soils with different void ratios. Bulletin of Engineering Geology and the Environment 79 (8), 4399–4409. https://doi.org/10.1007/s10064-020-01825-8.

Zhang, J., et al., 2022a. Bayesian estimation of soil-water characteristic curves. Canadian Geotechnical Journal 59 (4), 569–582. https://doi.org/10.1139/cgj-2021-0070.

Zhang, P., Jin, Y.F., Yin, Z.Y., 2021. Machine learning–based uncertainty modelling of mechanical properties of soft clays relating to time-dependent behavior and its application. International Journal for Numerical and Analytical Methods in Geomechanics 45 (11), 1588–1602. https://doi.org/10.1002/nag.3215.

Zhang, P., Yin, Z.Y., Jin, Y.F., 2022b. Bayesian neural network-based uncertainty modelling: application to soil compressibility and undrained shear strength prediction. Canadian Geotechnical Journal 59 (4), 546–557. https://doi.org/10.1139/cgj-2020-0751.

Zhou, A.N., Sheng, D., Carter, J.P., 2012. Modelling the effect of initial density on soil-water characteristic curves. Geotechnique 62 (8), 669–680. https://doi.org/10.1680/geot.10.P.120.