



Deposited via The University of Leeds.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/227950/>

Version: Accepted Version

Article:

Chen, W., Shi, C., Ding, J. et al. (2025) Data-driven sparse learning of three-dimensional subsurface properties incorporating random field theory. *Engineering Geology*, 349. 107972. ISSN: 0013-7952

<https://doi.org/10.1016/j.enggeo.2025.107972>

This is an author produced version of an article accepted for publication in *Engineering Geology* made available under the terms of the Creative Commons Attribution License (CC-BY), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Data-Driven Sparse Learning of Three-dimensional Subsurface Properties Incorporating Random Field Theory

Weihang Chen ^a, Chao Shi ^b, Jianwen Ding ^{a,*}, 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

1 **Abstract**

2 Geotechnical engineers rely on accurate soil property information for engineering analyses. However,
3 it is challenging for spatial learning of soil attributes because in-situ geotechnical testing is typically
4 performed sparsely at discrete locations, and soil properties also exhibit inherent spatial variability.
5 Traditional geostatistical methods for predicting spatial properties at these unsampled locations exhibit
6 high computational complexity and require pre-determination of hyper-parameters, while pure data-
7 driven methods fail to integrate geotechnical knowledge. In this study, a hybrid and parameter-free
8 framework that uses random field theory and machine learning is proposed to model 3D subsurface
9 field with reduced computational complexity. The framework constructs site-specific basis functions
10 for characterizing the spatial variations of soil properties by decomposing a correlation matrix through
11 principal component analysis. To further reduce the computational complexity involved in processing
12 high-dimensional correlation matrices, a sparse sampling strategy is adopted to map correlation matrix
13 onto lower-rank principal component space. A series of synthetic random field examples are generated
14 to illustrate the impact of scale of fluctuation and autocorrelation functions on the accuracy and
15 sensitivity of subsurface modeling. The performance of the proposed method is further validated using
16 both synthetic cases and two real case histories. It is demonstrated that the proposed method generally
17 achieves higher R^2 and lower root mean square error (RMSE) and mean absolute percentage error
18 (MAPE) compared to state-of-the-art methods, such as Kriging and Bayesian compressive sensing.
19 Moreover, the proposed method facilitates the explicit quantification of uncertainty associated with
20 the subsurface models, providing valuable insights for engineering design and analysis. The data and
21 code used in this study are available at <https://github.com/Data-Driven-RFT/Sparse-Learning>.

22

23 **Keywords:** Geotechnical spatial variability; Machine learning aided geotechnics; Random field theory;
24 Geotechnical site investigation; Principal component analysis

25 **1 Introduction**

26 The earth's historical geological and environmental processes have resulted in significant spatial
27 variability in near-surface soil deposits and thus geotechnical engineering properties (Gong et al., 2021;
28 Phoon et al., 2022; Shi and Wang, 2023). It is widely accepted in engineering geology and geotechnical
29 engineering that this spatial variability affects the interaction between geotechnical structures and the
30 soil, serving as a major source of engineering uncertainty (Jiang et al., 2024b; Liu et al., 2023; Wang
31 and Shi, 2023; Zhang et al., 2020). Accurate subsurface models which characterize the spatial
32 distribution of soil properties, can aid engineers in conducting more rational analyses and optimized
33 designs (Chen et al., 2023; Qiu et al., 2024; Wang et al., 2020; Zhao et al., 2020). Despite its importance,
34 time constraints, financial considerations, and site-specific limitations often restrict investigation to
35 sparsely located spot-testing. For example, cone penetration test (CPT) is widely used for its simplicity
36 and cost-effectiveness, but CPT tests are typically conducted at intervals of tens of meters along the
37 surface, providing only sparse information about the complex spatial distribution of soil properties
38 (Collico et al., 2024; Guan et al., 2020; Xie et al., 2022b). Consequently, constructing accurate and
39 reliable subsurface models using such sparse investigation information is often challenging (Shi and
40 Wang, 2021; Xie et al., 2024).

41 Currently, two-dimensional (2D) stratigraphic profiles are commonly used to guide the design and
42 construction of engineering projects (Guan et al., 2024a; Hu et al., 2024). Selecting representative
43 stratigraphic profiles to comprehensively reflect the spatial distribution of soil properties largely
44 depends on the engineer's experience. For large geotechnical projects with complex geological
45 conditions, detailed three-dimensional (3D) subsurface models are needed to thoroughly characterize

46 the spatial distribution of soil properties (Jiang et al., 2024a; Shi et al., 2023; Shi and Wang, 2022).
47 Additionally, detailed 3D subsurface models can help mitigate uncertainties caused by human factors.
48 Subsurface modelling using interpolation or data-driven methods is an active area of study (Hu et al.,
49 2024; Z. Z. Wang et al., 2023; Yang et al., 2023a, 2023b). The Kriging method, proposed in 1960, has
50 been widely applied for the interpolation of 2D and 3D soil properties (Zou et al., 2017). As a
51 geostatistical model, Kriging infers the properties of unknown points through the weighted averaging
52 of known properties within a certain range, making it an optimal linear unbiased estimator. However,
53 the Kriging method requires the estimation of model parameters to ensure reliability of predictions
54 (Nag et al., 2023). Moreover, when simulating large-scale or high-resolution 3D random fields, the
55 Kriging method may consume substantial computational resources in the storage and processing of
56 large correlation matrices. As a solution, Yang and Ching (2021) applied the conditional random field
57 method to 3D site modeling with reduced computational complexity based on the assumption of
58 separable autocorrelation function. Bayesian Compressive Sensing (BCS), a non-parametric and data-
59 driven method, has also been extensively used for modelling non-stationary processes and fields, but
60 BCS lacks site-specific basis functions for geotechnical modeling (Cami et al., 2020). Alternatively,
61 Xie et al. (2022b) proposed a data-driven modeling method based on geotechnical distance fields, but
62 it has yet to be extended to 3D subsurface modeling cases. Additionally, soil properties exhibit location
63 dependency, and the properties at a given location are only correlated with those at another location
64 within a certain scale of fluctuation (SoF) (Phoon et al., 2003). Therefore, predicting soil properties at
65 unsampled points solely based on correlation distance may overlook this fact.

66 Incorporating domain knowledge, such as random field theory into data-driven machine learning

67 models is a promising approach (Lyu et al., 2024; Wu et al., 2023). It addresses challenges related to
68 the weak interpretability, poor generalization, and physical inconsistency of machine learning models.
69 For example, Chen et al. (2023) used a large number of 2D synthetic samples derived from random
70 field theory as 'prior' information to generate deep learning models. These models, enriched with prior
71 information, then guided subsurface modeling for site-specific 2D applications. However, generating
72 and storing large numbers of 3D random field synthetic samples and inputting them into models for
73 training is challenging. Furthermore, Chen et al. (2024) used the correlation matrix from random field
74 theory to characterize the spatial positions of soil cells within a site, embedding random field theory
75 into a data-driven model. Nevertheless, calculating the correlation matrix requires additional random
76 field parameters. Estimating these parameters for 3D sites is particularly challenging when borehole
77 data is sparse (Qi et al., 2022; Xiao et al., 2018; J.-Z. Zhang et al., 2022; Zhang et al., 2021).
78 Additionally, the number of soil cells in a 3D site increases significantly compared to a 2D site, making
79 the storage and processing of large correlation matrices time-consuming (Z. Yang et al., 2022). It is
80 worth noting that in geotechnical subsurface modeling, 'prior' information encompasses not only the
81 domain knowledge (random field theory) but also measurement data from neighboring or similar sites
82 (Guan et al., 2024b) and multi-source measurement data (Xie et al., 2022a). This study primarily
83 focuses on embedding random field theory as 'prior' information into data-driven models, which proves
84 particularly beneficial in scenarios where similar measurement data are challenging to obtain.

85 To address the described issues above, this paper proposes a data-driven framework that uses
86 random field theory to achieve 3D subsurface modeling with reduced computational complexity and
87 no parameters. The framework employs a correlation matrix processed through principal component

88 analysis (PCA) to characterize the spatial positions of soil cells, providing a rich set of basis functions
89 for subsurface modeling. Additionally, to reduce the computational complexity of processing large
90 correlation matrices with PCA, a strategy is proposed that involves sparse sampling followed by
91 projection into the principal component space. A series of synthetic random field samples are
92 statistically analyzed to investigate the impact of scale of fluctuation on subsurface modeling accuracy,
93 aiming to eliminate correlated parameters. The performance of the proposed method is validated using
94 a large number of synthetic cases and two real case histories. The remainder of this study is organized
95 as follows: Section 2 introduces the proposed enhanced subsurface modeling framework. Section 3
96 discusses the improvements and simplification strategies of the proposed framework through a series
97 of synthetic cases. Section 4 provides a detailed description of the implementation procedures of the
98 proposed method. Sections 5 and 6 compare and validate the method using both 2D and 3D examples,
99 followed by the conclusion.

100 **2 Proposed Subsurface Modeling Methods**

101 The essence of data-driven subsurface modeling lies in using sparse measurement information to infer
102 soil properties at multiple unmeasured locations. This study aims to integrate random field theory into
103 data-driven models through Geotechnical Correlation Fields (GCFs) to reduce uncertainty in
104 subsurface modeling and enhance model reliability.

105 **2.1 Three-dimensional Geotechnical Correlation Field**

106 The geological and environmental processes during soil deposition result in spatial variability in
107 near-surface soils. Random Field Theory (RFT), as a powerful tool for evaluating the spatial variability
108 of soil properties, is widely used for modeling the inherent variability of soils ([Stuedlein et al., 2012a](#)).

109 RFT adheres to the fact that soil properties at a given site exhibit spatial dependency, meaning that soil
 110 properties are correlated only within a certain lag distance. In RFT, the inherent variability of soil
 111 properties is described not only by mean and variance but also by autocorrelation functions (ACFs)
 112 and scales of fluctuation (SoF) (Cami et al., 2020). Eq. (1) demonstrates a commonly used ACF—the
 113 Single Exponential (SNX) model. ACFs quantify the correlation $\rho_{i,j}$ between soil cells i and j , rather
 114 than relying solely on the distance between them. It is important to note that before implementing RFT,
 115 the site must be discretized into N soil cells (cubic elements) based on engineering requirements, with
 116 each cell assumed to be homogeneous internally.

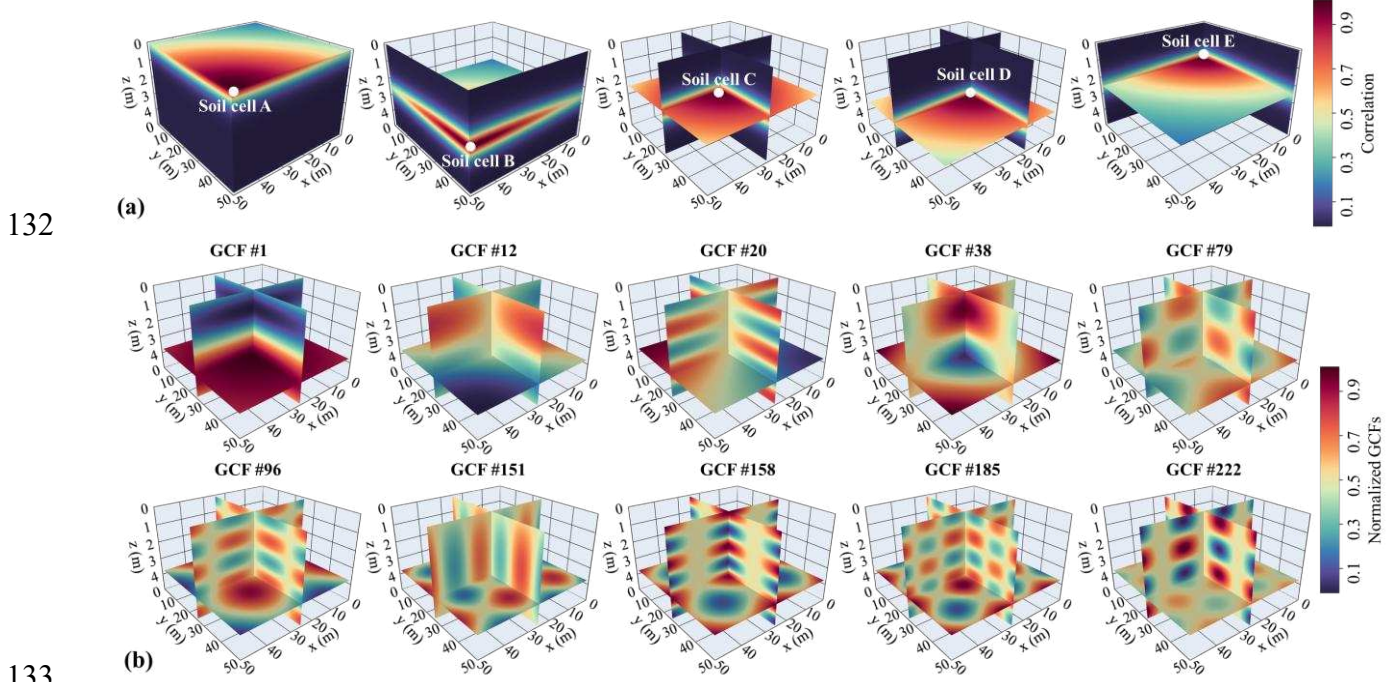
$$117 \quad \rho_{i,j} = \exp \left[-2 \left(\frac{|\tau_{i,j}^x|}{SO F_x} + \frac{|\tau_{i,j}^y|}{SO F_y} + \frac{|\tau_{i,j}^z|}{SO F_z} \right) \right] \quad (i, j \in [1, N]) \quad (1)$$

118 where $\tau_{i,j}^x$, $\tau_{i,j}^y$, and $\tau_{i,j}^z$ represent the distances between soil cells i and j along the x -, y -, and z -axis
 119 directions, respectively. $SO F_x$, $SO F_y$, and $SO F_z$ correspond to the scales of fluctuation for a specific site
 120 in the x -, y -, and z -axis directions. Table A1 in Appendix A presents seven common types of ACFs
 121 (Cami et al., 2020; Ching et al., 2019).

122 To calculate the correlation ρ between every pair of soil cells, a correlation matrix \mathbf{C} of dimension
 123 $N \times N$ is constructed, as shown in Eq. (2). In this matrix, $\rho_{i,j} = 1$ when $i=j$, and \mathbf{C} is symmetric ($\rho_{i,j} =$
 124 $\rho_{j,i}$). The first row of \mathbf{C} , denoted as \mathbf{CV}_1 , represents the correlation vector between the first soil cell
 125 and all other soil cells. Using \mathbf{CV}_1 , the relative spatial position of the first soil cell within the site can
 126 be expressed. Similarly, the relative spatial positions of the remaining soil cells can be represented as
 127 $\mathbf{CV}_i (i \in [1, N])$. As shown in Fig. 1(a), the correlation vectors for soil cells A–E are illustrated for the
 128 case where $SO F_x$ and $SO F_y=100$ m, and $SO F_z=1.0$ m. The correlation decreases as the distance from
 129 soil cells A–E increases, indicating that soil properties exhibit correlation only within the scale of

130 fluctuation.

$$131 \quad \mathbf{C}_{N \times N} = \begin{bmatrix} \rho_{1,1} & \rho_{1,2} & \cdots & \rho_{1,j} & \cdots & \rho_{1,N} \\ \rho_{2,1} & \rho_{2,2} & \cdots & \rho_{2,j} & \cdots & \rho_{2,N} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{i,1} & \rho_{i,2} & \cdots & \rho_{i,j} & \cdots & \rho_{i,N} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{N,1} & \rho_{N,2} & \cdots & \rho_{N,j} & \cdots & \rho_{N,N} \end{bmatrix} = \begin{bmatrix} \mathbf{CV}_1 \\ \mathbf{CV}_2 \\ \cdots \\ \mathbf{CV}_i \\ \cdots \\ \mathbf{CV}_N \end{bmatrix} \quad (i, j \in [1, N]) \quad (2)$$



133 Fig. 1 Visual representation of the 3D correlation matrix and geotechnical correlation fields: (a)
 134 Correlation of soil cells A–E with all other soil cells (SoF_x and SoF_y=100 m, and SoF_z=1.0 m); (b)
 135 Forms of geotechnical correlation fields (GCFs) corresponding to different principal components.
 136
 137

Note: Each GCF has been normalized to a range of 0–1.

138 Conventional subsurface modeling methods such as conditional random fields or Kriging often
 139 require operations like matrix inversion of the correlation matrix \mathbf{C} , which can be computationally
 140 intensive when N (the number of soil cells) is large. A simplified approach proposed by [Chen et al.](#)
 141 [\(2024\)](#) involves the following steps: ① Data Collection: Gather measured soil cell properties q_i ($i \in$
 142 [measured cells]) and their corresponding correlation vectors \mathbf{CV}_i ; ② Data-Driven Modeling: Use a
 143 data-driven model to establish a complex nonlinear relationship between \mathbf{CV}_i and q_i . ③ Prediction:

144 Apply the trained data-driven model to predict the properties of unmeasured soil cells.

145 It is important to note that soil properties are only correlated within the range of the scale of
 146 fluctuation, leading to a sparse matrix with many zero elements in \mathbf{C} . Therefore, directly using the
 147 correlation vector \mathbf{CV}_i as input features for the data-driven model is not ideal. This approach may result
 148 in excessively high input dimensions with a significant number of redundant zero features, which can
 149 substantially reduce both the training and prediction efficiency of the model. To address this issue, this
 150 study first applies dimensionality reduction to \mathbf{C} using Principal Component Analysis (PCA). The top
 151 k eigenvectors with the largest eigenvalues are used to represent $\mathbf{P}_{\mathbf{C}(N \times k)}$, as shown in Eq. (2). Since k
 152 is much smaller than N , $\mathbf{P}_{\mathbf{C}(N \times k)}$ is easier to store compared to the original correlation matrix $\mathbf{C}_{N \times N}$. In
 153 this study, $\mathbf{P}_{\mathbf{C}(N \times k)}$ is referred to as the geotechnical correlation field, which encapsulates the main
 154 information from the correlation matrix.

$$155 \quad \mathbf{P}_{\mathbf{C}(N \times k)} = \text{PCA}(\mathbf{C}_{N \times N}) = \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,k} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,k} \\ \cdots & \cdots & \cdots & \cdots \\ P_{i,1} & P_{i,2} & \cdots & P_{i,k} \\ \cdots & \cdots & \cdots & \cdots \\ P_{N,1} & P_{N,2} & \cdots & P_{N,k} \end{bmatrix} = \begin{bmatrix} \mathbf{PV}_1 \\ \mathbf{PV}_2 \\ \cdots \\ \mathbf{PV}_i \\ \cdots \\ \mathbf{PV}_N \end{bmatrix} (i \in [1, N]) \quad (2)$$

156 As shown in Fig. 1(b), ten geotechnical correlation fields (GCFs) of a specific site are presented,
 157 highlighting the significant differences in the forms of different GCFs. GCF #1 ($[P_{1,1}, P_{2,1}, \dots, P_{N,1}]^T$),
 158 representing the geotechnical correlation field corresponding to the first principal component, exhibits
 159 relatively smooth variations in the xz and yz planes. Compared to GCF #1, GCF #12 shows notable
 160 distinctions in the xy plane, while GCF #20 exhibits more complex fluctuations in the xz and yz planes
 161 relative to GCF #12. Overall, the complexity of GCFs increases with the growth of k . Consequently,
 162 GCFs effectively capture the spatial variability of soil properties within the site, encompassing both

163 low-frequency and high-frequency information. Additionally, the diverse forms of GCFs offer a robust
164 set of basis functions for accurately characterizing spatial variations of soil properties.

165 **2.2 Simplified Strategy for Generating 3D GCFs**

166 GCFs effectively address the challenges of low computational efficiency in data-driven models
167 caused by the sparsity and high dimensionality of the original correlation matrix. Additionally, GCFs
168 provide diverse basis functions for data-driven subsurface modeling and embed random field theory
169 knowledge into the input features of the model. However, for large-scale or high-resolution 3D sites,
170 the number of soil cells N can reach tens of thousands. Directly processing such massive correlation
171 matrices to obtain GCFs can be exceedingly time-consuming. Therefore, it is necessary to propose a
172 simplified approach to enhance the efficiency of GCF generation, thereby expanding the applicability
173 of GCF-based data-driven subsurface modeling methods.

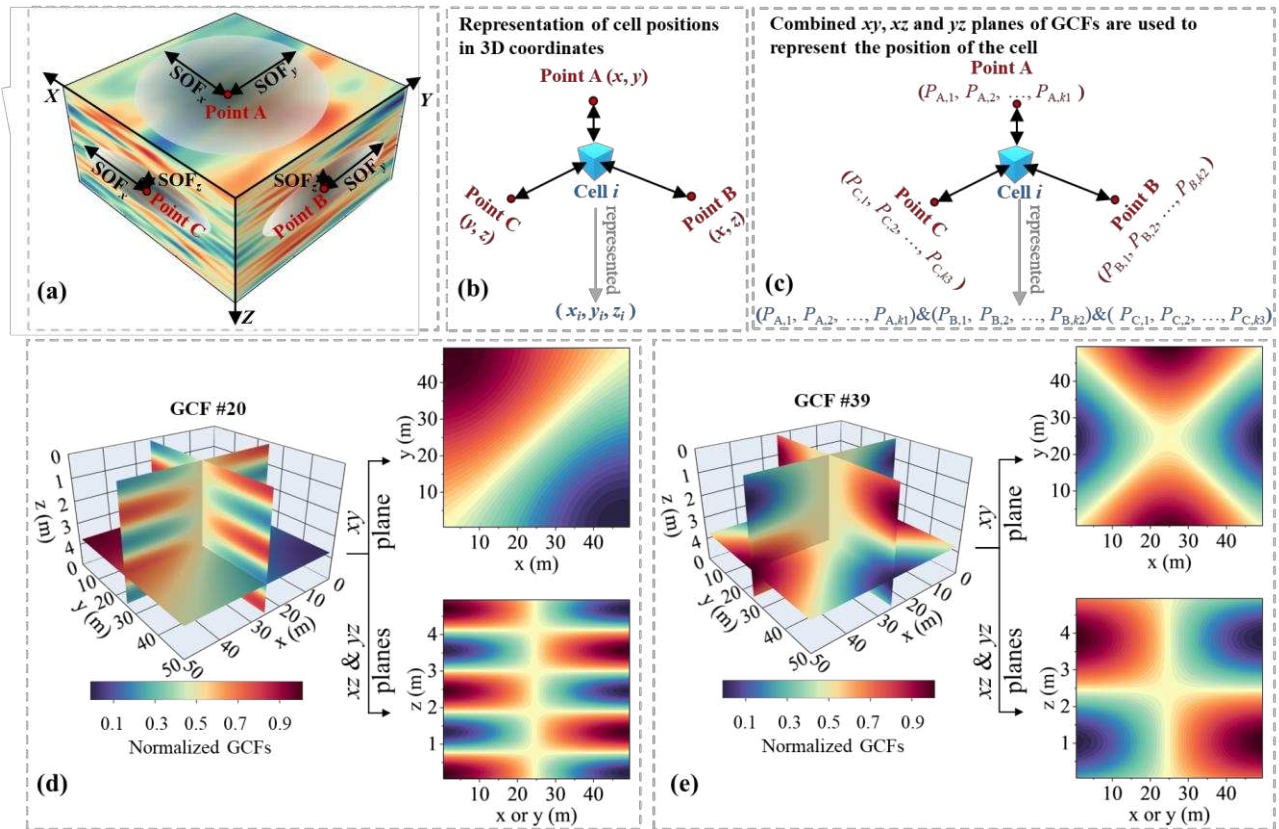
174 **(1) Simplifying 3D GCFs into a Combination of 2D GCFs on xy , xz , and yz Planes**

175 As shown in Fig. 2(a), horizontal scale of fluctuations in geotechnical sites are typically larger,
176 ranging from 5 to 105 m, while vertical scale of fluctuations are relatively smaller, between 0.1 and
177 3.1 m (Chen et al., 2023). Figs. 2(d)-(e) illustrate two 3D GCFs generated with SoF_x and $SoF_y=100$ m,
178 and $SoF_z=1.0$ m. Interestingly, directly calculating GCFs on the xy plane and the xz/yz 2D planes yields
179 GCFs with cross-sectional forms consistent with those of the 3D GCFs. This observation indicates that
180 3D GCFs essentially represent a superposition of a series of 2D GCFs.

181 Therefore, this study adopts a combination of GCFs from the xy , xz , and yz planes to represent the
182 spatial positions of soil cells in 3D space, referencing the representation of (x, y, z) coordinates in 3D
183 Euclidean space, as shown in Fig. 2(b). This approach eliminates the need for direct generation and

184 processing of large correlation matrices in 3D sites, substituting them with more manageable 2D planes.

185 To do so, the GCFs for the xy , yz , and xz planes are calculated separately. The projection points
 186 for soil cell i in these three planes (denoted as points A, B, and C in Fig. 2(c)) are identified. Combining
 187 the GCFs of the projection points from the three planes, the representation of the soil cell is obtained
 188 in 3D space as $\{[P_{A,1}, P_{A,2}, \dots, P_{A,k1}] \& [P_{B,1}, P_{B,2}, \dots, P_{B,k2}] \& [P_{C,1}, P_{C,2}, \dots, P_{C,k3}]\}$, where $k1$, $k2$,
 189 and $k3$ represent the number of principal components retained in the xy , yz , and xz planes. The values
 190 of k for different planes are discussed in Section 3.2.



191
 192 Fig. 2 Transforming 3D GCFs into a combination of 2D GCFs: (a) Schematic of a 3D geotechnical
 193 site and its SoFs; (b) Representation of soil cell spatial positions using 3D coordinates; (c)
 194 Representation of soil cell spatial positions using a combination of GCFs on the xy , xz , and yz planes;
 195 (d)-(e) Decomposition of 3D GCFs into GCFs on the xy , xz , and yz planes.

196 **(2) Reducing the Computational Complexity of GCFs via Sparse Sampling**

197 It is worth noting that simplifying 3D GCFs into a combination of GCFs on the xy , xz , and yz

198 planes, can effectively improve the efficiency of GCF generation. However, in large-scale or high-
 199 resolution sites, the number of soil cells in a 2D plane remains substantial, making subsequent PCA
 200 operations computationally challenging. Therefore, this study proposes a sparse sampling strategy to
 201 reduce the computational complexity of generating 2D GCFs. For clarity, Fig. 3 illustrates the process
 202 of generating GCFs and subsurface modeling using the xy -plane. Similar processes repeat for the xz
 203 and yz planes. The key steps are summarized as follows:

204 **Simplified Step 1:** Divide the site into N soil cells according to the requirements of the
 205 engineering project.

206 **Simplified Step 2:** A total of G soil cells are sampled at regular intervals from the N total cells,
 207 referred to as GCFs cells. The impact of the sampling interval is discussed in Section 3.1.

208 **Simplified Step 3:** As shown in Eqs. (3) and (4), the correlation matrix $\mathbf{C}_{N \times G}$ is computed for the
 209 N soil cells and the sampled G cells, resulting in an $N \times G$ dimensional matrix. Simultaneously, the
 210 correlation matrix $\mathbf{C}_{G \times G}$ for the G soil cells is calculated, yielding a $G \times G$ dimensional matrix. It is
 211 noteworthy that both $\mathbf{C}_{N \times G}$ and $\mathbf{C}_{G \times G}$ have lower dimensions than $\mathbf{C}_{N \times N}$, which facilitates storage.

$$212 \quad \mathbf{C}_{N \times G} = \begin{bmatrix} \rho_{1,1} & \rho_{1,2} & \cdots & \rho_{1,j} & \cdots & \rho_{1,G} \\ \rho_{2,1} & \rho_{2,2} & \cdots & \rho_{2,j} & \cdots & \rho_{2,G} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{i,1} & \rho_{i,2} & \cdots & \rho_{i,j} & \cdots & \rho_{i,G} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{N,1} & \rho_{N,2} & \cdots & \rho_{N,j} & \cdots & \rho_{N,G} \end{bmatrix} \quad (i \in [1, N], j \in [1, G]) \quad (3)$$

$$213 \quad \mathbf{C}_{G \times G} = \begin{bmatrix} \rho_{1,1} & \rho_{1,2} & \cdots & \rho_{1,j} & \cdots & \rho_{1,G} \\ \rho_{2,1} & \rho_{2,2} & \cdots & \rho_{2,j} & \cdots & \rho_{2,G} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{i,1} & \rho_{i,2} & \cdots & \rho_{i,j} & \cdots & \rho_{i,G} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{G,1} & \rho_{G,2} & \cdots & \rho_{G,j} & \cdots & \rho_{G,G} \end{bmatrix} \quad (i, j \in [1, G]) \quad (4)$$

214 where $\rho_{i,j}$ represents the correlation between the i -th and j -th soil cells, which is calculated using the
 215 ACFs and the corresponding SoFs. The detailed calculation formula can be found in Table A1 of

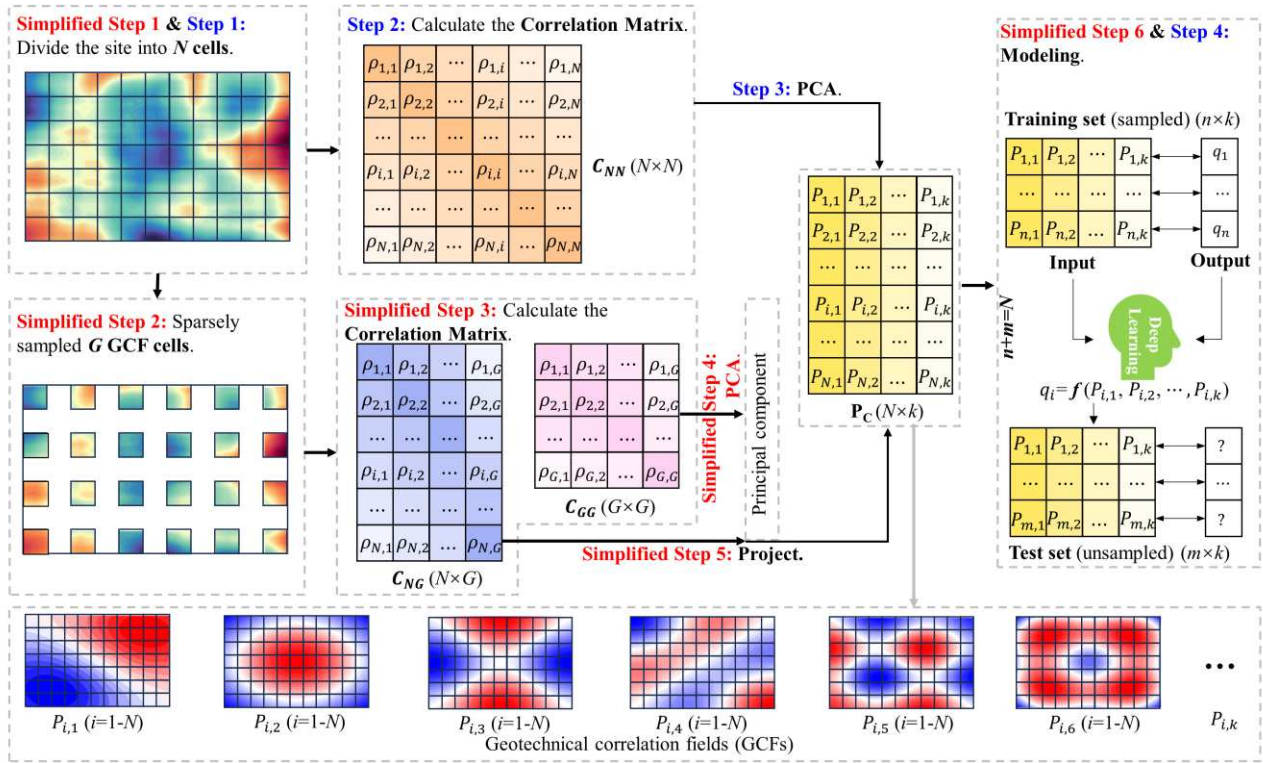
216 Appendix A. However, accurately determining these parameters is challenging in the presence of
 217 sparse survey information (Qi et al., 2022; Xiao et al., 2018; Zhang et al., 2021). Section 3.2 provides
 218 a detailed discussion on the impact of these parameters.

219 **Simplified Step 4 & 5:** PCA is applied to $\mathbf{C}_{G \times G}$ to obtain a projection matrix with k principal
 220 components. Subsequently, $\mathbf{C}_{N \times G}$ is projected onto the selected k principal components, resulting in an
 221 $N \times k$ dimensional $\mathbf{P}_{\mathbf{C}_{(N \times k)}}$ matrix, as shown in Eq. (5). Notably, this approach significantly reduces
 222 computational complexity compared to directly performing PCA on the $\mathbf{C}_{N \times N}$, especially when N is
 223 large. The PCA operations are implemented using Python's scikit-learn v1.1.3 (Pedregosa et al., 2018).
 224 Further discussion can be found in Section 3.1.

$$225 \quad \mathbf{P}_{\mathbf{C}_{(N \times k)}} = \mathbf{C}_{N \times G} \times (\text{PCA}(\mathbf{C}_{G \times G}))_{G \times k} \quad (i \in [1, N]) \quad (5)$$

226 **Simplified Step 6:** Determine the survey locations and collect the measured CPT results q (e.g.,
 227 cone tip resistance, sleeve friction) or other soil parameters (e.g., penetration pore water pressure,
 228 undrained shear strength). The measured properties of n soil cells are used to form the training set,
 229 while the remaining m ($m=N-n$) soil cells, with unknown properties, form the test set. The input
 230 features of the training set (i.e., $N \times k$ matrix) are the row vectors corresponding to n soil cells in the
 231 $\mathbf{P}_{\mathbf{C}}$ matrix, and the outputs are the measured values of the soil properties for n soil cells. A data-driven
 232 model is used to learn the nonlinear relationship between the input features and outputs in the training
 233 set. Once trained, the model can be used to predict the properties of m unsampled soil cells. This study
 234 integrates random field theory into the data-driven model through input features, making the proposed
 235 method highly flexible and applicable to any data-driven model. The Shortcut-Connected Neural
 236 Network (SCNN) and the Extra Trees (ET) models are adopted and compared. A detailed description

237 of the SCNN and ET models is provided in Appendix B.



238

239 Fig. 3 Process of generating GCFs and performing subsurface modeling, illustrated using the xy -
 240 plane as an example.

241 2.3 Characterization of Uncertainty in Subsurface Modeling

242 The task of subsurface modeling requires inferring soil properties at unmeasured locations based on
 243 sparse measurement data. Typically, the number of soil cells in unknown areas far exceeds the number
 244 of measured cells, leading to significant uncertainties in the subsurface modeling results. Accurate
 245 quantification of the uncertainty in subsurface modeling is crucial for guiding the investigation and
 246 design processes in geotechnical engineering (Yan et al., 2023; Zhang et al., 2024; C. Zhao et al., 2023).
 247 Prioritizing investigation points in areas of high uncertainty can effectively reduce the uncertainty of
 248 the subsurface model and mitigate the risk of geotechnical disasters caused by spatial variability in soil
 249 properties.

250 The data-driven models used in this study are the ET model and the SCNN model. The ET model,

251 as an ensemble learning method, approximates the true hypothesis by combining multiple weak
 252 learners. A common strategy to quantify the predictive uncertainty of ensemble learning models is to
 253 perform statistical analysis on the multiple weak learners (Xie et al., 2022a). According to the central
 254 limit theorem, the resampling distribution of the predicted values will approximate a gaussian
 255 distribution. The standard deviation (σ) of the predictions from all weak learners can be used to
 256 estimate the uncertainty of the prediction results:

$$257 \quad \sigma = \sqrt{\frac{\sum_{i=1}^n [f_{\text{ET}}^i(\mathbf{PV}) - \overline{f_{\text{ET}}}]^2}{n - 1}} \quad (6)$$

258 where $f_{\text{ET}}^i(\mathbf{PV})$ is the prediction result from the i -th weak learner; $\overline{f_{\text{ET}}}$ is the average prediction result
 259 from the n weak learners,

260 For the SCNN model, uncertainty can be quantified using Monte Carlo dropout (MC dropout) and
 261 random weight initialization methods to quantify uncertainty in model predictions. MC dropout
 262 operates by randomly disconnecting a certain proportion of neural connections, resulting in an altered
 263 model architecture (T. Wang et al., 2023; P. Zhang et al., 2022). Given that the model architecture
 264 changes each time, the standard deviation of the predictions after performing t Monte Carlo dropout
 265 sampling runs can be calculated as follows:

$$266 \quad \sigma = \sqrt{\frac{\sum_{i=1}^t [f_{\text{SCNN}}^i(\mathbf{PV}, \text{dropout}(\mathbf{W}, \mathbf{b})) - \overline{f_{\text{SCNN}}}]^2}{t - 1}} \quad (7)$$

267 where \mathbf{W} and \mathbf{b} are the weights and biases of the neural network; $\text{dropout}(\mathbf{W}, \mathbf{b})$ refers to randomly
 268 deactivating a portion of the neural connections during the t -th prediction, effectively altering the
 269 network's architecture and introducing variability; $f_{\text{SCNN}}^i(\mathbf{PV}, \text{dropout}(\mathbf{W}, \mathbf{b}))$ represents the
 270 prediction result from the i -th random forward pass of the network, where dropout is applied each time;

271 $\overline{f_{\text{SCNN}}}$ is the mean prediction result across all t forward passes (predictions). The MC dropout method
 272 is simple and efficient. However, the uncertainty in prediction results is often closely related to the
 273 design of the dropout structure, such as the dropout rate and the number of dropout layers.
 274 Consequently, selecting an appropriate dropout structure generally requires engineers to have
 275 substantial experience and time to conduct repeated trials.

276 The random weight initialization method does not require any additional user-defined parameters.
 277 It introduces variability into the model by assigning different initial trainable parameters to the neural
 278 network, causing the model to follow different optimization paths during the iterative optimization
 279 process. As a result, this method generates different predictions based on the distinct paths taken by
 280 the model. However, the random weight initialization method requires training m neural networks,
 281 each with different initial parameters. This can be computationally expensive, particularly for complex
 282 models that are time-consuming to train. The uncertainty in the prediction results can be represented
 283 as:

$$284 \quad \sigma = \sqrt{\frac{\sum_{i=1}^m [f_{\text{SCNN}}^i(\mathbf{PV}, \mathbf{W}_i, \mathbf{b}_i) - \overline{f_{\text{SCNN}}}]^2}{m - 1}} \quad (8)$$

285 where \mathbf{W}_i and \mathbf{b}_i are the weights and biases of the i -th neural network; $f_{\text{SCNN}}^i(\mathbf{PV}, \mathbf{W}_i, \mathbf{b}_i)$ represents
 286 the prediction from the i -th neural network, trained with different initial parameters; $\overline{f_{\text{SCNN}}}$ is the
 287 average prediction result from the m networks. It is important to note that the SCNN model structure
 288 employed in this study is relatively simple and easy to train. Therefore, random weight initialization
 289 is employed to quantify uncertainty in subsurface modeling.

290 **2.4 Model Interpretability**

291 Regarding deep learning for geotechnical engineering, the relationship between inputs and outputs

292 remains difficult to analyze and interpret. SHapley Additive exPlanations (SHAP) is a game-theory-
293 based method for interpreting black-box models. This study uses SHAP to interpret the ET and SCNN
294 models used for subsurface modeling, providing deeper insights into the model's internal mechanisms
295 and evaluating the contribution of each input feature (i.e., GCFs) to the random field predictions. SHAP
296 method avoids the necessity of heuristically selecting methods to linearize components. Instead, it
297 derives an effective linearization directly from the SHAP values calculated for each component
298 (Lundberg and Lee, 2017).

299 The TreeSHAP method can be used to interpret the output of the ET model and measure the
300 importance of each dimension of the GCFs in contributing to the subsurface modeling results. Similarly,
301 for the SCNN model, the DeepSHAP method, which is specifically designed for deep learning models,
302 can be applied to explain the model's predictions. In both cases, these methods provide a way to
303 quantify the contribution of each dimension of the GCFs to the final model output. This is particularly
304 useful for understanding the underlying factors driving the predictions of complex models, where
305 interpretability is crucial for ensuring model reliability and trustworthiness. By using TreeSHAP and
306 DeepSHAP for SCNN models, it is possible to gain insight into how different GCFs, as well as their
307 individual components, influence the model's predictions, and identify which features are most
308 important for making accurate subsurface predictions. This study uses the Python-based SHAP library
309 v0.46.0 (Lundberg and Lee, 2017) for model interpretation.

310 **2.5 Evaluation Metrics**

311 Three types of evaluation metric are employed to assess the accuracy of subsurface modeling: (1)
312 R-squared (R^2) is used to evaluate the goodness of fit, with values closer to 1 indicating better model

313 performance (see Eq. (7)); (2) the root mean square error (*RMSE*) is used to evaluate the model's error
 314 from the perspective of absolute error (see Eq. (8)); (3) the mean absolute percentage error (*MAPE*) is
 315 employed to assess model accuracy from the perspective of relative error (see Eq. (9)). It is important
 316 to note that when the measured value (e.g., cone tip resistance) of a soil cell is close to zero, *MAPE*
 317 may not adequately assess the model's performance. Therefore, the symmetric mean absolute
 318 percentage error (*sMAPE*) is introduced as a supplementary metric (see Eq. (10)).

$$319 \quad R^2 = 1 - \frac{\sum_{i=1}^N (q_i - \hat{q}_i)^2}{\sum_{i=1}^N (q_i - \bar{q})^2} \quad (7)$$

$$320 \quad RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (q_i - \hat{q}_i)^2} \quad (8)$$

$$321 \quad MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{q_i - \hat{q}_i}{q_i} \right| \times 100\% \quad (9)$$

$$322 \quad sMAPE = \frac{1}{N} \sum_{i=1}^N \frac{|q_i - \hat{q}_i|}{(q_i + \hat{q}_i)/2} \times 100\% \quad (10)$$

323 where N is the total number of soil cells; q_i and \hat{q}_i are the measured and predicted properties of soil
 324 cell i .

325 **3 Validation of two simplified strategies**

326 A set of synthetic zero-mean stationary gaussian random fields with different horizontal and
 327 vertical SoF are used to illustrate the performance and sensitivity of the proposed method with the
 328 special focus on the following aspects: (1) how to sample as few soil cells as possible to generate
 329 accurate GCFs, bypassing the storage of large correlation matrices and reducing the computational
 330 complexity of PCA; (2) the impact of RFT parameters on subsurface modeling. The synthetic 3D site
 331 spans 100 m horizontally (x and y) and has a depth (z) of 10 m. The site is simulated with resolutions

332 of 1.0 m horizontally and 0.1 m vertically. The mean (μ) and stand deviation (σ) of the synthetic
333 samples are set as 11 MPa and 3 MPa, respectively. The horizontal SoFs of the synthetic data are 10,
334 30, 50, 70, and 90 m, while the vertical SoFs are 0.1, 0.5, 1.0, 1.5, and 2.0 m. For each combination of
335 SoFs, 50 samples are randomly simulated using the Cholesky decomposition method (Y. Yang et al.,
336 2022), yielding a total of 3,750 synthetic samples.

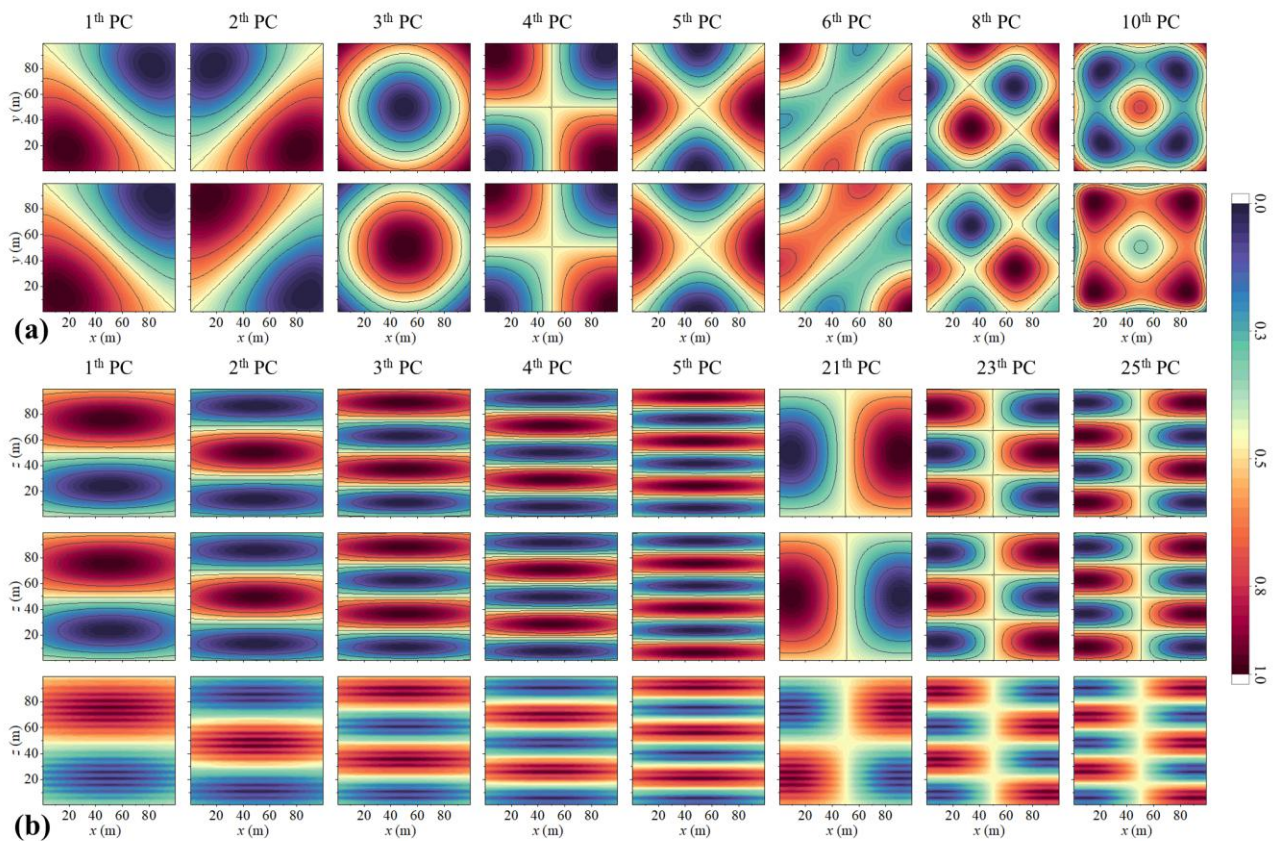
337 3.1 Reducing the Computational Complexity of GCF Generation

338 This study uses a combination of three 2D-plane GCFs (xy , xz , and yz) to represent the spatial
339 locations of soil cells within a 3D site. Since there is no essential difference between the yz and xz
340 planes, only the xy and xz planes are analyzed in subsequent steps. To avoid redundancy, the SoF values
341 in the x and y directions are set to 100 m, and the SoF in the z direction is set to 0.5 m for the
342 presentation. For detailed procedures, see Section 2.2.

343 As shown in Fig. 4(a), a comparison is made between the GCFs generated using all soil cells (first
344 row) and those generated with a sampling interval of 10 m (second row). The GCFs are arranged in
345 order of decreasing eigenvalue as part of the PCA operation. It is observed that the GCFs exhibit an
346 increasing trend in complexity from front to back. Furthermore, the GCFs generated using the sparse
347 sampling strategy demonstrate a high degree of consistency with the original GCFs. It is worth noting
348 that each GCF is normalized to ensure it remains within the gradient-sensitive range of the neural
349 network, thus accelerating the model's training process. Some of the simplified GCFs exhibit values
350 that are inversely related to those of the original GCFs. However, this does not affect the ML model's
351 performance.

352 As shown in Fig. 4(b), the GCFs generated using all soil cells (first row) closely resemble those

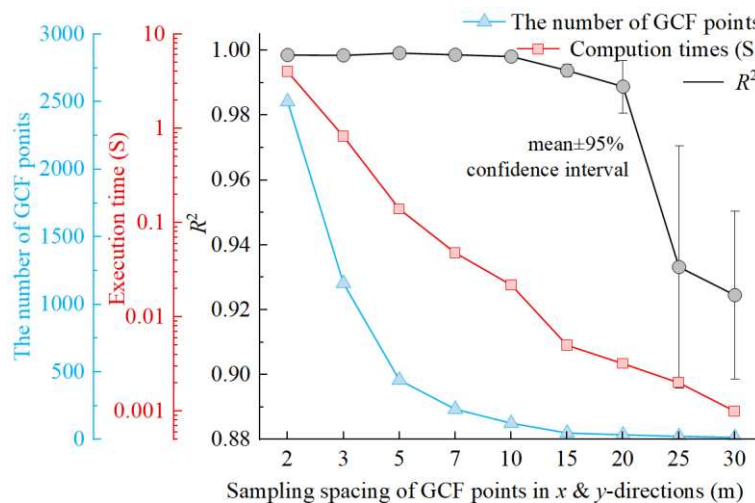
353 produced with a 0.2 m sampling interval in the z direction (second row). However, when a 0.5 m
 354 sampling interval is applied in the z direction, the generated GCFs exhibit noticeable striping. This
 355 phenomenon is primarily due to the site's SoF in the z direction being set at 0.5 m. Therefore, the
 356 sampling interval must be smaller than the SoF of the site. Additionally, there are significant
 357 differences in the morphology of GCFs between the xz and xy planes. The first 20 GCF components in
 358 the xz plane resemble vertical waves, with progressively shorter periods and higher frequencies as the
 359 components advance. Beyond the 21st component, the GCFs resemble overlapping transverse and
 360 vertical waves, with an increasing frequency. These diverse GCF components will serve as basis
 361 functions, providing a robust foundation for subsequent subsurface modeling.



362
 363 Fig. 4 Performance of generating GCFs using the simplified strategy: (a) Morphology of GCFs at
 364 different sampling intervals when the SoFs in the x and y directions of the xy plane are set to 100 m,
 365 with the first row showing the original GCFs and the second row showing the simplified GCFs at a

366 10 m sampling interval; (b) Morphology of GCFs when the SoFs in the x and z directions of the xz
 367 plane are set to 100 m and 0.5 m, respectively, with the first row showing the original GCFs and the
 368 second and third rows showing the simplified GCFs at sampling intervals of 0.2 m and 0.5 m.

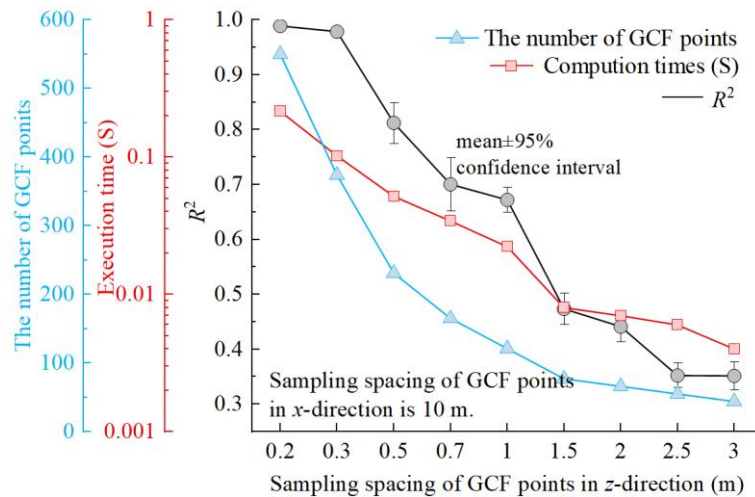
369 Fig. 5 illustrates the performance of generating GCFs using soil cells with different sampling
 370 intervals in the xy plane. It can be observed that when the sampling intervals in the x and y directions
 371 are set between 2 and 10 m, the GCFs generated using the simplified strategy show a high similarity
 372 to those generated using all soil cells, with an R^2 value approaching 1. When the sampling interval
 373 exceeds 15 m, the accuracy of the GCFs generated using the simplified strategy gradually decreases,
 374 leading to increased uncertainty. Therefore, it is recommended that the sparse sampling interval for
 375 generating the GCFs in the xy -plane should be smaller than 0.1 times the SoF_x or SoF_y . It is noteworthy
 376 that generating GCFs directly using all soil cells (10,000 in total) requires the PCA to process a 10,000
 377 \times 10,000 correlation matrix. However, when the sampling interval is set to 10 m, only 121 soil cells
 378 are available, requiring PCA to process a correlation matrix of size 121 \times 121. This significantly
 379 reduces the computational complexity of the PCA operation and confines the generation time of the
 380 GCFs to approximately 0.02 s.



381
 382 Fig. 5 Impact of different sampling intervals in the x and y directions on the performance of GCFs

383 generated in the xy plane.

384 As shown in Fig. 6, the performance of generating GCFs using soil cells with different sampling
385 intervals in the xz plane is evaluated. Given that sampling in the x direction at 10 m intervals in the xy
386 plane has minimal impact on the accuracy of the generated GCFs, the sampling interval in the x
387 direction is fixed at 10 m for this analysis. It can be observed that when the sampling interval in the z
388 direction is between 0.2 and 0.3 m, the GCFs generated closely resemble those produced using all soil
389 cells, with an R^2 value approaching 1. However, when the sampling interval exceeds 0.3 m, the
390 accuracy of the simplified GCFs declines rapidly. Therefore, it is recommended that the sparse
391 sampling interval for generating the GCFs in the xz and yz planes should be smaller than 0.6 times the
392 SoF_z . When the interval is smaller than the resolution in the z -direction of the site, no sparse sampling
393 should be performed. Notably, only 550 GCF cells are used when the sampling interval in the x
394 direction is 10 m and in the z direction is 0.2 m. This significantly reduces the size of the correlation
395 matrix and confines the generation time of the GCFs to approximately 0.2 s.



396
397 Fig. 6 Effect of different sampling intervals in the z direction on the performance of GCFs generated
398 in the xz plane

399 **3.2 Impact of RFT Parameters on Subsurface Modeling Results**

400 The GCFs generation process is controlled by random field theory. Therefore, it is necessary to
401 investigate the influence of RFT parameters (SoF and ACF type) on subsurface modeling. To achieve
402 this, a series of xy and xz sections with different combinations of SoFs are extracted from synthetic 3D
403 samples. Each SoF combination is randomly sampled 50 times. In the xy plane, 25 measurement points
404 are arranged at grid points corresponding to x and y coordinates of 0.5 m, 25.5 m, 50.5 m, 75.5 m, and
405 99.5 m. In the xz plane, 5 boreholes are arranged along sample paths corresponding to x coordinates
406 of 0.5 m, 25.5 m, 50.5 m, 75.5 m, and 99.5 m.

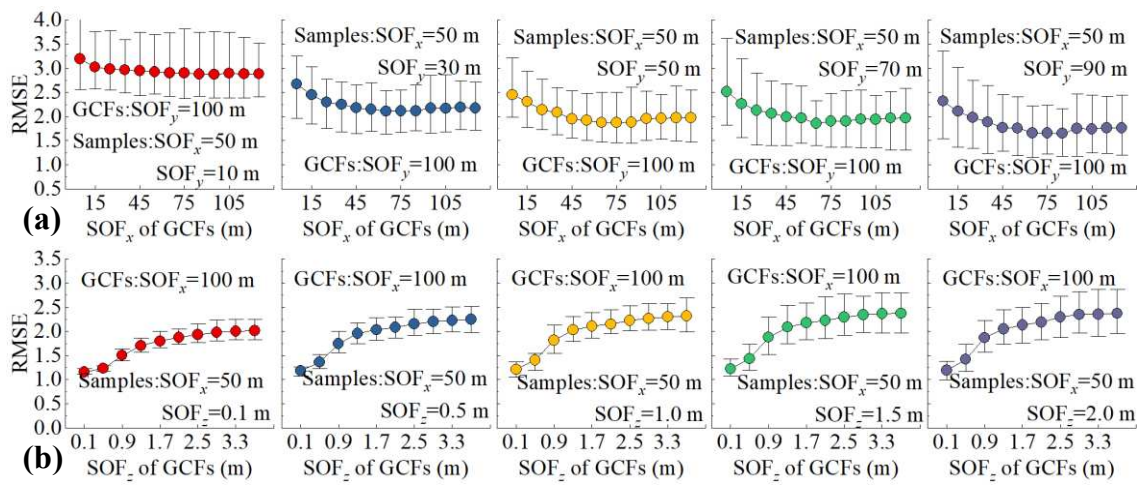
407 **(1) The impact of SoFs on modeling accuracy**

408 As shown in Figs. 7(a) and (b), different SoF_x and SoF_z values are used to generate GCFs and
409 perform subsurface modeling. For brevity, only 500 sets of the synthetic field with an SoF_x of 50 m
410 are presented. To eliminate the impact of other factors on modeling accuracy, the GCF dimensions
411 used for the xy and xz planes are set to 50 and 300, respectively.

412 For the xy plane, the modeling error (RMSE) decreases as the SoF_x used for generating GCFs
413 increases, up to the SoF_x of 50 m. This indicates that in the sparsely measured xy plane, using a larger
414 SoF to generate GCFs is beneficial, as it allows more measured data to be used for inferring the
415 properties of unsampled locations. Therefore, selecting a horizontal SoF as 2 times the borehole
416 spacing is a more general choice. Additionally, the horizontal SoF can be treated as a hyperparameter
417 and optimized using a cross-validation method to ensure the rationality of the generated GCFs.

418 Additionally, it can be observed that as the true scale of fluctuation at the site increases, the spatial
419 variability of soil properties becomes smoother, and the modeling error decreases. For the xz plane, the

420 modeling error increases as SoF_z increases. Using a larger SoF_z to generate GCFs results in predictions
 421 that reflect the average trend of the site, thus reducing prediction accuracy. However, given that
 422 measurement data in the z direction are abundant, using a smaller SoF_z to generate GCFs and perform
 423 subsurface modeling is reasonable. This approach helps to accurately reflect the spatial variability of
 424 soil properties with depth. Overall, setting SoF_z equal to the resolution of the site in the z direction is
 425 statistically the optimal choice for generating GCFs.



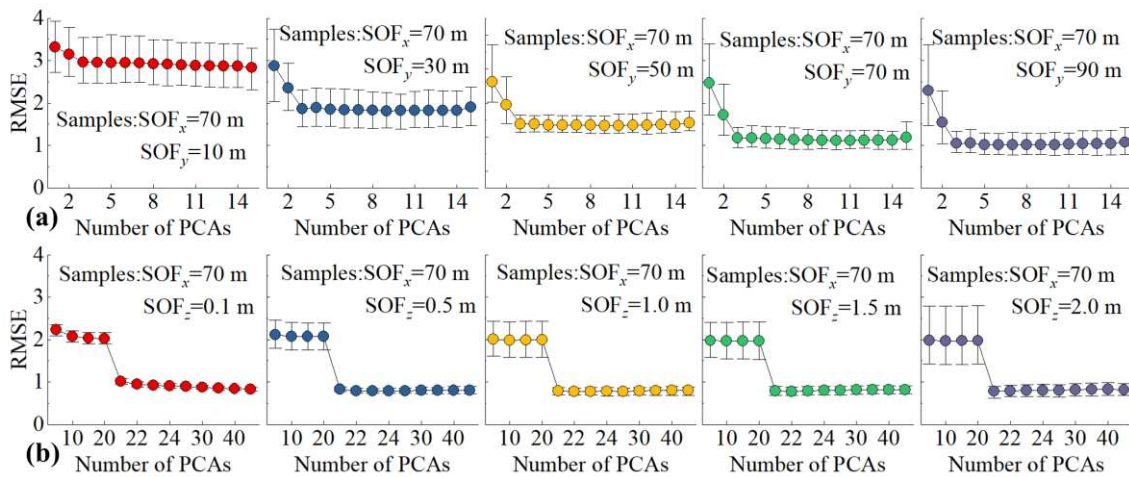
426

427 Fig.7 Impact of SoF used to generate GCFs on subsurface modelling results: (a) The impact of
 428 different SoF_x values on the modeling accuracy of the xy plane. (b) The impact of different SoF_z
 429 values on the modeling accuracy of the xz plane.

430 **(2) Impact of GCFs dimension on modeling accuracy**

431 Figs. 8(a) and (b) illustrate the impact of using different numbers of GCFs (i.e. PCAs) on the
 432 accuracy of subsurface modeling in the xy and xz planes. To avoid redundancy, only 500 synthetic
 433 cases with SoF_x of 70m are shown. For clarity and comparison with Fig. 4, the $SoFs$ used to generate
 434 GCFs are consistent with those in Section 3.1. As shown in Fig. 8(a), the error in subsurface modeling
 435 results decreases gradually with the increase in the number of GCFs. Specifically, when SoF_y is large,
 436 the accuracy of subsurface modeling is sufficiently high even with a small number of GCFs.

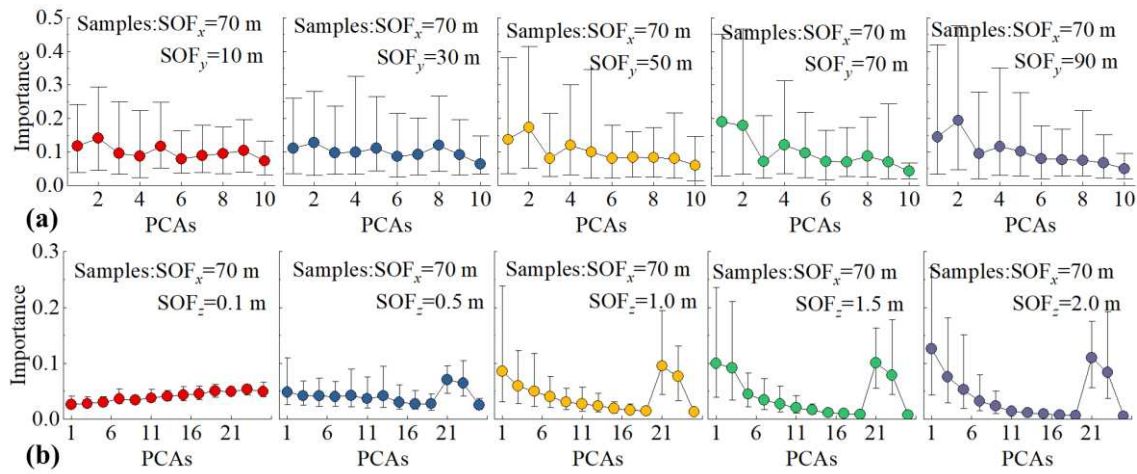
437 Conversely, when SoF_y is small, the spatial variability of the soil properties is higher, thus requiring
 438 more GCFs for accurate modeling. As shown in Fig. 9(a), when SoF_y is large, the first few GCFs have
 439 greater importance to the model. Overall, for the xy plane, choosing at least 10 GCFs as input features
 440 for the subsurface model is a reasonable choice. Increasing the number of GCFs helps to capture high-
 441 frequency information. In practical cases, the number of GCFs in the xy plane can be treated as a
 442 hyperparameter and optimized using cross-validation methods.



443
 444 Fig.8 The impact of retaining different numbers of GCFs (number of PCAs) on the modeling
 445 accuracy in the xy and xz planes.

446 As shown in Fig. 8(b), in the xz plane, the accuracy of subsurface modeling is low when the
 447 number of GCFs is fewer than 21. However, once the 21st GCF is incorporated into the modeling
 448 process, there is a significant improvement in model accuracy. As illustrated in Fig. 4(b), the first 20
 449 GCFs in the xz plane only describe the vertical correlation of the site, while the 21st GCF captures the
 450 horizontal correlation. The inclusion of the 21st GCF in subsurface modeling enhances accuracy
 451 significantly due to the combined effects of both horizontal and vertical correlations. As shown in Fig.
 452 9(b), the 21st GCF plays a significant role in the subsurface modeling results, confirming the validity
 453 of the previous analysis. Overall, for the xz plane, selecting at least 25 GCFs as input parameters for

454 the subsurface model is a reasonable choice. In practical applications, the number of GCFs in the xz
 455 plane should also be optimized using cross-validation methods.

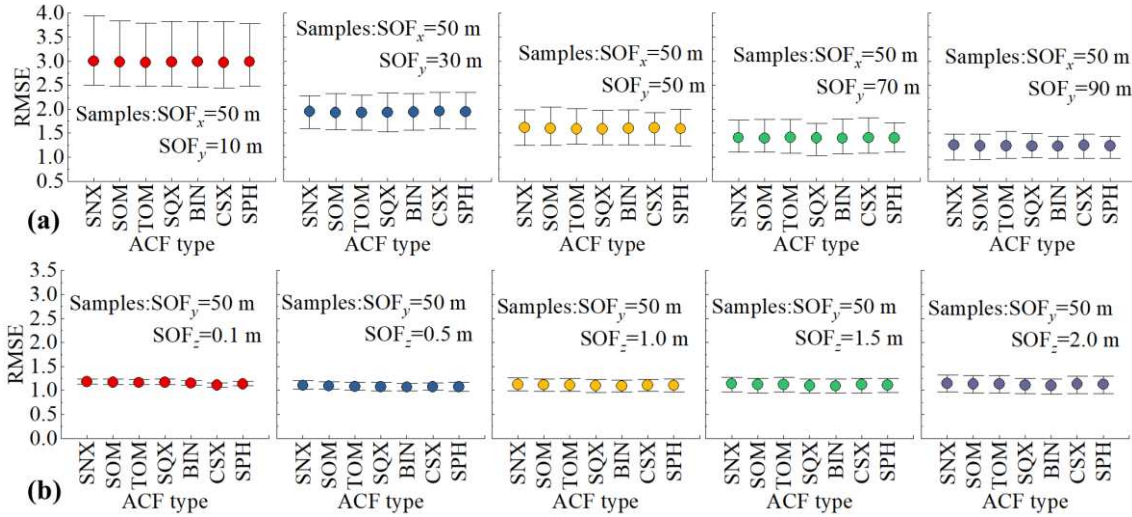


456
 457 Fig.9 The importance of different GCFs on the modeling results in the xy and xz planes.

458 (3) ACFs Type and Its Impact on Modeling Accuracy

459 As shown in Figs. 10(a) and (b), the impact of different ACF types on the accuracy of subsurface
 460 modeling in the xy and xz planes is compared. The primary distinction among the ACFs lies in the
 461 roughness of the synthetic sample paths. To avoid redundancy, only 500 synthetic cases with SoF_x of
 462 50 m are presented. In these cases, the SoF_x and SoF_y used to generate GCFs are both set at 100 m,
 463 while the SoF_z is set at 0.5 m. Statistical results indicate that the impact of different ACFs on modeling
 464 accuracy is negligible. Theoretically, when the number of GCFs actively participating in subsurface
 465 modeling is sufficiently large, the results should reflect the characteristics of the different ACFs.
 466 However, as shown in Figs. 8(a) and (b), even with an increased number of GCFs participating in
 467 subsurface modeling, the accuracy does not continuously improve. This is primarily due to the fact
 468 that the number of measured soil cells during the modeling process is significantly lower than that of
 469 the unmeasured cells. Therefore, even if more GCFs are included as input features, the model cannot
 470 learn the relationships between higher-order GCFs and soil cell properties from sparse data, leading to

471 the inevitable ignoring of higher-order GCFs by the model. Consequently, discussing the impact of
 472 ACF types on subsurface modeling results is only necessary when sufficient measurement data is
 473 available. However, in engineering practice, it is often not feasible to obtain copious measurement data,
 474 making the influence of different ACF types on subsurface modeling results negligible.



475
 476 Fig.10 The impact of using different types of ACFs on the subsurface modeling results in the xy and
 477 xz planes.

478 4 Implementation Procedures

479 A Python script has been developed to perform the subsurface modeling process based on GCFs.
 480 For detailed code, please refer to <https://github.com/Data-Driven-RFT/Sparse-Learning>. The proposed
 481 workflow is illustrated in Fig. 11, with the key steps summarized as follows:

482 (1) Data collection. Collect survey data, including locations and measurement data (q). Based on
 483 the site conditions, set reasonable resolution parameters and discretize the 3D site into N soil cells.

484 (2) Define the range of GCFs hyperparameters. For SoF_x and SoF_y , the range is set to $[x$ or y
 485 direction resolution, site x or y direction length]. For SoF_z , it is set to the z direction resolution. For the
 486 xy plane, the initial GCFs dimension should range between $[10, 100]$. For the xz and yz planes, the
 487 initial GCFs dimension should range between $[25, 300]$.

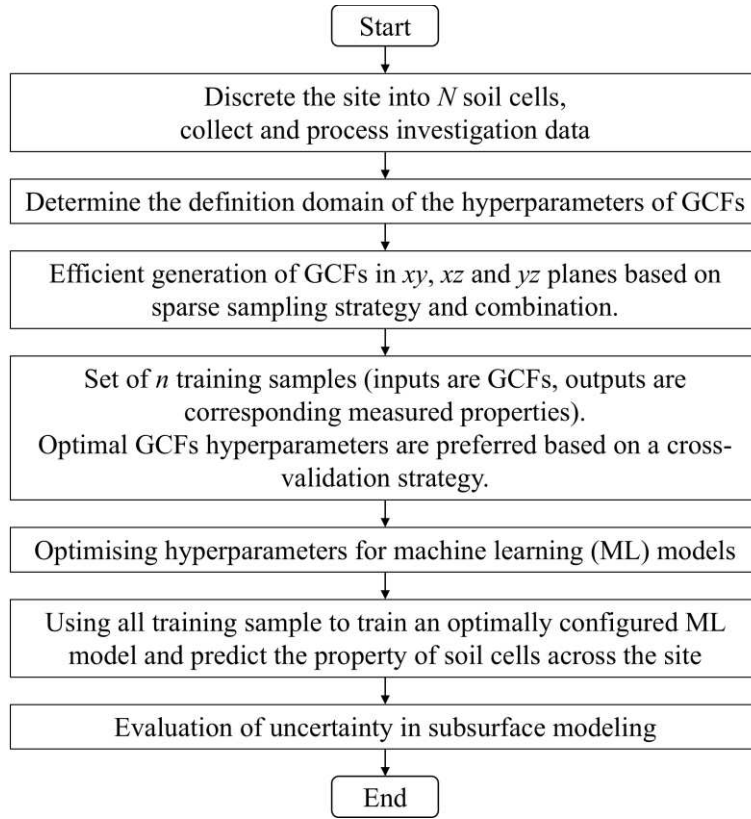
488 (3) Sparse sampling to generate GCFs. The corresponding GCFs are calculated according to Eqs.
489 (3)-(5) and then combined. It is important to note that the sampling intervals in the x and y directions
490 should be less than or equal to $0.1 \times \text{SoF}_{x/y}$, and in the z direction, the sampling interval should be
491 less than or equal to $0.6 \times \text{SoF}_z$. If the sampling interval is smaller than the resolution in the
492 corresponding direction, sparse sampling is not performed.

493 (4) Optimize the hyperparameters for GCFs. Pair the GCFs (input features) with the corresponding
494 soil properties (output) for the n measured soil cells, and organize them as training samples. Use cross-
495 validation strategies to train the data-driven model and optimize the best hyperparameters for GCFs.
496 It is important to select smaller GCFs dimensions that ensure modeling accuracy while accelerating
497 modeling efficiency.

498 (5) Optimize the hyperparameters of the ML model. Use grid search to optimize the
499 hyperparameters of the chosen machine learning model. Based on extensive synthetic case studies, it
500 was found that for the SCNN model, the main adjustments involve the number of neural network layers
501 and the number of neurons in each layer. For the ET model, the main adjustment involves the number
502 of trees, while other parameters remain default. Detailed configuration and description of the SCNN
503 and ET models can be found in Appendix B.

504 (6) Subsurface modeling. Use all measurement samples to train the machine learning model with
505 the best hyperparameters. Once the model is trained, it can directly be applied to predict the properties
506 of all soil cells across the entire site.

507 (7) Uncertainty estimation. Based on Eqs. (6)-(8), estimate the uncertainty in the subsurface
508 modeling.



509

510

Fig.11 Process flow of the proposed subsurface modeling method based on GCFs.

511

5 Illustrative Example

512

To validate the performance of the proposed method, the model is tested on a large number of 2D

513

stationary and nonstationary synthetic cases and compared with the alternative BCS and Kriging

514

methods. Subsequently, the proposed method is applied to a set of 3D synthetic cases to verify its

515

applicability for 3D sites. The BCS method is implemented using ASSD-BCS v1.2 Software, which

516

features a user-friendly visual interface (Lyu et al., 2023). The Kriging method is implemented using

517

the Python-based Gstools v1.5.0 (Müller et al., 2022).

518

5.1 Comparison with SOTA Methods for 2D Stationary Cases

519

From the 3D synthetic samples generated in Section 3, 50 xy and xz cross-sections with different

520

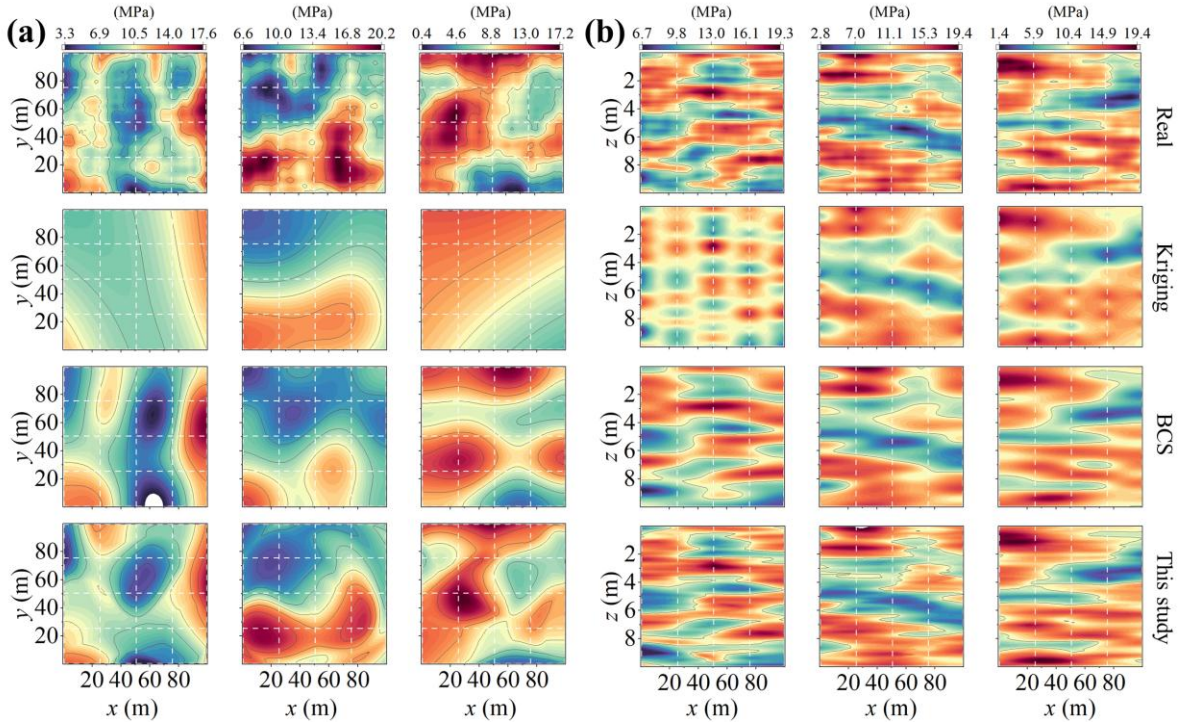
combinations of SoFs are extracted. Specifically, in the xy plane, the SoFs in both the x and y directions

521

are 50 m, while in the xz plane, the SoFs in the x and y directions are 70 m and 1.5 m, respectively. In

522 the xy plane, 25 measurement points are arranged at grid points corresponding to x and y coordinates
523 of 0.5 m, 25.5 m, 50.5 m, 75.5 m, and 99.5 m. In the xz plane, 5 boreholes are arranged along sample
524 paths corresponding to x coordinates of 0.5 m, 25.5 m, 50.5 m, 75.5 m, and 99.5 m.

525 It is worth noting that the Kriging method uses the same ACF as the synthetic samples, and the
526 model parameters are fitted using Gstools v1.5.0 based on measurement data. As shown in [Figs. 12\(a\)](#)
527 [and \(b\)](#), the prediction results from the Kriging method are smooth, approximating the average trend
528 within a local range. The Kriging method is a parametric method, and accurately estimating random
529 field parameters is crucial for its performance. However, in engineering, only sparse measurement
530 results are often available, making it difficult to accurately estimate random field parameters, which
531 can limit the performance of the Kriging method. In contrast, both the BCS method and the proposed
532 method do not require additional parameters and exhibit good performance, as they can also estimate
533 the spatial variability of soil properties smoothly. As shown in the third case of [Fig. 12\(a\)](#), the proposed
534 method accurately estimates soil connectivity, which is important in soil stratification processes ([T.](#)
535 [Zhao et al., 2023](#)).



536

537

538

539

540

Fig. 12 Comparison of the proposed method's subsurface modeling performance with SOTA methods: (a) Comparison of three synthetic cases in the xy plane with SoFs of 50 m in both x and y directions; (b) Comparison of three synthetic cases in the xz plane with SoFs of 70 m and 1.5 m in the x and z directions, respectively.

541

542

543

544

545

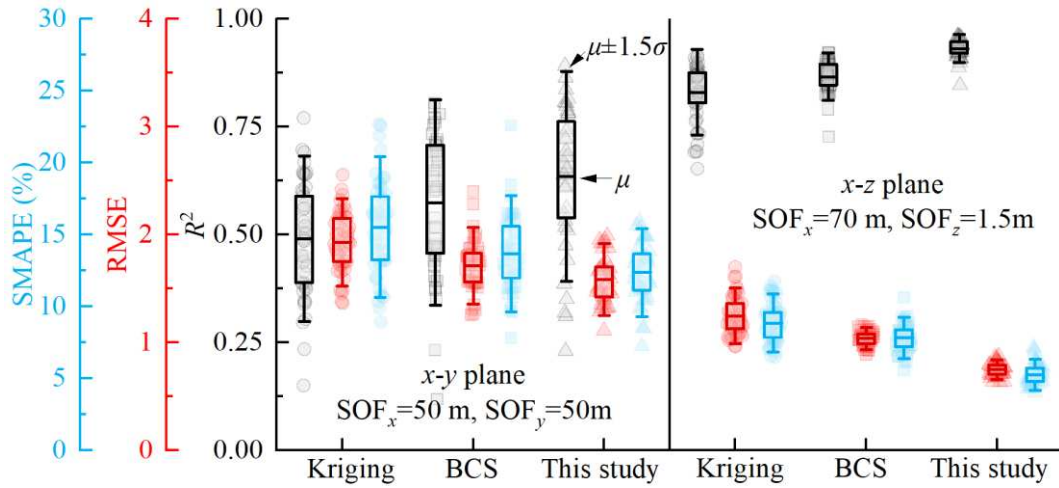
546

547

548

549

As shown in Fig. 13, the performance of the Kriging, BCS, and the proposed methods is summarized. In the xy plane, the evaluation metrics of the Kriging and BCS methods are close, with both capturing the average trend of the field. The proposed method achieves an average R^2 greater than 0.5, and $RMSE$ and $sMAPE$ less than 2 and 15%, respectively, outperforming both Kriging and BCS. In the xz plane, the R^2 values of the BCS and proposed methods are greater than those of the Kriging method, and their $RMSE$ and $sMAPE$ are lower. This indicates the BCS and the proposed methods have significant advantages when more sample points are known. Moreover, the proposed method achieves an R^2 close to 1, and $RMSE$ and $sMAPE$ less than 1 and 7%, respectively, further validating its superior performance.



550

551

Fig. 13 Statistical analysis of R^2 , $RMSE$, and $sMAPE$ for different modeling methods.

552

As shown in Figs. 14(a) and (c), the model uncertainty for the first case in Figs. 12(a) and (b) is

553

displayed. The uncertainty of the proposed method is characterized by the standard deviation of the

554

prediction results from 50 randomly initialized SCNN models, and the detailed calculation process is

555

provided in Section 2.3. It can be observed that the model uncertainty is zero at the measurement

556

locations, and the uncertainty increases as the distance from the measurement location grows.

557

Additionally, the model uncertainty is also related to the spatial variability of the surrounding soil

558

properties. Moreover, the uncertainty in the xy plane is significantly higher than in the xz plane,

559

indicating that an increase in measurement data can substantially reduce the model uncertainty. As

560

shown in Figs. 14(b) and (d), the model uncertainty of the proposed method is compared with that of

561

the BCS method. The BCS method shows high uncertainty at the measurement locations because the

562

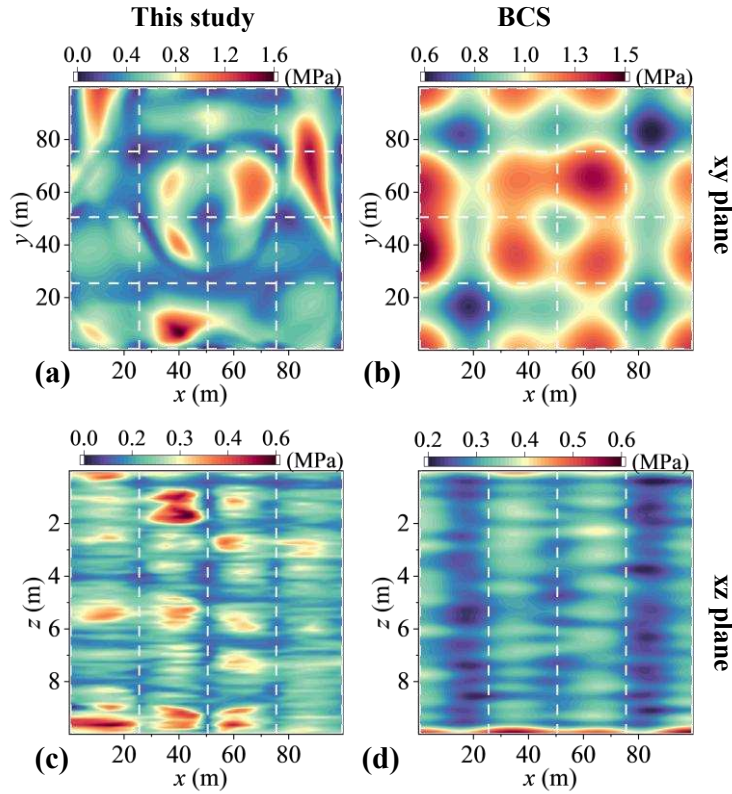
predictions from the BCS method do not match the measured data exactly at the measurement locations,

563

which clearly contradicts the reality. Overall, the uncertainty range of the proposed method is in close

564

agreement with the BCS method, validating the effectiveness of the proposed approach.



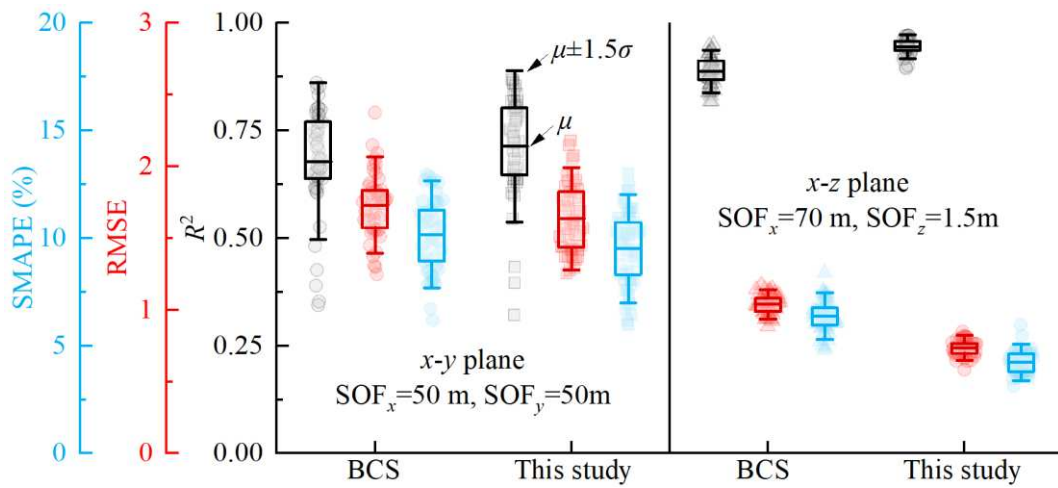
565

566 Fig. 14 Comparison of modeling uncertainty (standard deviation) between the proposed method and
 567 the BCS method in the (a) & (b) xy plane, (c) & (d) xz plane.

568 **5.2 Comparison with BCS Method for 2D Nonstationary Cases**

569 To more comprehensively validate the performance of the proposed method, trend terms are
 570 introduced into the 100 stationary random fields used in Section 5.1. A random linear trend $\alpha \times x$ (or y)
 571 is assumed in the x and y directions, where the random coefficient α ranges from 0.02 to 0.05. Similarly,
 572 a random linear trend $\beta \times z$ is assumed in the z direction, with the random coefficient β ranging from
 573 0.2 to 0.5. This section primarily evaluates the ability of the proposed method to perform subsurface
 574 modeling directly based on non-stationary data. This approach helps reduce the uncertainty caused by
 575 manual detrending and enhances the efficiency of automated modeling. Since Kriging requires
 576 additional detrending when applied to non-stationary random fields, only the proposed method and the
 577 BCS method are compared here. As shown in Fig. 15, the R^2 values of BCS and the proposed method

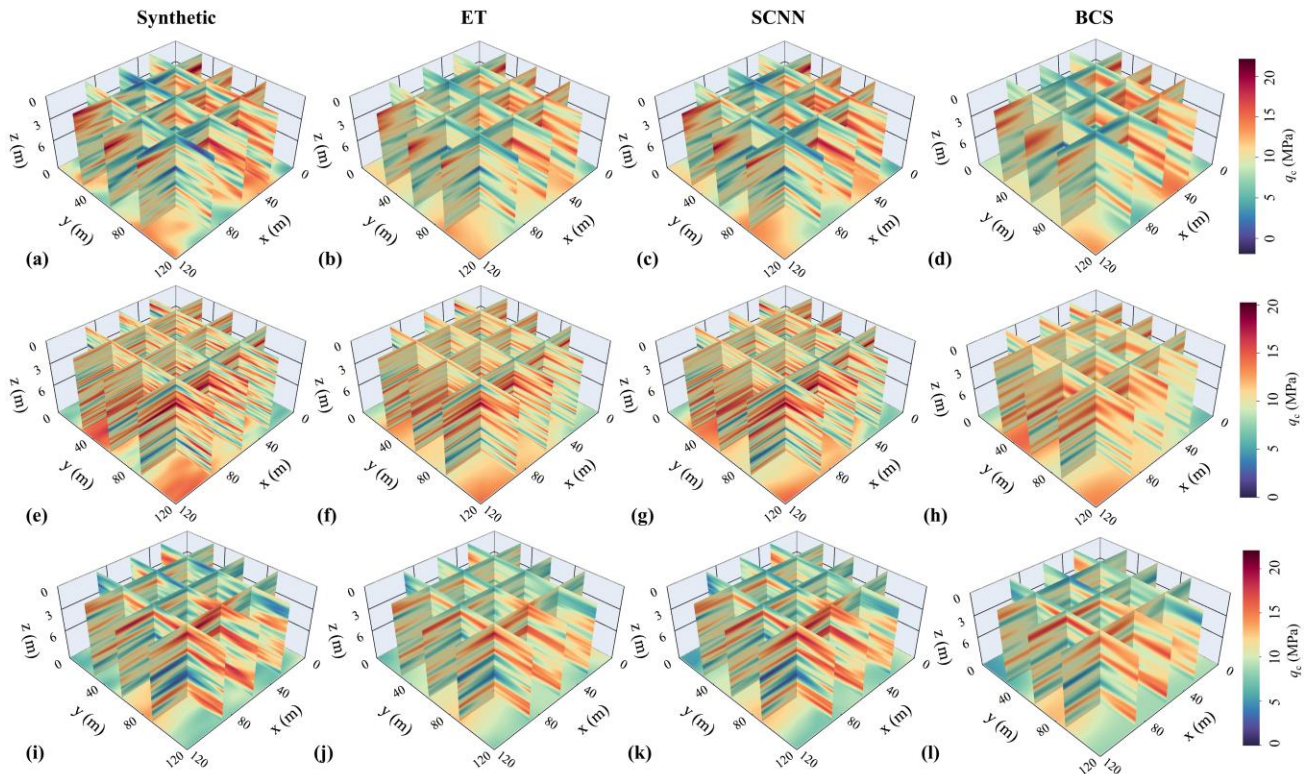
578 fluctuate between 0.4 and 0.85 in the xy -plane. This variation is primarily due to the limited number
 579 of known data points (25 points), which account for only 0.25% of the total soil cells. In the xz -plane,
 580 where the number of known data points increases significantly (500 points), both BCS and the
 581 proposed method perform well, with R^2 values exceeding 0.8. However, the proposed method
 582 demonstrates more consistent performance, with smaller fluctuations in R^2 , $RMSE$, and $SMAPE$
 583 compared to the BCS method. This indicates that the proposed method improves modeling accuracy
 584 and is suitable for both stationary and non-stationary data with mild trends.



585
 586 Fig. 15 Performance comparison between the BCS method and the proposed method in nonstationary
 587 cases.

588 5.3 Performance of the Proposed Method for 3D Cases

589 Three 3D cases are extracted from the synthetic random field samples in Section 3 to validate the
 590 proposed method: Case #1 (SoF_x : 50 m, SoF_y : 50 m, SoF_z : 1 m), Case #2 (SoF_x : 90 m, SoF_y : 50 m,
 591 SoF_z : 0.5 m), and Case #3 (SoF_x : 90 m, SoF_y : 50 m, SoF_z : 1 m), as illustrated in Figs. 16(a), (e), and
 592 (i). For each synthetic site, the measurement points are established at grid points corresponding to x
 593 and y values of 0.5 m, 25.5 m, 50.5 m, 75.5 m, and 99.5 m, resulting in a total of 25 sample paths used
 594 as training data.



595

596 Fig.16 Performance of the proposed method in synthetic 3D cases: (a)-(d) show the synthetic site for
 597 Case #1, along with the modeling results using ET, SCNN and BCS; (e)-(h) present the modeling
 598 results for Case #2; (i)-(l) depict the modeling results for Case #3.

599

It is important to note that the framework proposed is flexible, meaning that after generating the
 600 GCFs, various machine learning models can be employed for subsurface modeling. The primary
 601 subsurface modeling approach employed in this study is the SCNN model. To conduct a more
 602 comprehensive comparison, the ET approach is used due to its strong fitting capabilities. Figs. 16(a)
 603 to (d) illustrate the true distribution of soil properties for synthetic Case #1, the results obtained using
 604 ET, SCNN, and BCS modeling, respectively. Figs. 16(e) to (h) and Figs. 16(i) to (l) present the
 605 prediction results for Case #2 and Case #3, respectively. It is observed that using the proposed method
 606 yields subsurface modeling results consistent with the true distribution, regardless of whether ET or
 607 SCNN is used. However, the results obtained from ET and BCS modeling tend to approximate the

608 average more closely. In contrast, the results from SCNN modeling align more closely with the true
609 distribution. As shown in Table 1, the predictive results of the SCNN model demonstrate a higher R^2
610 compared to the ET and BCS models, further validating the superiority of the SCNN approach. The
611 $RMSE$ and $MAPE$ values for the ET and SCNN models are relatively similar. Therefore, averaging the
612 predictive results of the ET and SCNN models reveals that while the R^2 of the average model decreases
613 relative to the SCNN model, the $RMSE$ and $MAPE$ for the average model achieve lower values in Case
614 #1 and Case #2. Thus, adopting the average model may be a more robust choice, especially when faced
615 with sparse known information, as stacking multiple algorithms often enhances the model's reliability.
616 Furthermore, both the ET and SCNN models take only 2-3 minutes for modeling and prediction, with
617 the ET model being slightly faster than the SCNN model. Therefore, for engineering applications, it is
618 acceptable to integrate multiple algorithms for subsurface modeling to obtain more robust prediction
619 results.

620 Table 1 Comparison of subsurface modeling performance of different models in 3D synthetic cases

	Model	R^2	RMSE (MPa)	MAPE (%)
Case #1 (SoFx=50, SoFy=50, SoFz=1)	BCS	0.571	1.903	17.437
	ET	0.499	1.745	14.513
	SCNN	0.750	1.598	14.909
	Mean	0.683	1.576	13.574
Case #2 (SoFx=90, SoFy=50, SoFz=0.5)	BCS	0.489	1.937	15.133
	ET	0.545	1.350	10.249
	SCNN	0.730	1.350	10.769
	Mean	0.689	1.267	9.738
Case #3 (SoFx=90, SoFy=50, SoFz=1)	BCS	0.473	1.742	17.797
	ET	0.445	1.573	11.975
	SCNN	0.772	1.344	10.469
	Mean	0.682	1.373	10.507

621 Note: 'Mean' refers to the average of the prediction results from the ET and SCNN models.

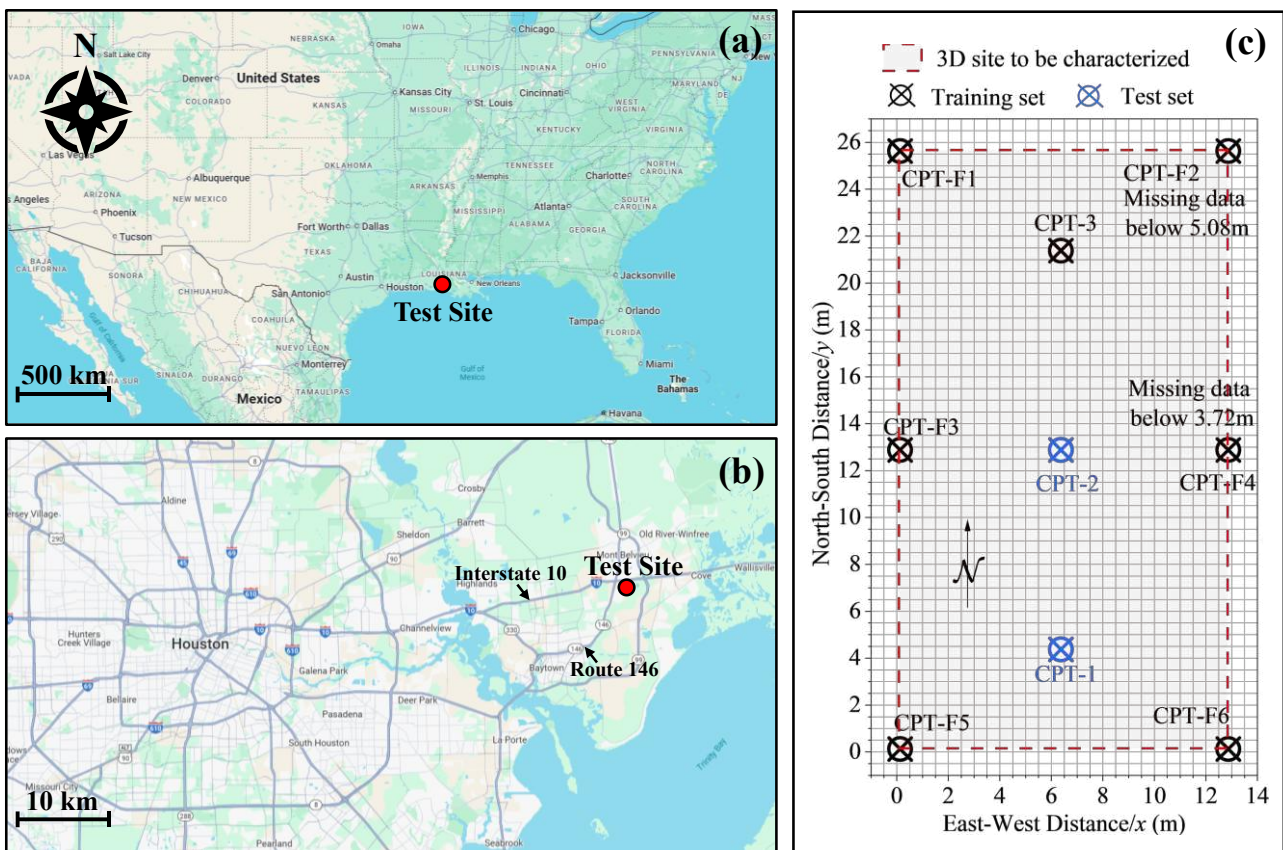
622 **6 Real Data Case Study**

623 **6.1 Baytown, TX, USA**

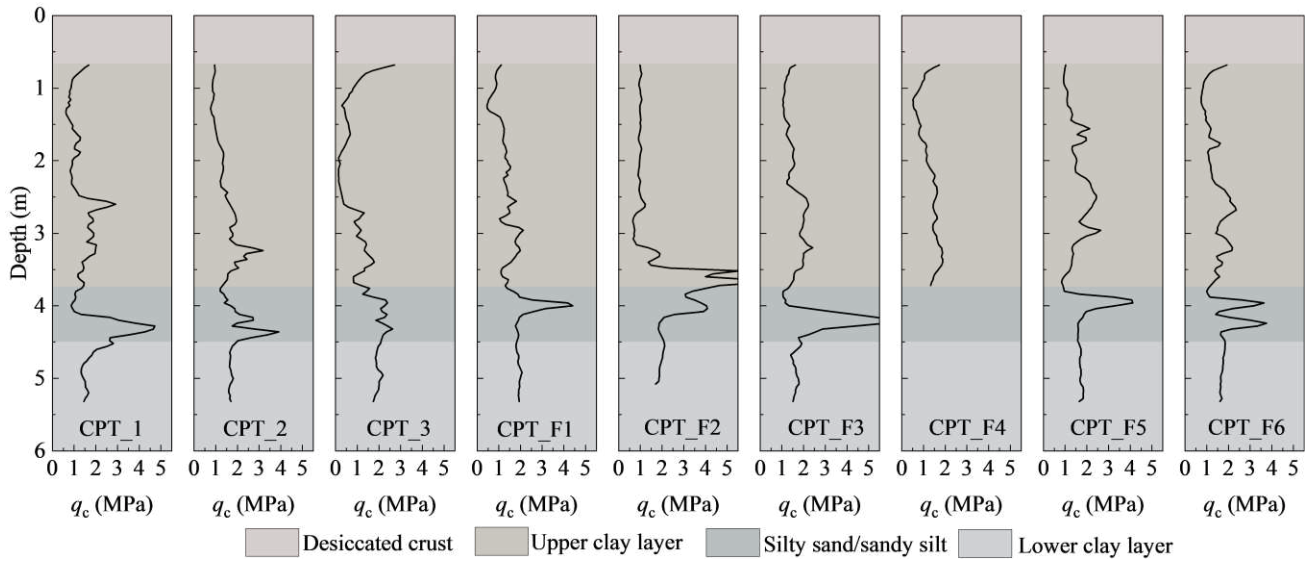
624 As shown in [Fig. 17](#), the rectangular test site is located in the eastern region of the intersection of
625 Interstate-10 and Route-146 in Baytown, Texas ([Stuedlein, 2008](#)). Baytown is situated to the north of
626 Upper Galveston Bay, which is part of the San Jacinto River, with the San Jacinto River flowing south
627 into Galveston Bay and north into Lake Houston. The test site is located in the Quaternary coastal plain
628 of Texas, which forms a band-shaped zone about 110-150 km wide, parallel to the Gulf of Mexico
629 coastline ([Stuedlein et al., 2012a](#)). The soil at the test site consists of the Beaumont Clay Layer, which
630 is composed of brownish-red and brownish-gray clay, occasionally interspersed with fine sand and silt
631 layers. The Beaumont Clay Layer was deposited in the floodplains at the beginning of the Wisconsin
632 glacial period. Additionally, the terrace has been influenced by the long-term lowering of the Gulf of
633 Mexico during the post-depositional glacial period, causing global preconsolidation ([O'Neill, 2014](#)).
634 As the Beaumont Clay Layer dried, it developed a series of slickensides, fissures, and a few joints.

635 A series of in-situ tests, including the standard penetration test (SPT), thin-walled tube sampling,
636 and cone penetration test (CPTu), were conducted at the test site to characterize the subsurface
637 conditions ([Stuedlein et al., 2012a](#)). As shown in [Fig. 18](#), this study focuses on the results from 9 sets
638 of CPTu tests, using the cone tip resistance measurements as an example for subsurface modeling.
639 According to the survey results, the test site was divided into four different soil layers. The very stiff
640 desiccated clay crust extends to an average depth of about 0.66 m from the surface. The second layer,
641 the upper clay layer, extends to an average depth of 3.74 m and consists of medium stiff lean clay.
642 Notably, CPT-3 reveals a 1-meter-thick soft zone at a depth of 1.66 m. Samples recovered from the

643 upper clay layer typically contain many cracks and occasional slickensides. The third layer consists of
 644 loose silty sand/sandy silt (SM/ML), extending from an average depth of 3.74 to 4.5 m. Below 4.5 m,
 645 the lower clay layer is present, consisting of stiff, slightly silty, and fat clay (CH) (Stuedlein et al.,
 646 2012b, 2012a). CPTu data can be directly downloaded from the ISSMGE TC304 database
 647 (<http://140.112.12.21/issmge/tc304.htm>).



648
 649 Fig. 17 Geographical layout of the test site in Baytown, Texas, USA: (a) The location of the test site
 650 in Texas, USA; (b) The location of samples in Baytown; (c) The distribution of the collocated
 651 borehole and CPTs. Map data from Google Earth.



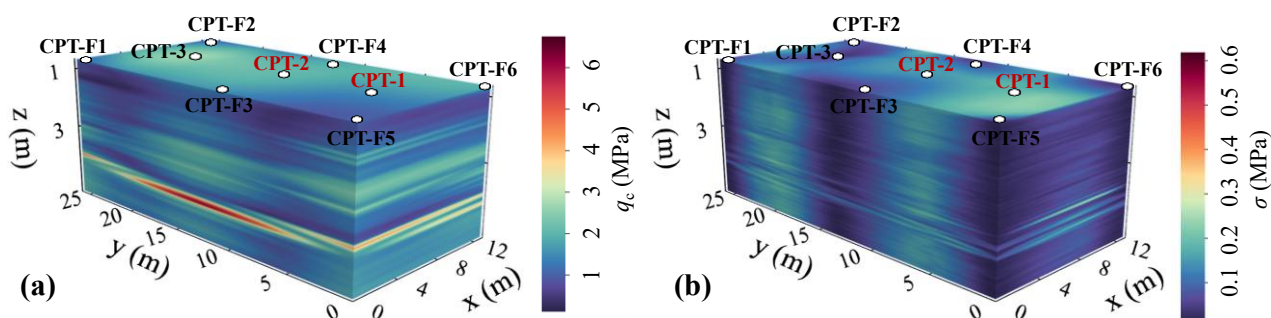
652

653 Fig. 18 Nine CPT measurements and the average stratigraphy of the site (raw data can be accessed
 654 from <http://140.112.12.21/issmge/tc304.htm>).

655 The test site has a length of 13 m in the x (east-west) direction and 26 m in the y (north-south)
 656 direction. In the z direction, the focus is mainly on the clay layer and silty sand/sandy silt layer at
 657 depths between 0.68 m and 5.32 m, consistent with [Stuedlein et al. \(2012a\)](#). The resolution is set at
 658 0.25 m in both the x and y directions, and 0.04 m in the z direction. As shown in [Fig. 17](#), a total of 7
 659 sets of measurements, including CPT_3 and CPT_F1-F6, are used to construct the subsurface model,
 660 while the remaining data from CPT-1 and CPT-2 are used for model validation. As shown in [Fig. 18](#),
 661 some of the measurement paths are incomplete. For example, CPT_F2 is missing data below a depth
 662 of 5.08 m, and CPT_F4 is missing data below a depth of 3.72 m. In engineering practice, measurement
 663 data often have missing sections due to limitations in equipment, operation, and site conditions. It is
 664 worth noting that the proposed method does not require complete measurement data, making it widely
 665 applicable to situations with missing data.

666 As shown in [Fig. 19](#), the subsurface modeling results and their uncertainty based on the SCNN
 667 model are presented. For detailed configuration of the SCNN model and the subsurface modeling code,

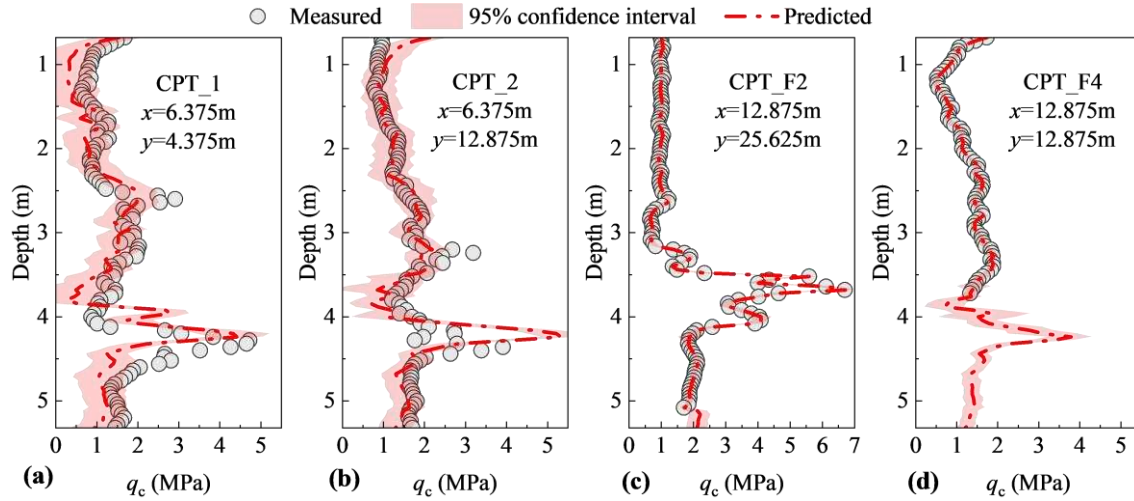
668 please refer to Appendix B. Fig. 19(a) shows the average of the subsurface modeling results from 50
669 SCNN models with different initialized trainable parameters, while Fig. 19(b) shows the standard
670 deviation of the 50 modeling results. It can be observed that there is a thin interlayer at a depth of 4 m,
671 which perfectly aligns with the field exploration results (SM/ML is distributed within the 3.74–4.5 m
672 depth range). Furthermore, the model exhibits small uncertainty at the locations of the training samples.
673 Since CPT_2 and CPT_1 are not involved in the training process, there is higher uncertainty at their
674 respective locations. Additionally, compared to CPT_2, the location of CPT_1 is farther from the
675 training samples, resulting in greater uncertainty in the model’s prediction of the CPT_1 sample path.



676
677 Fig. 19 Subsurface modeling results of the proposed method in the Baytown, TX, USA case: (a)
678 Subsurface modeling results based on 6 CPT measurement data; (b) Uncertainty of subsurface
679 modeling results.

680 As shown in Fig. 20, model prediction performance and uncertainty on the test set are presented.
681 Due to missing sample paths in CPT_F2 and CPT_F4, these missing positions are excluded from the
682 model training and considered as part of the test set. In the upper parts of CPT_F2 and CPT_F4
683 (training data), the model's predictions exactly match the measured data, indicating that the model has
684 successfully learned the high-dimensional nonlinear mapping between GCFs (input features) and
685 corresponding q_c (output) at these positions. Furthermore, the model shows good agreement with the
686 measured data at the CPT_1 and CPT_2 positions, with most of the measurement data included within

687 the 95% confidence interval. Additionally, the model successfully predicted the presence of the silty
 688 sand/sandy silt layer below 3.72m in CPT_F4.

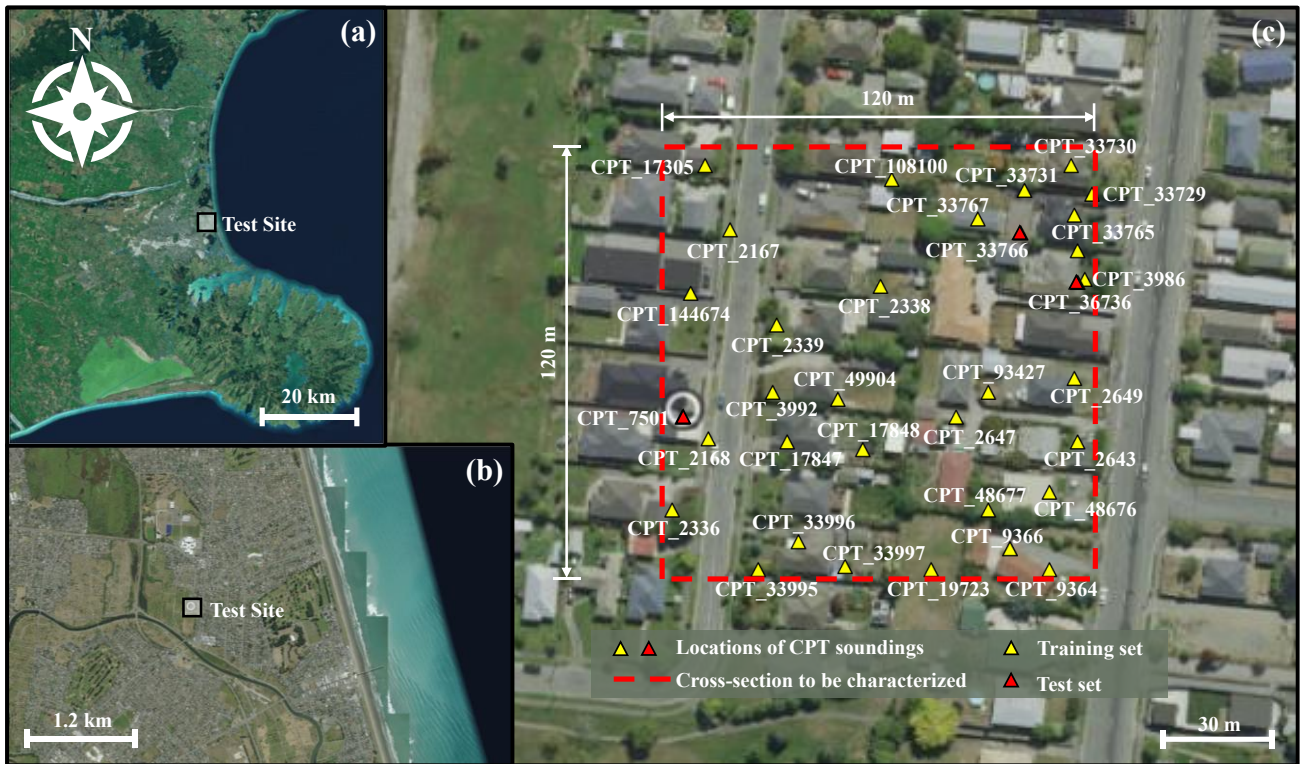


689

690 Fig.20 Prediction results and uncertainty of the proposed method at test locations.

691 **6.2 Christchurch, New Zealand**

692 As shown in Fig. 21, the test site is located in Christchurch, New Zealand, where 34 sets of CPT
 693 data are collected from a 120 m × 120 m square site. The measurement data can be accessed directly
 694 from the New Zealand Geotechnical Database (NZGD) (NZGD, 2023). According to the Robertson
 695 method (Robertson, 1990; Robertson and Wide, 1998) for soil classification, the soil behavior type
 696 index (I_c) generally ranges from 1.1 to 2.6. The surface layer is primarily composed of dense sand to
 697 gravelly sand, while the layer beneath is mainly clean sand to silty sand, occasionally interspersed with
 698 silty sand to sandy silt. It is important to note that there are significant data gaps near the surface in the
 699 34 CPT measurements. To better validate the proposed method, test data from depths of 5-15m below
 700 the surface are collected for subsurface modeling. The site resolution is set at 1 m in both the x and y
 701 directions, and 0.1 m in the z direction.

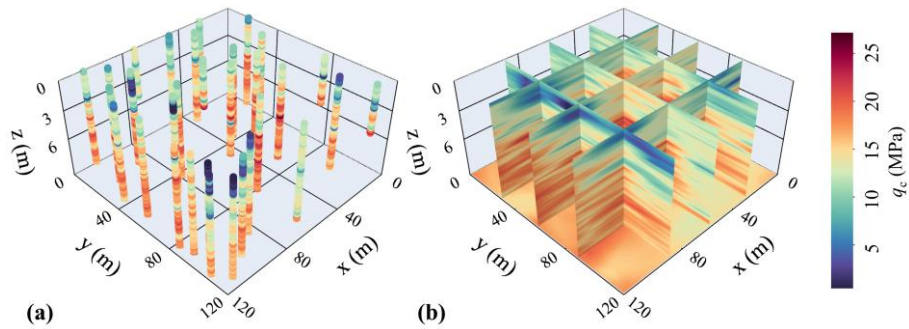


702

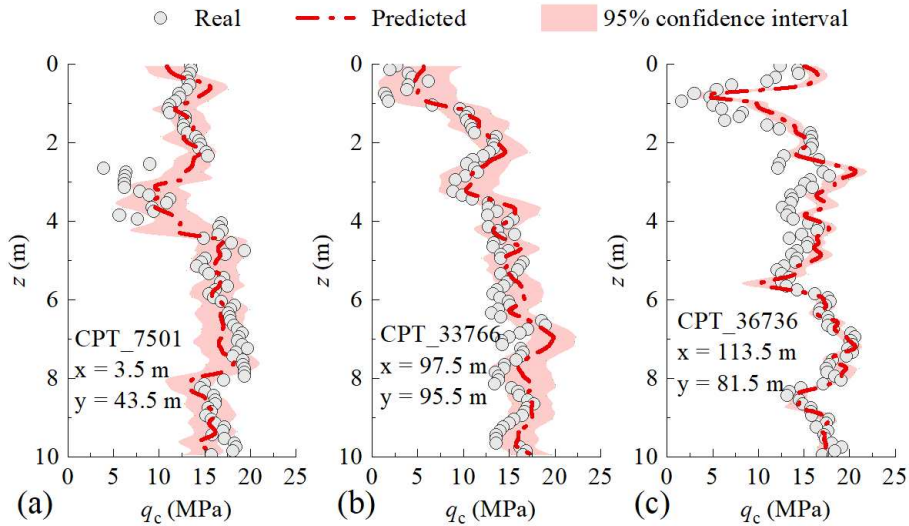
703 Fig.21 Geographical layout of 34 CPT surveys in Christchurch, New Zealand: (a) The location of the
 704 test site in New Zealand; (b) The location of samples in Christchurch; (c) The distribution of the
 705 collocated CPTs. (Note: CPT codes used are consistent with the New Zealand Geotechnical
 706 Database)(NZGD, 2023).

707 Fig. 22(a) shows the CPT sample paths used for subsurface modeling in the Christchurch case,
 708 where some sample paths are incomplete. It is noteworthy that the proposed method does not require
 709 complete data at sampling locations, allowing it to be applied with more flexibility in practice.
 710 CPT_7501, CPT_33766, and CPT_36736 are reserved for the model testing, while the remaining 31
 711 CPT soundings are used for model training. Fig. 22(b) presents the average subsurface modeling results
 712 of the SCNN and ET models, where the q_c values are generally low at depths above 3 m and increase
 713 with depth in the 3-10 m range, indicating the trend in soil distribution. Figs. 23(a)-(c) show the
 714 predicted results of the model at the three test locations: CPT_7501, CPT_33766, and CPT_36736,
 715 along with their 95% confidence intervals. It can be observed that the predicted sample paths align

716 well with the measured data, with most test data falling within the 95% confidence interval.
 717 Additionally, the model exhibits greater uncertainty at CPT_7501 and CPT_33766 due to the limited
 718 number of boreholes in their vicinity. In contrast, the location of CPT_36736 is in close proximity to
 719 CPT_3986 used for subsurface modeling, resulting in lower uncertainty at this position. This further
 720 validates the rationale behind the proposed method for assessing subsurface modeling uncertainty.



721
 722 Fig.22 Performance of the proposed method in real 3D cases: (a) represents the measured CPT
 723 sample paths from the New Zealand case; (b) shows the average modeling results of ET and SCNN
 724 based on 31 training data.



725
 726 Fig.23 The subsurface modeling results and uncertainties at three test locations.

727 7 Conclusions

728 This study proposes an enhanced data-driven framework that uses random field theory to recover
 729 subsurface geotechnical properties in the presence of sparse in-situ test data. Additionally, statistical

730 analyses are conducted to reduce the computational complexity of the proposed method and eliminate
731 additional parameters. Finally, the superiority of the proposed method is validated through a series of
732 synthetic cases and two real cases. Based on the findings, the following conclusions can be drawn:

733 (1) The proposed subsurface modeling framework is capable of integrating random field theory
734 into data-driven models through geotechnical correlation fields, providing a variety of basis functions
735 for subsurface modeling. Validation through 2D and 3D synthetic cases, as well as two real cases,
736 shows that the method generally exhibits higher R^2 and lower $RMSE$ and $MAPE$ compared to the
737 alternative Kriging and BCS methods.

738 (2) The strategy of first performing sparse sampling and then projecting into the principal
739 component space effectively reduces the computational complexity of PCA operations in the process
740 of obtaining GCFs. This approach also avoids the challenges associated with storing and processing
741 large correlation matrices for large or high-resolution sites.

742 (3) The influence of random field parameters (SoF and ACF) on subsurface modeling results is
743 investigated. Since geotechnical measurement data are generally sparse in the horizontal direction and
744 denser in the vertical direction, using larger horizontal SoF (greater than 2 times the borehole spacing)
745 and smaller vertical SoF (site vertical resolution) to generate GCFs can significantly improve modeling
746 accuracy. Additionally, in cases of sparse measurement data, the type of ACFs has a negligible
747 influence on the modeling results.

748 (4) The framework proposed is flexible, allowing for the application of different machine learning
749 models for subsurface modeling after generating GCFs. Validation against a series of 3D sites revealed
750 that the modeling results of the proposed SCNN model exhibit a higher R^2 compared to an extreme

751 random tree model. Furthermore, the average predictions from both the extreme random tree and
752 shortcut-connected neural network models demonstrated lower *RMSE* and *MAPE*. Therefore, it is
753 recommended to employ a stacking strategy using multiple algorithms to enhance the robustness of
754 underground modeling.

755 It should be noted that when subsurface conditions involve weak interlayers or soil layers with
756 significant differences in properties, the spatial distribution of the subsurface stratigraphic boundaries
757 can be predefined. Then, the proposed method can be used to model the spatial variability of soil
758 properties within each layer. Additionally, measurement data for weak interlayers often constitute only
759 a small portion of the total data, leading to potential prediction errors due to data imbalance. Further
760 exploration of improvement strategies for data-driven subsurface modeling methods under conditions
761 of data imbalance is needed.

762 **CRedit authorship contribution statement**

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

768 **Acknowledgments**

769 This study was supported by the National Natural Science Foundation of China (Grant No.
770 52378330), the Natural Science Foundation of Sichuan Province (Grant No. 2023NSFSC0391), the
771 111 Project (Grant No. B21011), the Postgraduate Research & Practice Innovation Program of Jiangsu

772 Province (Grant No. SJCX23_0084), and the SEU Innovation Capability Enhancement Plan for
 773 Doctoral Students (Grant No. CXJH_SEU 24182).

774 Appendix A.

775 Table A1. Frequently Used Autocorrelation Functions (ACFs) - Adapted from Cami et al. (2020)

Model	Autocorrelation function $\rho_{i,j}$	Frequency of usage
Single exponential (SNX)	$\exp\left[-2\left(\frac{ \tau_{i,j}^x }{SOF_x} + \frac{ \tau_{i,j}^y }{SOF_y}\right)\right]$	47 %
Spherical (SPH)	$\begin{cases} \left[1 - \frac{9 \tau_{i,j}^x }{8SOF_x} + \frac{27}{128}\left(\frac{ \tau_{i,j}^x }{SOF_x}\right)^3\right] \left[1 - \frac{9 \tau_{i,j}^y }{8SOF_y} + \frac{27}{128}\left(\frac{ \tau_{i,j}^y }{SOF_y}\right)^3\right] & \tau_{i,j}^x \leq \frac{4}{3}SOF_x \text{ and } \tau_{i,j}^y \leq \frac{4}{3}SOF_y \\ 0 & \text{otherwise} \end{cases}$	15 %
Squared exponential (SQX)	$\exp\left[-\pi\left(\frac{\tau_{i,j}^{x^2}}{SOF_x^2} + \frac{\tau_{i,j}^{y^2}}{SOF_y^2}\right)\right]$	15 %
Cosine exponential (CSX)	$\cos\left(\frac{ \tau_{i,j}^x }{SOF_x}\right) \cos\left(\frac{ \tau_{i,j}^y }{SOF_y}\right) \exp\left(-\left(\frac{ \tau_{i,j}^x }{SOF_x} + \frac{ \tau_{i,j}^y }{SOF_y}\right)\right)$	10 %
Binary noise (BIN)	$\begin{cases} \left(1 - \frac{ \tau_{i,j}^x }{SOF_x}\right) \left(1 - \frac{ \tau_{i,j}^y }{SOF_y}\right) & \tau_x \leq SOF_x \text{ and } \tau_y \leq SOF_y \\ 0 & \text{otherwise} \end{cases}$	9 %
Second-order Markov (SOM)	$\left(1 + 4\frac{ \tau_{i,j}^x }{SOF_x}\right) \left(1 + 4\frac{ \tau_{i,j}^y }{SOF_y}\right) \exp\left[-4\left(\frac{ \tau_{i,j}^x }{SOF_x} + \frac{ \tau_{i,j}^y }{SOF_y}\right)\right]$	4 %
Third-order Markov (TOM)	$\left(1 + \frac{16 \tau_{i,j}^x }{3SOF_x} + \frac{256}{27}\left(\frac{ \tau_{i,j}^x }{SOF_x}\right)^2\right) \left(1 + \frac{16 \tau_{i,j}^y }{3SOF_y} + \frac{256}{27}\left(\frac{ \tau_{i,j}^y }{SOF_y}\right)^2\right) \exp\left(-\frac{16}{3}\left(\frac{ \tau_{i,j}^x }{SOF_x} + \frac{ \tau_{i,j}^y }{SOF_y}\right)\right)$	New

776 **Note:** $\rho_{i,j}$ represents the correlation between soil cells i and j . $\tau_{i,j}^x$ and $\tau_{i,j}^y$ are the spacing of soil cell i and j in the x
 777 and y directions. SOF_x and SOF_y are the scales of fluctuation in the x and y directions, respectively.

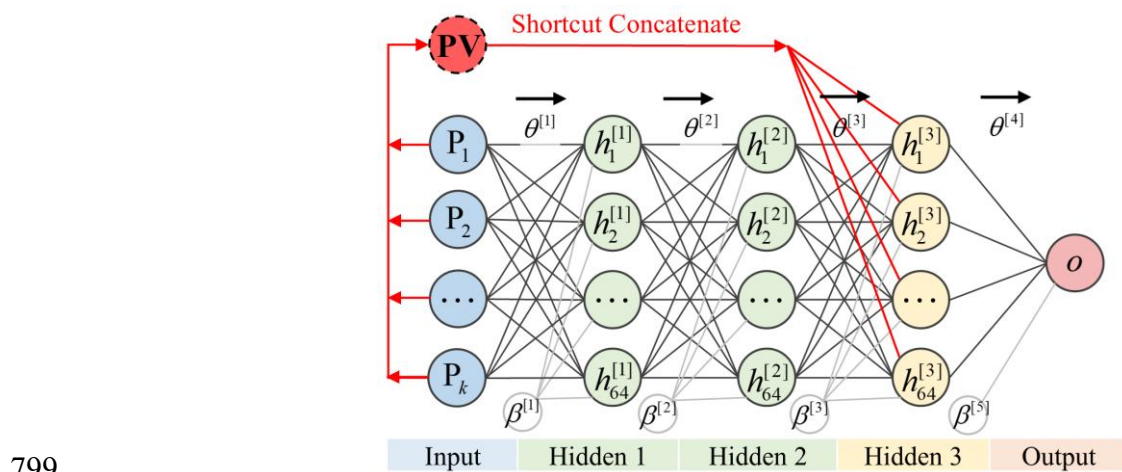
778 Appendix B.

779 The code used in this study can be found in <https://github.com/Data-Driven-RFT/Sparse-Learning>,
 780 which includes: ① a case for generating GCFs, and ② the complete subsurface modeling process
 781 based on the Baytown, TX, USA case.

782 Recently, ensemble learning algorithms with powerful nonlinear regression capabilities have been
 783 widely used in geotechnical engineering (Xie et al., 2024), such as Random Forest (RF), Extra Trees

784 (ET), and Gradient Boosting (GB). Both ET and RF algorithms are based on the BAGGING (Bootstrap
785 Aggregating) ensemble technique. The RF algorithm constructs multiple subsets using random
786 sampling with replacement, and each subset is used to build a corresponding regression tree. In contrast,
787 the ET algorithm trains each regression tree on the entire dataset, which reduces prediction bias.
788 Additionally, the ET algorithm introduces extra randomness in the tree-building process by randomly
789 selecting split thresholds to compute the split points for each variable, and then selecting the best split
790 point based on the scoring criterion to reduce prediction bias. Compared to RF, when the number of
791 regression trees is sufficiently large, ET can generate continuous and smooth prediction results.
792 Therefore, the ET algorithm is better suited for modeling continuous geotechnical sites or random
793 fields. A detailed introduction to the ET algorithm can be found in [Simm et al. \(2014\)](#). In this study,
794 the ET model is constructed using the Python-based Scikit-learn v1.1.3 library ([Pedregosa et al., 2018](#)).

795 As shown in [Fig. A1](#), this study uses a shortcut-connected neural network model (SCNN) for
796 subsurface modeling. The SCNN integrates input features with high-level features extracted by hidden
797 layers, thus mitigating the risk of gradient vanishing and explosion, and improving training efficiency
798 and model performance. The SCNN model formulation is as follows:



799 Fig. A1 The SCNN model architecture for subsurface modeling (Note: P_1 to P_k correspond to the k
800

801 principal components in the \mathbf{PV} vector.)

$$802 \quad \mathbf{h}^1 = \text{ReLU}(\mathbf{PV}, \boldsymbol{\theta}^1, \boldsymbol{\beta}^1) \quad (\text{A1})$$

$$803 \quad \mathbf{h}^2 = \text{ReLU}(\mathbf{h}^1, \boldsymbol{\theta}^2, \boldsymbol{\beta}^2) \quad (\text{A2})$$

$$804 \quad \mathbf{h}^3 = \text{ReLU}(\text{concatenate}(\mathbf{PV}, \mathbf{h}^2), \boldsymbol{\theta}^3, \boldsymbol{\beta}^3) \quad (\text{A3})$$

$$805 \quad \mathbf{o} = \text{Linear}(\mathbf{h}^3, \boldsymbol{\theta}^4, \boldsymbol{\beta}^4) \quad (\text{A4})$$

806 where \mathbf{h} represents the hidden layer feature vector of the neural network. θ and β are the feature weights
807 and bias terms for each layer, respectively, with no trainable θ and β parameters in the input layer. The
808 concatenation operation in the third hidden layer directly combines the input feature \mathbf{PV} with the
809 feature vector \mathbf{h}^2 from the second hidden layer. The output of the model is denoted as \mathbf{o} . The model
810 consistently uses the ReLU activation function due to its computational simplicity and rapid
811 convergence rate (P. Zhang et al., 2022). Model training is based on the Nadam optimizer, an extension
812 of the Adam optimizer that incorporates Nesterov momentum and RMSprop. The hyperparameters for
813 the model are determined using grid search. The number of neurons in the hidden layers is set to 64,
814 the learning rate is 0.001, and the number of iterations is 500. In this study, the SCNN model is
815 constructed using the Python-based Tensorflow v2.8.0 library (Abadi et al., 2016).

816 **References**

- 817 Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G.,
818 Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker,
819 P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: a system for
820 large-scale machine learning, in: Proceedings of the 12th USENIX Conference on Operating
821 Systems Design and Implementation, OSDI'16. USENIX Association, USA, pp. 265–283.
- 822 Cami, B., Javankhoshdel, S., Phoon, K.-K., Ching, J., 2020. Scale of Fluctuation for Spatially Varying
823 Soils: Estimation Methods and Values. ASCE-ASME J. Risk Uncertain. Eng. Syst. Part Civ.
824 Eng. 6, 03120002. <https://doi.org/10.1061/AJRUA6.0001083>
- 825 Chen, W., Ding, J., Shi, C., Wang, T., Connolly, D.P., 2024. Geotechnical correlation field-informed
826 and data-driven prediction of spatially varying geotechnical properties. Comput. Geotech. 171,
827 106407. <https://doi.org/10.1016/j.compgeo.2024.106407>

828 Chen, W., Ding, J., Wang, T., Connolly, D.P., Wan, X., 2023. Soil property recovery from incomplete
829 in-situ geotechnical test data using a hybrid deep generative framework. *Eng. Geol.* 326,
830 107332. <https://doi.org/10.1016/j.enggeo.2023.107332>

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

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

837 Gong, W., Zhao, C., Juang, C.H., Zhang, Y., Tang, H., Lu, Y., 2021. Coupled characterization of
838 stratigraphic and geo-properties uncertainties – A conditional random field approach. *Eng.*
839 *Geol.* 294, 106348. <https://doi.org/10.1016/j.enggeo.2021.106348>

840 Guan, Z., Wang, Y., Cao, Z., Hong, Y., 2020. Smart sampling strategy for investigating spatial
841 distribution of subsurface shallow gas pressure in Hangzhou Bay area of China. *Eng. Geol.*
842 274, 105711. <https://doi.org/10.1016/j.enggeo.2020.105711>

843 Guan, Z., Wang, Y., Phoon, K.-K., 2024a. Data-driven geotechnical site recognition using machine
844 learning and sparse representation. *Eng. Geol.* 107893.
845 <https://doi.org/10.1016/j.enggeo.2024.107893>

846 Guan, Z., Wang, Y., Phoon, K.-K., 2024b. Dictionary Learning of Spatial Variability at a Specific Site
847 Using Data from Other Sites. *J. Geotech. Geoenvironmental Eng.* 150, 04024072.
848 <https://doi.org/10.1061/JGGEFK.GTENG-12408>

849 Hu, Y., Wang, Z.Z., Guo, X., Kek, H.Y., Ku, T., Goh, S.H., Leung, C.F., Tan, E., Zhang, Y., 2024.
850 Three-dimensional reconstruction of subsurface stratigraphy using machine learning with
851 neighborhood aggregation. *Eng. Geol.* 337, 107588.
852 <https://doi.org/10.1016/j.enggeo.2024.107588>

853 Jiang, Q.-H., Zhang, J.-Z., Zhang, D.-M., Huang, H.-W., 2024a. Simulation of geological uncertainty
854 based on improved three-dimensional coupled Markov chain model. *Eng. Geol.* 340, 107647.
855 <https://doi.org/10.1016/j.enggeo.2024.107647>

856 Jiang, Q.-H., Zhang, J.-Z., Zhang, D.-M., Huang, H.-W., Shi, J.-K., Li, Z.-L., 2024b. Influence of
857 geological uncertainty on longitudinal deformation of tunnel based on improved coupled
858 Markov chain. *Eng. Geol.* 337, 107564. <https://doi.org/10.1016/j.enggeo.2024.107564>

859 Liu, H., Luo, Q., El Naggar, M.H., Zhang, L., Wang, T., 2023. Centrifuge Modeling of Stability of
860 Embankment on Soft Soil Improved by Rigid Columns. *J. Geotech. Geoenvironmental Eng.*
861 149, 04023069. <https://doi.org/10.1061/JGGEFK.GTENG-11314>

862 Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions, in: *Proceedings*
863 *of the 31st International Conference on Neural Information Processing Systems, NIPS'17.*
864 *Curran Associates Inc., Red Hook, NY, USA, pp. 4768–4777.*

865 Lyu, B., Hu, Y., Wang, Y., 2023. Data-Driven Development of Three-Dimensional Subsurface Models
866 from Sparse Measurements Using Bayesian Compressive Sampling: A Benchmarking Study.
867 *ASCE-ASME J. Risk Uncertain. Eng. Syst. Part Civ. Eng.* 9, 04023010.
868 <https://doi.org/10.1061/AJRUA6.RUENG-935>

869 Lyu, B., Wang, Y., Shi, C., 2024. Multi-scale generative adversarial networks (GAN) for generation of

870 three-dimensional subsurface geological models from limited boreholes and prior geological
871 knowledge. *Comput. Geotech.* 170, 106336. <https://doi.org/10.1016/j.compgeo.2024.106336>

872 Müller, S., Schüler, L., Zech, A., Heße, F., 2022. GStools v1.3: a toolbox for geostatistical modelling
873 in Python. *Geosci. Model Dev.* 15, 3161–3182. <https://doi.org/10.5194/gmd-15-3161-2022>

874 Nag, P., Sun, Y., Reich, B.J., 2023. Spatio-temporal DeepKriging for interpolation and probabilistic
875 forecasting. *Spat. Stat.* 57, 100773. <https://doi.org/10.1016/j.spasta.2023.100773>

876 NZGD, 2023. New Zealand Geotechnical Database (NZGD). Available at <https://www.nzgd.org.nz>.

877 O’Neill, M.W., 2014. National Geotechnical Experimentation Site: University of Houston. *Natl.*
878 *Geotech. Exp. Sites* 72–101. <https://doi.org/10.1061/9780784404843.ch04>

879 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller,
880 A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos,
881 A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2018. Scikit-learn: Machine
882 Learning in Python. <https://doi.org/10.48550/arXiv.1201.0490>

883 Phoon, K.-K., Cao, Z.-J., Ji, J., Leung, Y.F., Najjar, S., Shuku, T., Tang, C., Yin, Z.-Y., Ikumasa, Y.,
884 Ching, J., 2022. Geotechnical uncertainty, modeling, and decision making. *Soils Found.* 62,
885 101189. <https://doi.org/10.1016/j.sandf.2022.101189>

886 Phoon, K.-K., Quek, S.-T., An, P., 2003. Identification of Statistically Homogeneous Soil Layers Using
887 Modified Bartlett Statistics. *J. Geotech. Geoenvironmental Eng.* 129, 649–659.
888 [https://doi.org/10.1061/\(ASCE\)1090-0241\(2003\)129:7\(649\)](https://doi.org/10.1061/(ASCE)1090-0241(2003)129:7(649))

889 Qi, X., Wang, H., Chu, J., Chiam, K., 2022. Effect of autocorrelation function model on spatial
890 prediction of geological interfaces. *Can. Geotech. J.* 59, 583–600. <https://doi.org/10.1139/cgj-2020-0644>

891

892 Qiu, Y., Zhang, N., Yin, Z., Wang, Y., Xu, C., Zhang, P., 2024. Novel multi-spatial receptive field
893 (MSRF) XGBoost method for predicting geological cross-section based on sparse borehole
894 data. *Eng. Geol.* 338, 107604. <https://doi.org/10.1016/j.enggeo.2024.107604>

895 Robertson, P.K., 1990. Soil classification using the cone penetration test. *Can. Geotech. J.* 27, 151–
896 158. <https://doi.org/10.1139/t90-014>

897 Robertson, P.K., Wride, C. (Fear), 1998. Evaluating cyclic liquefaction potential using the cone
898 penetration test. *Can. Geotech. J.* 35, 442–459. <https://doi.org/10.1139/t98-017>

899 Shi, C., Wang, Y., 2023. Data-driven spatio-temporal analysis of consolidation for rapid reclamation.
900 *Géotechnique* 1–21. <https://doi.org/10.1680/jgeot.22.00016>

901 Shi, C., Wang, Y., 2022. Data-driven construction of Three-dimensional subsurface geological models
902 from limited Site-specific boreholes and prior geological knowledge for underground digital
903 twin. *Tunn. Undergr. Space Technol.* 126, 104493. <https://doi.org/10.1016/j.tust.2022.104493>

904 Shi, C., Wang, Y., 2021. Non-parametric machine learning methods for interpolation of spatially
905 varying non-stationary and non-Gaussian geotechnical properties. *Geosci. Front.* 12, 339–350.
906 <https://doi.org/10.1016/j.gsf.2020.01.011>

907 Shi, C., Wang, Y., Kamchoom, V., 2023. Data-driven multi-stage sampling strategy for a three-
908 dimensional geological domain using weighted centroidal voronoi tessellation and IC-
909 XGBoost3D. *Eng. Geol.* 325, 107301. <https://doi.org/10.1016/j.enggeo.2023.107301>

910 Simm, J., Magrans De Abril, I., Sugiyama, M., 2014. Tree-Based Ensemble Multi-Task Learning
911 Method for Classification and Regression. *IEICE Trans. Inf. Syst.* E97.D, 1677–1681.

912 <https://doi.org/10.1587/transinf.E97.D.1677>

913 Stuedlein, A.W., 2008. Bearing capacity and displacement of spread footings on aggregate pier
914 reinforced clay (Ph.D.). University of Washington.

915 Stuedlein, A.W., Kramer, S.L., Arduino, P., Holtz, R.D., 2012a. Geotechnical characterization and
916 random field modeling of desiccated clay. *J. Geotech. Geoenvironmental Eng.* 138, 1301–1313.
917 [https://doi.org/10.1061/\(ASCE\)GT.1943-5606.0000723](https://doi.org/10.1061/(ASCE)GT.1943-5606.0000723)

918 Stuedlein, A.W., Kramer, S.L., Arduino, P., Holtz, R.D., 2012b. Reliability of Spread Footing
919 Performance in Desiccated Clay. *J. Geotech. Geoenvironmental Eng.* 138, 1314–1325.
920 [https://doi.org/10.1061/\(ASCE\)GT.1943-5606.0000706](https://doi.org/10.1061/(ASCE)GT.1943-5606.0000706)

921 Wang, T., Chen, W., Li, T., Connolly, D.P., Luo, Q., Liu, K., Zhang, W., 2023. Surrogate-assisted
922 uncertainty modeling of embankment settlement. *Comput. Geotech.* 159, 105498.
923 <https://doi.org/10.1016/j.compgeo.2023.105498>

924 Wang, Y., Hu, Y., Zhao, T., 2020. Cone penetration test (CPT)-based subsurface soil classification and
925 zonation in two-dimensional vertical cross section using Bayesian compressive sampling. *Can.*
926 *Geotech. J.* 57, 947–958. <https://doi.org/10.1139/cgj-2019-0131>

927 Wang, Y., Shi, C., 2023. Data-driven analysis of soil consolidation with prefabricated vertical drains
928 considering stratigraphic variation. *Comput. Geotech.* 161, 105569.
929 <https://doi.org/10.1016/j.compgeo.2023.105569>

930 Wang, Z.Z., Hu, Y., Guo, X., He, X., Kek, H.Y., Ku, T., Goh, S.H., Leung, C.F., 2023. Predicting
931 geological interfaces using stacking ensemble learning with multi-scale features. *Can. Geotech.*
932 *J.* 60, 1036–1054. <https://doi.org/10.1139/cgj-2022-0365>

933 Wu, X., Ma, J., Si, X., Bi, Z., Yang, J., Gao, H., Xie, D., Guo, Z., Zhang, J., 2023. Sensing prior
934 constraints in deep neural networks for solving exploration geophysical problems. *Proc. Natl.*
935 *Acad. Sci.* 120, e2219573120. <https://doi.org/10.1073/pnas.2219573120>

936 Xiao, T., Li, D.-Q., Cao, Z.-J., Zhang, L.-M., 2018. CPT-Based Probabilistic Characterization of Three-
937 Dimensional Spatial Variability Using MLE. *J. Geotech. Geoenvironmental Eng.* 144,
938 04018023. [https://doi.org/10.1061/\(ASCE\)GT.1943-5606.0001875](https://doi.org/10.1061/(ASCE)GT.1943-5606.0001875)

939 Xie, J., Huang, J., Lu, J., Burton, G.J., Zeng, C., Wang, Y., 2022a. Development of two-dimensional
940 ground models by combining geotechnical and geophysical data. *Eng. Geol.* 300, 106579.
941 <https://doi.org/10.1016/j.enggeo.2022.106579>

942 Xie, J., Huang, J., Zeng, C., Huang, S., Burton, G.J., 2022b. A generic framework for geotechnical
943 subsurface modeling with machine learning. *J. Rock Mech. Geotech. Eng.* 14, 1366–1379.
944 <https://doi.org/10.1016/j.jrmge.2022.08.001>

945 Xie, J., Zeng, C., Huang, J., Zhang, Y., Lu, J., 2024. A back analysis scheme for refined soil
946 stratification based on integrating borehole and CPT data. *Geosci. Front.* 15, 101688.
947 <https://doi.org/10.1016/j.gsf.2023.101688>

948 Yan, W., Shen, P., Zhou, W.-H., Ma, G., 2023. A rigorous random field-based framework for 3D
949 stratigraphic uncertainty modelling. *Eng. Geol.* 323, 107235.
950 <https://doi.org/10.1016/j.enggeo.2023.107235>

951 Yang, H.-Q., Chu, J., Qi, X., Wu, S., Chiam, K., 2023a. Stochastic simulation of geological cross-
952 sections from boreholes: A random field approach with Markov Chain Monte Carlo method.
953 *Eng. Geol.* 327, 107356. <https://doi.org/10.1016/j.enggeo.2023.107356>

- 954 Yang, H.-Q., Chu, J., Qi, X., Wu, S., Chiam, K., 2023b. Bayesian evidential learning of soil-rock
955 interface identification using boreholes. *Comput. Geotech.* 162, 105638.
956 <https://doi.org/10.1016/j.compgeo.2023.105638>
- 957 Yang, Y., Wang, P., Brandenberg, S.J., 2022. An algorithm for generating spatially correlated random
958 fields using Cholesky decomposition and ordinary kriging. *Comput. Geotech.* 147, 104783.
959 <https://doi.org/10.1016/j.compgeo.2022.104783>
- 960 Yang, Z., Ching, J., 2021. Simulation of three-dimensional random field conditioning on incomplete
961 site data. *Eng. Geol.* 281, 105987. <https://doi.org/10.1016/j.enggeo.2020.105987>
- 962 Yang, Z., Li, X., Qi, X., 2022. Efficient simulation of multivariate three-dimensional cross-correlated
963 random fields conditioning on non-lattice measurement data. *Comput. Methods Appl. Mech.*
964 *Eng.* 388, 114208. <https://doi.org/10.1016/j.cma.2021.114208>
- 965 Zhang, J.-Z., Phoon, K.K., Zhang, D.-M., Huang, H.-W., Tang, C., 2021. Novel approach to estimate
966 vertical scale of fluctuation based on CPT data using convolutional neural networks. *Eng. Geol.*
967 294, 106342. <https://doi.org/10.1016/j.enggeo.2021.106342>
- 968 Zhang, J.-Z., Zhang, D.-M., Huang, H.-W., Phoon, K.K., Tang, C., Li, G., 2022. Hybrid machine
969 learning model with random field and limited CPT data to quantify horizontal scale of
970 fluctuation of soil spatial variability. *Acta Geotech.* 17, 1129–1145.
971 <https://doi.org/10.1007/s11440-021-01360-0>
- 972 Zhang, P., Yin, Z.-Y., Jin, Y.-F., 2022. Bayesian neural network-based uncertainty modelling:
973 application to soil compressibility and undrained shear strength prediction. *Can. Geotech. J.*
974 59, 546–557. <https://doi.org/10.1139/cgj-2020-0751>
- 975 Zhang, S., Tao, Y., Geng, X., 2024. A generic random field approach for stratification uncertainty
976 quantification. *Eng. Geol.* 343, 107800. <https://doi.org/10.1016/j.enggeo.2024.107800>
- 977 Zhang, W., Zhang, R., Wu, C., Goh, A.T.C., Lacasse, S., Liu, Z., Liu, H., 2020. State-of-the-art review
978 of soft computing applications in underground excavations. *Geosci. Front.* 11, 1095–1106.
979 <https://doi.org/10.1016/j.gsf.2019.12.003>
- 980 Zhao, C., Gong, W., Juang, C.H., Tang, H., Hu, X., Wang, L., 2023. Optimization of site exploration
981 program based on coupled characterization of stratigraphic and geo-properties uncertainties.
982 *Eng. Geol.* 317, 107081. <https://doi.org/10.1016/j.enggeo.2023.107081>
- 983 Zhao, T., Wang, Y., Lu, S.-F., Xu, L., 2023. Fast stratification of geological cross-section from CPT
984 results with missing data using multitask and modified Bayesian compressive sensing. *Can.*
985 *Geotech. J.* 60, 1812–1834. <https://doi.org/10.1139/cgj-2022-0131>
- 986 Zhao, T., Xu, L., Wang, Y., 2020. Fast non-parametric simulation of 2D multi-layer cone penetration
987 test (CPT) data without pre-stratification using Markov Chain Monte Carlo simulation. *Eng.*
988 *Geol.* 273, 105670. <https://doi.org/10.1016/j.enggeo.2020.105670>
- 989 Zou, H., Liu, S., Cai, G., Bheemasetti, T.V., Puppala, A.J., 2017. Mapping probability of liquefaction
990 using geostatistics and first order reliability method based on CPTU measurements. *Eng. Geol.*
991 218, 197–212. <https://doi.org/10.1016/j.enggeo.2017.01.021>

992