Article:
Chen, W., Ding, J., Shi, C. et al. (2 more authors) (2024) Geotechnical correlation field-informed and data-driven prediction of spatially varying geotechnical properties. Computers and Geotechnics, 171. 106407. ISSN 0266-352X

https://doi.org/10.1016/j.compgeo.2024.106407

Geotechnical Correlation Field-informed and Data-Driven Prediction of Spatially Varying Geotechnical Properties

Weihang Chen [a], **Jianwen Ding** [a,*], **Chao Shi** [b], **Tengfei Wang** [c,d], **David P. Connolly** [e]

[a] School of Transportation, Southeast University, Nanjing 210096, China

[b] School of Civil and Environmental Engineering, Nanyang Technological University, Singapore 639798, Singapore

[c] School of Civil Engineering, Southwest Jiaotong University, Chengdu 610031, China

[d] MOE Key Laboratory of High-Speed Railway Engineering, Southwest Jiaotong University, Chengdu 610031, China

[e] School of Civil Engineering, University of Leeds, Leeds LS2 9JT, UK

* Corresponding author

Email: jwding@seu.edu.cn

Abstract

Geotechnical measurements are often limited, leading to the use of interpolation techniques for interpreting spatial variations in geotechnical properties from sparse geo-data. Traditional geostatistical methods suffer from significant computational complexity. On the other hand, data-driven approaches often lack integration with geotechnical domain knowledge, potentially oversimplifying or complicating predictions related to the spatial variability of geotechnical properties. This study introduces a novel framework that combines geotechnical knowledge with data-driven methods to model inherent soil spatial variability incorporating Geotechnical Correlation Field (GCF) that reflects domain knowledge. The GCF, influenced by Autocorrelation Function (ACF) types and Scale of Fluctuation (SOF), provides a flexible basis for accurately representing spatially varying geotechnical properties. Using a large synthetic database comprising known ACF types and SOFs, we constructed a series of specialized neural networks. These networks identify random field parameters at different sites based on sparse data, and the estimated parameters can be directly used to calculate GCFs for a given site. The performance of the proposed method is validated using a set of synthetic data and a real case history in New Zealand. The results demonstrate the model can accurately predict random field parameters for irregularly spaced geo-data, even with limited information. Significantly, the GCFs offer improved physical interpretations and enhance the performance of subsurface modeling. The computational complexity of this method is independent of the number of soil cells, making it highly efficient and scalable.

Keywords: Spatial variability; Data-driven Method; Random field theory; Site investigation; Neural Network

1 Introduction

The spatial variability of soil properties is a significant source of uncertainty in geotechnical engineering (Phoon et al., 2022; Shi and Wang, 2023; Uzielli et al., 2005). This challenge is prevalent across various fields, such as accurately assessing the extent and concentration of contaminated sites in environmental engineering or predicting mineral reserves and distribution densities in mining engineering (Gu et al., 2023; Wang and Shi, 2023; Zhang et al., 2020; Zhao et al., 2018). Precise and high-resolution geographic information can assist engineers in analysis and design optimization (Chen et al., 2023; Wang et al., 2020; Zhao et al., 2020). However, obtaining high-density in-situ measurements using expensive and time-consuming testing equipment is often impractical. Engineering projects typically retrieve sparse measurements from limited locations. For instance, it is common to use Cone Penetration Tests (CPT) at intervals of 25-100 m along the ground surface (Guan et al., 2020) to characterize subsurface ground conditions (Collico et al., 2024). Although CPT tests provide continuous measurements with depth, the interpretation of spatial distribution of soil properties is challenging, particularly in the horizontal direction (Xie et al., 2022b, 2022a; W. Zhang et al., 2022).

The use of interpolation or data-driven methods to predict soil properties at unsampled locations is an active area of study (Shi and Wang, 2021a; Wang et al., 2018; Xie et al., 2024). In this context, several techniques, including Geostatistical methods, Distance-based methods, Bayesian analysis, and data-driven methods, have found wide application (Hu and Wang, 2024; Wang and Chen, 2023; Zou et al., 2017). For instance, Kriging provides direct estimates and uncertainty assessments of soil properties at unsampled locations, but when geotechnical measurements are sparse and exhibit significant variations, Kriging predictions often tend to capture the average trend and disregard the

43 spatial variability of soil properties (Nag et al., 2023). Notably, when simulating large-scale or fine-
44 resolution random fields, traditional conditional random field methods may suffer from low
45 computation due to excessively large correlation matrices (Yang et al., 2021). Recently, Yang and
46 Ching (2021) proposed an efficient method for simulating conditional random fields (CRFs) by
47 utilizing the Kronecker-product to decompose the large correlation matrix, and further extended this
48 approach to simulate multivariate cross-correlated CRFs (Z. Yang et al., 2022). Bayesian Compressive
49 Sensing (BCS) is a non-parametric and data-driven method that can be directly applied to non-
50 stationary random fields (Wang and Zhao, 2017). However, BCS does not incorporate specific basis
51 function forms relevant to geotechnical modeling (Cami et al., 2020; Phoon and Wang, 2019). Inverse
52 Distance Weighting (IDW) and the Geotechnical Distance Field (GDF) method both attempt to recover
53 the soil properties at unsampled points based on the "distance" between unsampled points and sampled
54 points (Xie et al., 2022b). It's important to note that soil properties exhibit location-specific
55 dependencies, implying that soil properties within a certain scale of fluctuation (SOF, δ) in a particular
56 subsurface stratigraphy are correlated (Phoon et al., 2003). Therefore, relying solely on relative
57 distances to infer soil properties at unsampled points may overlook this fact.

58 Although data-driven methods can be straightforward to use, they tend to overlook the
59 geotechnical expertise, such as random field theory. This may lead to overly simplified or complex
60 subsurface modeling results. These 'black-box' models may impede effective collaboration between
61 engineers and the models, compromising the prediction performance. The incorporation of
62 geotechnical knowledge can steer the prediction towards correct solutions. Therefore, a random field-
63 informed and data-driven model is introduced to predict spatially varying soil property from sparse

site-specific measurements. In this approach, random field theory is embedded in the data-driven model through Geotechnical Correlation Field (GCF). Essentially, GCF is derived from the decomposition of the correlation matrix ($\boldsymbol{C}$) used in random field theory to describe the correlation ρ between soil cells. For instance, Fig. 1 (a)-(c) illustrate three GCFs for point A, where darker colors correspond to a higher correlation between the respective soil cell and point A. The advantages of employing GCFs are threefold: (1) Soil properties at point A are only correlated with those within a specific range, strictly adhering to the fact that soil properties exhibit location-specific dependencies; (2) The correlation between point A and other points is affected by the type of autocorrelation function (ACF)(Ching et al., 2019), providing a flexible tool for accurately describing soil properties; (3) The impact of point A on other points is constrained by the Scale of Fluctuation (SOF), facilitating the geotechnical engineer to enhance further collaboration with the model by controlling the SOF. In contrast, Fig. 1(d)-(f) represent three Geotechnical Distance Fields (GDFs) corresponding to the distances from the sampling point to the ground surface (Line A), the corner point (Point B), and the exploration location (Line B). Compared to methods using two-dimensional coordinates as input features, GDFs significantly enhance feature dimensions and modeling accuracy. However, GDFs overlook the fact that soil properties exhibit location-specific dependencies. Furthermore, in comparison to BCS and GDFs, GCFs can comprehensively consider multiple ACF types, providing flexible input features, as opposed to relying on a single base function or fixed feature type.

It is noteworthy that the generation of GCF depends on the specific site's ACF types and corresponding SOFs. Estimating random field parameters based on sparsely distributed measurement data is challenging (Dasaka and Zhang, 2012; Qi and Liu, 2019; Yan et al., 2023). Traditional methods

include the method of moments (Lloret-Cabot et al., 2014; Onyejekwe et al., 2016), maximum likelihood estimation (Xiao et al., 2018) and Bayesian analysis(Cami et al., 2020; Ching et al., 2018). However, these methods are influenced by human experience and involve certain assumptions about describing SOFs, mainly used for estimating vertical SOFs. Recently, some machine learning-based methods for estimating horizontal and vertical SOFs have been proposed, with prediction models based on Convolutional Neural Networks (CNN) demonstrating good performance and efficiency (Zhang et al., 2021; 2022). Nevertheless, current CNN methods cannot assess the optimal ACF type based on measurement data.

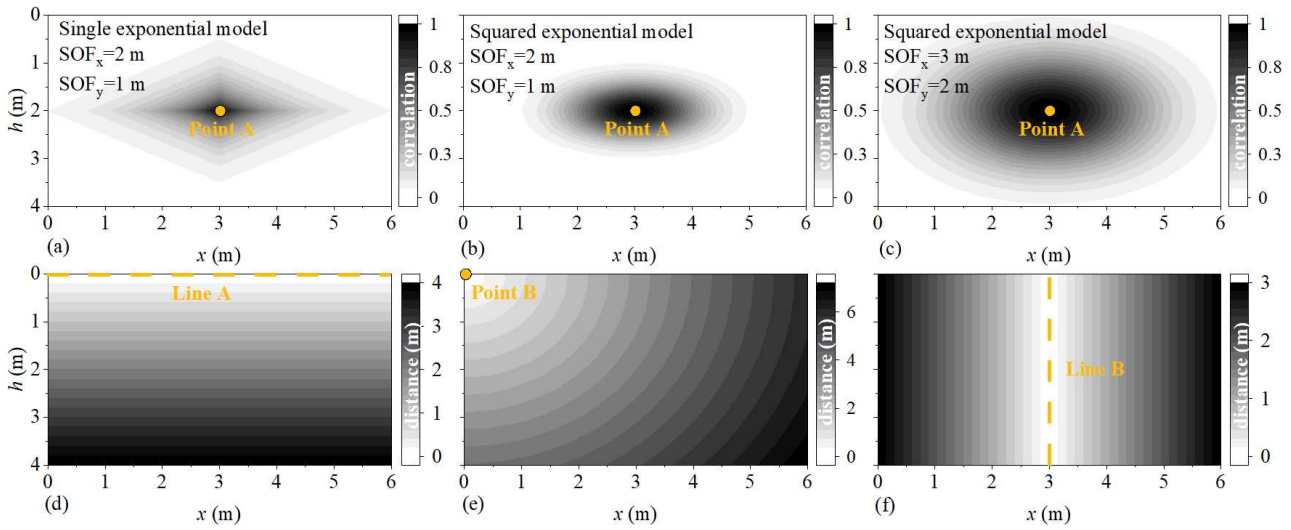


Fig. 1 Illustration of GCFs and GDFs: GCFs (a-c); GDFs (d-f)

To address these challenges, this study draws inspiration from the generation process of traditional conditional random field to construct a framework for the data-driven model. This framework provides support for the proposed subsurface modeling method, by allowing the developed data-driven model to integrate geotechnical knowledge. The framework employs random field samples generated from specified GCFs serve as training data to train a neural network that is used for forward prediction of spatially varying soil properties and estimation of random field parameters. The performance of the

proposed method is illustrated using a set of synthetic data and a real case study in New Zealand. The remainder of this study is organized as follows: In section two, the proposed subsurface modeling framework is introduced. The construction and validation of the random field parameter estimation model is in section three. Subsequently, the performance of the proposed GCF-based subsurface modeling approach is validated by a set of synthetic data. In section five, further validation is conducted using a real case study in New Zealand, followed by conclusions.

2 Proposed Methods for Predicting Spatially Varying Geotechnical Properties

The essence of data-driven subsurface modeling lies in utilizing machine learning (ML) techniques to learn the correlation between sampling positions (coordinates) and the corresponding soil properties, and then using the trained ML model to infer soil properties at unsampled positions. Notably, this study effectively integrates random field theory into the process of subsurface modeling using GCF, aiming to enhance modeling reliability and reduce modeling uncertainty.

In GCF, instead of using 2D or 3D coordinates to represent the positions of soil cells, it employs the correlation between each soil cell and every sampled soil cell. Therefore, the framework proposed in this study first estimates the random field parameters of the site (e.g., types of ACFs, horizontal SOF, and vertical SOF). Subsequently, based on random field theory and predicted parameters, corresponding GCFs are computed. Finally, a machine learning (ML) model is employed to learn the complex relationship between GCFs and observed soil properties, enabling predictions of soil properties at unsampled locations. As illustrated in Fig. 2, the geotechnical subsurface modeling process in this study involves six key steps, using the case of three CPTs as an illustrative example:

(1) Collect CPT data (e.g., cone tip resistance $\boldsymbol{q_c}$) along with corresponding horizontal coordinates

(***L***).

(2) Utilize the ACFs type prediction model (Model #1) to assess the probability *P* of measured CPT data belonging to each ACF type. This study considers seven common ACF types, ensuring that $\sum P_i$=1, *i*=1-7.

(3) Utilizing the horizontal and vertical SOFs prediction models (Model #2 and Model #3) to accurately estimate the measured CPT data's horizontal and vertical SOFs. Notably, SOFs predicted based on different ACF types can vary. Therefore, both Model #2 and Model #3 incorporate seven sub-models (corresponding to the seven ACF types), resulting in seven sets of predicted horizontal and vertical SOFs corresponding to different ACF types.

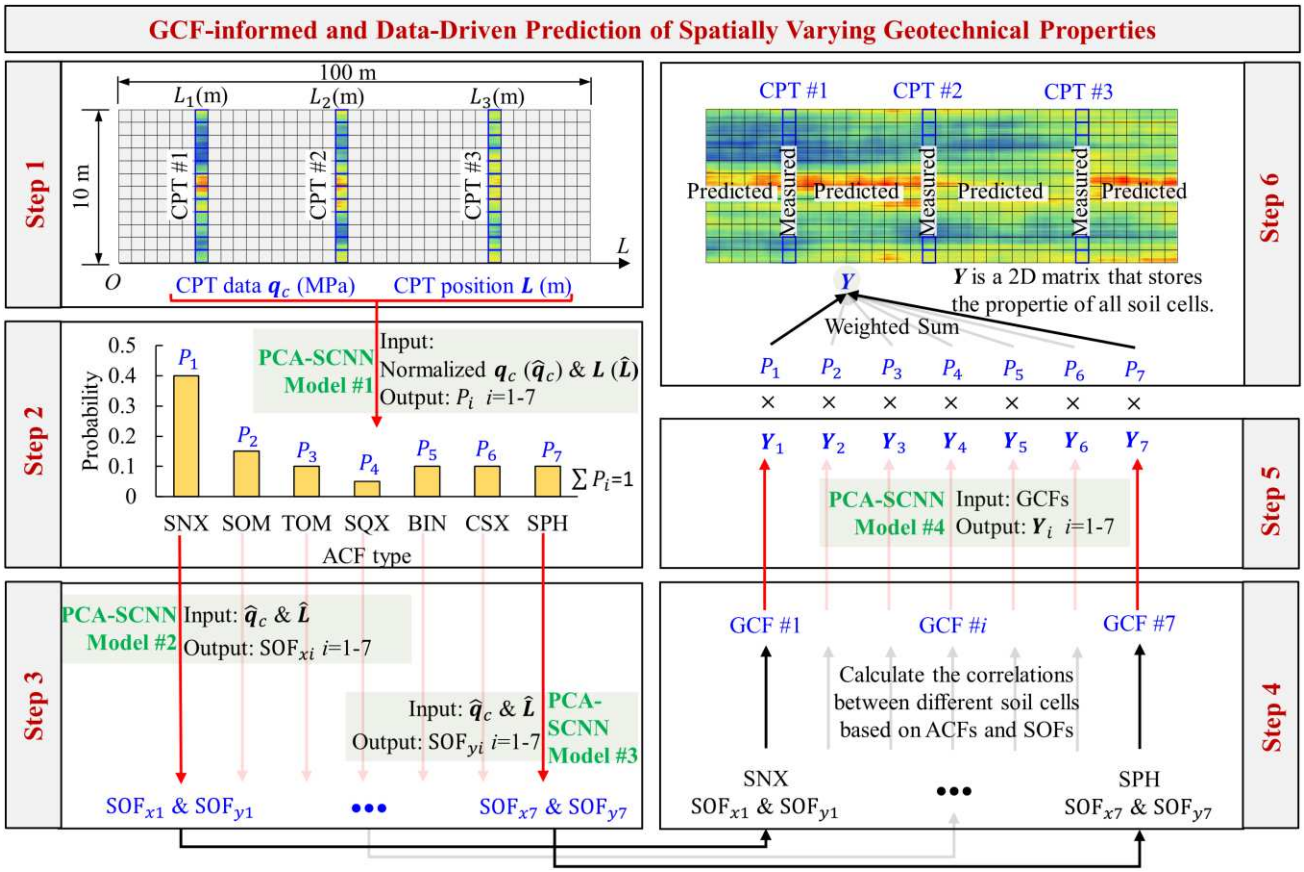
(4) Employing seven sets of random field parameters (ACFs and corresponding horizontal and vertical SOFs) to calculate the seven sets of GCFs, as outlined in Section 2.3.

(5) Constructing the subsurface prediction model (Model #4) involves using GCFs as input and the geotechnical properties of soil cells as output. Once Model #4 is well-trained, it can be utilized to predict soil properties at unexplored locations. Notably, since Step (4) generates seven sets of GCFs, it allows for the creation of seven sub-models within Model #4.

(6) Weighted summation of the predicted results ($\boldsymbol{Y}_i$) from the seven sub-Model #4, based on the probabilities (P_i), yields the subsurface modeling outcome ($\boldsymbol{Y} = \sum P_i \boldsymbol{Y}_i$) closely correlated with observed data.

Notably, in Steps (2), (3), and (5), independent data-driven models (PCA-SC neural networks) are employed for predicting ACFs types, estimating horizontal and vertical SOFs, and subsurface modeling. Specifically, the ACFs types and SOFs prediction models are trained using a large synthetic

q_c database containing known ACFs types and SOFs. Detailed descriptions of Models #1-4 are provided in the subsequent sections.



Note: $\hat{q}_c$ and $\hat{L}$ represent the standardized cone tip resistance q_c and horizontal coordinates L. Standardization of data is necessary for neural network models to expedite the training process and enhance model performance. In Step (2), the seven types of ACFs can be referenced from Table 1. Since the values of GCF range between 0 and 1, larger GCF values indicate stronger correlation between two soil cells.

Fig. 2. Framework for Proposed Soil Property Recovery for 3 CPTs

2.1 PCA–Shortcut Connection Neural Network

It is noteworthy that a universal PCA-SCNN structure, as depicted in Fig. 3, is utilized for ACFs classification (Model #1), horizontal SOF estimation (Model #2), vertical SOF estimation (Model #3), and subsurface modeling (Model #4). While the training datasets fed into the PCA-SCNN differ across tasks, resulting in different inputs and outputs for each model, the remaining structure of the models is consistent.

In Fig. 2, Models #1-3 take standardized $\boldsymbol{q}_c$ and $\boldsymbol{L}$ as input, while Model #4 utilizes GCFs. It's worth noting that CPT tests provide almost continuous soil information in the vertical direction, leading to higher-dimensional $\boldsymbol{q}_c$ and GCFs. However, handling high-dimensional input features increases model complexity, extends training time, and requires substantial memory. This study proposes a solution: the PCA-NN structure. PCA, an unsupervised dimensionality reduction method, transforms multiple potentially correlated features into a smaller set of linearly uncorrelated principal components (PCs). This process eliminates redundancy in input data and reduces the number of neurons in the neural network's input layer. By simplifying the data-driven model structure, this enhances model convergence speed (Bai et al., 2023; He et al., 2016). The dimension of PCs is set to 100 in this study.

As the depth of a neural network increases, its non-linear representation strengthens (Hong et al., 2024; Zhang et al., 2023). But deeper networks introduce challenges like gradient vanishing and exploding. To tackle these issues, shortcut connections are incorporated into the model. Research by Li et al. (2018) indicates that these connections promote loss surface minimization and prevent chaotic behavior. Therefore, a shortcut connection is established between the input layer and the last hidden layer, concatenating input features with those of the final hidden layer to improve training efficiency and overall performance.

Fig. 3 displays the model structure. Notably, Monte–Carlo dropout (MC dropout) is applied after the hidden layers. Dropout randomly deactivates a portion of neuron connections during training, preventing the model from becoming too dependent upon specific neurons and thus reducing overfitting (P. Zhang et al., 2022). MC dropout extends this concept by randomly deactivating neuron

connections not only during training but also during prediction. While the network structure and weight parameters are fixed after training, MC dropout introduces randomness to the model structure. Through multiple predictions, an output distribution is obtained, assisting in the assessment of prediction uncertainty. The model formulation is as follows:

$$\boldsymbol{x} = \mathrm{PCA}(\mathrm{Input}) \tag{1}$$

$$\boldsymbol{h}^i = \mathrm{Dropout}\left[\mathrm{ReLU}\left(\boldsymbol{x}, \boldsymbol{\theta}^i, \boldsymbol{\beta}^i\right)\right] \quad i = 1, 2, 3 \tag{2}$$

$$\boldsymbol{h}^4 = \mathrm{Dropout}[\mathrm{ReLU}(\mathrm{concatenate}(\boldsymbol{x}, \boldsymbol{h}^3), \boldsymbol{\theta}^4, \boldsymbol{\beta}^4)] \tag{3}$$

$$\boldsymbol{o} = \mathrm{Dropout}[\mathrm{ReLU}(\boldsymbol{h}^4, \boldsymbol{\theta}^4, \boldsymbol{\beta}^4)] \tag{4}$$

where $\boldsymbol{x}$ denotes the input feature vector following PCA preprocessing; $\boldsymbol{h}$ is the hidden layer feature vector; $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are the weight and bias vectors for each layer, respectively. concatenate refers to the direct merging of two vectors. For activation functions, ReLU is frequently employed due to its rapid convergence, computational simplicity, and absence of gradient vanishing, defined as $\mathrm{ReLu}(v) = \max(0, v)$. It is essential to note that, the model is used for both regression and classification, and in classification models, the activation function in Eq. (4) should be changed as Softmax to ensure that the sum of probabilities across all classifications equals 1.

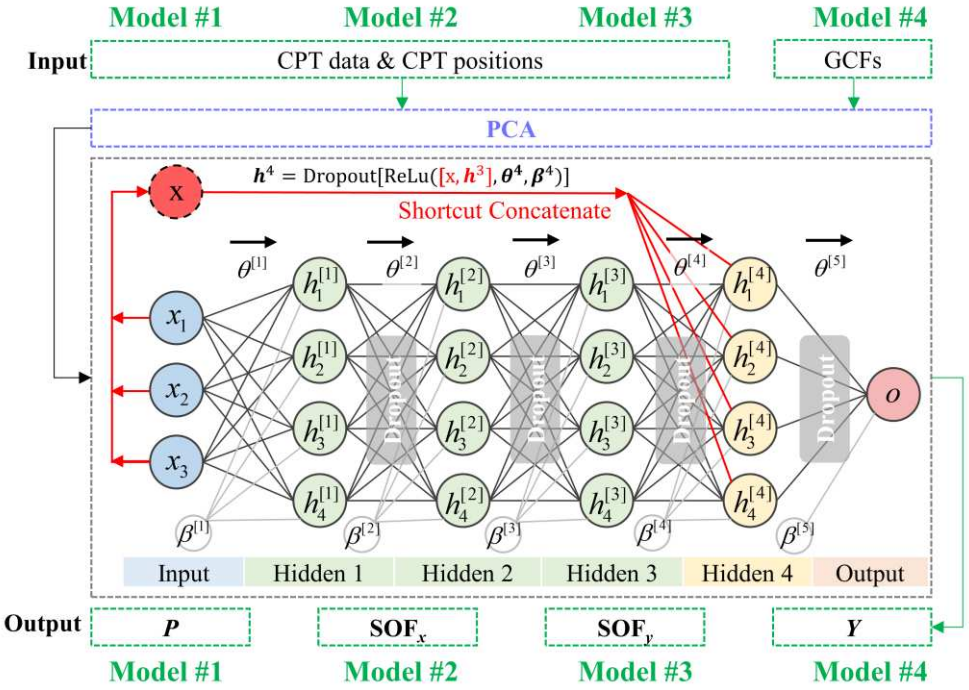


Fig. 3 Architecture of the PCA–Shortcut Connection Neural Network (PCA-SCNN)

2.2 Classification of ACF Types and Estimation of SOFs

In this study, the training samples for Models #1-3 consist of synthetic zero-mean stationary normal random fields $\boldsymbol{q}_c$ (Zhang et al., 2021). The CPT data depth is 10 m with a resolution of 0.05 m thus ensuring sufficient information in the synthetic sample path. The CPT ranges from 0 to 100 m horizontally with a resolution of 0.25 m. Given that the number of CPTs within a site is variable, four scenarios of CPT quantities are considered (2, 3, 4, and 5). The analytical modeling process for different CPT quantities follows a similar approach. To avoid redundancy, an example with 3 CPTs is illustrated in Fig. 4. Apart from the CPT quantity, this study considers three other variables:

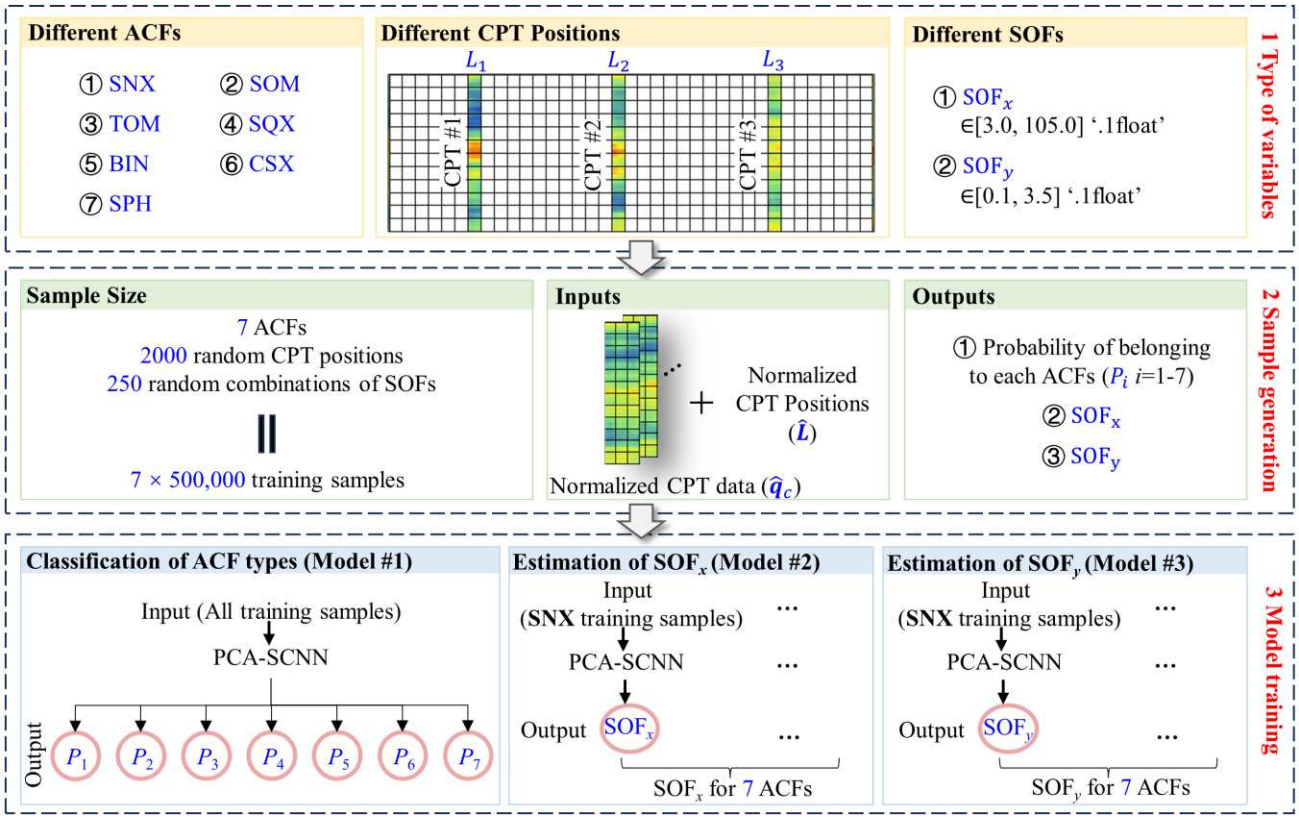


Fig. 4 Flow of ACF Type Classification and SOFs Estimation for 3 CPTs

(1) Different ACF types. Table 1 presents expressions for common ACFs (Cami et al., 2020), where τ_h and τ_v represent the horizontal and vertical distances of two soil cells, δ_h and δ_v are the horizontal and vertical SOFs. This study opts for seven commonly used one-parameter ACFs prevalent in geotechnical engineering. While some two-parameter ACFs that allow defining smoothness have been developed (Ching et al., 2018), their application in practical cases is relatively limited.

Table 1. Frequently Used Autocorrelation Functions (ACFs) - Adapted from Cami et al. (2020)

Autocorrelation model	Autocorrelation function $\rho(\tau_h, \tau_v)$	Frequency of usage
Single exponential (SNX)	$\exp\left[-2\left(\frac{\lvert\tau_h\rvert}{\delta_h}+\frac{\lvert\tau_v\rvert}{\delta_v}\right)\right]$	47 %
Second–order Markov (SOM)	$\left(1+4\frac{\lvert\tau_h\rvert}{\delta_h}\right)\left(1+4\frac{\lvert\tau_v\rvert}{\delta_v}\right)\exp\left[-4\left(\frac{\lvert\tau_h\rvert}{\delta_h}+\frac{\lvert\tau_v\rvert}{\delta_v}\right)\right]$	4 %
Third–order Markov (TOM)	$\left(1+\frac{16}{3}\frac{\lvert\tau_h\rvert}{\delta_h}+\frac{256}{27}\left(\frac{\lvert\tau_h\rvert}{\delta_h}\right)^2\right)\left(1+\frac{16}{3}\frac{\lvert\tau_v\rvert}{\delta_v}+\frac{256}{27}\left(\frac{\lvert\tau_v\rvert}{\delta_v}\right)^2\right)\exp\left(-\frac{16}{3}\left(\frac{\lvert\tau_h\rvert}{\delta_h}+\frac{\lvert\tau_v\rvert}{\delta_v}\right)\right)$	New

Squared exponential (SQX)	$\exp\left[-\pi\left(\frac{\tau_h^2}{\delta_h^2}+\frac{\tau_v^2}{\delta_v^2}\right)\right]$	15 %
Binary noise (BIN)	$\begin{cases}\left(1-\frac{\lvert\tau_h\rvert}{\delta_h}\right)\left(1-\frac{\lvert\tau_v\rvert}{\delta_v}\right) & \lvert\tau_h\rvert \le \delta_h \text{ and } \lvert\tau_v\rvert \le \delta_v \\ 0 & \text{otherwise}\end{cases}$	9 %
Cosine exponential (CSX)	$\cos\left(\frac{\lvert\tau_h\rvert}{\delta_h}\right)\cos\left(\frac{\lvert\tau_v\rvert}{\delta_v}\right)\exp\left(-\left(\frac{\lvert\tau_h\rvert}{\delta_h}+\frac{\lvert\tau_v\rvert}{\delta_v}\right)\right)$	10 %
Spherical (SPH)	$\begin{cases}\left[1-\frac{9}{8}\frac{\lvert\tau_h\rvert}{\delta_h}+\frac{27}{128}\left(\frac{\lvert\tau_h\rvert}{\delta_h}\right)^3\right]\left[1-\frac{9}{8}\frac{\lvert\tau_v\rvert}{\delta_v}+\frac{27}{128}\left(\frac{\lvert\tau_v\rvert}{\delta_v}\right)^3\right] & \lvert\tau_h\rvert \le \frac{4}{3}\delta_h \text{ and } \lvert\tau_v\rvert \le \frac{4}{3}\delta_v \\ 0 & \text{otherwise}\end{cases}$	15 %

(2) Different distributions of CPT positions. Due to constraints in equipment, site conditions, and costs, the positions of CPT measurements often exhibit irregular distribution. Existing data-driven models solely rely on CPT measurement data as input (Zhang et al., 2022), potentially overlooking the actual distances between different CPT data points. As depicted in Fig. 4, this study enhances the model's adaptability in unevenly spaced scenarios by incorporating the coordinates ($\boldsymbol{L}$) of each CPT test as an additional input.

(3) Different horizontal and vertical SOFs. The horizontal SOF range from 3 to 105 m, and the vertical SOF range from 0.1 to 3.5 m (Cami et al., 2020; Zhang et al., 2021).

Notably, assessing the ACF type and SOFs for unevenly spaced CPT data poses a more intricate challenge. This study addresses this by (1) utilizing a PCA-SCNN with robust nonlinear fitting capabilities to construct a data-driven model and (2) augmenting the training data volume. Consequently, in the process of constructing training samples, seven common ACFs, 2000 sets of random CPT positions, and 250 sets of random horizontal and vertical SOFs combinations are considered. Therefore, the total training dataset comprises 7 × 500,000 samples, with 500,000 samples for each ACF type. Detailed model information for 3 CPTs is provided in Table 2.

Table 2. Overview of the Architecture of the ACF Classification Model and the SOFs Estimation Model for 3 CPTs

Model	Number of models	Inputs	Outputs	Activation function & (number of neurons) in the output layer	Sample size
Model #1 (Classification of ACFs)	1	CPT data ($\hat{\boldsymbol{q}}_{\mathbf{c}}$) & positions ($\hat{\boldsymbol{L}}$)	P_i	Softmax (7)	7×500,000
Model #2 (Estimation of SOF$_x$)	7		SOF$_x$	ReLU (1)	500,000
Model #3 (Estimation of SOF$_y$)	7		SOF$_y$	ReLU (1)	500,000

Note: Model #1 is utilized to determine the probability of CPT data belonging to the 7 ACF types. Therefore, the training samples for Model #1 comprise all samples of the 7 ACF types (7 × 500,000). The 7 neurons in the output layer of Model #1 correspond to the probabilities P_i (i=1-7) of CPT data belonging to each ACF type. The activation function used is Softmax, ensuring $\sum P_i = 1$. Model #2 and Model #3 are used to estimate the horizontal and vertical SOFs of CPT data, respectively. Therefore, both Model #2 and Model #3 consist of 7 independent sub-models corresponding to the 7 ACF types. Each sub-model's training samples only include samples of one ACF type (500,000 samples). Each sub-model's output layer contains only 1 neuron, and the activation function used is ReLU.

After training, Model #1-3 can be utilized to predict the probabilitiesP_i (i=1-7) of measured CPT data $\boldsymbol{q}_{\mathrm{cm}}$belonging to the 7 ACF types, as well as the corresponding SOF$_x$ and SOF$_y$ for each ACF type. It is noteworthy that during the training process, the input of the Model #1-3 is [$\hat{\boldsymbol{q}}_{\mathbf{c}}$, $\hat{\boldsymbol{L}}$], where $\hat{\boldsymbol{q}}_{\mathbf{c}}$ is the normalized $\boldsymbol{q}_{\mathbf{c}}$; $\hat{\boldsymbol{L}}$ is the normalized position coordinate, $\hat{\boldsymbol{L}} = \boldsymbol{L}/100$. To ensure the accuracy of the prediction results, the real measurement data $\boldsymbol{q}_{\mathrm{cm}}$ should closely resemble the training data $\hat{\boldsymbol{q}}_{\mathbf{c}}$. Therefore, the following steps are needed: ① First, apply the Box-Cox method to transform $\boldsymbol{q}_{\mathrm{cm}}$ into a normal distribution (Zou et al., 2017), as shown in Eq. (5); ② Calculate the mean and standard deviation of the normal distribution, and transform the data into a standard normal distribution; ③ Apply Min-Max normalization to transform the data back into $\hat{\boldsymbol{q}}_{\mathbf{c}}$.

$$\boldsymbol{q}_{\mathrm{c}} = \begin{cases} \dfrac{\boldsymbol{q}_{\mathrm{cm}}^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln(\boldsymbol{q}_{\mathrm{cm}}) & \lambda = 0 \end{cases} \tag{5}$$

where λ is the power parameter that needs to be estimated. The optimal λ value can be efficiently determined using common statistical software such as the Scipy library implemented in Python.

2.3 Geotechnical Subsurface Modeling Using GCFs

Currently, spatial positions of sampling points are commonly represented using Euclidean distance-based coordinates. However, these coordinates have low-dimensional features and lack effective constraints from random field theory. To address this issue, this study proposes a high-dimensional GCF that conforms to random field theory to characterize the spatial positions of sampling points: (1) The GCF aligns with random field theory, where the correlation between two soil cells within the site is calculated using random field theory rather than simple Euclidean distance. (2) The GCF has higher-dimensional features. Instead of using 2D or 3D relative coordinates to represent the positions of soil cells, it employs the correlation between each soil cell and every sampled soil cell, as illustrated in Fig. 5.

In this study, GCFs are employed to characterize the correlation between soil cells (calculated based on ACFs, SOFs, and relative distances), effectively integrating random field theory into the data-driven model. The process of generating GCFs is illustrated in Fig. 5. Assuming discretization of the geological site into a 4 × 6 grid, where the second and fifth columns represent sampled points, and the other columns represent unsampled points. Taking soil cell e_5 as an example, the GCF of e_5 is defined as the correlation matrix ($\mathrm{GCF_{e5}}$) between e_5 and other soil cells within the site:

$$\mathrm{GCF_{e5}} = \begin{bmatrix} \rho_{5,1} & \rho_{5,5} & \rho_{5,9} & \rho_{5,13} & \rho_{5,17} & \rho_{5,21} \\ \rho_{5,2} & \rho_{5,6} & \rho_{5,10} & \rho_{5,14} & \rho_{5,18} & \rho_{5,22} \\ \rho_{5,3} & \rho_{5,7} & \rho_{5,11} & \rho_{5,15} & \rho_{5,19} & \rho_{5,23} \\ \rho_{5,4} & \rho_{5,8} & \rho_{5,12} & \rho_{5,16} & \rho_{5,20} & \rho_{5,24} \end{bmatrix} \tag{6}$$

where $\rho_{5,j}$ represents the correlation between soil cell e_5 and e_j, calculated using different ACFs from Table 1, and $\rho_{5,j} = \rho_{j,5}$. Considering that only the soil properties of cells e_5 to e_8 and e_{17} to e_{20} are known, a total of 8 GCFs can be generated: $\mathrm{GCF_{e5}}$ to $\mathrm{GCF_{e8}}$ and $\mathrm{GCF_{e17}}$ to $\mathrm{GCF_{e20}}$. Therefore, the spatial position of soil cell e_i (i=1-24) in the geological correlation field can be represented by an 8-

dimensional coordinate vector: [$\rho_{i,5}, \rho_{i,6}, \rho_{i,7}, \rho_{i,8}, \rho_{i,17}, \rho_{i,18}, \rho_{i,19}, \rho_{i,20}$], which indicates the correlation between e_i and the 8 sampled soil cells.

The essence of subsurface modeling based on GCFs is to utilize a data-driven model to learn the relationship between the coordinate vectors of e_i and their corresponding geological properties, represented as $\hat{q}_c^i = f([\rho_{i,5}, \rho_{i,6}, \rho_{i,7}, \rho_{i,8}, \rho_{i,17}, \rho_{i,18}, \rho_{i,19}, \rho_{i,20}])$, where f represents a complex implicit function. In this study, PCA-SCNN is employed to solve *f*, which corresponds to Model #4 in Fig. 2. The model takes an 8-dimensional coordinate vector as input and outputs the corresponding geological properties of the soil cells. As shown in Fig. 5, the input and output of the training set for Model #4 can be represented as $x^i = [\rho_{i,5}, \rho_{i,6}, \rho_{i,7}, \rho_{i,8}, \rho_{i,17}, \rho_{i,18}, \rho_{i,19}, \rho_{i,20}]$, $y^i = [\hat{q}_c^i]$, where i = 5-8 and 17-20. The test set can be represented as $x^j = [\rho_{j,5}, \rho_{j,6}, \rho_{j,7}, \rho_{j,8}, \rho_{j,17}, \rho_{j,18}, \rho_{j,19}, \rho_{j,20}]$, where j = 1-4, 9-16, and 21-24. Therefore, the training set consists of 8 samples, while the test set consists of 16 samples, all with a feature dimension of 8. If there are *N* discrete soil cells within the site, among which *M* cells have been measured, then the training set will have *M* samples, and the test set will have *N*−*M* samples, all with a feature dimension of *M*.

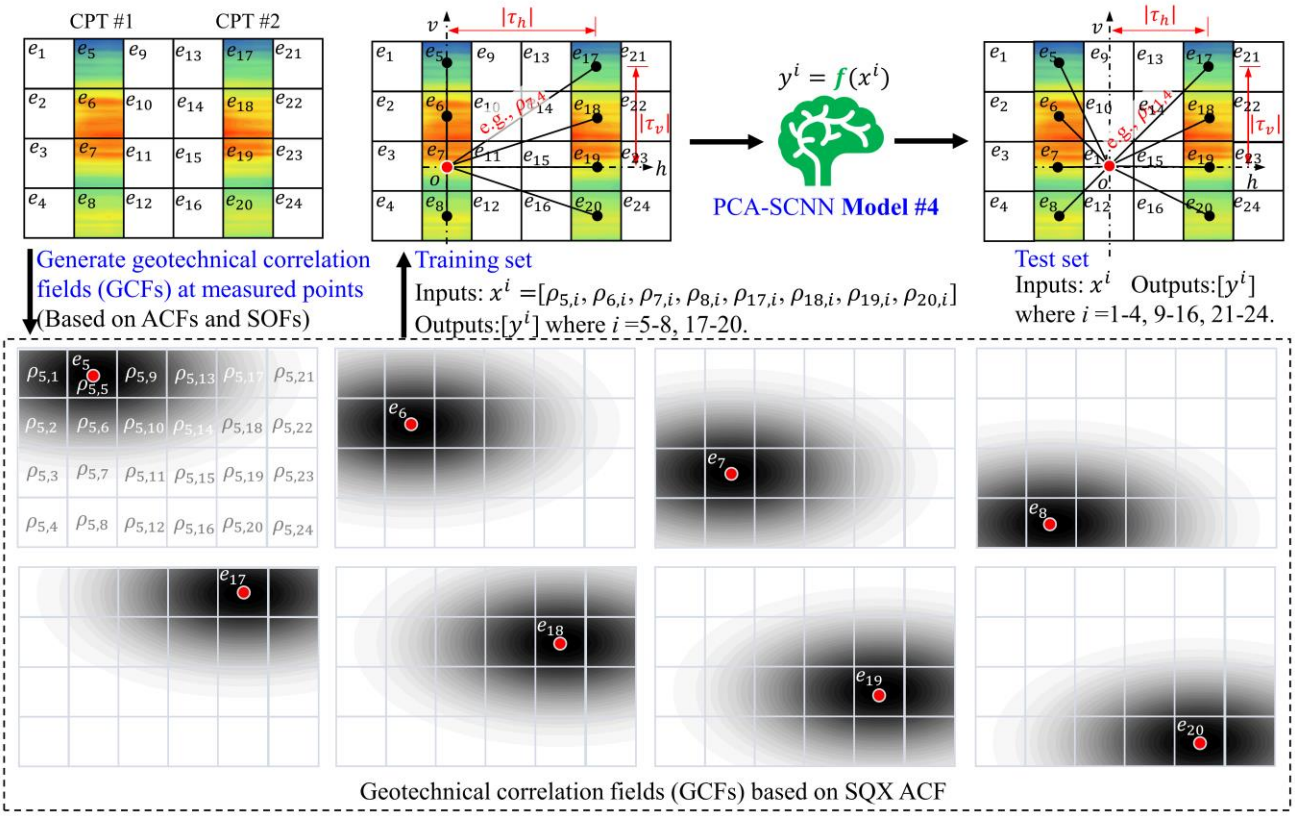


Fig. 5. Illustration of Geotechnical Correlation Fields and Associated Training and Test Set (Note: When the resolution of CPT does not align with the size of the soil cell, the average property of the soil cell is the mean value of the included measurement data.)

As depicted in Fig. 6 (a)-(g), using the seven different ACFs from Table 1, GCFs for soil cells at the coordinate origin (0 m, 0 m) are generated under horizontal and vertical SOFs of 2 m and 1 m, respectively. The contour lines of different GCFs exhibit distinct shapes, indicating significant variations in the influence range of the soil cells around the coordinate origin. Fig. 6 (h) and (g) illustrate one-dimensional correlation curves for different GCFs along the y=0 m section. While these curves demonstrate similar trends, GCFs generated by SNX, BIN, CSX, and SPH are not differentiable at zero lag, leading to lower smoothness of the corresponding sample paths (Ching et al., 2019). Real CPT data are often challenging to interpret with a single ACF. Vanmarcke (1983) proposed overlaying two or more ACFs to create a more flexible ACF. As shown in Fig. 2, this study adopts this overlay approach, generating seven sets of GCFs using seven sets of ACFs, followed by constructing seven

independent Model #4 for subsurface modeling. Finally, the predictions of the seven subsurface models are weighted by P_i to obtain a modeling result closely related to the measured data.

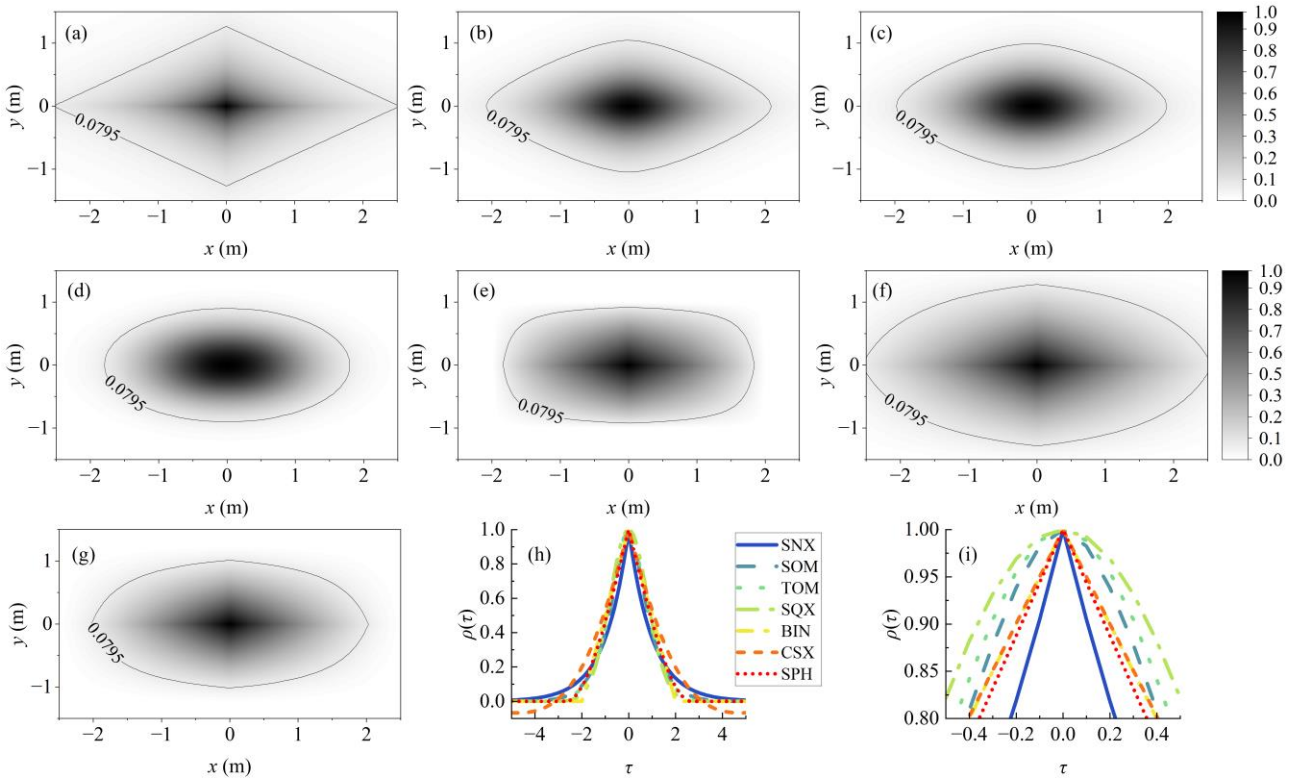


Fig. 6. Correlations of Various GCFs Under Horizontal and Vertical Scale of Fluctuation (SOF) Set at 2 m and 1 m Respectively: (a) SNX, (b) SOM, (c) TOM, (d) SQX, (e) BIN, (f) CSX, (g) SPH; (h) Cross-Sectional Trends at y=0 m for ACFs; (i) zoom-in view of (h)

2.4 Uncertainty Estimation of the Model

In this study, the uncertainty of the subsurface models is evaluated using the Monte Carlo Dropout structure of neural networks (P. Zhang et al., 2022). The MC dropout structure randomly deactivates a certain percentage of connections between neurons, introducing randomness into the model's architecture. Through repeated predictions (e.g., 100 times), the uncertainty of the output results can be directly obtained. As shown in Fig. 3, the models used for ACFs classification (Model #1), horizontal SOF estimation (Model #2), vertical SOF estimation (Model #3), and subsurface modeling (Model #4) all incorporate the MC dropout structure. Therefore, the output results of Model #1-4 are

associated with uncertainty.

It is noteworthy that different horizontal and vertical SOFs generate distinct GCFs, necessitating the repetitive construction of different Model #4 for subsurface reconstruction. To minimize computational costs, the most robust SOFs predictions are considered, obtained by running the SOFs prediction model 100 times and averaging the results. Therefore, the primary source of uncertainty in subsurface modeling stems from the combined contributions of ACFs classification (Model #1) and subsurface modeling (Model #4). As shown in Fig. 2, the one subsurface modeling result $\boldsymbol{Y}$ considers the influence of 7 ACFs, represented as $\boldsymbol{Y} = \sum P_i \boldsymbol{Y}_i$, where i ranges from 1 to 7. To obtain uncertainty in the prediction results, Model #1 and Model #4 need to predict 100 times each, yielding 100 sets of P_i and $\boldsymbol{Y}_i$. Computing according to the aforementioned formula yields 100 subsurface modeling results.

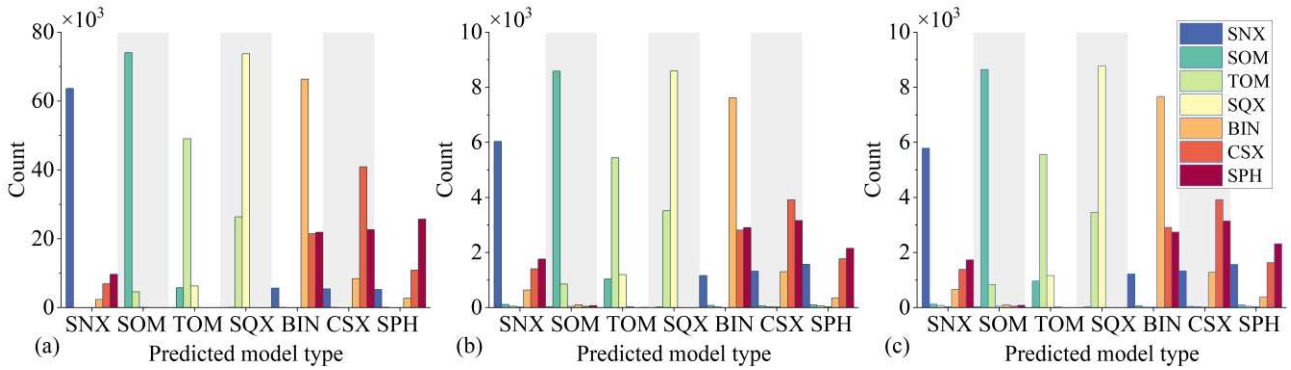
3 Classification and SOF Prediction Models

The PCA operation is implemented using Scipy v1.9.3, and the construction and training of the SCNN are carried out using Tensorflow-GPU v2.8.0—both are open-source packages developed in Python. The neural network consists of four hidden layers, each containing 512 neurons, with a dropout rate set to 0.3. For SCNN, a batch size of 512 is chosen to expedite the model training process. The initial learning rate is set at 0.001, with a 0.5 reduction if the loss on the validation set does not decrease for 15 consecutive iterations. Early stopping is employed to control the number of model iterations, terminating the training process if the model's loss on the validation set does not improve for 30 consecutive iterations. These hyperparameters are determined through a grid search approach. The training uses the Nadam optimizer, an extension of the Adam optimizer with RMSprop and Nesterov momentum.

3.1 Training and Validation of Classification Models

The classification of ACF is a multi-class problem. Hence, the classification cross-entropy loss function is used and the activation function for the output layer is set to Softmax. One-hot encoding is employed to represent ACF categories. Each dimension of the one-hot encoding represents the probability that the measured data belongs to a specific category. For example, the SNX category can be represented as [1, 0, 0, 0, 0, 0], while the SOM category is represented as [0, 1, 0, 0, 0, 0], and so on.

To avoid redundancy, Fig. 7 illustrates the performance of the classification model on the training, validation, and test sets using only three CPT quantities. It can be observed that the classification model exhibits high prediction accuracy on the training set, and its performance on the validation and test sets is similar. This indicates the classification model is able to capture the complex relationship between CPT data and ACF categories and does not suffer from overfitting. Furthermore, the model can distinguish between SNX, SOM, and SQX categories. The TOM sample paths exhibit a degree of smoothness between SOM and SQX, and the model occasionally misclassifies TOM as SOM or SQX. In the case of BIN, CSX, and SPH sample paths, their smoothness is quite similar, leading to potential misclassifications among the three models. However, these misclassifications are acceptable, as even in cases of misclassification, the model provides similar predictions.



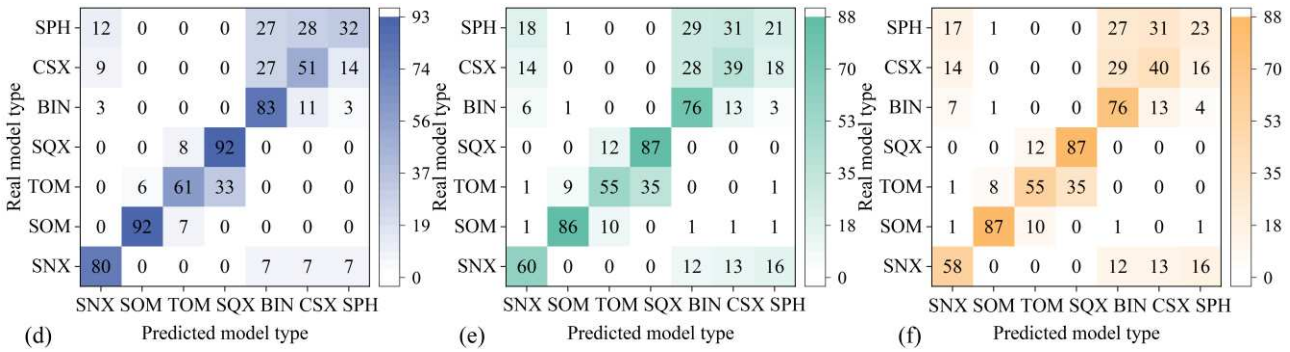


Fig. 7. Classification Model Performance with 3 Sets of CPT Measurements: Training (a, d), Validation (b, e), Test (c, f) Sets

As shown in Fig. 8, the classification model exhibits strong predictive performance on the test set across different numbers of CPT data. The classification model maintains a high level of prediction accuracy even with just two CPT curves. As the measurement data increases, the model's accuracy in predicting BIN significantly improves. As depicted in Fig. 6, for a given SOF, the influence range of the BIN model varies considerably from that of the CSX and SPH models. Therefore, with an increasing amount of measurement data, the model can not only consider the smoothness of sample paths for classification but also effectively account for the influence range of SOF.

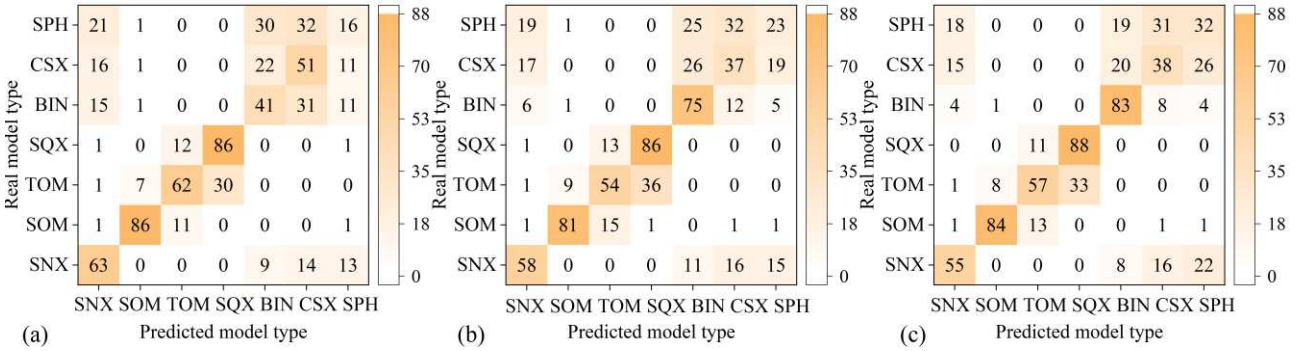


Fig. 8. Test Set Performance of the Classification Model: (a) 2CPT, (b) 4CPT, (c) 5CPT

3.2 Training and Validation of SOF Models

The prediction of SOF is a regression problem. The model employs mean squared error (MSE) as the loss function and applies the ReLU activation function in the output layer. All other configurations

363 remain consistent with the classification model. Root mean square error (RMSE), mean absolute
364 percentage error (MAPE), and the coefficient of determination (R^2) are used to assess the differences
365 between the predicted values ($\hat{y}_i$) and the measured values (y_i) and can be expressed as follows:

366 $$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

367 $$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \quad (7)$$

368 $$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (8)$$

369 where $\bar{y}$ is the mean q_c value; n is the total number of samples. When R^2 is close to 1, and RMSE and
370 MAPE are relatively small, the predictive performance of the model is better.

371 Usually, the training set exhibits higher accuracy compared to the validation and test sets.
372 Therefore, only the prediction results of the models for different ACFs in the validation and test sets
373 are displayed. As shown in Fig. 9 (a)–(g), the horizontal SOF prediction models all demonstrate good
374 performance, with the SQX model exhibiting the highest prediction accuracy. Its test set's R^2 is close
375 to 1, with RMSE and MAPE values of only 2.49% and 3.29%, respectively. Although the SNX model
376 has marginally lower accuracy, it still performs well on the test set, with an R^2 of 0.91 and RMSE and
377 MAPE values of 7.93 and 11.60%, respectively.

378 As shown in Fig. 9 (h)–(n), the vertical SOF prediction models all exhibit improved performance.
379 Among them, the SQX model shows the highest prediction accuracy. Its test set's R^2 is 1.00, with
380 RMSE and MAPE values of 0.05 and 3.68%, respectively. The SNX model has marginally lower
381 prediction accuracy, with an R^2 of 0.98 and RMSE and MAPE values of 0.17 and 10.09%, respectively.

382 It can be observed that the model's prediction performance is related to the smoothness of the

sample path. The smoother the sample path, the higher the prediction accuracy of the model. When the sample path is very rough, it becomes more challenging to distinguish whether the fluctuations in the measurement data are caused by the SOF or the roughness of the sample path itself. However, overall, the established prediction models demonstrate good accuracy.

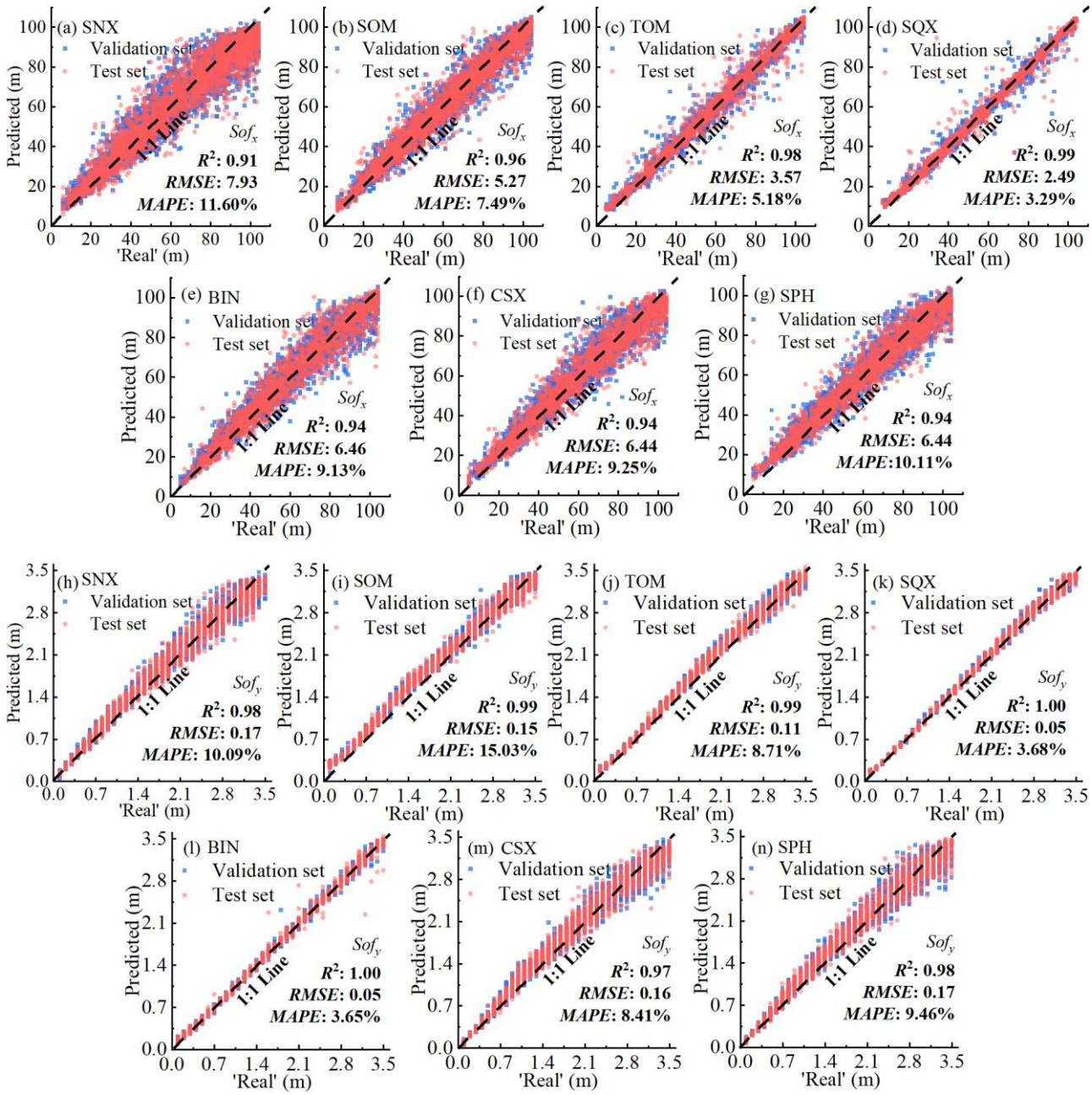


Fig. 9. SOF Prediction Model Performance: Horizontal SOF (a–g); Vertical SOF (h–n)

4 Subsurface Modeling and Validation

4.1 Synthetic Case Study

It is uncommon to record high-resolution test data along the surface of a site. Therefore, this section illustrates the proposed method using a set of synthetic two-dimensional Gaussian q_c field, as depicted in Fig. 10. The two-dimensional cross-section has a depth (h) of 10 m and extends 100 m along the surface. The q_c field is simulated with a resolution of 0.05 m and 1.0 m along the depth and horizontal directions, respectively. In this example, the mean (μ) and standard deviation (σ) are taken as 10 MPa and 4 MPa, respectively. The horizontal and vertical SOF are set as 50 m and 2 m. As shown in Fig. 10, under fixed random seeds, simulation results for two types of ACF, SNX, and SOM, are generated using the matrix decomposition method. The sample paths of SNX exhibit more roughness compared to SOM. The results of SNX and SOM are averaged to obtain synthetic data with sample path smoothness between the two, as shown in Fig. 10 (c).

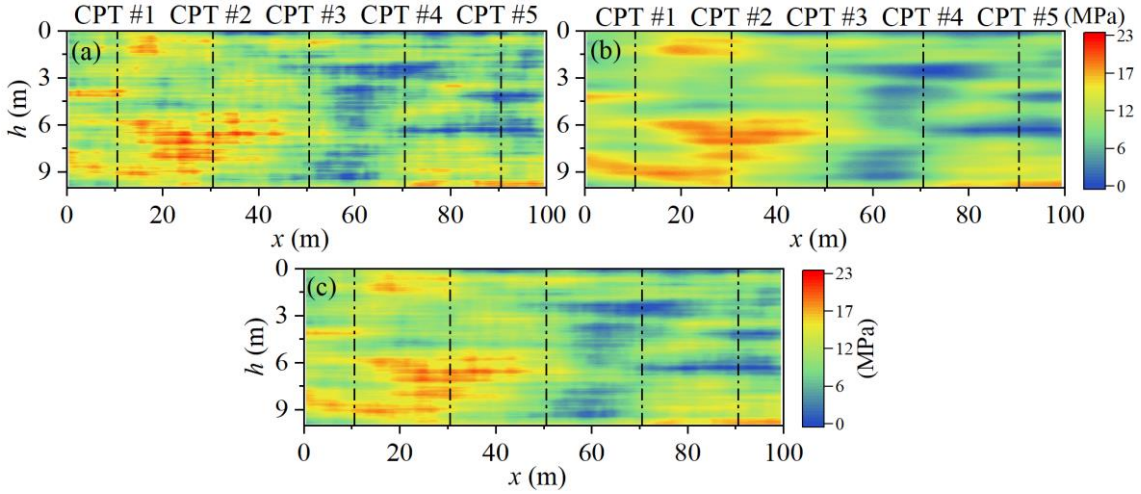


Fig. 10. Synthetic q_c Stratigraphy Depictions: (a) SNX, (b) SOM, (c) Averaged

4.2 Model Construction and Validation

To validate the influence of different CPT quantities on modeling accuracy, the measurement

locations are x = 10.5 m, 30.5 m, 50.5 m, 70.5 m, and 90.5 m. CPT #1 and CPT #5 are not placed at boundary positions, aiming to assess the extrapolation capability of the proposed method. As illustrated in Fig. 11, when the CPT quantity is two, the horizontal distance between CPT #1 and CPT #5 is significantly greater than the actual horizontal SOF. This leads to challenges in accurately predicting the horizontal SOF of the model. However, with more than three CPT measurements, the predicted mean values of the SOF closely align with the actual values.

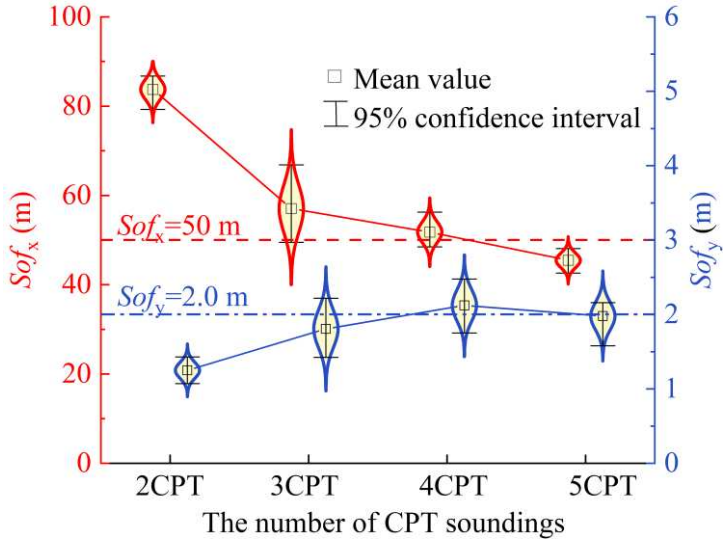


Fig. 11. Predictions of Horizontal and Vertical SOFs for Different CPT Quantities

As shown in Fig. 12 (a)–(d), subsurface modeling is performed using the proposed method with 2, 3, 4, and 5 CPTs, respectively. It is evident that, with an increase in the number of CPTs, the modeling results become progressively finer. When only 2 CPTs are used, the predicted horizontal SOF is relatively large, indicating the model has limited capability to predict the spatial distribution of soil properties near x=50 m. However, it accurately predicts the soil properties near the measurement locations. With 3 CPTs, the model's predictions near the measurement locations closely resemble those with more CPT records. The predictions corresponding to 4 and 5 CPTs are in close agreement, indicating that the information from CPT #4 is already well-predicted with just 4 CPTs.

As shown in Fig. 12 (d)–(f), the proposed method, GDF–ET (Xie et al., 2022b), BCS (Wang et al., 2020; Zhao et al., 2020), and Kriging (Zou et al., 2017) are presented for the case of 5 CPTs. Both the Kriging and the proposed methods necessitate the estimation of random field parameters. This is typically accomplished through the method of moments, maximum-likelihood estimation, and Bayesian analysis (Ching et al., 2020; Ching and Phoon, 2019; Liu et al., 2017; Liu and Leung, 2018; Xiao et al., 2018, 2016). The maximum-likelihood estimation further facilitates the selection of the optimal ACF model using either the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) (Chang et al., 2021). As illustrated in Fig.12 (g), the Kriging method employs the SNX model, which aligns with the ACF model used for generating synthetic cases in Fig. 10, and exhibits smaller AIC and BIC values.

As shown in Fig. 12 (d)–(f), GDF–ET excels at predicting the mean values of the soil properties within the stratigraphy. BCS produces more intricate predictions compared to synthetic stratigraphy. In cases where soil property variations are relatively gradual, BCS often yields accurate predictions. Notably, when the subsurface modeling process lacks the constraints of geotechnical knowledge, the complexity of predictions in different data-driven methods is often related to the algorithm's basis functions or the complexity of input features. This can lead to over-simplification or over-complication of predictions in some specific stratigraphy. Embedding geotechnical knowledge can help prevent such occurrences. The Kriging method exhibits high precision in areas close to CPT boreholes, but tends to estimate using the mean of nearby areas when positioned between two boreholes (Zou et al., 2017). Notably, the Kriging method exhibits cubic complexity (Y. Yang et al., 2022) which may consume excessive computational resources and result in slow solutions when dealing with large amounts of

soil cells. The method proposed in this study benefits from the control of random field theory, ensuring a good consistency between the prediction results and the actual site. Simultaneously, the proposed method is modeled based on a data-driven approach, allowing it to be applied to situations with large amounts of data at a lower computational complexity, as detailed in Section 4.4.

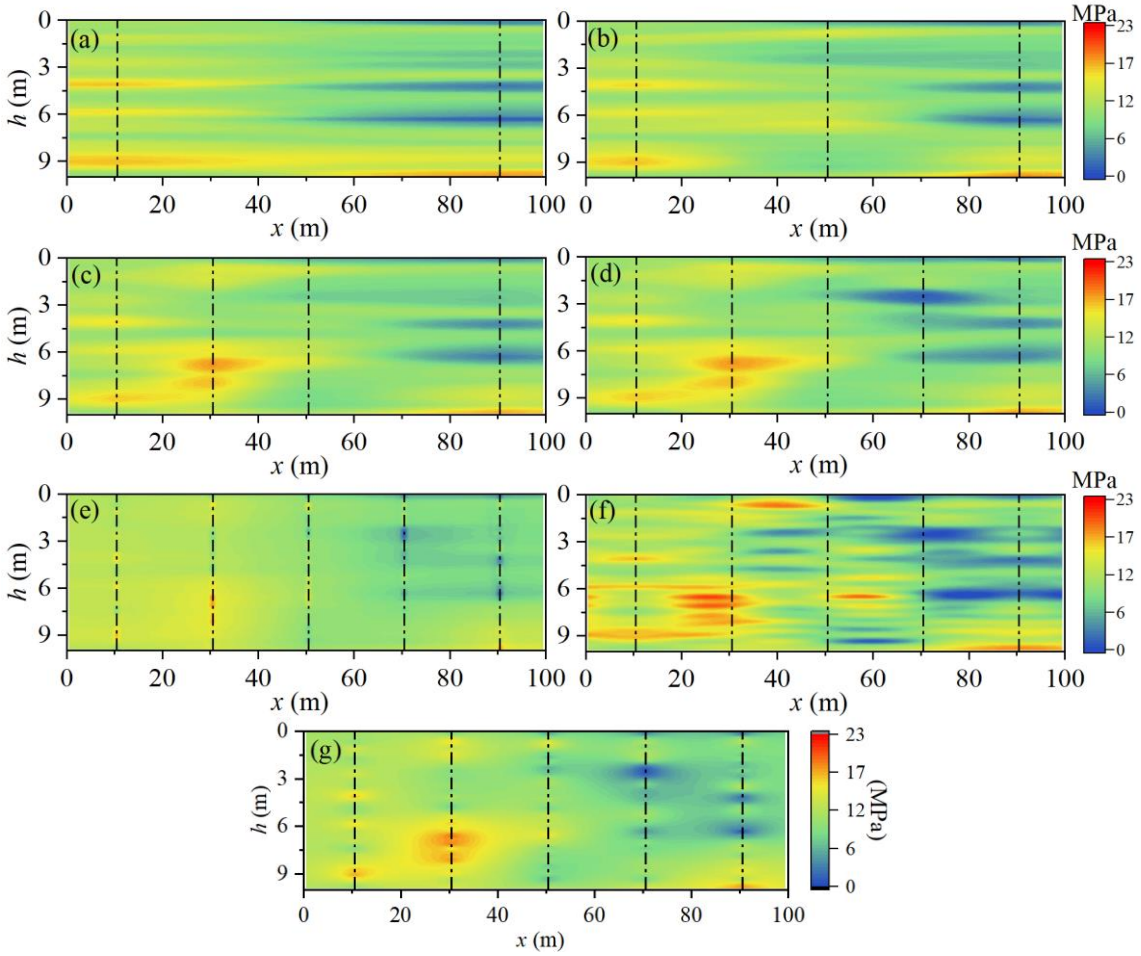


Fig. 12. Comparative q_c Modeling Results: Proposed Method (a–d) at 2, 3, 4, and 5 CPTs; GDF-ET Method (e), BCS Method (f) and Kriging Method (g) at 5 CPTs

As illustrated in Fig. 13 (a)–(d), with an increasing number of CPTs, R^2 of the predictions consistently improves, while RMSE and MAPE decrease, indicating improved prediction accuracy. Notably, since the synthetic random field follows a normal distribution, a substantial amount of data is around the mean value (10 MPa). As the volume of measurement data increases, the predictions gradually approach the 1:1 line. In Fig. 13 (d)–(f), with 5 CPTs, the R^2 values for the GDF–ET and

BCS models are close, with the BCS model having a slightly larger RMSE compared to the GDF–ET, and GDF–ET model exhibiting a higher MAPE than BCS. The predictions by the GDF-ET method mostly fall within the range of mean ± one standard deviation, while the BCS method exhibits overall better consistency across the entire range. Benefiting from accurate random field parameters, the Kriging method's predictive results surpass those of the GCF-ET and BCS methods. The proposed method shows higher R^2 and lower RMSE and MAPE with better accuracy compared to the other methods.

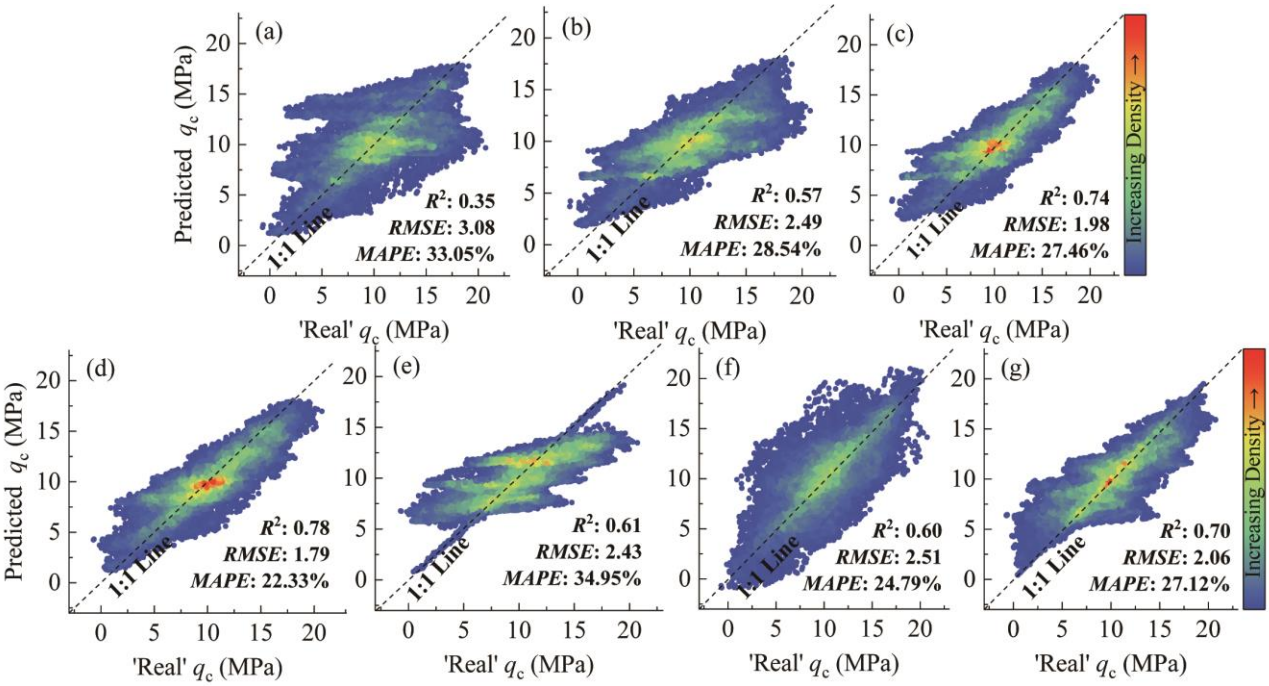


Fig. 13. Comparison of Modeling Results: Proposed Method (a–d) at 2, 3, 4, and 5 CPTs; GDF-ET Method (e), BCS Method (f) and Kriging Method (g) at 5 CPTs

As shown in Fig. 14, the predicted results and their confidence intervals at x = 40.5 m and 80.5 m are extracted. It is evident that the overall trend of the predicted results closely aligns with the actual values, and the majority of observed data falls within the 95% confidence interval of the predictions. Additionally, the predictions from the GDF and kriging models tend to converge toward the mean, while the BCS method exhibits unexplained fluctuations at certain locations.

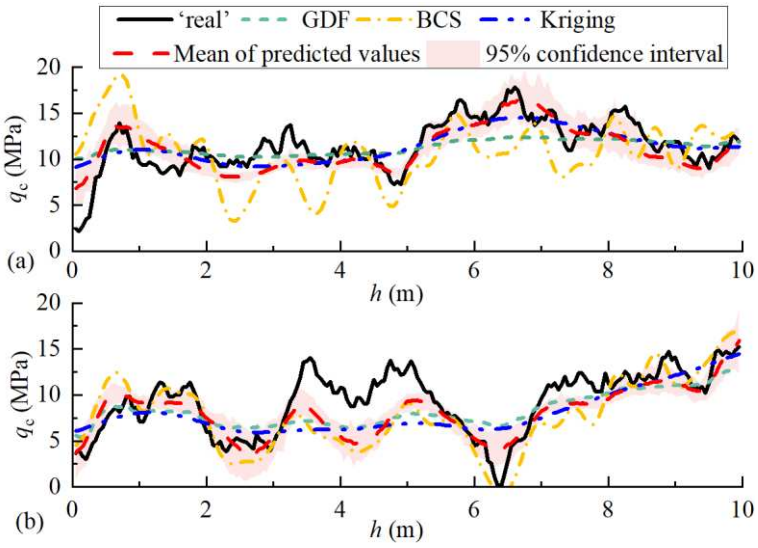


Fig. 14. Model Uncertainty Assessment at Locations: (a) x=40.5 m, (b) x=80.5 m

4.3 Nonstationary Synthetic Case Study

It is worth noting that when evaluating random field parameters, the input for the ACF classification model and SOF prediction model is the CPT data after detrending. This section primarily evaluates the capability of the proposed method to perform subsurface modeling directly on non-stationary data (measured CPT data) after obtaining random field parameters. This approach helps to improve modeling efficiency and reduce uncertainties caused by detrending. Therefore, a non-linear trend is introduced into the stationary random field depicted in Fig. 10(c). The synthetic data assumed a simple second-order increasing trend in the depth direction: $0.05h^2$, followed by Xie et al. (2022b). Fig. 15(a) illustrates the synthesized non-stationary random field. Fig. 15(b)-(d) show predictive results based on five CPT datasets using the proposed method, GDF, and BCS. It can be observed that all three methods capture the trends of the site well. Among them, GDF-ET and BCS show some simplification or complication compared to the synthetic stratigraphy. Although the proposed method's predictive results exhibit some simplification, due to the constraints of random field information, it effectively recovers most of the information from the synthetic stratigraphy. Combining these results

with Fig. 16 reveals consistently higher R^2 values for the proposed method compared to the GDF and BCS methods, with relatively minimal RMSE and MAPE. Compared to modeling results for stationary data using GCFs, as shown in Fig. 13 (d), direct modeling results for non-stationary data show higher R^2 and lower MAPE values, with RMSE values close. This indicates the proposed method achieves improved modeling accuracy and is applicable for both stationary and non-stationary data with a mild trend.

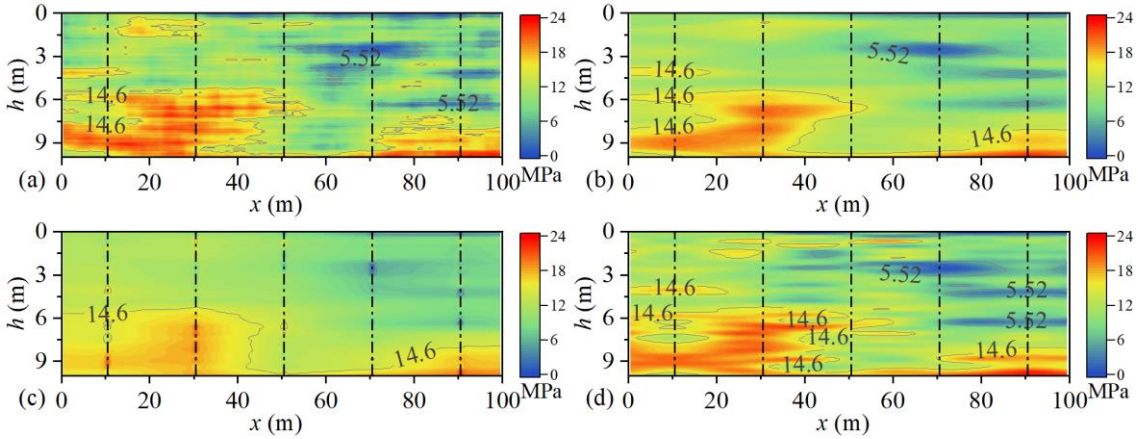


Fig. 15. Modeling q_c Results Comparison: (a) Actual, (b) Proposed Method, (c) GDF-ET, (d) BCS

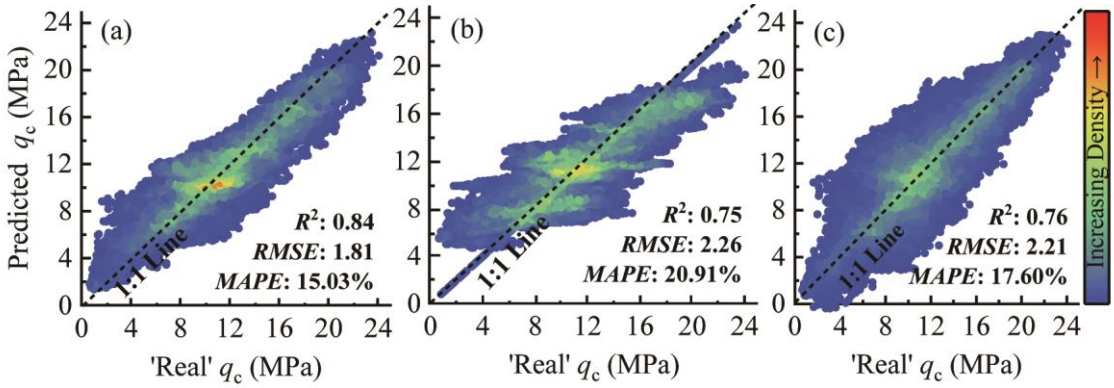


Fig. 16. Comparative Analysis of Different Modeling Methods: (a) Proposed Method, (b) GDF-ET, (c) BCS

4.4 Computational Complexity of Subsurface Modelling

Illustrating the computational efficiency and complexity of the proposed subsurface modeling method using the non-stationary case described in Section 4.3. As shown in Fig. 15 (a), the synthetic

site consists of 20,000 discrete soil cells arranged in a 200×100 grid. Each CPT borehole includes 200 soil cells. Considering the case of five CPTs, a total of 1,000 soil cells are measured (5×200). Consequently, the corresponding training and testing datasets consist of 1,000 and 19,000 samples, respectively. Each sample has 1,000 features (denoted as (M)), as described in Section 2.3.

As depicted in Fig. 17(a), a comparison is made between the total trainable parameters in the subsurface model with and without the PCA operation module. Since CPT tests provide nearly continuous information in the vertical direction, the input feature dimension (M) is typically large. As the number of CPTs increases, not applying PCA to the input data leads to an exponential growth in the neural network's parameter count, significantly reducing its computational efficiency. The PCA-SCNN model proposed in this study preprocesses the input data with PCA dimensionality reduction before feeding it into the neural network. Hence, the feature dimensionality M of the input data does not affect the model's complexity.

As depicted in Fig. 17 (b), after applying PCA processing to the input data, the training time for an individual subsurface model stabilizes at around 20 seconds. The training time remains independent of the number of CPTs. Furthermore, PCA processing effectively reduces the training time, especially when dealing with a larger quantity of CPTs. Notably, if PCA processing is not performed, when the number of CPTs reaches 40, the model consumes a significant amount of memory, increase training difficulty, and may even become untrainable.

Fig. 17(c) illustrates the training process of the subsurface modeling model using 5 CPTs as an example. Models with PCA-processed input not only converge more easily but also exhibit generally lower validation losses. The PCA operation significantly reduces the training complexity of the model.

In Fig. 17 (d)-(f), we evaluate the model's prediction results using R^2, RMSE, and MAPE metrics. As the number of CPTs increases, the model's prediction accuracy steadily improves, with the PCA-processed model consistently outperforming the non-PCA model.

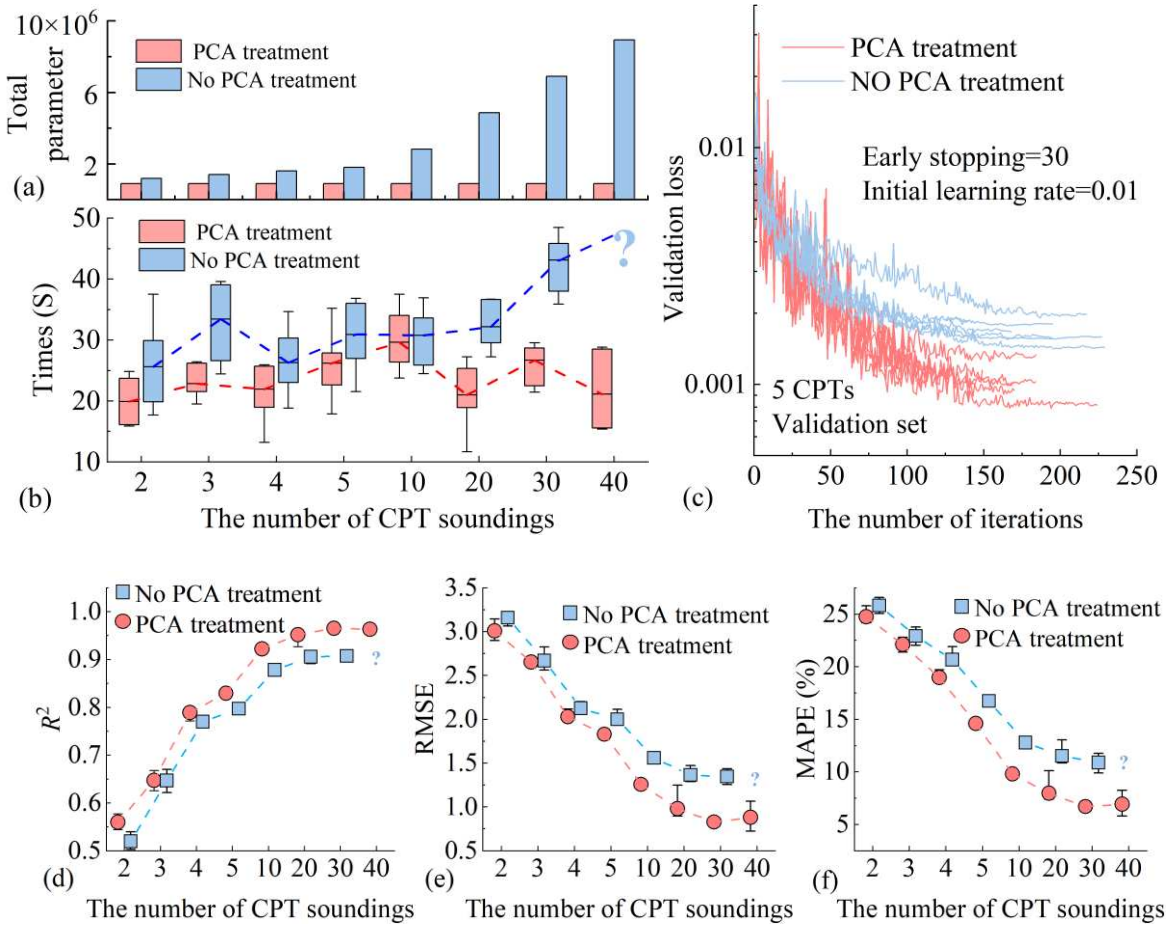


Fig. 17 The Impact of PCA Operations and the Number of CPTs on the Complexity of Subsurface Model (Model #4): (a) Increasing the Number of CPTs Affects the Trainable Parameter Count in Model #4. (b) The Effect of Increasing CPT Numbers on the Training Time of Model #4. (c) The Training Process of Model #4 with 5 CPTs. (d)-(f) Evaluation of the Subsurface Modeling Results Using R^2, *RMSE*, and *MAPE* Metrics.

5 Real Data Case Study

A set of CPT data from the Christchurch region in New Zealand is used to further demonstrate the proposed method. As shown in Fig. 18, a total of seven sets of q_c data are selected for subsurface modeling, taken from the New Zealand Geotechnical Database (NZGD) (NZGD, 2023). It is important to note that the measurement locations of these seven sets of CPT data are not strictly aligned along a

straight line. Therefore, CPT soundings are projected onto the two-dimensional vertical cross-section represented by the red dashed line in Fig. 18. The two-dimensional vertical cross-section extends approximately 82 m along the ground surface, with q_c data typically collected to depths of around 20 m below the surface, as illustrated in Fig. 19. The codes and positions of the seven datasets in NZGD are as follows: CPT_4047 (10.5 m), CPT_4556 (33.5 m), CPT_2715 (42.5 m), CPT_99651 (47.5 m), CPT_2353 (63.5 m), CPT_2358 (70.5 m), and CPT_14412 (92.5 m). For this case, the q_c data have spatial resolutions of 1 m along the ground surface and 0.05 m in the depth direction.

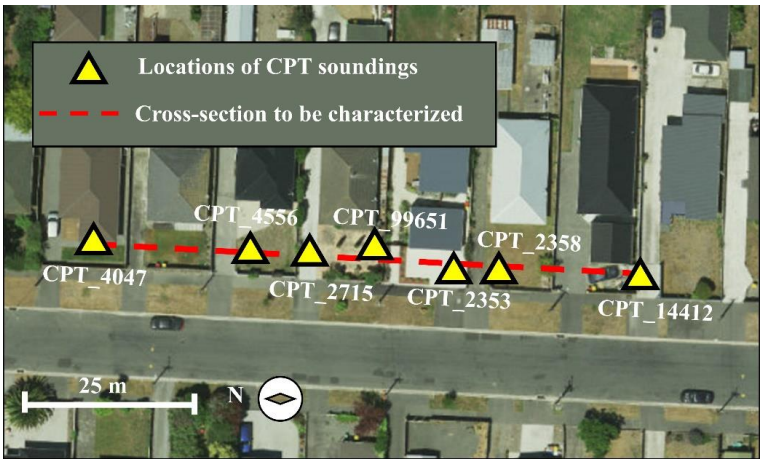


Fig. 18. Geographical Layout and Cross-Section of 7 CPT Soundings in Christchurch, New Zealand (NZGD, 2023)

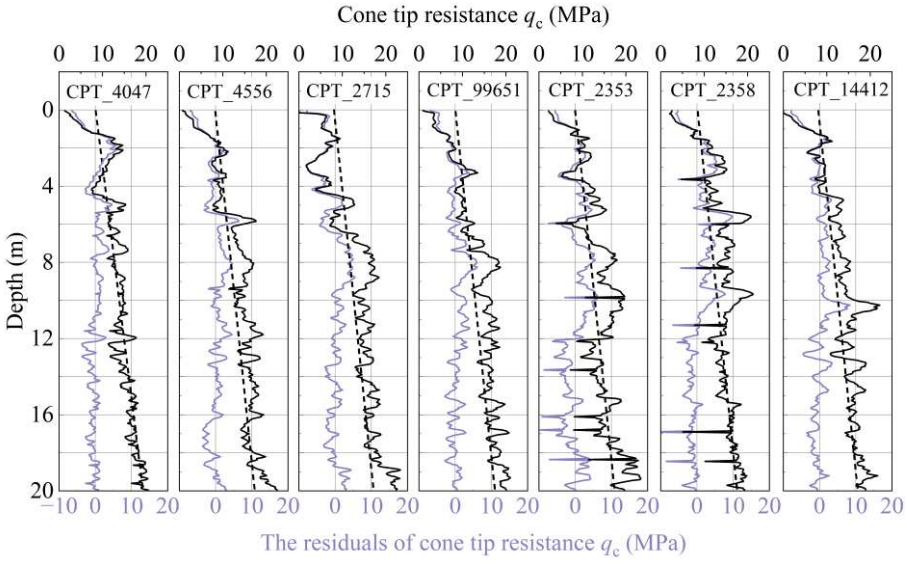


Fig. 19. Set of Seven Measured CPT Data

546 Given that actual CPT data consists of trend and residual terms, this study employs a linear
547 function for detrending, as illustrated in Fig. 19. The modified Bartlett statistical test (Phoon et al.,
548 2003) is employed to assess the stationarity of the residual term after detrending. It is worth noting that
549 the critical Bartlett peak value, B_{crit}, used in the modified Bartlett statistical test to determine the
550 stationarity of the sample sequence, depending on the ACF type. Using a classification model for
551 prediction, it is found that the actual CPT data has the highest probability of belonging to the SOM
552 and CSX models. $B_{\mathrm{crit\_SOM}}$ and $B_{\mathrm{crit\_CSX}}$ values of 71.85 and 48.87, detailed calculation steps are
553 available in (Phoon et al., 2003). The B_{stat} values for the residual terms from all CPT measurements
554 are mostly below 71.85, with the majority falling below 48.87. Therefore, when using the SOM ACF,
555 the null hypothesis of weak stationarity for the residual terms cannot be rejected at a significance level
556 of 5%. In the case of CSX, the assumption of weak stationarity holds for most scenarios. Furthermore,
557 it is worth noting that within the 7 sets of measured data, there are noticeable outliers, such as
558 CPT_2353 and CPT_2358. These outliers are common in real-world data, and in this study, no special
559 treatment is applied to them. This is done to further test the robustness of the proposed method.
560 Section 4.3 validates the proposed method's applicability to both stationary and non-stationary
561 data. Therefore, in this case, a detrending operation is solely applied during the assessment of random
562 field parameters. Subsequently, the GCFs, based on the obtained random field parameters, are used as
563 the model's inputs, with non-stationary observed data serving as output for subsurface modeling. As
564 shown in Fig. 20, subsurface modeling is conducted using different numbers of CPTs, with the
565 remaining CPTs serving as the test set for validation. With only 2 CPTs, the model struggles to provide
566 a detailed prediction of soil properties in the horizontal direction. As more CPT datasets are

incorporated, the modeling results become progressively more refined. Notably, near x=70m, the modeling results exhibit a lens-shaped (hole effect) distribution.

As depicted in Fig. 21, the model's predictions on the training set are nearly identical to the observed values, indicating that using correlation as inputs for the model provides enhanced non-linear expressive capabilities. On the test set, the model predictions closely match the observed values, suggesting the model can predict soil properties in unknown areas based on the correlation relationships established in the training set. It's worth noting that the model does not appear to be affected by the outlier data from CPT_2353 and CPT_2358. This further validates the robustness of the method, showing it can provide accurate predictions, even when dealing with anomalies. This feature reduces the complexity of applying the method for subsurface modeling on a large scale, thus making it more accessible to geotechnical engineers.

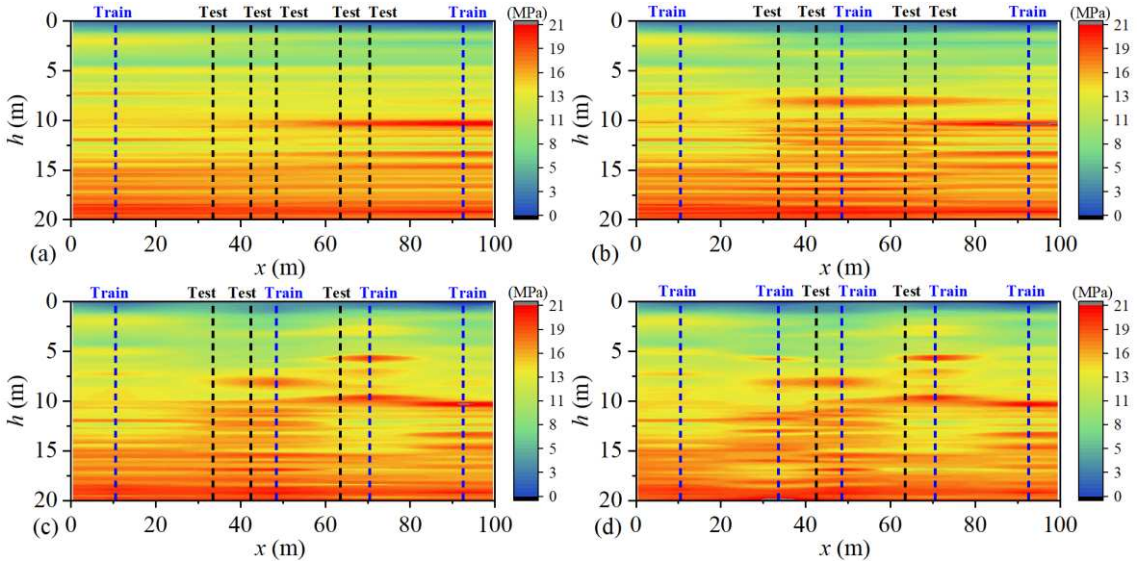


Fig. 20. Subsurface Modeling of Varied Numbers of CPT data: (a) 2; (b) 3; (c) 4; (d) 5

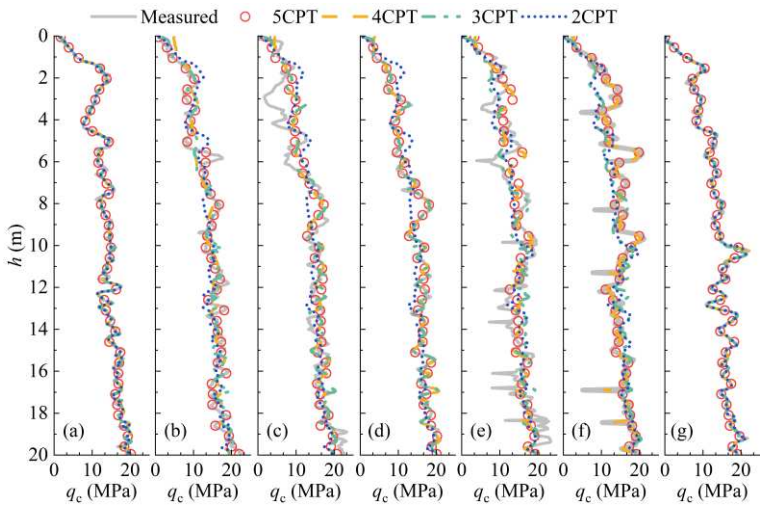


Fig. 21. Predictions of Measurement Locations of Varied Numbers of CPT Data: (a) CPT_4047, (b) CPT_4556, (c) CPT_2715, (d) CPT_99651, (e) CPT_2353, (f) CPT_2358, and (g) CPT_14412

6 Conclusions

This study introduces an innovative data-driven framework for soil property recovery which integrates geotechnical knowledge. This framework attempts to recover soil properties at unsampled points using sparse geotechnical measurements. Based on results from this study, the following conclusions can be drawn:

(1) The use of geotechnical correlation fields as inputs for the subsurface reconstruction model align with the fact that soil properties exhibit location-specific dependencies. This integration of random field theory into the data-driven model fosters enhanced collaboration between the model and geotechnical engineers. The superiority of the data-driven model has been validated through experiments with two synthetic random fields and a real-world case study.

(2) The subsurface modeling method proposed in this study benefits from PCA-SCNN model's dimensionality reduction of the input. The computational complexity of the model is independent of the number of soil cells, resulting in stable modeling efficiency and excellent scalability.

(3) A robust model for identifying types of autocorrelation functions is proposed, and this model

explicitly estimates the probability of observed data belonging to a specific type of autocorrelation function, even in cases with limited measurement data.

(4) The addition of location labels to measurement data addresses the challenge of predicting SOFs in irregularly spaced CPT locations, and the developed SOF prediction models exhibit superior performance, as indicated by R^2, MAPE, and RMSE.

The proposed geotechnical knowledge-based data-driven framework is promising for geotechnical engineering applications and bridges the gap between data-driven modeling and domain-specific knowledge, thereby enhancing the accuracy and reliability of estimating spatially varying geotechnical properties. It should be noted that: ① The proposed framework is flexible, where each model can be replaced with common methods according to the user's preference. For example, the Maximum Likelihood Estimation method can be used to estimate random field parameters. Then, the estimated random field parameters can be used to calculate GCFs. Subsequently, user-friendly machine learning models can be employed to establish the relationship between GCFs and the measured soil properties. ② The construction process of the proposed random field parameter prediction model is relatively complex. Once the model is trained, it can be directly applied to the target site, making the prediction process highly efficient and straightforward. Moreover, it does not require users to have a background in mathematical statistics. ③ Compared to the Kriging model, which has a cubic computational complexity, the computational complexity of the data-driven subsurface model proposed in this study is independent of the number of soil cells, resulting in stable modeling efficiency and excellent scalability.

Notably, when subsurface conditions involve multiple soil layers with significant variations in

properties, and each layer has distinct random field parameters, researchers can employ either manual or IC-XGBoost methods (Shi and Wang, 2023, 2021b) to delineate the spatial distribution of subsurface stratigraphic boundaries. Then, our proposed method can be used to model the spatially varying soil properties within each soil layer. Furthermore, due to the sparsity of CPT data in the horizontal direction, the detection data for weak thin layers often constitute only a small proportion of the overall dataset. Therefore, data-driven subsurface modeling methods should further investigate their applicability and improvement strategies under conditions of data imbalance.

Notation

The following terms and notations are used in this paper:

Terms	
CPT	Cone penetration test
CRF	Conditional random field
BCS	Bayesian compressive sensing
IDW	Inverse distance weighting
GDF	Geotechnical distance field
GCF	Geotechnical correlation field
SOF	Scale of fluctuation
ACF	The autocorrelation function
CNN	Convolutional neural network
SNX	Single exponential
SOM	Second–order Markov
TOM	Third–order Markov
SQX	Squared exponential
BIN	Binary noise
CSX	Cosine exponential
SPH	Spherical
PCA-SCNN	Principal component analysis–shortcut connection neural network
Model #1	The ACFs classification model
Model #2	The horizontal SOF estimation model
Model #3	The vertical SOF estimation model
Model #4	The subsurface modeling model
MSE	Mean squared error
RMSE	Root mean square error
MAPE	Mean absolute percentage error
R^2	The coefficient of determination
Notations	
$\boldsymbol{C}$	The correlation matrix
$\boldsymbol{q}_c$	Cone tip resistance (MPa)
$\boldsymbol{L}$	Horizontal coordinates (m)
P	The probability of measured CPT data belonging to each ACF type.
$\boldsymbol{Y}$	The subsurface modeling outcome (MPa)

$\boldsymbol{q}_{\mathrm{cm}}$	The measured $\boldsymbol{q}_{\mathrm{c}}$ (MPa)
$\widehat{\boldsymbol{q}}_c$	The standardized cone tip resistance
$\widehat{\boldsymbol{L}}$	The standardized horizontal coordinates
$\boldsymbol{x}$	The input feature vector following PCA preprocessing
$\boldsymbol{h}$	The hidden layer feature vector
$\boldsymbol{\theta}$	The weight vectors
$\boldsymbol{\beta}$	The bias vectors
$\boldsymbol{o}$	The output of the model
τ_h and τ_v	The horizontal and vertical distances of soil properties at two discrete points (m)
δ_h and δ_v	The horizontal and vertical SOFs (m)
λ	The power parameter of the Box-Cox method
e_i	The i-th soil cell
GCF_{ei}	The GCF of e_i
$\rho_{i,j}$	The correlation between soil cell e_i and e_j
f	The complex implicit function
N	Total number of soil cells
M	The number of samples in the training set, the feature dimension of the model #4
$\bar{y}$	The mean value
$\hat{y}_i$	The predicted values
y_i	The measured values
n	Total number of samples
h	The depth of the site
μ	The mean value of the random field (MPa)
σ	The standard deviation of the random field (MPa)
B_{crit}	The critical Bartlett peak value

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant No. 52378330), the Natural Science Foundation of Sichuan Province (Grant No. 2023NSFSC0391), the 111 Project (Grant No. B21011), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. SJCX23_0084). The authors would like to extend their most sincere gratitude to the Editors and Reviewers who provided help to improve this paper.

CRediT authorship contribution statement

Weihang Chen: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Jianwen Ding:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing. **Chao Shi:** Conceptualization, Investigation, Supervision, Writing – review & editing. **Tengfei Wang:** Conceptualization, Investigation, Writing – review & editing. **David P. Connolly:** Conceptualization,

Investigation, Writing – review & editing.

References

Bai, L., Song, C., Zhou, X., Tian, Y., Wei, L., 2023. Assessing project portfolio risk via an enhanced GA-BPNN combined with PCA. Eng. Appl. Artif. Intell. 126, 106779. https://doi.org/10.1016/j.engappai.2023.106779

Cami, B., Javankhoshdel, S., Phoon, K.-K., Ching, J., 2020. Scale of Fluctuation for Spatially Varying Soils: Estimation Methods and Values. ASCE-ASME J. Risk Uncertain. Eng. Syst. Part Civ. Eng. 6, 03120002. https://doi.org/10.1061/AJRUA6.0001083

Chang, Y.-C., Ching, J., Phoon, K.-K., Yue, Q., 2021. On the Hole Effect in Soil Spatial Variability. ASCE-ASME J. Risk Uncertain. Eng. Syst. Part Civ. Eng. 7, 04021039. https://doi.org/10.1061/AJRUA6.0001168

Chen, W., Ding, J., Wang, T., Connolly, D.P., Wan, X., 2023. Soil property recovery from incomplete in-situ geotechnical test data using a hybrid deep generative framework. Eng. Geol. 326, 107332. https://doi.org/10.1016/j.enggeo.2023.107332

Ching, J., Huang, W.-H., Phoon, K.-K., 2020. 3D Probabilistic Site Characterization by Sparse Bayesian Learning. J. Eng. Mech. 146, 04020134. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001859

Ching, J., Phoon, K.-K., 2019. Constructing Site-Specific Multivariate Probability Distribution Model Using Bayesian Machine Learning. J. Eng. Mech. 145, 04018126. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001537

Ching, J., Phoon, K.-K., Stuedlein, A.W., Jaksa, M., 2019. Identification of sample path smoothness in soil spatial variability. Struct. Saf. 81, 101870. https://doi.org/10.1016/j.strusafe.2019.101870

Ching, J., Wu, T.-J., Stuedlein, A.W., Bong, T., 2018. Estimating horizontal scale of fluctuation with limited CPT soundings. Geosci. Front., Reliability Analysis of Geotechnical Infrastructures 9, 1597–1608. https://doi.org/10.1016/j.gsf.2017.11.008

Collico, S., Arroyo, M., Devincenzi, M., 2024. A simple approach to probabilistic CPTu-based geotechnical stratigraphic profiling. Comput. Geotech. 165, 105905. https://doi.org/10.1016/j.compgeo.2023.105905

Dasaka, S. m., Zhang, L. m., 2012. Spatial variability of in situ weathered soil. Géotechnique 62, 375–384. https://doi.org/10.1680/geot.8.P.151.3786

Gu, X., Wang, L., Ou, Q., Zhang, W., 2023. Efficient stochastic analysis of unsaturated slopes subjected to various rainfall intensities and patterns. Geosci. Front. 14, 101490. https://doi.org/10.1016/j.gsf.2022.101490

Guan, Z., Wang, Y., Cao, Z., Hong, Y., 2020. Smart sampling strategy for investigating spatial distribution of subsurface shallow gas pressure in Hangzhou Bay area of China. Eng. Geol. 274, 105711. https://doi.org/10.1016/j.enggeo.2020.105711

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

679 Hong, L., Wang, X., Zhang, W., Li, Y., Zhang, R., Chen, C., 2024. System reliability-based robust
680 design of deep foundation pit considering multiple failure modes. Geosci. Front. 15, 101761.
681 https://doi.org/10.1016/j.gsf.2023.101761
682 Hu, Y., Wang, Y., 2024. Evaluating statistical homogeneity of cone penetration test (CPT) data profile
683 using auto-correlation function. Comput. Geotech. 165, 105852.
684 https://doi.org/10.1016/j.compgeo.2023.105852
685 Li, H., Xu, Z., Taylor, G., Studer, C., Goldstein, T., 2018. Visualizing the Loss Landscape of Neural
686 Nets.
687 Liu, W.F., Leung, Y.F., 2018. Characterising three-dimensional anisotropic spatial correlation of soil
688 properties through in situ test results. Géotechnique 68, 805–819.
689 https://doi.org/10.1680/jgeot.16.P.336
690 Liu, W.F., Leung, Y.F., Lo, M.K., 2017. Integrated framework for characterization of spatial variability
691 of geological profiles. Can. Geotech. J. 54, 47–58. https://doi.org/10.1139/cgj-2016-0189
692 Lloret-Cabot, M., Fenton, G.A., Hicks, M.A., 2014. On the estimation of scale of fluctuation in
693 geostatistics. Georisk Assess. Manag. Risk Eng. Syst. Geohazards 8, 129–140.
694 https://doi.org/10.1080/17499518.2013.871189
695 Nag, P., Sun, Y., Reich, B.J., 2023. Spatio-temporal DeepKriging for interpolation and probabilistic
696 forecasting. Spat. Stat. 57, 100773. https://doi.org/10.1016/j.spasta.2023.100773
697 NZGD, 2023. World Wide Web Address [WWW Document]. URL https://www.nzgd.org.nz (accessed
698 10.18.23).
699 Onyejekwe, S., Kang, X., Ge, L., 2016. Evaluation of the scale of fluctuation of geotechnical
700 parameters by autocorrelation function and semivariogram function. Eng. Geol. 214, 43–49.
701 https://doi.org/10.1016/j.enggeo.2016.09.014
702 Phoon, K.-K., Cao, Z.-J., Ji, J., Leung, Y.F., Najjar, S., Shuku, T., Tang, C., Yin, Z.-Y., Ikumasa, Y.,
703 Ching, J., 2022. Geotechnical uncertainty, modeling, and decision making. Soils Found. 62,
704 101189. https://doi.org/10.1016/j.sandf.2022.101189
705 Phoon, K.-K., Quek, S.-T., An, P., 2003. Identification of Statistically Homogeneous Soil Layers Using
706 Modified Bartlett Statistics. J. Geotech. Geoenvironmental Eng. 129, 649–659.
707 https://doi.org/10.1061/(ASCE)1090-0241(2003)129:7(649)
708 Phoon, K.-K., Wang, Y., 2019. Chicken (method) and egg (data)— Which comes first? Presented at
709 the Int. Symp. on Reliability of Multidisciplinary Engineering Systems under Uncertainty
710 (ISRMES2019), Ministry of Education and Ministry of Science and Technology, Da'an, Taipei.
711 Qi, X.-H., Liu, H.-X., 2019. Estimation of autocorrelation distances for in-situ geotechnical properties
712 using limited data. Struct. Saf. 79, 26–38. https://doi.org/10.1016/j.strusafe.2019.02.003
713 Shi, C., Wang, Y., 2023. Data-driven spatio-temporal analysis of consolidation for rapid reclamation.
714 Géotechnique 1–21. https://doi.org/10.1680/jgeot.22.00016
715 Shi, C., Wang, Y., 2021a. Non-parametric machine learning methods for interpolation of spatially
716 varying non-stationary and non-Gaussian geotechnical properties. Geosci. Front. 12, 339–350.
717 https://doi.org/10.1016/j.gsf.2020.01.011
718 Shi, C., Wang, Y., 2021b. Development of Subsurface Geological Cross-Section from Limited Site-
719 Specific Boreholes and Prior Geological Knowledge Using Iterative Convolution XGBoost. J.
720 Geotech. Geoenvironmental Eng. 147, 04021082. https://doi.org/10.1061/(ASCE)GT.1943-

5606.0002583
Uzielli, M., Vannucchi, G., Phoon, K.K., 2005. Random field characterisation of stress-nomalised cone penetration testing parameters. Géotechnique 55, 3–20. https://doi.org/10.1680/geot.2005.55.1.3
Vanmarcke, E.H., 1983. Random Fields: Analysis and Synthesis. MIT Press.
Wang, F., Chen, J.E., 2023. Efficient modeling of random fields by using Gaussian process inducing-point approximations. Comput. Geotech. 157, 105304. https://doi.org/10.1016/j.compgeo.2023.105304
Wang, H., Wang, X., Wellmann, J.F., Liang, R.Y., 2018. Bayesian Stochastic Soil Modeling Framework Using Gaussian Markov Random Fields. ASCE-ASME J. Risk Uncertain. Eng. Syst. Part Civ. Eng. 4, 04018014. https://doi.org/10.1061/AJRUA6.0000965
Wang, Y., Hu, Y., Zhao, T., 2020. Cone penetration test (CPT)-based subsurface soil classification and zonation in two-dimensional vertical cross section using Bayesian compressive sampling. Can. Geotech. J. 57, 947–958. https://doi.org/10.1139/cgj-2019-0131
Wang, Y., Shi, C., 2023. Data-driven analysis of soil consolidation with prefabricated vertical drains considering stratigraphic variation. Comput. Geotech. 161, 105569. https://doi.org/10.1016/j.compgeo.2023.105569
Wang, Y., 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
Xiao, T., Li, D.-Q., Cao, Z.-J., Au, S.-K., Phoon, K.-K., 2016. Three-dimensional slope reliability and risk assessment using auxiliary random finite element method. Comput. Geotech. 79, 146–158. https://doi.org/10.1016/j.compgeo.2016.05.024
Xiao, T., Li, D.-Q., Cao, Z.-J., Zhang, L.-M., 2018. CPT-Based Probabilistic Characterization of Three-Dimensional Spatial Variability Using MLE. J. Geotech. Geoenvironmental Eng. 144, 04018023. https://doi.org/10.1061/(ASCE)GT.1943-5606.0001875
Xie, J., Huang, J., Lu, J., Burton, G.J., Zeng, C., Wang, Y., 2022a. Development of two-dimensional ground models by combining geotechnical and geophysical data. Eng. Geol. 300, 106579. https://doi.org/10.1016/j.enggeo.2022.106579
Xie, J., Huang, J., Zeng, C., Huang, S., Burton, G.J., 2022b. A generic framework for geotechnical subsurface modeling with machine learning. J. Rock Mech. Geotech. Eng. 14, 1366–1379. https://doi.org/10.1016/j.jrmge.2022.08.001
Xie, J., Zeng, C., Huang, J., Zhang, Y., Lu, J., 2024. A back analysis scheme for refined soil stratification based on integrating borehole and CPT data. Geosci. Front. 15, 101688. https://doi.org/10.1016/j.gsf.2023.101688
Yan, W., Shen, P., Zhou, W.-H., Ma, G., 2023. A rigorous random field-based framework for 3D stratigraphic uncertainty modelling. Eng. Geol. 323, 107235. https://doi.org/10.1016/j.enggeo.2023.107235
Yang, Y., Wang, P., Brandenberg, S.J., 2022. An algorithm for generating spatially correlated random fields using Cholesky decomposition and ordinary kriging. Comput. Geotech. 147, 104783. https://doi.org/10.1016/j.compgeo.2022.104783
Yang, Z., Ching, J., 2021. Simulation of three-dimensional random field conditioning on incomplete

site data. Eng. Geol. 281, 105987. https://doi.org/10.1016/j.enggeo.2020.105987
Yang, Z., Li, X., Qi, X., 2022. Efficient simulation of multivariate three-dimensional cross-correlated random fields conditioning on non-lattice measurement data. Comput. Methods Appl. Mech. Eng. 388, 114208. https://doi.org/10.1016/j.cma.2021.114208
Yang, Z., Nie, J., Peng, X., Tang, D., Li, X., 2021. Effect of random field element size on reliability and risk assessment of soil slopes. Bull. Eng. Geol. Environ. 80, 7423–7439. https://doi.org/10.1007/s10064-021-02422-z
Zhang, J.-Z., Phoon, K.K., Zhang, D.-M., Huang, H.-W., Tang, C., 2021. Novel approach to estimate vertical scale of fluctuation based on CPT data using convolutional neural networks. Eng. Geol. 294, 106342. https://doi.org/10.1016/j.enggeo.2021.106342
Zhang, J.-Z., Zhang, D.-M., Huang, H.-W., Phoon, K.K., Tang, C., Li, G., 2022. Hybrid machine learning model with random field and limited CPT data to quantify horizontal scale of fluctuation of soil spatial variability. Acta Geotech. 17, 1129–1145. https://doi.org/10.1007/s11440-021-01360-0
Zhang, P., Yin, Z.-Y., Jin, Y.-F., 2022. Bayesian neural network-based uncertainty modelling: application to soil compressibility and undrained shear strength prediction. Can. Geotech. J. 59, 546–557. https://doi.org/10.1139/cgj-2020-0751
Zhang, W., Gu, X., Tang, L., Yin, Y., Liu, D., Zhang, Y., 2022. Application of machine learning, deep learning and optimization algorithms in geoengineering and geoscience: Comprehensive review and future challenge. Gondwana Res. 109, 1–17. https://doi.org/10.1016/j.gr.2022.03.015
Zhang, W., Wu, C., Tang, L., Gu, X., Wang, L., 2023. Efficient time-variant reliability analysis of Bazimen landslide in the Three Gorges Reservoir Area using XGBoost and LightGBM algorithms. Gondwana Res., Data driven models 123, 41–53. https://doi.org/10.1016/j.gr.2022.10.004
Zhang, W., Zhang, R., Wu, C., Goh, A.T.C., Lacasse, S., Liu, Z., Liu, H., 2020. State-of-the-art review of soft computing applications in underground excavations. Geosci. Front. 11, 1095–1106. https://doi.org/10.1016/j.gsf.2019.12.003
Zhao, T., Hu, Y., Wang, Y., 2018. Statistical interpretation of spatially varying 2D geo-data from sparse measurements using Bayesian compressive sampling. Eng. Geol. 246, 162–175. https://doi.org/10.1016/j.enggeo.2018.09.022
Zhao, T., Xu, L., Wang, Y., 2020. Fast non-parametric simulation of 2D multi-layer cone penetration test (CPT) data without pre-stratification using Markov Chain Monte Carlo simulation. Eng. Geol. 273, 105670. https://doi.org/10.1016/j.enggeo.2020.105670
Zou, H., Liu, S., Cai, G., Bheemasetti, T.V., Puppala, A.J., 2017. Mapping probability of liquefaction using geostatistics and first order reliability method based on CPTU measurements. Eng. Geol. 218, 197–212. https://doi.org/10.1016/j.enggeo.2017.01.021