



RESEARCH ARTICLE

# Uncertainty quantification in tree structure and polynomial regression algorithms toward material indices prediction

Geng-Fu He<sup>1</sup> , Pin Zhang<sup>2,3</sup>  and Zhen-Yu Yin<sup>1</sup>

<sup>1</sup>Department of Civil and Environmental Engineering, The Hong Kong, Polytechnic University, Hong Kong, China

<sup>2</sup>Department of Engineering, University of Cambridge, Cambridge, UK

<sup>3</sup>Department of Civil and Environmental Engineering, National University of Singapore, Singapore

**Corresponding authors:** Pin Zhang and Zhen-Yu Yin; Emails: [pinzhang@nus.edu.sg](mailto:pinzhang@nus.edu.sg); [zhenyu.yin@polyu.edu.hk](mailto:zhenyu.yin@polyu.edu.hk)

**Received:** 23 May 2024; **Revised:** 25 October 2024; **Accepted:** 05 January 2025

**Keywords:** evolutionary polynomial regression; quantile; random forest; uncertainty quantification; variational inference

## Abstract

Machine learning's integration into reliability analysis holds substantial potential to ensure infrastructure safety. Despite the merits of flexible tree structure and formulable expression, random forest (RF) and evolutionary polynomial regression (EPR) cannot contribute to reliability-based design due to absent uncertainty quantification (UQ), thus hampering broader applications. This study introduces quantile regression and variational inference (VI), tailored to RF and EPR for UQ, respectively, and explores their capability in identifying material indices. Specifically, quantile-based RF (QRF) quantifies uncertainty by weighting the distribution of observations in leaf nodes, while VI-based EPR (VIEPR) works by approximating the parametric posterior distribution of coefficients in polynomials. The compression index of clays is taken as an exemplar to develop models, which are compared in terms of accuracy and reliability, and also with deterministic counterparts. The results indicate that QRF outperforms VIEPR, exhibiting higher accuracy and confidence in UQ. In the regions of sparse data, predicted uncertainty becomes larger as errors increase, demonstrating the validity of UQ. The generalization ability of QRF is further verified on a new creep index database. The proposed uncertainty-incorporated modeling approaches are available under diverse preferences and possess significant prospects in broad scientific computing domains.

## Impact Statement

The proposed uncertainty-incorporated approaches extend the application range of popular random forest and evolutionary polynomial regression and ensure their practicability for reliability-based design in various industries. The proposed approaches can meet diverse requirements due to their respective merits in strong fitting capability and formulable expressions and hold substantial potential in the field of computational science.

## 1. Introduction

The identification of mechanical indices of materials is vital in engineering design. Experiments in materials can reveal these indices as design parameters but take much effort and expense (He et al., 2024a). Alternatively, some readily measured physical characteristics are often employed to estimate these indices through certain transformation models (Phoon et al., 2022). Numerous empirical relations and data-driven approaches were proposed to identify mechanical indices of materials in the past (Karimi et al., 2023; Rezanian et al., 2011; Vázquez-Escobar et al., 2021). However, the majority of these methods

neglected underlying uncertainties and merely generated point estimations (Li et al., 2022; Nassr et al., 2018; Pei et al., 2023; Rezaia et al., 2011; Wang et al., 2020; Zhang et al., 2019, 2020, 2021b). The direct application of these methods without uncertainty quantification (UQ) may pose huge risks in engineering practice.

Uncertainty is an inevitable factor contributed by various sources in geotechnical engineering. In general, uncertainty can be categorized into epistemic and aleatoric uncertainties (Zhang et al., 2021a), which are also known as reducible and irreducible uncertainties (Olivier et al., 2021). The former arises in modeling the studied phenomena of interest and is dominated by a lack of knowledge, while the latter is induced by slightly intrinsic randomness and is hard to mitigate such as measurement error in laboratory tests. Thereafter, numerous research works focus on reducing prominently epistemic uncertainty if there is no specific demonstration. More specifically, epistemic uncertainty is primarily contributed by intrinsically spatial variability due to widely spaced materials (Ching and Phoon, 2019; Guan and Wang, 2022; Lyu et al., 2023), transformation uncertainty in estimating design parameters (Phoon and Kulhawy, 1999), and statistical uncertainty due to limited site investigation data in geotechnics (Ching and Phoon, 2017; Guan and Wang, 2020). The current study focuses on the second source of epistemic uncertainty, that is, transformation uncertainty associated with data-driven models and aims to tailor traditionally deterministic approaches for broader applications.

To capture substantial uncertainty for reliability-based design, many scholars have been equipping traditionally deterministic approaches with UQ to improve their reliability and practicability. For example, Zhang et al. (2021b) enhanced the reliability of neural networks (NNs) by incorporating Monte Carlo dropout (Gal and Ghahramani, 2016), which adopts random dropout of weights to generate mean and variance for predicting soil maximum dry density. Zhang et al. (2022b) further integrated NN with variational inference (VI), which approximates the parametric posterior distribution of weights for interval estimates of soil compression index  $C_c$  and creep index  $C_a$ . The bootstrapping technique has also been used to quantify uncertainty in ensemble learning models, such as NN (Olivier et al., 2021; Psaros et al., 2023; Zhao et al., 2023) and support vector regression (Li et al., 2023). Apart from the preceding integrations, Wang and Zhao (2017) proposed a novel Bayesian compressive sampling approach that weighs over prespecified basis functions to capture the uncertainty of the tip resistance profile. Ching et al. (2022) also designed a hierarchical multivariate normal probability density functions (PDFs) to generate three-dimensional (3D) probabilistic predictions of  $C_c$ .

Nevertheless, some popular approaches with unique structures, such as random forest (RF) and evolutionary polynomial regression (EPR), have not been equipped with UQ in numerous geotechnical applications (Nassr et al., 2018; Pei et al., 2023; Rezaia et al., 2011; Zhang et al., 2021b). Through incorporating bootstrap aggregating (Breiman, 1996) and random subspace (Ho, 1998), RF characterizes flexible tree structures built upon dichotomy and exhibits strong fitting capability (Karimi et al., 2023; Wang et al., 2020; Zhang et al., 2021d). In contrast, EPR works by combining symbolic and numerical regressions to search for a polynomial function, known as user-friendly and applicable expressions in vast scenarios (Ghorbani and Hasanzadehshooiili, 2018; Gomes et al., 2021). Based on these characteristics, RF has been widely used in soil indices prediction (Li et al., 2022), 3D reconstruction of granular grains (Zhang et al., 2022a), landslide susceptibility mapping (Sun et al., 2021), settlement estimation (Zhang et al., 2019), and slope stability evaluation (Pei et al., 2023), while EPR has been applied in constitutive modeling (Nassr et al., 2018; Zhang et al., 2021e), liquefaction evaluation (Ghorbani and Eslami, 2021; Rezaia et al., 2010), settlement prediction (Shahin, 2016), and so forth. Nonetheless, absent UQ hampers their scenarios, and hence, the current study aims to tailor them for broader applications.

A key issue behind RF is that the output of each decision tree only relies on the average observation in a certain leaf and neglects the distribution information inside. The distribution information typically can be interpreted by PDF, cumulative distribution function or quantiles (Hao and Naiman, 2007). Owing to this neglect, uncertainty was not involved in the aforementioned studies of RF because it merely averages the generated conditional mean of each tree. Some efforts have been made to address this issue, but they only

pay attention to classification tasks (Reis et al., 2018; Vázquez-Escobar et al., 2021). Regarding this issue, quantile regression holds the potential to utilize the distribution information in each leaf because it can estimate arbitrary quantiles (Koenker and Bassett, 1978) such as the quartile or median. Therefore, RF deserves to be integrated with quantile regression to enhance reliability for broader applications in scientific computing domains.

On the other hand, the coefficients of polynomials in EPR are scalars, simply fitted by the least-squares method. In this aspect, Jin and Yin (2020) discussed the potential of equipping EPR with uncertainty evaluation by replacing the least-squares method with a Bayesian approximation through the Markov chain Monte Carlo (MCMC) (Marasco et al., 2022). Nonetheless, MCMC comes with several limitations: sampling directly from a target PDF becomes intractable when the prior and likelihood PDFs are quite different (Phoon et al., 2022); it suffers from the curse of dimensionality and becomes inefficient if the number of target parameters is large (Lykkegaard et al., 2021); it tends to be stuck in local convergence for some tasks (Lee and Song, 2017). VI has recently become an effective alternative to the MCMC method and has already been incorporated into NN for probabilistic shear strength predictions of clays (Zhang et al., 2022b). The integration of EPR and VI is hence worth investigating for designing an efficient modeling approach with UQ.

To this end, this study proposed two uncertainty-incorporated baseline modeling approaches, that is, quantile-based RF and VI-based EPR (QRF and VIEPR) to enrich the current data-driven modeling framework and identify material indices. As a vital design parameter, the  $C_c$  of clays was taken as an exemplar to develop models and verify the feasibility of the proposed approaches. Therefore, two models denoted by QRF and VIEPR with three input variables (i.e., initial void ratio  $e_0$ , liquid limit  $LL$  and plasticity index  $I_p$ ) were developed and applied to predict  $C_c$  values. Their performances were fairly compared in terms of accuracy and reliability and also compared with results of deterministic counterparts. The dependency of error and uncertainty on input variables was also investigated. The better-performing method was further applied to predict  $C_\alpha$  for exploring its generalization ability.

## 2. Methodology

### 2.1. QRF

As a type of nonparametric algorithm, RF incorporates bagging (Breiman, 1996) and random subspace (Ho, 1998) to generate trees and leaf nodes for making predictions. Specifically, bagging generates a bootstrap set by sampling with replacement for training each tree, while random features are considered for node split. In this context, each tree is an independent predictor, in which the output for a given dataset is determined by the average observation in a certain leaf. RF further averages outputs generated by each tree to control overfitting (Zhang et al., 2021b). However, such a strategy only considers the average observation in leaf nodes and neglects the distribution information inside, offering predictions without reliability.

Figure 1 illustrates RF using similar notations as Breiman (2001), let  $\theta_k$  be learned splits that determine the growth of the  $k$ th tree  $T_k$  based on a bootstrap set  $D_k$ . Each bootstrap set is sampled from original training datasets  $D = \{(X_i, Y_i) | X_i \in \mathbb{R}^m, Y_i \in \mathbb{R}, i = 1, 2, \dots, n\}$ .

As can be seen from Figure 1, dropping arbitrary input  $x$  from the top of  $T_k$  will ultimately fall in a unique leaf  $\lambda(x, \theta_k)$ . To calculate the output, observations in this leaf are assigned the same weight  $w(x, \theta_k) = 1/C[\lambda(x, \theta_k)]$ , where  $C[\cdot]$  counts the number of observations inside. Conversely, observations outside the leaf are assigned zero weight. Therefore, the output of a single tree  $T_k$  can be calculated by weighting each observation  $Y_i$  as follows:

$$\hat{Y}^k = \sum_{i=1}^n w_i(x, \theta_k) Y_i \quad (1)$$

where  $w_i(x, \theta_k)$  denotes the weight of  $i$ th observation in  $T_k$ . When it comes to RF with  $K$  trees, the weight of each observation  $w_i(x)$  becomes the average of weights over all trees as follows:

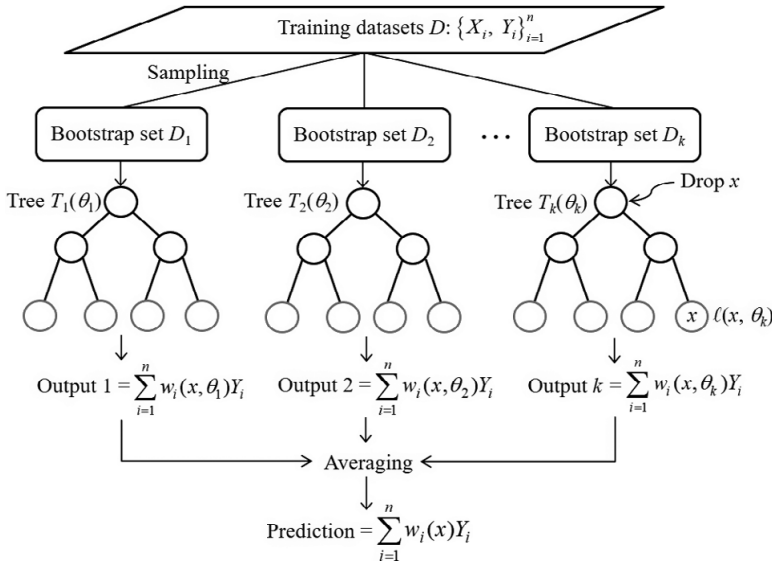


Figure 1. Schematic of RF.

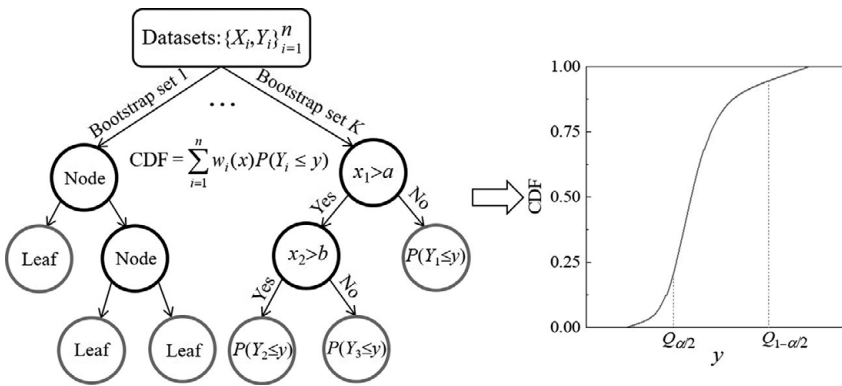


Figure 2. Schematic of QRF.

$$w_i(x) = \frac{1}{K} \sum_{k=1}^K w_i(x, \theta_k) \tag{2}$$

Consequently, the prediction of RF can be expressed by

$$\hat{Y} = \sum_{i=1}^n w_i(x) Y_i \tag{3}$$

Note that  $\hat{Y}$  is merely an average of the conditional mean generated by leaf nodes and neglects distribution information inside. To address this issue, the integration of RF with quantile regression (Koenker and Bassett, 1978; Meinshausen and Ridgeway, 2006), QRF is used to extract distribution information in leaf nodes for reliability evaluation as shown in Figure 2.

Different from RF that weights over each observation  $Y_i$  in Equation (3), QRF replaces  $Y_i$  with a conditional distribution  $P(Y_i \leq y)$  to estimate the CDF of predictions by:

$$\text{CDF}(y) = \sum_{i=1}^n w_i(x) P(Y_i \leq y) \quad (4)$$

where  $P(Y_i \leq y)$  represents the probability of  $Y_i \leq y$ , which equals 1 if  $Y_i \leq y$  or 0 otherwise. As a result, a schematic of predicted CDF can be drawn in the right plot of Figure 2. By performing the inverse of Equation (4), arbitrarily specified quantile  $Q_\alpha$  of predictions can be calculated as follows:

$$Q_\alpha = \inf\{y : \text{CDF}(y) \geq \alpha\} \quad (5)$$

where  $\inf\{\cdot\}$  denotes the infimum of a set and is applied to a set of  $y$  that allows  $\text{CDF}(y) \geq \alpha$ .

To sum up, QRF provides a nonparametric distribution of  $y$  as a function of  $x$ , instead of merely a conditional mean generated by RF. For QRF, the median is employed for point estimations due to its robustness toward outliers, similar to previous research works (Ching et al., 2022; Löfman and Korkiala-Tanttu, 2021). For interval estimate, the commonly used 95% credible interval (CI) can be formulated by quantiles  $Q_{2.5\%}(x)$  and  $Q_{97.5\%}(x)$  as the lower and upper bounds, respectively. It is noteworthy that QRF keeps the same structure as RF when absorbing distribution information over numerous trees. Compared with existing Bayesian methods (Ching et al., 2019; Guan et al., 2024; Wang and Zhao, 2017; Yoshida et al., 2021; Zhang et al., 2022b), the prominent difference of QRF is its frequentist manner to generate interval estimates simply using votes inside trees without involving any prior information, rather than update prior distributions to posterior distributions. On the other hand, QRF only involves the simplest dichotomy for tree growth and avoids complex sampling. As a baseline algorithm, QRF is generic and can be applied in broad applications, for example, 3D reconstruction of granular grains (Zhang et al., 2022a) and landslide susceptibility mapping (Sun et al., 2021), not limited to material indices identification. Given these promising characteristics, such an approach holds the potential of further fusing data with limited features to refine the generated CDF for interval estimates.

## 2.2. VIEPR

Inspired by numerous empirical polynomials used for estimating material indices (Nagaraj and Srinivasa Murthy, 1986; Zhu and Yin, 2000), EPR has been proposed to automatically search for a polynomial function (Giustolisi and Savic, 2006). EPR works by combining symbolic and numerical regressions in two phases: structure identification and coefficient estimation. For structure identification, evolutionary algorithms are used to search for an exponent matrix  $\mathbf{E} \in \mathbb{R}^{n \times m}$  such that  $m$  original input variables can be combined with interior exponents to form  $nt$  new transformed variables  $XT$  as follows:

$$XT_j = x_1^{E(j,1)} \cdot x_2^{E(j,2)} \cdot x_i^{E(j,i)} \dots \cdot x_m^{E(j,m)} \quad (6)$$

where  $x_i = i$ th original input variable;  $XT_j = j$ th transformed variable;  $E(j, m) = (j, m)$ -entry of  $\mathbf{E}$ ; Next, coefficient estimation is implemented by the least-squares method to obtain a general EPR expression as follows:

$$y = \sum_{j=1}^{nt} w_j \cdot XT_j + w_0 \quad (7)$$

where  $y =$  predicted output;  $w_j =$  an adjustable coefficient of  $XT_j$ ;  $w_0 =$  an additional coefficient as bias. Although EPR provides efficient search and explicit solutions, it only generates predictions without reliability.

To address this issue, all coefficients  $\mathbf{w}$  in Equation (7) can be denoted by probability distributions, such as normal distributions instead of scalars. Suppose that they comply with a certain prior distribution  $P(\mathbf{w})$ , their posterior distribution then can be formulated by Bayes' theorem as follows:

$$P(\mathbf{w}|D) = \frac{P(D|\mathbf{w})P(\mathbf{w})}{P(D)} \quad (8)$$

where  $P(D|\mathbf{w})$  is the likelihood of observing  $D$  given  $\mathbf{w}$ ;  $P(D)$  is model evidence, which can be regarded as a constant. From the view of Bayesian, Jin and Yin (2020) discussed the potential of equipping EPR with uncertainty by approximating the distribution of coefficients through the MCMC method (Al-Bittar and Soubra, 2014). However, solutions may be intractable due to large computational costs and convergence problems, especially for high-dimensional tasks. Thus, VI is introduced to address this issue.

As an approximation method, VI utilizes a parametric distribution  $q$  to approximate target posterior  $P(\mathbf{w}|D)$ . Specifically, VI learns variational parameters  $\theta$  in  $q$  to make parametric probability density  $q(\mathbf{w}|\theta)$  and target posterior PDF  $P(\mathbf{w}|D)$  as close as possible. This objective is mathematically achieved by minimizing the distance between two distributions, measured by Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951). Given two distributions  $p$  and  $q$ , the KL divergence between  $p$  and  $q$  can be formulated by

$$KL(q||p) = E_q[\ln q - \ln p] = \int q(\mathbf{w}) \ln \frac{q(\mathbf{w})}{p(\mathbf{w})} d\mathbf{w} \quad (9)$$

where  $q(\mathbf{w})$  and  $p(\mathbf{w})$  denote the probability densities given certain  $\mathbf{w}$ .

Therefore, the variational parameters  $\theta$  are found by minimizing the KL divergence between parameterized  $q(\mathbf{w}|\theta)$  and target posterior  $P(\mathbf{w}|D)$  as follows:

$$\begin{aligned} \theta &= \arg \min_{\theta} KL[q(\mathbf{w}|\theta)||P(\mathbf{w}|D)] \\ &= \arg \min_{\theta} \int q(\mathbf{w}|\theta) \ln \frac{q(\mathbf{w}|\theta)}{P(\mathbf{w})P(D|\mathbf{w})} d\mathbf{w} \\ &= \arg \min_{\theta} KL[q(\mathbf{w}|\theta)||P(\mathbf{w})] - E_{q(\mathbf{w}|\theta)}[\ln P(D|\mathbf{w})] \end{aligned} \quad (10)$$

Using MC sampling to evaluate the expectation term in Equation (10), the KL divergence is approximated and taken as the loss function as follows (Olivier et al., 2021; Zhang et al., 2022b):

$$L(D, \theta) \approx \frac{1}{s} \sum_i^s \left( \ln q(\mathbf{w}^{(i)}|\theta) - \ln P(\mathbf{w}^{(i)}) - \ln P(D|\mathbf{w}^{(i)}) \right) \quad (11)$$

where  $\mathbf{w}^{(i)}$  denotes the  $i$ th MC sample drawn from the variational posterior  $q(\mathbf{w}^{(i)}|\theta)$  and  $s$  represents the number of samples.

Overall, the development of VIEPR follows procedures shown in Figure 3. For brevity, the variational posterior of  $\theta$  is assumed to comply with normal distribution with only two parameters: mean and standard deviation denoted  $u$  and  $\rho$ . The training process aims to optimize  $u$  and  $\rho$  by gradient descent, instead of a scalar in EPR, meaning that the number of parameters merely doubles in comparison with EPR. Compared with other Bayesian methods such as Bayesian NNs (Zhang et al., 2022b), VIEPR has the same computational principle with gradient descent to estimate parametric posterior distributions of coefficients/weights. For practical applications, VIEPR facilitates the understanding of engineers by directly organizing related variables as polynomials in vast scenarios although with limited structures and fitting capability.

### 3. Data source

The database used in this study is extracted from (Zhang et al., 2021d) and includes 331 datasets, in which the  $C_c$  of various clays was measured with three widely observed physical indices:  $e_0$ ,  $LL$ , and  $I_p$ . Numerous empirical relations were built on these indices to identify  $C_c$ , primarily because  $e_0$  directly affects the available space to be compressed, while  $LL$  and  $I_p$  reflect the range of soil moisture content associated with compression-induced plastic deformation. Nonetheless, these equations neglected underlying uncertainties and merely generated point estimation. Table 1 summarizes the statistics of indices in the database, which covers wide ranges of initial void ratio ( $0.66 \leq e_0 \leq 4.64$ ), liquid limit



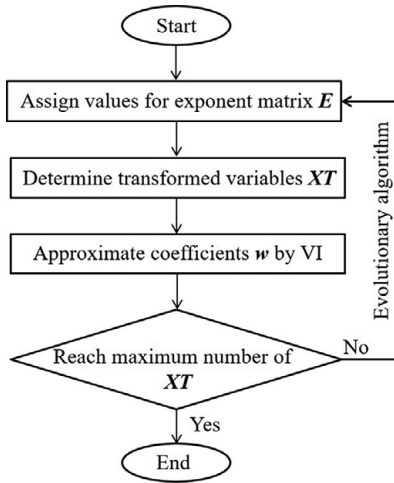


Figure 3. Schematic of VIEPR.

Table 1. Statistics of soil indices

Variable	Min.	Max.	Mean	STD
$C_c$	0.12	1.34	0.45	0.22
$e_0$	0.66	4.64	2.18	0.92
$LL$	25.00	166.2	67.95	24.87
$I_p$	8.00	113.9	34.72	18.75

Note. STD = standard deviation.

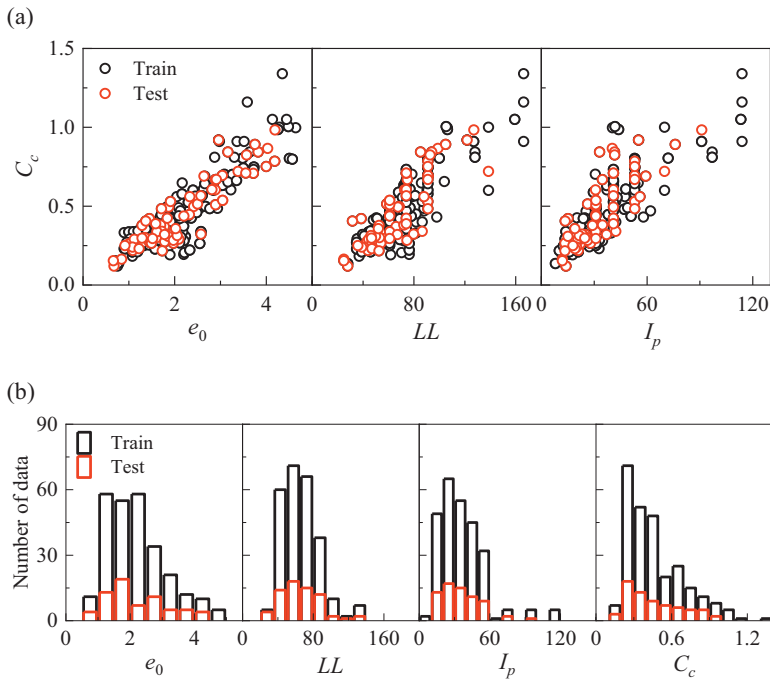
(25% ≤ LL ≤ 166.2%), and plasticity index (8% ≤ Ip ≤ 113.9%). They were then rescaled into the range (0, 1) using min–max normalization to eliminate scale effects before training.

Figure 4a presents the scatterplots of these soil indices, in which dense and sparse clusters can be easily distinguished. In this case, reliability evaluation becomes significant because accurate point estimations are elusive, due to insufficient knowledge of sparse clusters. On the other hand, the distributions of the indices are illustrated by the histograms in Figure 4b, in which they exhibit notably inhomogeneous distribution with positive values, similar to most observations (Camós et al., 2016; Jin and Yin, 2020; Shi and Wang, 2022; Tang and Phoon, 2019). Therefore, the natural logarithm is applied to the indices before training to ensure nonnegative predictions, similar to other published works (Ching et al., 2022; Löfman and Korkiala-Tanttu, 2021). This operation can be mathematically inverted in closed form. For instance, the median value of the predictions corresponds to the inverse logarithmic value of the mean of the normal distribution (Ching et al., 2022).

Among the database, 80% was randomly drawn to train the model and the remaining 20% was employed to test the model. For comparison, all models adopt the same training and testing sets such that their performances can be fairly compared. To assess model performance, the commonly used mean absolute error (MAE) and the coefficient of determination (R<sup>2</sup>) are computed as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i^m - y_i^p| \tag{12}$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i^m - \bar{y}_i)^2}{\sum_{i=1}^n (y_i^m - \bar{y}_i^a)^2} \tag{13}$$



**Figure 4.** Statistics of soil indices. (a) Scatterplots. (b) Histograms.

where  $n$  is the total number of data points;  $y_i^p$ ,  $y_i^m$ , and  $\bar{y}_i^m$  are predicted, measured, and average measured quantity of interest, respectively. For probabilistic approaches, model reliability can be assessed by the ratio of the target quantity of interest falling inside the predicted 95% CI. Specifically, this ratio is expected to reach 95% for a desirable 95% CI in statistics, implying that a model would be unreliable if the ratio is less than 95% (He et al., 2024a, 2024b; Lyu et al., 2023; Zhao and Wang, 2018). The generalization refers to the ability to produce accurate and reliable predictions on unseen data and can be reflected by model performance on the testing set (Zhang et al., 2021c).

## 4. Development and evaluation of uncertainty models

### 4.1. Configurations for model development

The training of data-driven models was conducted using an algorithm under a given configuration. In this section, QRF and VIEPR were proposed as baseline algorithms for predicting  $C_c$ , and their performances related to respective hyper-parameters. Nowadays, MAE integrated with 10-fold cross-validation (Stone, 1974) has been commonly used as the loss function when optimizing hyperparameters in several data-driven modeling platforms (Zhang et al., 2021b, 2022b) and has also been used in the current study.

During the training process of QRF, each validation set is evaluated by trees that are trained on the remaining datasets to prevent overfitting (Breiman, 2001). Subsequently, the generated MAE is adopted as the loss function as shown in Equation (12). Similarly, the KL divergence with regularization term is employed as the loss function of VIEPR to optimize its sole hyperparameter  $nt$ . The hyperparameters of QRF and VIEPR are summarized in Table 2. For QRF, the numbers of trees, the candidate features, and the minimum number of samples for split ( $ntree$ ,  $mtry$ , and  $min\_sample\_split$ ) are optimized because they determine the size of forests and the growth of trees (Pei et al., 2023). As a result, the depth of each tree can be automatically determined once they are fixed. As a widely used evolutionary algorithm, particle swarm optimization (PSO) is employed to optimize these hyper-parameters and search for the elements of the exponent matrix in VIEPR. More details of PSO are introduced in Appendix A.



**Table 2.** Hyperparameters of QRF and VIEPR

Algorithm	Hyperparameters	Description	Optimization method
QRF	<i>ntree</i>	Number of decision trees	PSO
	<i>mtry</i>	Number of candidate features for split	PSO
	<i>min_sample_split</i>	Minimum number of samples for split	PSO
	<i>max_depth</i>	Maximum depth	Automatic determined
VIERP	<i>nt</i>	Number of transformed variables	PSO

**Table 3.** Configuration of QRF and VIEPR based models

Configuration	QRF	VIEPR
Loss function	MAE	KL divergence
Bounds of <i>ntree</i>	[1, 200]	—
Bounds of <i>mtry</i>	[1, 3]	—
Bounds of <i>min_sample_split</i>	[1, 5]	—
Bounds of <i>nt</i>	—	[1, 10]
Training strategy	Dichotomy	Gradient descent
Optimizer	—	Adam
Epoch	100	20,000
Prior of coefficients	—	$N(0, 1)$
Number of MC samples for training	—	1000
Number of MC samples for predicting	—	1000

The configuration of QRF and VIEPR is summarized in Table 3, where sufficient ranges of hyperparameters can be arbitrarily assigned and optimized to find the optimal one for predicting any quantity of interest. QRF determines the split by dichotomy to minimize MAE, while VIEPR is trained by gradient descent to minimize the KL divergence. The Adam optimizer is employed because of its effectiveness in handling nonstationary objectives (Kingma and Ba, 2014; Zhang et al., 2022b). QRF utilizes efficient dichotomy and only requires 500 epochs for hyperparameter optimization, whereas VIEPR updates the posterior coefficients over 20,000 epochs. Since the coefficients in VIEPR do not involve physical knowledge, their prior distributions are assigned a standard normal distribution. Finally, a total of 1000 MC samples are drawn to evaluate the KL divergence of VIEPR and are also employed for predictions.

#### 4.2. Training processes of models

According to the configuration and data source described in the preceding sections, the training results of QRF and VIEPR are presented in Figure 5, illustrating the evolution of their losses. In Figure 5a, the loss value of QRF drops rapidly at the beginning and stabilizes after around 100 epochs, reaching its minimum. This corresponds to the optimized hyperparameters, that is,  $ntree = 180$ ,  $mtry = 1$ , and  $min\_sample\_split = 1$ . Figure 6 presents one of its trees to illustrate the optimized QRF, where each node attaches the split value of the feature, the MAE value, and the number of samples inside. Starting with an initial MAE of 0.45 at the root node, the error gradually decreases as the tree branch undergoes random splits. This indicates that tree structures effectively minimize errors by distributing datasets to leaf nodes with similar features.

Similarly, Figure 5b illustrates the loss values of VIEPR, using 1, 4, 7, and 10 transformed variables as examples. The VIEPR model with only one transformed variable quickly reaches a

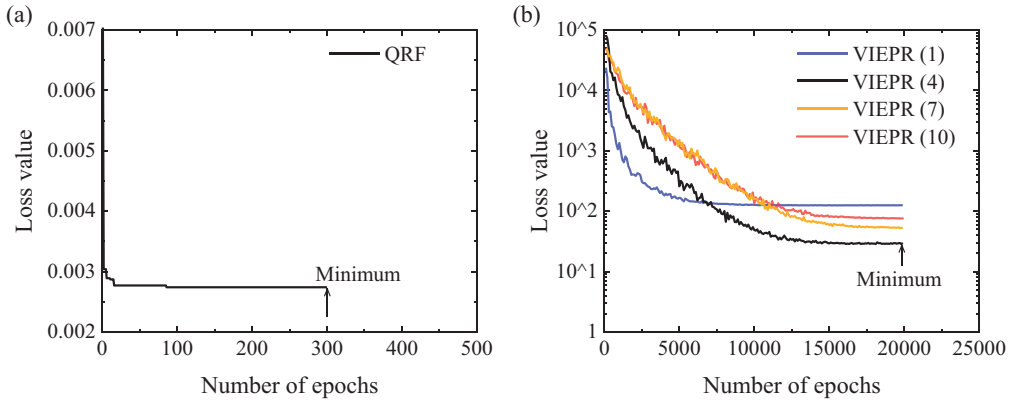


Figure 5. Evolution of loss value. (a) QRF. (b) VIEPR (number of transformed variables).

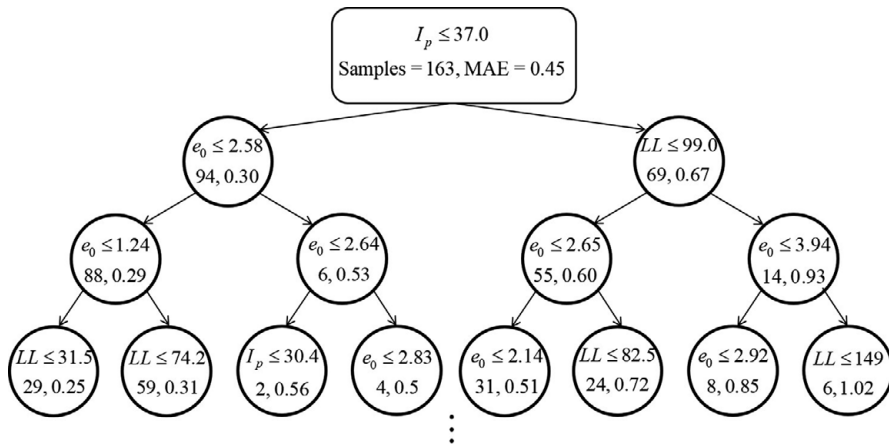


Figure 6. Structure of one tree of QRF.

relatively higher convergence value at around 1000 epochs primarily because of limited fitting capability. By contrast, VIEPR models with multiple transformed variables attain relatively lower loss values through more epochs, due to larger search space. The loss values for models with 7 and 10 variables are close to each other, while the minimum loss value is attained by the model with four variables. This is mainly because apart from likelihood, the KL divergence also includes a regularization term to penalty model complexity, as shown in the first term of Equation (10). Consequently, the VIEPR model with four transformed variables reaches the minimum loss, and its expression is formulated by

$$y_{\text{VIEPR}} = w_0 + w_1 e_0 LL + w_2 \frac{LL}{I_p} + w_3 LL + w_4 e_0 \tag{14}$$

where each coefficient in the VIEPR model follows a normal distribution and can be expressed by  $w_0 \sim N(-1.364, 0.062)$ ;  $w_1 \sim N(-0.001, 0.001)$ ;  $w_2 \sim N(-0.253, 0.026)$ ;  $w_3 \sim N(0.006, 0.001)$ ;  $w_4 \sim N(0.324, 0.027)$ . Figure 7 illustrates the prior and posterior of coefficients in VIEPR, in which the posterior distribution of coefficients was derived in Equation (14). Utilizing the QRF and VIEPR models with the minimum losses, material indices can be predicted. Note that Equation (14) represents the expression in interior computations and requires the addition of a natural logarithmic base to generate nonnegative predictions, due to data preprocessing before training.

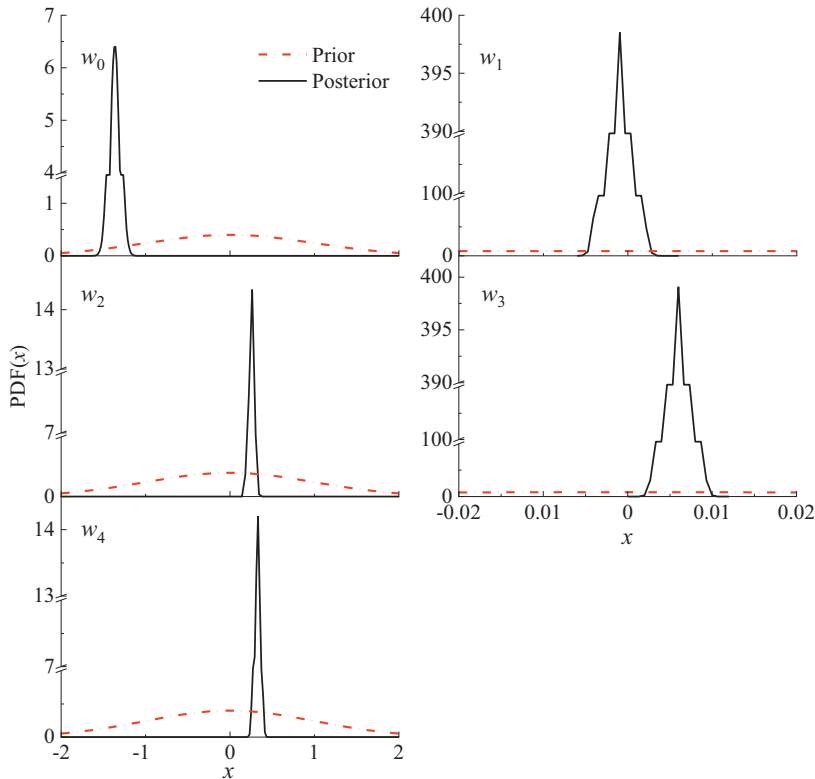
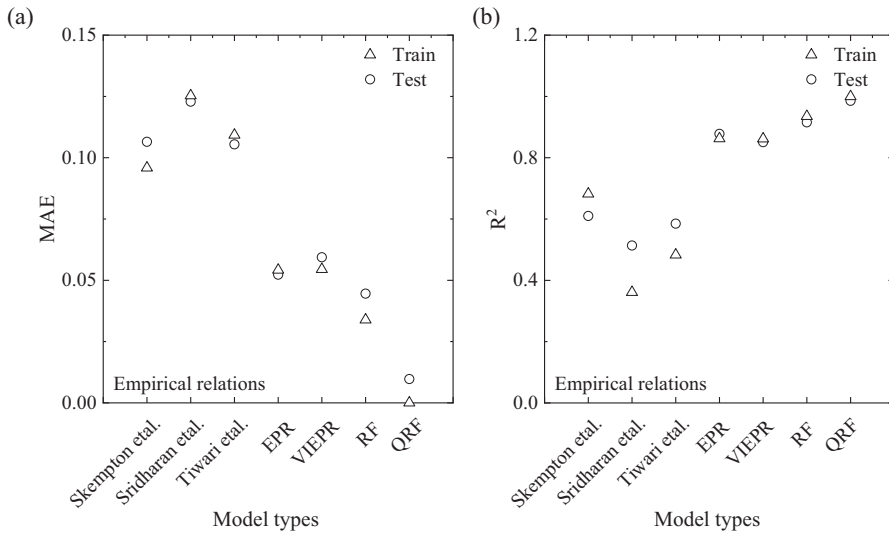


Figure 7. Prior and posteriors of coefficients in VIEPR.

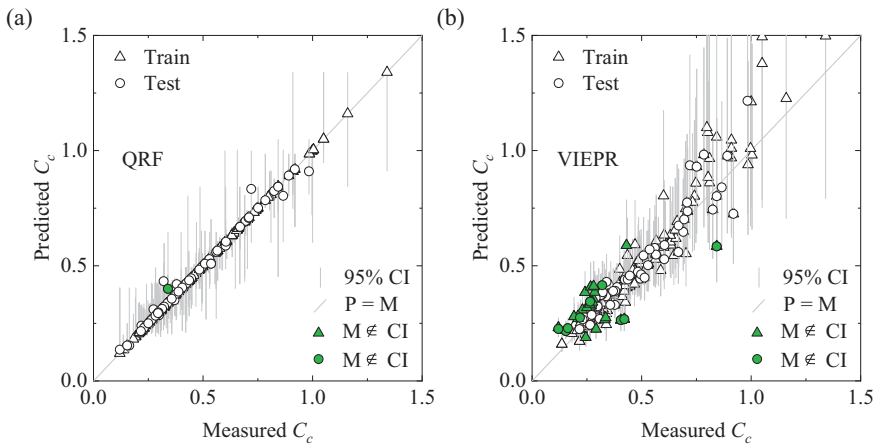
#### 4.3. Accuracy and reliability evaluation of models

Based on the above training results, developed QRF and VIEPR models were applied to predict  $C_c$  values and were also compared with RF and EPR using the same training and testing sets (Zhang et al., 2021d). By comparison, QRF makes inferences based on the generated CDF, while RF relies on the average output of each tree. On the other hand, EPR followed the same procedure as VIEPR, but its coefficients are approximated by the least-squares method.  $C_c$  is also computed by three commonly used empirical relations (Skempton and Jones, 1944; Sridharan and Nagaraj, 2000; Tiwari and Ajmera, 2012), denoted three relations in Figure 8 as a benchmark.

To assess model accuracy, Figure 8a and 8b presents the performance of the four models as well as three relations, in terms of MAE and  $R^2$  values, respectively. Notably, these empirical relations showed significantly larger errors and less agreement with measurements, compared with others. For machine learning-based models, QRF and RF attained notably lower errors, reaching less than one-third of those generated by VIEPR and EPR on both training and testing sets. This difference can be attributed to model structures. Specifically, EPR-based models are fixed expressions using several variables, while RF-based models allow flexible random trees and splits to minimize errors. Compared to the results generated by RF (Zhang et al., 2021d), QRF observed significantly improved accuracy on both training and testing sets. This is primarily because QRF makes predictions based on the generated CDF incorporating distribution information in each leaf node, while RF merely relies on the average output of each tree. On the other hand, VIEPR showed a similar level of accuracy as EPR with a minor increase of error in the testing set, which primarily relates to the regularization term on model complexity in VIEPR. Overall, the trends observed in the  $R^2$  values align with the discussion on the MAE values. Among the four models, QRF generated the highest  $R^2$  and showed the greatest agreement with the measured  $C_c$  values, demonstrating its superior capability in capturing correlations from limited data.

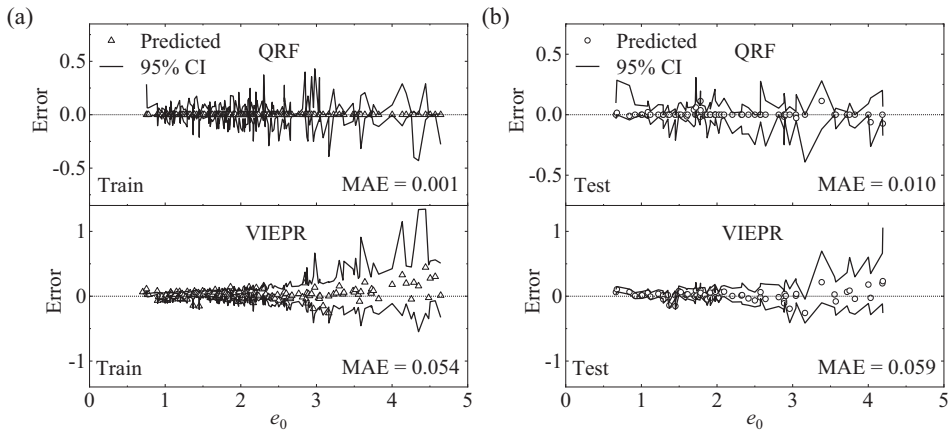


**Figure 8.** Performance of models in terms of (a) MAE and (b)  $R^2$ .



**Figure 9.** Comparison between measured and predicted  $C_c$  with 95% CI generated by (a) QRF and (b) VIEPR.

Regarding reliability evaluation, Figure 9a and 9b illustrates the results of QRF and VIEPR through a measured versus predicted  $C_c$  (median) relationship, along with 95% CIs to examine model reliabilities. Specifically, the measured  $C_c$  values outside 95% CIs on the training and testing sets are plotted as green triangles and circles, of which the number can reveal the reliability of models. Ideally, a 95% CI approximately covers 95% of measurements, and therefore, excessive measurements outside the interval indicate poor reliability and vice versa. Figure 9a shows that nearly all measured  $C_c$  values fell in the 95% CI predicted by QRF, apart from one measurement in the testing set. Such a performance relates to the weighted sum of distribution information over an entire forest, therefore offering desirable reliability and generalization ability. In contrast, Figure 9b shows that VIEPR observes around 10% of measurements falling outside the 95% CIs, showing a slightly underestimated uncertainty due to limited model structures. Also, predictions of QRF are distributed denser around the diagonal than VIEPR, and this observation also aligns with what has been discussed earlier about  $R^2$ .



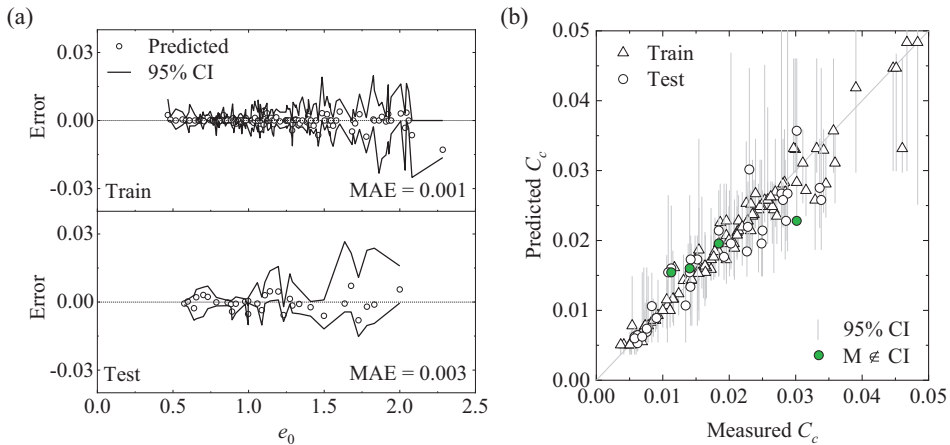
**Figure 10.** Relationship of predicted error and uncertainty on  $e_0$ . (a) Training set. (b) Testing set.

To investigate the dependency of model accuracy and uncertainty on input variables,  $e_0$  is taken as an example to illustrate its relationship between error and 95% CI in Figure 10. The results show that QRF obtained less error and uncertainty than explicit VIEPR along the distribution of  $e_0$ . Therefore, QRF is expected to perform more competitively than VIEPR to optimize cost-effectiveness in reliability-based design, due to reliable control of design parameters. For QRF, smaller uncertainty was found to appear in the densely distributed range of  $e_0$ , that is, 1.0–2.0, indicating that more datasets refine predicted CDF and enhance confidence level. In this regard, the understandable CDF using a simple weighted form allows engineers to easily understand the principle of interval estimates in tree structure-based models. By comparison, the uncertainty of VIEPR relies on the contribution of each variable in Equation (14), in which increased  $e_0$  enlarges the predicted uncertainty of  $C_c$  since they are positively related. Such a relationship is uniform with physics phenomena because a larger  $e_0$  implies broader space to be compressed.

With the regularization term, VIEPR generated a consistent error and mean standard deviation on the training set (0.054 and 0.075) than on the testing set (0.059 and 0.072). Since model accuracy and uncertainty relate to the mean and variance of polynomials, respectively, it was possible to observe such a result on randomly sampled training and testing sets. A key advantage of EPR models is their explicit expression, making them highly accessible and easy to apply in practice (Gomes et al., 2021; Nassr et al., 2018). Although not as accurate as QRF, the developed VIEPR can be readily formulated and even applied through Excel. For instance, Equation (14) can be used by engineers without expertise in data-driven modeling and also delivered into other scientific computing software. For small datasets, these baseline approaches can fuse data from multiple sources to generate predictions for material indices of interest, similar to other research works (Bozorgzadeh et al., 2019; Ching and Phoon, 2020; He et al., 2024a; Wu et al., 2022).

## 5. Verification of generalization with application to creep index prediction

The previous results have demonstrated the capabilities of QRF and VIEPR with UQ, improving their applicability in engineering practice compared to deterministic methods. The findings indicate that QRF generates more reliable predictions, exhibiting superior accuracy and reliability using the same database. Moreover, its flexible structures and easily adjustable hyperparameters contribute to a simpler training process and application. Several platforms have already been developed to integrate RF with automatic hyperparameter optimization and model development (Zhang et al., 2021b, 2022b), ensuring its availability. Considering these highly proposing characteristics, QRF is suggested to further verify its generalization ability through additional material index prediction.



**Figure 11.** Performance of QRF in predicting  $C_\alpha$ . (a) Accuracy. (b) Reliability.

As a common design parameter,  $C_\alpha$  is associated with multiple soil physical indices and involves substantial uncertainty.  $C_\alpha$  holds significant importance in engineering practice as it affects the long-term performance of infrastructure such as settlement. Therefore, it is worthwhile to explore  $C_\alpha$  for reliability-based design. The database was drawn from Zhang et al. (2020) and includes 151 datasets, where the  $C_\alpha$  of various clays was measured with three commonly observed variables:  $e_0$ , clay content  $CI$  and  $I_p$ . For these datasets, 80% was randomly drawn for model training, while the remaining 20% was reserved for testing. The same configuration of QRF used earlier is employed.

Figure 11a presents the predicted error of  $C_\alpha$  along the distribution of  $e_0$  using QRF. The results show that QRF observed smaller MAE values, compared to 0.0012 and 0.0039 on the training and testing sets given by RF in the previous study (Zhang et al., 2020), demonstrating superior accuracy and generalization ability. Apart from point estimation, the 95% CI was provided in Figure 11a, which became broader in the sparsely distributed range of  $e_0$ , especially the range from 1.5 to 2.0 in the testing set. This means that QRF can perceive unknown information and generate a reasonable level of confidence when applied to reliability-based design. Figure 11b compares the predicted and measured  $C_c$  values, in which the majority of measured  $C_\alpha$  values fell in the predicted 95% CI with merely four outliers out of 31 datasets in the testing set. It is worth noting that the ratio of outliers did not strictly align with the expected 5%, which relates to insufficient datasets. These results showcase the effectiveness of QRF in enhancing accuracy and reliability on both training and testing sets, thus validating its generalization ability.

## 6. Conclusions

This study has tailored two generic uncertain modeling approaches based on RF and EPR. Specifically, RF is integrated with quantile regression to extract distribution information in leaf nodes for UQ, while VI is employed to approximate parametric posterior distributions of coefficients in EPR. Thereafter, QRF and VIEPR were developed and applied to predict the soil compression index of clays.

The results indicated that QRF effectively reduced errors and covered almost all observations in the predicted 95% CI by integrating tree structures and distribution information inside. In contrast, VIEPR efficiently found parametric posterior distributions of coefficients to generate explicit solutions but observed greater errors and uncertainty due to restricted model structures. Therefore, QRF is suggested to optimize cost-effectiveness for reliability-based design due to efficient control of design accuracy and range. VIEPR is recommended to find user-friendly solutions even for those without expertise in data-driven modeling, offering enhanced reliability and wide applicability.

Compared with deterministic counterparts, QRF offered slightly improved accuracy than RF mainly because QRF generated a robust inference incorporating the distribution information in each leaf node, rather than merely relying on the average of each tree. Conversely, VIEPR showed a minor decrease in accuracy than EPR, primarily due to an additional penalty in model complexity. The dependency of model accuracy and uncertainty on input variables has been also investigated. Both QRF and VIEPR observed larger error and uncertainty in sparsely distributed ranges of input variables, meaning that they can recognize where their drawbacks exist and provide reasonable confidence in predictions.

Taking into account superior performance in predicting the compression index, QRF was further utilized to predict additional soil creep index using a new database. The accuracy and generalization of QRF were validated by comparing to RF in the previous study, while also incorporating interval estimates. It is worth noting that the proposed approach to quantify uncertainty is not limited to RF but also various tree structure-based models such as gradient boosting and extremely randomized trees. These highly promising characteristics strongly suggest that the proposed approaches have immense potential to enhance the data-driven modeling framework and can be further extended to a broader range of applications in scientific computing domains.

**Data availability statement.** Data are well reported in this study and will be made available from the corresponding author upon reasonable request.

**Acknowledgements.** The authors want to thank anonymous reviewers for their calibrations and suggestions.

**Author contribution.** Geng-Fu He: Writing—review and editing, Writing—original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. Pin Zhang: Writing—review and editing, Visualization, Formal analysis, Methodology, Conceptualization. Zhen-Yu Yin: Writing—review and editing, Software, Supervision, Project administration, Funding acquisition.

**Funding statement.** This research was financially supported by the Research Grants Council (RGC) of Hong Kong Special Administrative Region Government (HKSARG) of China (Grant Nos.: 15220221, 15227923, 15229223, T22-607/24-N, E-PolyU501/24). The second author is supported by the Royal Society under the Newton International Fellowship.

**Competing interest.** The authors declare there are no competing interests.

## References

- Al-Bittar T and Soubra A-H** (2014) Probabilistic analysis of strip footings resting on spatially varying soils and subjected to vertical or inclined loads. *Journal of Geotechnical and Geoenvironmental Engineering* 140, 04013043. [https://doi.org/10.1061/\(ASCE\)GT.1943-5606.0001046](https://doi.org/10.1061/(ASCE)GT.1943-5606.0001046).
- Bozorgzadeh N, Harrison JP and Escobar MD** (2019) Hierarchical Bayesian modelling of geotechnical data: Application to rock strength. *Géotechnique* 69, 1056–1070. <https://doi.org/10.1680/jgeot.17.P.282>.
- Breiman L** (1996). Bagging predictors. *Machine Learning* 24, 123–140. <https://doi.org/10.1007/BF00058655>.
- Breiman L** (2001) Random forests. *Machine Learning* 45, 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Camós C, Špačková O, Straub D and Molins C** (2016) Probabilistic approach to assessing and monitoring settlements caused by tunneling. *Tunnelling and Underground Space Technology*, 51, 313–325. <https://doi.org/10.1016/j.tust.2015.10.041>.
- Ching J and Phoon KK** (2017) Characterizing uncertain site-specific trend function by sparse Bayesian learning. *Journal of Engineering Mechanics* 143, 04017028. [https://doi.org/10.1061/\(ASCE\)em.1943-7889.0001240](https://doi.org/10.1061/(ASCE)em.1943-7889.0001240).
- Ching J and Phoon K-K** (2019) Constructing site-specific multivariate probability distribution model using Bayesian machine learning. *Journal of Engineering Mechanics* 145, 04018126. [https://doi.org/10.1061/\(asce\)em.1943-7889.0001537](https://doi.org/10.1061/(asce)em.1943-7889.0001537).
- Ching J and Phoon K-K** (2020) Measuring similarity between site-specific data and records from other sites. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* 6, 04020011. <https://doi.org/10.1061/ajrua6.0001046>.
- Ching J, Phoon K-K, Li K-H and Weng M-C** (2019) Multivariate probability distribution for some intact rock properties. *Canadian Geotechnical Journal* 56, 1080–1097. <https://doi.org/10.1139/cgj-2018-0175>.
- Ching J, Phoon KK and Wu CT** (2022) Data-centric quasi-site-specific prediction for compressibility of clays. *Canadian Geotechnical Journal* 59, 2033–2049. <https://doi.org/10.1139/cgj-2021-0658>.
- Gal Y and Ghahramani Z** (2016) Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. arXiv preprint:1506.02142. Available at <https://arxiv.org/abs/1506>.
- Ghorbani A and Eslami A** (2021) Energy-based model for predicting liquefaction potential of sandy soils using evolutionary polynomial regression method. *Computers and Geotechnics* 129, 103867. <https://doi.org/10.1016/j.compgeo.2020.103867>.



- Ghorbani A and Hasanzadehshooilli H** (2018). Prediction of UCS and CBR of microsilica-lime stabilized sulfate silty sand using ANN and EPR models; application to the deep soil mixing. *Soils and Foundations* 58, 34–49. <https://doi.org/10.1016/j.sandf.2017.11.002>.
- Giustolisi O and Savic DA** (2006) A symbolic data-driven technique based on evolutionary polynomial regression. *Journal of Hydroinformatics* 8, 207–222. <https://doi.org/10.2166/hydro.2006.020b>.
- Gomes GJC, Gomes RGdS and Vargas EdA** (2021) A dual search-based EPR with self-adaptive offspring creation and compromise programming model selection. *Engineering with Computers* 38, 2155–2173. <https://doi.org/10.1007/s00366-021-01313-x>.
- Guan Z and Wang Y** (2020) Statistical charts for determining sample size at various levels of accuracy and confidence in geotechnical site investigation. *Géotechnique* 70, 1145–1159. <https://doi.org/10.1680/jgeot.18.P.315>.
- Guan Z and Wang Y** (2022) Assessment of liquefaction-induced differential ground settlement and lateral displacement using standard penetration tests with consideration of soil spatial variability. *Journal of Geotechnical and Geoenvironmental Engineering* 148, 04022018. [https://doi.org/10.1061/\(ASCE\)GT.1943-5606.000277](https://doi.org/10.1061/(ASCE)GT.1943-5606.000277).
- Guan Z, Wang Y and Phoon K-K** (2024) Fusion of sparse non-co-located measurements from multiple sources for geotechnical site investigation. *Canadian Geotechnical Journal* 61, 1574–1592. <https://doi.org/10.1139/cgj-2023-0289>.
- Hao L and Naiman DQ** (2007) *Quantile Regression*. London: Sage.
- He G-F, Zhang P and Yin Z-Y** (2024a) Active learning inspired multi-fidelity probabilistic modelling of geomaterial property. *Computer Methods in Applied Mechanics and Engineering* 432, 117373. <https://doi.org/10.1016/j.cma.2024.117373>.
- He G-F, Zhang P, Yin Z-Y and Goh SH** (2024b) Multifidelity-based Gaussian process for quasi-site-specific probabilistic prediction of soil properties. *Canadian Geotechnical Journal* 61, 2304–2322. <https://doi.org/10.1139/cgj-2023-0641>.
- Ho TK** (1998) The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 832–844. <https://doi.org/10.1109/34.709601>.
- Jin YF and Yin ZY** (2020) An intelligent multi-objective EPR technique with multi-step model selection for correlations of soil properties. *Acta Geotechnica* 15, 2053–2073. <https://doi.org/10.1007/s11440-020-00929-5>.
- Karimi J, Rahmani F and Jia SB** (2023) Improving the detection efficiency of IRAND based on machine learning. *Computer Physics Communications* 291, 108833. <https://doi.org/10.1016/j.cpc.2023.108833>.
- Kingma DP and Ba J** (2014) Adam: A method for stochastic optimization. arXiv preprint:1412.6980. Available at <https://arxiv.org/abs/1412.6980>.
- Koenker R and Bassett G, Jr** (1978) Regression quantiles. *Econometrica* 46, 33–50. <https://doi.org/10.2307/1913643>.
- Kullback S and Leibler RA** (1951) On information and sufficiency. *Annals of Mathematical Statistics* 22, 79–86. <https://doi.org/10.1214/aoms/1177729694>.
- Lee S-H and Song J** (2017) System identification of spatial distribution of structural parameters using modified transitional Markov chain Monte Carlo method. *Journal of Engineering Mechanics* 143, 04017099. [https://doi.org/10.1061/\(ASCE\)EM.1943-7889.0001316](https://doi.org/10.1061/(ASCE)EM.1943-7889.0001316).
- Li KQ, Yin ZY and Liu Y** (2023) A hybrid SVR-BO model for predicting the soil thermal conductivity with uncertainty. *Canadian Geotechnical Journal* 61, 258–274. <https://doi.org/10.1139/cgj-2023-0105>.
- Li Y, Rahardjo H, Satyanaga A, Rangarajan S and Lee DT-T** (2022) Soil database development with the application of machine learning methods in soil properties prediction. *Engineering Geology* 306, 106769. <https://doi.org/10.1016/j.enggeo.2022.106769>.
- Löfman MS and Korkiala-Tanttu LK** (2021) Transformation models for the compressibility properties of Finnish clays using a multivariate database. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards* 16, 330–346. <https://doi.org/10.1080/17499518.2020.1864410>.
- Lykkegaard MB, Dodwell TJ and Moxey D** (2021) Accelerating uncertainty quantification of groundwater flow modelling using a deep neural network proxy. *Computer Methods in Applied Mechanics and Engineering* 383, 113895. <https://doi.org/10.1016/j.cma.2021.113895>.
- Lyu B, Hu Y and Wang Y** (2023) Data-driven development of three-dimensional subsurface models from sparse measurements using Bayesian compressive sampling: A benchmarking study. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* 9, 04023010. <https://doi.org/10.1061/ajrua6.Rueng-935>.
- Marasco S, Marano GC and Cimellaro GP** (2022) Evolutionary polynomial regression algorithm combined with robust Bayesian regression. *Advances in Engineering Software* 167, 103101. <https://doi.org/10.1016/j.advengsoft.2022.103101>.
- Meinshausen N and Ridgeway G** (2006) Quantile regression forests. *Journal of Machine Learning Research* 7, 983–999.
- Nagaraj T and Srinivasa Murthy B** (1986) A critical reappraisal of compression index equations. *Géotechnique* 36, 27–32. <https://doi.org/10.1680/geot.1986.36.1.27>.
- Nassar A, Javadi A and Faramarzi A** (2018) Developing constitutive models from EPR-based self-learning finite element analysis. *International Journal for Numerical and Analytical Methods in Geomechanics* 42, 401–417. <https://doi.org/10.1002/nag.2747>.
- Olivier A, Shields MD and Graham-Brady L** (2021) Bayesian neural networks for uncertainty quantification in data-driven materials modeling. *Computer Methods in Applied Mechanics and Engineering* 386, 114079. <https://doi.org/10.1016/j.cma.2021.114079>.
- Pei T, Qiu T and Shen C** (2023) Applying knowledge-guided machine learning to slope stability prediction. *Journal of Geotechnical and Geoenvironmental Engineering* 149, 04023089. <https://doi.org/10.1061/jggefck.Gteng-11053>.
- Phoon KK, Cao ZJ, Ji J, Leung YF, Najjar S, Shuku T, Tang C, Yin ZY, Ikumasa Y and Ching J** (2022) Geotechnical uncertainty, modeling, and decision making. *Soils and foundations* 62, 101189. <https://doi.org/10.1016/j.sandf.2022.101189>.

- Phoon K-K and Kulhawey FH** (1999) Characterization of geotechnical variability. *Canadian Geotechnical Journal* 36, 612–624. <https://doi.org/10.1139/t99-038>.
- Psaros AF, Meng X, Zou Z, Guo L and Karniadakis GE** (2023) Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *Journal of Computational Physics* 477, 111902. <https://doi.org/10.1016/j.jcp.2022.111902>.
- Reis I, Baron D and Shahaf S** (2018) Probabilistic random forest: A machine learning algorithm for noisy data sets. *Astronomical Journal* 157, 16. <https://doi.org/10.3847/1538-3881/aaf101>.
- Rezania M, Faramarzi A and Javadi AA** (2011) An evolutionary based approach for assessment of earthquake-induced soil liquefaction and lateral displacement. *Engineering Applications of Artificial Intelligence* 24, 142–153. <https://doi.org/10.1016/j.engappai.2010.09.010>.
- Rezania M, Javadi AA and Giustolisi O** (2010) Evaluation of liquefaction potential based on CPT results using evolutionary polynomial regression. *Computers and Geotechnics* 37, 82–92. <https://doi.org/10.1016/j.compgeo.2009.07.006>.
- Shahin MA** (2016) State-of-the-art review of some artificial intelligence applications in pile foundations. *Geoscience Frontiers* 7, 33–44. <https://doi.org/10.1016/j.gsf.2014.10.002>.
- Shi C and Wang Y** (2022) Assessment of reclamation-induced consolidation settlement considering stratigraphic uncertainty and spatial variability of soil properties. *Canadian Geotechnical Journal*, 59(7), 1215–1230. <https://doi.org/10.1139/cgj-2021-0349>.
- Skempton AW and Jones O** (1944) Notes on the compressibility of clays. *Quarterly Journal of the Geological Society* 100, 119–135. <https://doi.org/10.1144/GSL.JGS.1944.100.01-04.08>.
- Sridharan A and Nagaraj H** (2000) Compressibility behaviour of remoulded, fine-grained soils and correlation with index properties. *Canadian Geotechnical Journal* 37, 712–722. <https://doi.org/10.1139/t99-128>.
- Stone M** (1974) Cross-validators choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)* 36, 111–133. <https://doi.org/10.1111/j.2517-6161.1974.tb00994.x>.
- Sun D, Xu J, Wen H and Wang D** (2021) Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: A comparison between logistic regression and random forest. *Engineering Geology* 281, 105972. <https://doi.org/10.1016/j.enggeo.2020.105972>.
- Tang C and Phoon K-K** (2019) Characterization of model uncertainty in predicting axial resistance of piles driven into clay. *Canadian Geotechnical Journal*, 56(8), 1098–1118. <https://doi.org/10.1139/cgj-2018-0386>.
- Tiwari B and Ajmera B** (2012) New correlation equations for compression index of remolded clays. *Journal of Geotechnical and Geoenvironmental Engineering* 138, 757–762. [https://doi.org/10.1061/\(ASCE\)GT.1943-5606.0000639](https://doi.org/10.1061/(ASCE)GT.1943-5606.0000639).
- Vázquez-Escobar J, Hernández JM and Cárdenas-Montes M** (2021) Estimation of machine learning model uncertainty in particle physics event classifiers. *Computer Physics Communications* 268, 108100. <https://doi.org/10.1016/j.cpc.2021.108100>.
- Wang HL, Yin ZY, Zhang P and Jin YF** (2020) Straightforward prediction for air-entry value of compacted soils using machine learning algorithms. *Engineering Geology* 279, 105911. <https://doi.org/10.1016/j.enggeo.2020.105911>.
- Wang Y and Zhao T** (2017) Statistical interpretation of soil property profiles from sparse data using Bayesian compressive sampling. *Géotechnique* 67, 523–536. <https://doi.org/10.1680/jgeot.16.P.143>.
- Wu S, Ching J and Phoon K-K** (2022) Quasi-site-specific soil property prediction using a cluster-based hierarchical Bayesian model. *Structural Safety* 99, 102253. <https://doi.org/10.1016/j.strusafe.2022.102253>.
- Yoshida I, Tomizawa Y and Otake Y** (2021) Estimation of trend and random components of conditional random field using Gaussian process regression. *Computers and Geotechnics* 136, 104179. <https://doi.org/10.1016/j.compgeo.2021.104179>.
- Zhang P, Jin YF and Yin ZY** (2021a) 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, 1588–1602. <https://doi.org/10.1002/nag.3215>.
- Zhang P, Yin ZY and Chen Q** (2022a) Image-based 3D reconstruction of granular grains via hybrid algorithm and level set with convolution kernel. *Journal of Geotechnical and Geoenvironmental Engineering* 148, 04022021. [https://doi.org/10.1061/\(ASCE\)GT.1943-5606.0002790](https://doi.org/10.1061/(ASCE)GT.1943-5606.0002790).
- Zhang P, Yin ZY and Jin YF** (2021b) Machine learning-based modelling of soil properties for geotechnical design: review, tool development and comparison. *Archives of Computational Methods in Engineering* 29, 1229–1245. <https://doi.org/10.1007/s11831-021-09615-5>.
- Zhang P, Yin ZY and Jin YF** (2021c) State-of-the-art review of machine learning applications in constitutive modeling of soils. *Archives of Computational Methods in Engineering* 28, 3661–3686. <https://doi.org/10.1007/s11831-020-09524-z>.
- Zhang P, Yin ZY and Jin YF** (2022b) Bayesian neural network-based uncertainty modelling: Application to soil compressibility and undrained shear strength prediction. *Canadian Geotechnical Journal* 59, 546–557. <https://doi.org/10.1139/cgj-2020-0751>.
- Zhang P, Yin ZY, Jin YF and Chan THT** (2020) A novel hybrid surrogate intelligent model for creep index prediction based on particle swarm optimization and random forest. *Engineering Geology* 265, 105328. <https://doi.org/10.1016/j.enggeo.2019.105328>.
- Zhang P, Yin ZY, Jin YF, Chan THT and Gao FP** (2021d). Intelligent modelling of clay compressibility using hybrid meta-heuristic and machine learning algorithms. *Geoscience Frontiers* 12, 441–452. <https://doi.org/10.1016/j.gsf.2020.02.014>.
- Zhang P, Yin ZY, Jin YF and Liu XF** (2021e) Modelling the mechanical behaviour of soils using machine learning algorithms with explicit formulations. *Acta Geotechnica* 17, 1403–1422. <https://doi.org/10.1007/s11440-021-01170-4>.
- Zhang W, Wu C, Li Y, Wang L and Samui P** (2019) Assessment of pile drivability using random forest regression and multivariate adaptive regression splines. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards* 15, 27–40. <https://doi.org/10.1080/17499518.2019.1674340>.

- Zhao S, Tan D, Lin S, Yin Z and Yin J** (2023) A deep learning-based approach with anti-noise ability for identification of rock microcracks using distributed fibre optic sensing data. *International Journal of Rock Mechanics and Mining Sciences* 170, 105525. <https://doi.org/10.1016/j.ijrmms.2023.105525>.
- Zhao T and Wang Y** (2018) Simulation of cross-correlated random field samples from sparse measurements using Bayesian compressive sensing. *Mechanical Systems and Signal Processing* 112, 384–400. <https://doi.org/10.1016/j.ymsp.2018.04.042>.
- Zhu JG and Yin JH** (2000) Strain-rate-dependent stress-strain behavior of overconsolidated Hong Kong marine clay. *Canadian Geotechnical Journal* 37, 1272–1282. <https://doi.org/10.1139/t00-054>.

## Appendix A. Particle swarm optimization (PSO)

As an evolutionary algorithm, PSO searches for a globally optimal solution by iteratively updating a population of particles within a series of generations. Each particle represents a candidate solution and contains three components including position ( $X$ ), velocity ( $V$ ), and an objective function value. For the  $i$ th particle at the  $(k + 1)$ th generation, it is updated by (Zhang et al., 2022a).

$$V_i^{k+1} = wV_i^k + c_1r_1(P_i^k - X_i^k) + c_2r_2(P_g^k - X_i^k) \quad (\text{A1})$$

$$X_i^{k+1} = X_i^k + V_i^{k+1} \quad (\text{A2})$$

where  $c_1$  and  $c_2$  = acceleration coefficients;  $w$  = an inertia weight;  $r_1$  and  $r_2$  = random numbers, distributed within the range [0, 1];  $P_i$  denotes the optimal location of the  $i$ th particle, and  $P_g$  denotes the globally optimal location among all particles. Table A1 summarizes the configuration of PSO used in this study, in which  $w$ ,  $c_1$ , and  $c_2$  are 1.0, 1.5, and 1.5, respectively, within commonly used ranges. PSO is assigned 20 particles and 100 generations to guarantee its exploration ability. Considering hyperparameters should be integers in this study,  $X$  is always approximated to the closest integer.

**Table A1.** Configuration of PSO

$w$	$c_1$	$c_2$	Popsiz	Maxgen
1.0	1.5	1.5	20	100

Note. Popsiz = size of population; Maxgen = maximum number of generations.