Scientific Research Publishing

World Journal of Engineering and Technology, 2024, 12, 895-913
https://www.scirp.org/journal/wjet
ISSN Online: 2331-4249
ISSN Print: 2331-4222

Optimization of the Number and Location of Boreholes for Gassy Soil Site Investigation Considering the Statistical Uncertainty

Shaolin Ding[1], Quanhong Li[2]

[1]Key Laboratory of Geotechnical Mechanics and Engineering of the Ministry of Water Resources, Yangtze River Scientific Research Institute, Wuhan, China
[2]Mid-Route Source of South-to-North Water Transfer Corp., Ltd., Danjiangkou, China
Email: 2403749231@qq.com

How to cite this paper: Ding, S.L. and Li, Q.H. (2024) Optimization of the Number and Location of Boreholes for Gassy Soil Site Investigation Considering the Statistical Uncertainty. *World Journal of Engineering and Technology*, **12**, 895-913.
https://doi.org/10.4236/wjet.2024.124055

Received: July 31, 2024
Accepted: October 13, 2024
Published: October 16, 2024

Abstract

The research addresses the prevalence of gassy soil, containing methane (CH_4), within the soil particles of southeast coastal areas of China, such as the Quaternary deposit in the Hangzhou Bay area. This soil exhibits spatial variability in the distribution of gas pressure, posing a potential threat of engineering disasters, including fire outbreaks and blasting, during the construction of underground projects. Consequently, it is crucial to assess the risk state of gas pressure, involving accurate identification and reduction of associated uncertainty, through site investigation. This is indispensable prior to the commencement of underground projects. However, during the site investigation stage, the random field parameters that quantify the spatial variability distribution of gas pressure (e.g., mean value, standard deviations, and scale of fluctuation) are unknown, introducing corresponding statistical uncertainty. Therefore, the most significant consideration for planning site investigation from an engineering perspective involves determining the risk state of gas pressure while considering the statistical uncertainty of these random field parameters. This consideration heavily relies on the engineering experience gained from current site investigation practices. To address this challenge, the study introduces a probabilistic site investigation optimization method designed for planning the site investigation scheme for gassy soils, including determining the number and locations of boreholes. The method is based on the expected state-identification probability, representing the probability of identifying the risk state of gas pressure, and takes into account the statistical uncertainty of random field parameters. The proposed method aims to determine an optimal investigation scheme before conducting the site investigation, leveraging prior knowledge. This optimal scheme is identified using Subset Simulation Optimization (SSO) in the space of candidate site investigations,

maximizing the value of the expected state-identification probability at the minimal value point. Finally, the paper illustrates the proposed approach through a case study.

Keywords

Gassy Soils, Site Investigation, Subset Simulation Optimization (SSO), Uncertainty

1. Introduction

The prevalence of gassy soils is widely distributed in the eastern coastal areas of China, particularly in the Hangzhou Bay area, Zhejiang province, as shown in Figure 1. Gassy soils, originating from the anaerobic decomposition of organic materials [1], are predominantly methane-dominated, with CH_4 constituting over 90% of the samples in the Hangzhou Bay area in Table 1 and Table 2. The spatial variability in the distribution of gas pressure in these soils poses potential risks, such as fire outbreaks and blasting, during underground construction projects [2]. To mitigate these risks during underground construction, a site investigation scheme is imperative. The scheme, specifying the number and locations of boreholes, strategically places them to measure gas pressure values using a modified Cone Penetration Test (CPT) device, demonstrated in Figure 2.

Figure 1. The enrichment area of gassy soils in the eastern coastal area of China.

Table 1. Gas composition of tunnel across the Qiantang River [3].

Borehole number	Number of gas sample	CH_4/%	N_2/%	CO_2/%	CO/10^{-6}
C-02	1	91.6	5.7	2.69	110
C-04	2	95.2	2.2	2.58	85
Additional 1	3	94.6	1.9	3.44	96

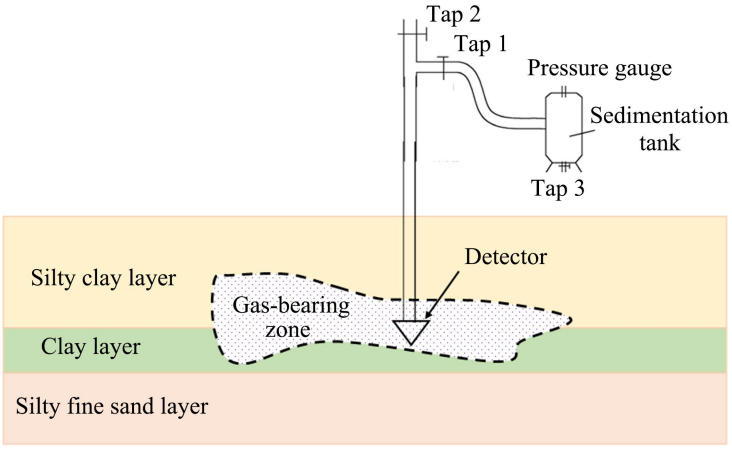

Figure 2. Site investigation of gassy soil with modified CPT device [3].

Table 2. Gas composition in wells at the south of the Qiantang River [3].

Borehole number	Number of gas sample	CH_4/%	N_2/%	CO_2/%	$CO/10^{-6}$
C26	1	90.4	7.67	1.92	230
C31	2	92.8	5.31	1.88	125
C35	3	91.5	6.96	1.53	180

However, due to the substantial cost and human commitments associated with site investigations for gassy soils, the data obtained are limited in engineering practice. Predicting the state of gas pressure (safe or dangerous) at unknown points based on the acquired data becomes essential. However, such predictions relying on limited data, introduce uncertainty, particularly considering that the random field parameters (e.g., mean value, standard deviations, and scale of fluctuation) characterizing the spatial variability distribution of gas pressure remain unknown during the site investigation stage and result in corresponding statistical uncertainty. In light of these challenges, determining the optimal site investigation scheme, including the optimal number of boreholes and their corresponding locations, becomes a pertinent and open question. This optimization is crucial for effectively identifying the risk state and reducing associated uncertainty at unknown locations before the construction of underground projects.

As previously discussed, it is crucial to carefully determine the number and locations of boreholes to effectively identify the risk state and reduce corresponding uncertainty at unknown locations. This task is challenging, particularly considering the statistical uncertainty associated with random field parameters that quantify the distribution of gas pressure. While some studies have explored gassy soils in the Hangzhou Bay area, these predominantly focused on aspects such as the formation and composition of biogenic gas [2], features and distributions of gas pools [3], and exploration methods [4] [5]. Some researchers have discussed planning investigation schemes for gassy soil, primarily focusing on decreasing uncertainty in gas pressure distribution at unknown points, often overlooking the

identification of the risk state at these locations and disregarding the statistical uncertainty of random field parameters [6]-[12].

This study introduces a probabilistic site investigation optimization method to determine the optimal scheme for investigating gassy soils. The method utilizes the expected state-identification probability to recognize the risk state of gas pressure and quantify corresponding uncertainty at unknown points. To decrease the uncertainty of the identified risk state, the site investigation scheme seeks the larger value of the expected state-identification probability at each unknown point. The scheme with the maximal value of expected state-identification probability at the minimal value location (*i.e.*, the expected state-identification probability) is then identified using SSO in the space of candidate site investigation schemes generated through discretization [13]-[15]. The candidate scheme satisfying the condition that its maximal expected state-identification probability at the minimal value location exceeds a given threshold probability is determined as the optimal scheme.

The research is structured with an introduction, followed by a demonstration of the proposed framework. Subsequently, the generation of the space of candidate site investigation schemes, quantification of expected state-identification probability, and optimization of the optimal scenario using SSO are covered in detail. Lastly, the implementation procedure of the proposed approach is presented and illustrated through a case study in the Hangzhou Bay area.

2. Framework for Probabilistic Site Investigation Optimization for Gassy Soils

Accurately identifying the risk state (safe or dangerous) of gas pressure and quantifying the corresponding uncertainty before construction is crucial to prevent engineering disasters caused by gassy soils. Typically, site investigations of gassy soils are conducted to estimate the risk state of gas pressure at unknown locations, relying on a limited amount of investigation data. To address this, an effective investigation scheme is significant that not only accurately identifies the risk state at unknown locations but also reduces the uncertainty associated with the identified gas pressure risk state. This study introduces a probabilistic site investigation approach for gassy soils to fulfill this purpose. It is important to note that this study focuses on the one-dimensional spatial variability of gas pressure in the horizontal direction, ignoring consideration of vertical spatial variability, which may be explored in future studies.

The proposed framework, illustrated in **Figure 3**, comprises three key steps: generation of the space of candidate site investigation schemes, quantification of expected state-identification probability, and optimization of borehole locations using SSO. The approach commences with the generation of all possible candidate site investigation schemes, achieved through a discretization procedure based on the site investigation range of gassy soils and a given discretization interval. It is crucial to emphasize that the determination of the discretization interval should

align with the specific requirements and accuracy standards of the site investigation. After obtaining the space of candidate site investigation schemes, the expected state-identification probability is employed to identify the risk state and quantify corresponding uncertainty at unknown locations. This is calculated using simulation data, given that real gas pressure data cannot be obtained at the scheme design stage. To reduce the uncertainty of the risk state at each unknown point, the candidate site investigation scheme must ensure that the expected state-identification probability at the minimum value point has the maximum value. This optimization problem can be addressed using SSO. The candidate scheme that guarantees the value of the expected state-identification probability at the minimum value point surpasses a given probability threshold is determined as the optimal scheme.

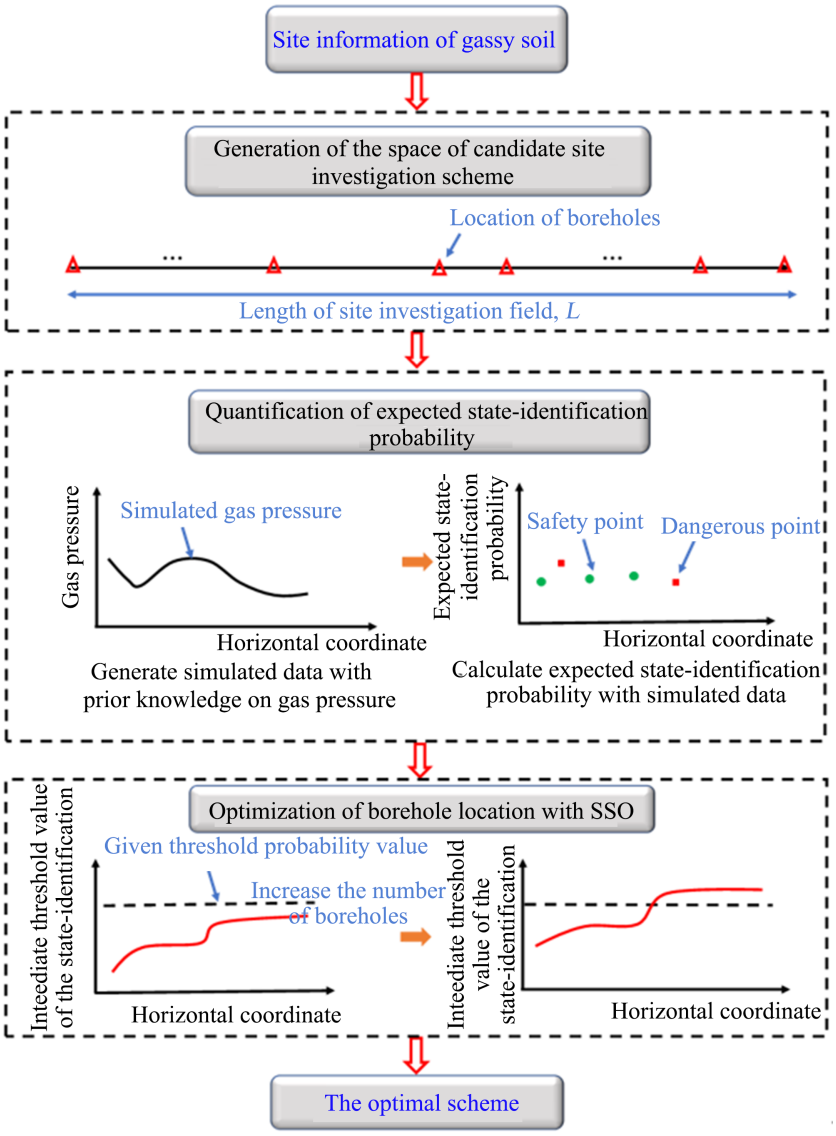

Figure 3. The framework of proposed probabilistic site investigation approach for gassy soils.

3. Space of Candidate Site Investigation Schemes

The determination of candidate site investigation schemes, relying on the number and placement of boreholes, is achieved through a discretization process. Consider the length, L, of the site investigation field. The points of interest, denoted as L_m (where $m = 1, 2, 3, \ldots, N$), adhere to $L_m = (m - 1)$ with a given interval ΔL. Here, N is calculated as INT $[L/\Delta L]$, where INT $[\cdot]$ denotes the rounding function returning the integer part of $L/\Delta L$. All values of L_m ($m = 1, 2, 3, \ldots, N$) can be represented as a vector $\boldsymbol{L_N} = [L_1, L_2, \ldots, L_N]$, as shown in Figure 4, encompassing a total of N possible values of L_m.

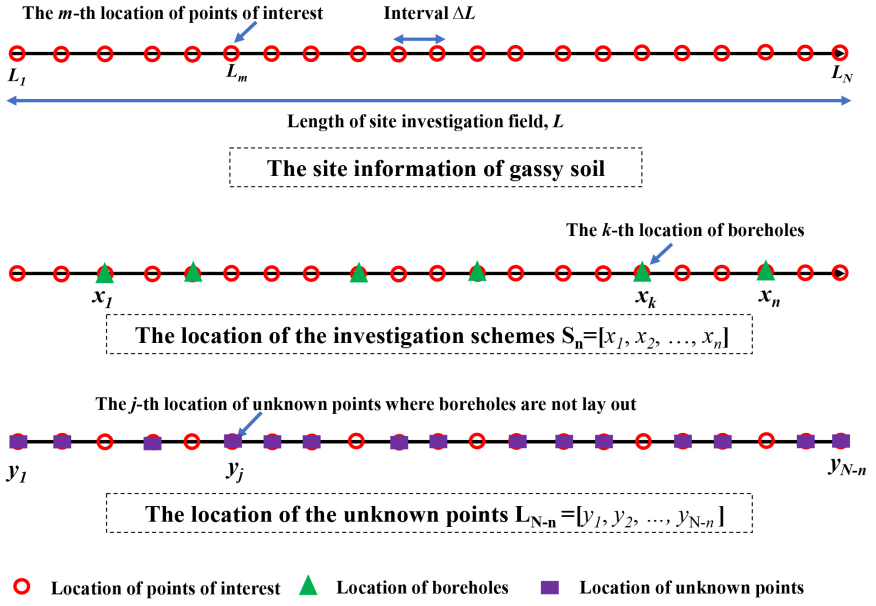

Figure 4. Site investigation scheme $S_n = [x_1, x_2, \ldots, x_k, \ldots, x_n]$.

Assuming that investigation schemes are denoted by a vector $\mathbf{S_n} = [x_1, x_2, \ldots, x_k, \ldots, x_n]$, representing borehole locations horizontally. x_k signifies the location of the k-th borehole, and n denotes the number of boreholes. The potential value of x_k should correspond to an element (*i.e.*, a feasible discretization point L_m ($m = 1, 2, 3, \ldots, N$)) in $\boldsymbol{L_N}$. Based on this, each possible value of x_1-x_n constitutes the candidate scheme $\mathbf{S_n}$, and it can be deduced that there is a total of C_N^n candidate site investigation schemes.

In practical engineering scenarios, based on the data of scheme $\mathbf{S_n}$, engineers need to predict the risk state, denoted as the expected state-identification probability, at unknown locations. These locations are represented by the vector $\boldsymbol{L_{N\text{-}n}} = [y_1, y_2, \ldots, y_j, \ldots, y_{N\text{-}n}]$, where boreholes are not placed to measure gas pressure. The value of y_j should belong to the set Ω_o, representing feasible values of L_m, while not being identical to any values among $x_1, x_2, \ldots, x_k, \ldots, x_n$. The number of unknown points y_k, representing the difference between the total number of points of interest (*i.e.*, L_m ($m = 1, 2, 3, \ldots, N$)) and the number of points (x_i ($i = 1, 2, \ldots, n$)) corresponding to scheme $\mathbf{S_n}$, is determined as N-n.

In the context of engineering practice, the primary focus is determining the risk state and associated uncertainty of the unknown point, y_j. Identification of the risk state at an unknown location and the effective reduction of uncertainty related to the identified gas pressure risk state are pivotal considerations in the site investigations from an engineering perspective. These objectives can be accomplished by maximizing the expected state-identification probability. The specifics regarding the quantification of the expected state-identification probability will be covered in the subsequent Section 4.

4. Definition of Expected State-Identification Probability

4.1. Simulated Data with Prior Knowledge of Gas Pressure

To assess the expected probability, $E\left(p_a\left(y_j\right)|\mathbf{S_n}\right)$, at the point y_j, simulated data is employed. This data is generated based on prior knowledge of gas pressure, mean values μ, standard deviations σ, and the scale of fluctuation λ. Given that real gas pressure data (*i.e.*, $\boldsymbol{Z_{br}}$) is unavailable at the scheme design stage, simulated data becomes crucial. For instance, when considering mean values μ, standard deviations σ, and the scale of fluctuation λ varying within their respective typical ranges $[\mu_{\min}, \mu_{\max}]$, $[\sigma_{\min}, \sigma_{\max}]$, and $[\lambda_{\min}, \lambda_{\max}]$, these parameters can be treated as uniform random variables defined by their typical ranges. Prior knowledge of random field parameters can be derived from historical data available in global databases as well as data specific to the site under consideration. In cases where no prevailing knowledge exists, the potential ranges of random field parameters can be determined based on their typical values reported in the literature. This approach provides a relatively uninformative prior knowledge, allowing for the incorporation of parameter uncertainty in the analysis. Random samples of μ, σ, and λ can be generated, denoted as $\mu_{s,i}$, $\sigma_{s,i}$, and $\lambda_{s,i}$ (where $i = 1, 2, 3, \ldots, N_e$), representing N_e sets of random samples. For each set of $\mu_{s,i}$, $\sigma_{s,i}$, and $\lambda_{s,i}$, the simulated data at discretization point L_m (where $m = 1, 2, 3, \ldots, N$) can be expressed as $\mathbf{Z_{s,i}(L_N)} = [Z_{s,i}(L_1), \ldots, Z_{s,i}(L_m), \ldots, Z_{s,i}(L_N)] (i = 1, 2, 3, \ldots, N_e)$. In this study, $\mathbf{Z_{s,i}(L_N)}$ is simulated using Karhunen-Loeve (K-L) expansion [16] [17], and the formulation is as follows:

$$Z_{s,i}\left(L_m\right) = \mu_{s,i} + \sum_{j=1}^{\infty} \sigma_{s,i} \sqrt{v_j} f_j\left(L_m\right) \zeta\left(\theta\right) \tag{5}$$

where $Z_{s,i}(L_m) (i = 1, 2, 3, \ldots, N_e)$ is the gas pressure data simulated using the sample $\mu_{s,i}$, $\sigma_{s,i}$ and $\lambda_{s,i}$; L_m is the discretization points with the given the length, L, of site investigation field concerned and corresponding interval ΔL. $\zeta(\theta)$ is independent standard normal random variable; v_j and $f_j(x)$ are the eigenvalues and eigenfunctions of the covariance function, which is taken as a squared exponential correlation function in this study:

$$\rho\left(\tau\right) = \exp\left[-\pi\left(\tau/\lambda_{s,i}\right)^2\right] \tag{6}$$

where τ is the separate distance between two locations in the horizontal direction; $\rho(\tau)$ is the autocorrelation coefficient between the gas pressures at the two

locations. For the sake of conciseness, details of the random field simulation based on K-L expansion are not provided here. Interested readers may refer to related reference [16] [17].

4.2. Prediction of the Gas Pressure Values with Gaussian Process

The simulated gas pressures at borehole locations (*i.e.*, x_1, x_2, …, x_k, …, x_n) of scheme $\mathbf{S_n}$ are denoted as vector $\mathbf{Z_{br,i}(S_n)} = [Z_{br,i}(x_1), \ldots, Z_{br,i}(x_k), \ldots, Z_{br,i}(x_n)]$. Employing $\mathbf{Z_{br,i}(S_n)}$, Gaussian Process (GP) is applied to predict the gas pressure values at the unknown location $[y_1, y_2, \ldots, y_j, \ldots, y_{N-n}]$, denoted as $\mathbf{Z_{c,i}(L_{N-n})} = [Z_{c,i}(y_1), \ldots, Z_{c,i}(y_j), \ldots, Z_{c,i}(y_{N-n})]$. $\mathbf{Z_{c,i}(L_{N-n})}$ comprises random variables with a joint Gaussian distribution, expressed as

$$\mathbf{Z_{c,i}}\left(\mathbf{L_{N-n}}\right)\left|\mathbf{L_{N-n}}, \mathbf{S_n}, \mathbf{Z_{br,i}}\left(\mathbf{S_n}\right) \sim N\left(\mu\left(\mathbf{Z_{c,i}}\left(\mathbf{L_{N-n}}\right)\right), \operatorname{cov}\left(\mathbf{Z_{c,i}}\left(\mathbf{L_{N-n}}\right), \mathbf{Z_{c,i}}\left(\mathbf{L_{N-n}}\right)\right)\right)\right.$$

[18], $\mu\left(\mathbf{Z_{c,i}}\left(\mathbf{L_{N-n}}\right)\right) = \left[\mu_{y_1}, \mu_{y_2}, \cdots, \mu_{y_{N-n}}\right]$ and

$$\operatorname{cov}\left(\mathbf{Z_{c,i}}\left(\mathbf{L_{N-n}}\right), \mathbf{Z_{c,i}}\left(\mathbf{L_{N-n}}\right)\right) = \begin{bmatrix} \sigma_{y_1y_1}, \sigma_{y_1y_2}, \cdots, & \sigma_{y_1y_{N-n}} \\ \sigma_{y_2y_1}, \sigma_{y_2y_2}, \cdots, & \sigma_{y_2y_{N-n}} \\ \\ \sigma_{y_{N-n}y_1}, \sigma_{y_{N-n}y_2}, \cdots, & \sigma_{y_{N-n}y_{N-n}} \end{bmatrix}$$

respectively. μ_{y_j} (j = 1, 2, 3, …, N-n) is the expectation of the gas pressure value $Z_{c,i}(y_j)$ at the location y_j. $\sigma_{y_jy_k}$ (j = 1, 2, 3, .., N-n; k = 1, 2, 3, …, N-n) is the covariance between $Z_{c,i}(y_j)$ and $Z_{c,i}(y_k)$.

4.3. Calculation of Expected State-Identification Probability with Simulated Data

Given that the multi-dimensional variable $\mathbf{Z_{c,i}}\left(\mathbf{L_{N-n}}\right)$ represents a joint Gaussian distribution with an expectation $m\left(\mathbf{Z_{c,i}}\left(\mathbf{L_{N-n}}\right)\right)$ and covariance $\operatorname{cov}\left(\mathbf{Z_{c,i}}\left(\mathbf{L_{N-n}}\right), \mathbf{Z_{c,i}}\left(\mathbf{L_{N-n}}\right)\right)$ [18], it follows that the marginal distribution $Z_{c,i}(y_j)$ (j = 1, 2, 3,…, N-n) is also a Gaussian distribution. The probability of E_s and E_d can be achieved using Equations (7) and (8), respectively.

$$p_s^i\left(y_j\right) = p\left(Z_{c,i}\left(y_j\right) < R\right) = \Phi\left(\frac{R - \mu_{y_j}}{\sigma_{y_jy_j}}\right) \tag{7}$$

$$p_d^i\left(y_j\right) = p\left(Z_{c,i}\left(y_j\right) \geq R\right) 1 - p_s^i\left(y_j\right) \tag{8}$$

where $p_s^i\left(y_j\right)$ and $p_d^i\left(y_j\right)$ are the probability of E_s and E_d respectively, given data $\mathbf{Z_{br,i}(S_n)}$. $Z_{c,i}(y_j)$ is the gas pressure at y_j that is a Gaussian random variable with expectation μ_{y_j} and standard deviation $\sigma_{y_jy_j}$. It is worth pointing out that $\sigma_{y_jy_j}$ is the diagonal elements of $\operatorname{cov}\left(\mathbf{Z_{c,i}}\left(\mathbf{L_{N-n}}\right), \mathbf{Z_{c,i}}\left(\mathbf{L_{N-n}}\right)\right)$.

To assess the uncertainty in gas pressure distribution, Monte Carlo simulation is employed for the repetitive prediction of gas pressure using GP based on the N_e simulated data $\mathbf{Z_{br,i}(S_n)} = [Z_{br,i}(x_1), \ldots, Z_{br,i}(x_k), \ldots, Z_{br,i}(x_n)]$ (i = 1, 2, 3, …, N_e). This results in N_e sets of expected values of predicted gas pressure, denoted as $\mathbf{Z_{c,i}}\left(\mathbf{L_{N-n}}\right)$ (i = 1, 2, 3, …, N_e). With each set of simulated data $\mathbf{Z_{br,i}(S_n)}$ (i = 1, 2, …, N_e), the probabilities of E_s and E_d are computed as $p^i_s(y_j)$ and $p^i_d(y_j)$ (i = 1,

2, …, N_e) using Equations (7) and (8). Subsequently, the mean values of $p^i_s(y_j)$ and $p^i_d(y_j)$ corresponding to the N_e sets of simulated data $Z_{br,i}$ (i = 1, 2, …, N_e) are determined with Equations (9) and (10):

$$p_{se}\left(y_j\right)=\frac{1}{N_e}\sum_{i=1}^{N_e}p_s^i\left(y_j\right) \tag{9}$$

$$p_{de}\left(y_j\right)=1-p_{se}\left(y_j\right) \tag{10}$$

where $p_{se}(y_j)$ and $p_{de}(y_j)$ are the mean values of $p^i_s(y_j)$ and $p^i_d(y_j)$ corresponding to the N_e sets of simulated data $Z_{br,i}$ (i = 1, 2, …, N_e). N_e is the total number of simulated data $\mathbf{Z_{br,i}(S_n)}$ (i = 1, 2, …, N_e).

Substitute Equations (9)-(10) into Equation (4), $E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n}\right)$ can be expressed as Equation (11).

$$\begin{aligned}E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n}\right)=\min\Bigg\{&\max\left\{\frac{1}{N_e}\sum_{i=1}^{N_e}p_s^i\left(y_1\right),1-p_{se}\left(y_1\right)\right\},\\&\max\left\{\frac{1}{N_e}\sum_{i=1}^{N_e}p_s^i\left(y_2\right),1-p_{se}\left(y_2\right)\right\},\cdots,\\&\max\left\{\frac{1}{N_e}\sum_{i=1}^{N_e}p_s^i\left(y_{N-n}\right),1-p_{se}\left(y_{N-n}\right)\right\}\Bigg\}\end{aligned} \tag{11}$$

The next section makes uses of SSO to identify the optimal scheme $\mathbf{S_n^*}$ among the candidate site investigation scheme space.

5. Definition of Expected State-Identification Probability

As discussed in the "Space of Candidate Site Investigation Schemes" section, a total of C_N^n candidate schemes are generated by randomly selecting n discretization points from Ω_o. The process of identifying the scheme $\mathbf{S_n^*}$ with the highest value of $E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n}\right)$ at the location $y_{\min}$ can be expressed as the optimization problems in Equation (12):

$$\begin{aligned}&\max_{\mathbf{S_n}}\left(E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n}\right)\right)\\&\mathbf{S_n}=\left\{x_1,x_2,\cdots,x_k,\cdots,x_n\right\}\end{aligned} \tag{12}$$

As demonstrated in Equation (12), the optimization of the borehole locations is carried out with the expected state-identification probability, $E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n}\right)$, at the $y_{\min}$ location as the objective function. Solving the optimization problem (Equation (12)) to determine the scheme $\mathbf{S_n^*}$ and its corresponding $E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n^*}\right)$ can be challenging due to the potentially large number (C_N^n) of candidate schemes. In this study, SSO, a well-established global optimization algorithm, is employed to address Equation (12). Within the SSO framework, the optimal scheme, $\mathbf{S_n^*}$, characterized by the maximum $E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n^*}\right)$, is identified by exploring the design space of candidate schemes in a stochastic manner. Theoretically, $\mathbf{S_n^*}$ can be found among the candidate schemes by solving the following reliability analysis problem in Equation (13) [13]:

$$P(F) = P\left(E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n}\right) > E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n^*}\right)\right) \tag{13}$$

where $F = \left\{E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n}\right) > E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n^*}\right)\right\}$ is an auxiliary failure event. $P(F)$ represents the probability that event F occurs, which becomes to zero as scheme $\mathbf{S_n}$ is equal to $\mathbf{S_n^*}$.

A number of conditional samples of a series of nested intermediate failure events satisfying $F_1 \supset F_2 \supset F_3 \supset \cdots \supset F_{N_s} = F$ is generated with SSO, with which $P(F)$ is expressed as Equation (14):

$$P(F) = P\left(F_{N_s}\right) = P\left(F_1\right)\prod_{m=2}^{N_s} P\left(F_m\middle|F_{m-1}\right) \tag{14}$$

where $F_m = \left\{E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n}\right) > E_m\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n}\right)\right\}$, m = 1, 2, 3, …, N_s. $P\left(F_1\right)$ is equal to $P\left(E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n}\right) > E_1\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n}\right)\right)$;
$E_1\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n}\right) < E_2\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n}\right) < \cdots < E_{N_s}\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n}\right) = E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n^*}\right)$ are an increasing sequence of N_s intermediate threshold values, which are determined adaptively with simulated samples so that the sample estimates of $P(F_1)$ and $P(F_m|F_{m-1})$ are always equivalent to a specific value of conditional probability p_0 (e.g., 0.1). For a given number of boreholes, each set of random samples of feasible locations constitutes a random candidate scheme. The Subset Simulation approach begins with direct Monte Carlo simulation to generate a specified number, N_L, of random schemes. Subsequently, the expected state-identification probability values of these random schemes are calculated and ranked in ascending order to identify a number, p_0N_L, of seed schemes. These seed schemes define the first threshold, F_1, and another N_L - p_0N_L random schemes satisfying F_1 are simulated using Markov Chain Monte Carlo simulation (MCMCS). Similar procedures are then iterated to progressively explore m = 2, 3, ..., N_s levels, level by level. The implementation of SSO involved with related parameters (e.g., conditional probability p_0 and N_s) setting, details of which can refer to related reference [15].

For various values of n, representing the number of boreholes in scheme $\mathbf{S_n}$, employ the SSO approach mentioned earlier to identify the corresponding $\mathbf{S_n^*}$. If the value of $E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n^*}\right)$ at the $y_{\min}$ point associated with $\mathbf{S_n^*}$ exceeds a predefined threshold probability value, denoted as p^*, then $\mathbf{S_n^*}$ is designated as the optimal scheme. The specific procedure for determining $\mathbf{S_n^*}$ using the proposed method will be covered in the subsequent section.

6. Illustrative Example

6.1. Candidate Investigation Schemes

To illustrate the application of the proposed approach in this study, an example of site investigation for gassy soils from the literature is adopted [6], focusing solely on the horizontal spatial variability of gas pressure. The cross-sectional length, L, in this example is 1023 m, discretized at 5 m intervals, resulting in a total of 205 discretization points, L_m (m = 1, 2, 3, …, 205), where $L_m = 5(m - 1)$ (m = 1, 2, 3, …, 205). For varying investigation schemes, the number, n, of boreholes ranges from 10 to 30 at intervals of 5, *i.e.*, 10, 15, 20, 25, 30.

Consider the case of n = 25, and the corresponding scheme $\mathbf{S_{25}}$ = [x_1, x_2, …, x_k, …, x_{25}]. The space of candidate schemes encompasses C_{205}^{25} instances of $\mathbf{S_{25}}$, randomly selected from the 205 discretization points (*i.e.*, L_m (m = 1, 2, 3, …, 205)). The set of unknown points is denoted as $\mathbf{L_{180}}$ = [y_1, y_2, …, y_j, …, y_{180}]. Each y_j (j = 1, 2, 3, …, 180) must belong to the feasible value set of L_m, without being equal to any values among x_1, x_2, …, x_k, …, x_{25}.

The scheme $\mathbf{S}_{25}^{*}$, maximizing the value of $E\left(p_a\left(y_{\min}\right)|\mathbf{S}_{25}\right)$ at the $y_{\min}$ point can be determined from the space of candidate schemes. This determination relies on simulated data generated with prior knowledge of gas pressure. The given prior knowledge in the literature assumes μ = 0.278 MPa, σ = 0.097 MPa, and λ = 50 m. However, since precise knowledge during the site investigation stage is unavailable, the mean values μ, standard deviations σ, and scale of fluctuation λ are considered as uniform random variables within their typical ranges specifically, $\mu \in$ (0 kPa, 300 kPa], $\sigma \in$ (0 kPa, 125 kPa], and $\lambda \in$ (0 m, 100 m] in this study, covering the prior knowledge assumption (*i.e.*, μ = 278 kPa, σ = 97 kPa, and λ = 50 m) used in the literature [6].

The impact of gassy soils is contingent on the gas pressure's specific value. Generally, gassy soils are deemed hazardous if the gas pressure is greater than or equal to 100 kPa; otherwise, the risk associated with gassy soils is considered negligible. Consequently, R is set at 100 kPa in this study, defining E_s (Safe event) and E_d (Dangerous event) as the event with gas pressure less than 100 kPa and its complementary event, as shown in Table 3.

Table 3. Definition of ignorable and risky events.

Events	Gas pressure (kPa)	Notes
E_s	(0, 100)	Safe event
E_d	[100, +∞)	Dangerous event

As outlined in the "Definition of expected state-identification probability" section, the proposed approach uses the value of $E\left(p_a\left(y_j\right)|\mathbf{S_n}\right)$ to ascertain the presence of gas pressure risk at certain locations. In general, if the value of $E\left(p_a\left(y_j\right)|\mathbf{S_n}\right)$ is substantial, indicating the likely risk, the uncertainty associated with the presence of risk can be disregarded. Employing verbal probability descriptors in Table 4, the threshold value (*i.e.*, $p^{\star}$) for determining the presence of risk is set at 0.9 (very likely) in this example. Since $p^{\star}$ exceeds 0.9, E_s or E_d is highly likely to occur based on whether $E\left(p_a\left(y_j\right)|\mathbf{S_n}\right) = p_{se}\left(y_j\right)$ or $p_{de}\left(y_j\right)$, respectively. The optimal scheme chosen by the proposed approach must ensure that the $E\left(p_a\left(y_j\right)|\mathbf{S_n}\right)$ values at each unknown location are all greater than 0.9, regardless of whether $E\left(p_a\left(y_j\right)|\mathbf{S_n}\right) = p_{se}\left(y_j\right)$ or $p_{de}\left(y_j\right)$.

Table 4. Verbal descriptors and their probability equivalents [19].

Verbal descriptor	Virtually impossible	Very unlikely	Equally likely	Very likely	Virtually certain
Probability equivalent	0.01	0.10	0.50	0.90	0.99

6.2. Expected State-Identification Probability Given Different Investigation Schemes

For instance, consider $\mathbf{S_{25}} = [x_1, x_2, \ldots, x_k, \ldots, x_{25}]$ with the borehole number of $n = 25$. To determine the optimal scheme, $\mathbf{S}_{25}^{*}$, from the array of candidate schemes (*i.e.*, C_{205}^{25}), the value of $E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_{25}}\right)$ at the $y_{\min}$ point must be calculated. Initially, N_e ($N_e = 500$), of random field parameters μ, σ, and λ were generated from prior knowledge (*i.e.*, uniform distribution within typical ranges for μ, σ, and λ). Using each set of μ, σ, and λ samples, $Z_{s,i}$ is simulated using Equation (5) and (6), where the gas pressures at borehole locations in $\mathbf{S_{25}}$ constitute a set of simulated data denoted as $\mathbf{Z_{br,i}(S_{25})}$. For each $\mathbf{Z_{br,i}(S_{25})}$, the mean $\mu\left(\mathbf{Z_{c,i}}\left(\mathbf{L}_{180}\right)\right)$ and covariance $\operatorname{cov}\left(\mathbf{Z_{c,i}}\left(\mathbf{L}_{180}\right), \mathbf{Z_{c,i}}\left(\mathbf{L}_{180}\right)\right)$ of the predicted gas pressure at the unknown points $\mathbf{L_{180}}$ are obtained with GP. $p_{se}(y_j)$ and $p_{de}(y_j)$) corresponding to the N_e sets of simulated data $\mathbf{Z_{br,i}(S_{25})}$ are then used to calculate $E\left(p_a\left(y_j\right)\middle|\mathbf{S}_{25}\right)$ at each $\mathbf{L_{180}}$ location.

As discussed in Section 5 titled "Optimization of borehole location with SSO", SSO is employed to locate the optimal scheme $\mathbf{S}_{25}^{*}$ that maximizes the value of $E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_{25}}\right)$ at the $y_{\min}$ point in the space of candidate schemes, where p_0 and N_s are set as 0.1 and 30, respectively, and 1000 samples are simulated in each level. Figure 5 illustrates the intermediate threshold value of $E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_{25}}\right)$ at different simulation levels as m increases. With an increase in the number of simulation levels (*i.e.*, m), $E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_{25}}\right)$ increases and reaches a value of 0.910 at $m = 4$. In this example, the SSO is executed until the 20th level to ensure the convergence of $E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_{25}}\right)$, after which the $E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_{25}}\right)$ at $m = 4$ is considered as the estimate of the maximal $E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_{25}}\right)$ values.

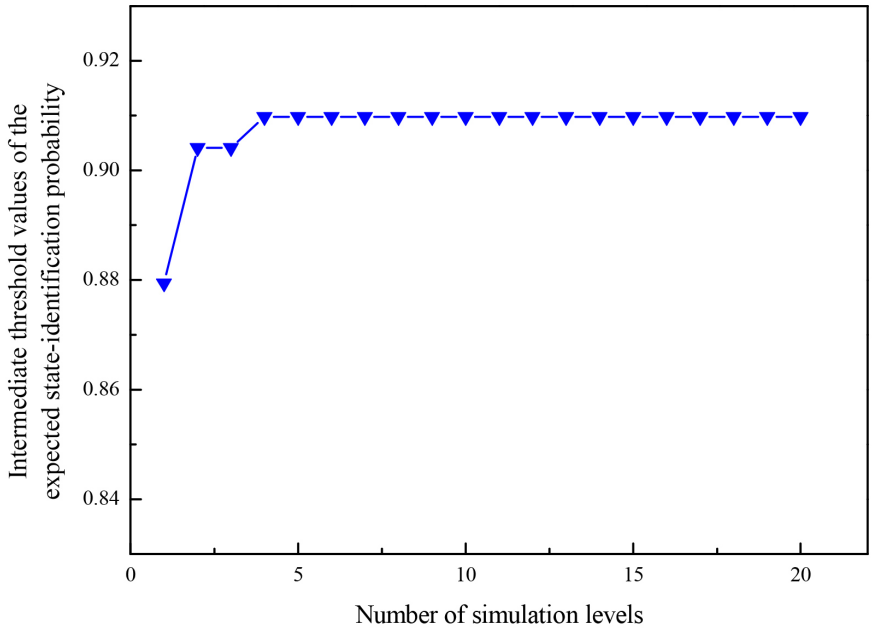

Figure 5. Evolution of the intermediate threshold value of the state-identification probability during SSO of S_{25}.

For varying borehole number (*i.e.*, n), Figure 6 displays the optimized maximal $E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S}_n\right)$ values using SSO for n = 10, 15, 20, 25, and 30, respectively. The results indicate that as the value of n increases to 20 (*i.e.*, $\mathbf{S}_{20}$), the maximal $E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S}_{20}\right)$ values surpass 0.9. This suggests that all $E\left(p_a\left(y_j\right)\middle|\mathbf{S}_{20}\right)$ values at each location $\mathbf{L}_{185}$ (*i.e.*, y_1, y_2,…, and y_{185}) are greater than 0.9. For schemes with a larger number of boreholes (e.g., $\mathbf{S}_{25}$ and $\mathbf{S}_{30}$), the proposed approach identifies optimal schemes, such as $\mathbf{S}_{25}^*$ and $\mathbf{S}_{30}^*$, where all $E\left(p_a\left(y_j\right)\middle|\mathbf{S}_n\right)$ values exceed 0.9, as demonstrated in Figure 7. However, it's important to note that the optimal schemes (*i.e.*, $\mathbf{S}_{25}^*$, $\mathbf{S}_{30}^*$) corresponding to $\mathbf{S}_{25}$ and $\mathbf{S}_{30}$, as optimized by the proposed method, require more investigation efforts due to the larger number of boreholes compared to $\mathbf{S}_{20}$. Therefore, considering the investigation effort, n = 20 is determined to be the optimal number of boreholes, and the corresponding scheme $\mathbf{S}_{20}^*$ is selected as the optimal scheme.

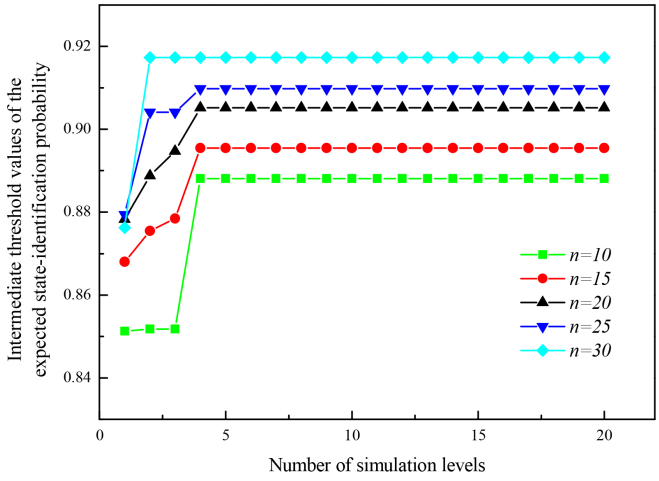

Figure 6. Evolution of the intermediate threshold value of the expected state-identification probability during SSO for different numbers of boreholes.

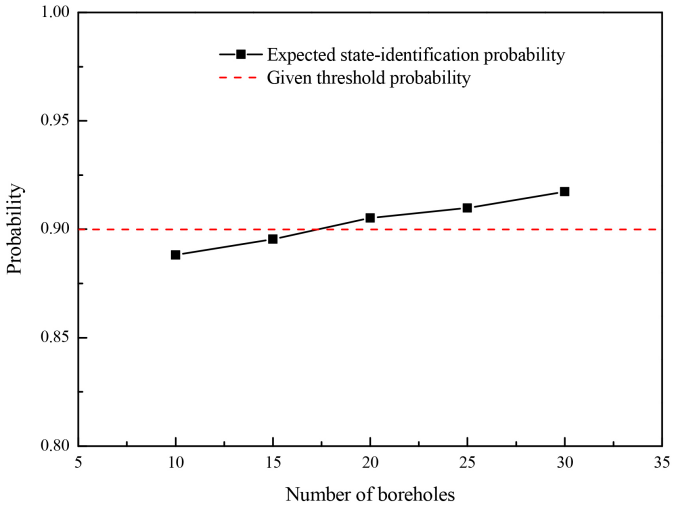

Figure 7. Expected state-identification of the optimal experimental schemes with different number of boreholes.

The specific horizontal coordinates of the optimal scheme $\mathbf{S}_{20}^{*}$ are illustrated in Figure 8 with blue-filled circles. The expected state identification for all unknown locations (*i.e.*, $\mathbf{L_{N\text{-}n}}$) along the horizontal direction is obtained and depicted in Figure 8. It's noteworthy that some of $\mathbf{L_{185}}$ correspond to $E\left(p_a\left(y_j\right)|\mathbf{S}_{20}\right)=p_{de}\left(y_j\right)$ (indicated by green triangles representing safe gas pressures), while others correspond to $E\left(p_a\left(y_j\right)|\mathbf{S}_{20}\right)=p_{re}\left(y_j\right)$ (depicted by red squares denoting dangerous gas pressures). For locations where $p_a = p_{de}(y_j)$ and $p_a \geq 0.9$, it is highly likely that the gas pressure is not risky. Conversely, for locations where $p_a = p_{re}(y_j)$ and $p_a \geq 0.9$, there is a high likely that the gas pressure is risky.

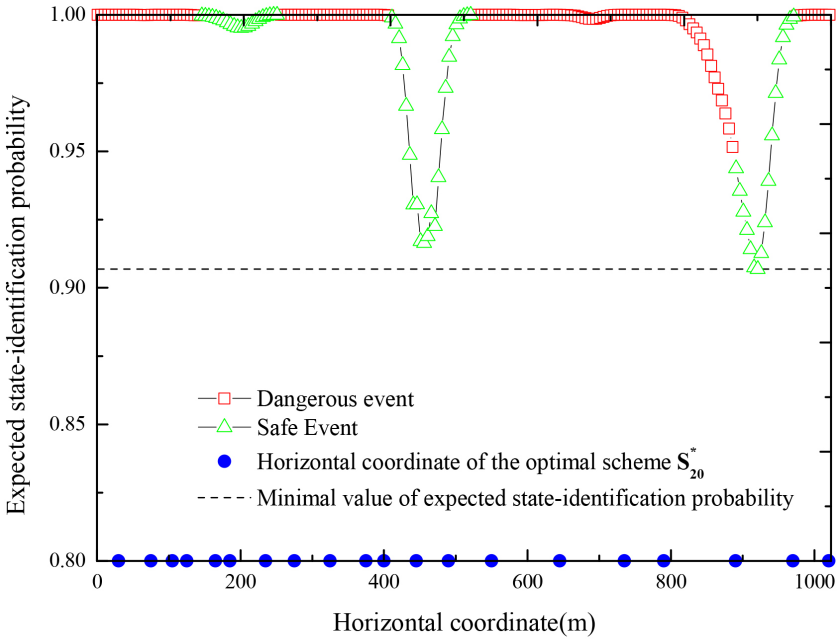

Figure 8. Expected state-identification probability of location $L_{\text{N-n}}$ along the horizontal direction corresponding to optimal scheme $\mathbf{S}_{20}^{*}$.

6.3. Comparison with Bayesian Compressive Sampling

To assess the effectiveness of the optimal scheme determined by the proposed method, a comparison with Guan *et al.*'s approach for planning a site investigation scheme is crucial. Guan *et al.*'s approach utilizes Bayesian compressive sampling (BCS) and information entropy to automatically determine sample size and optimal sampling locations for predicting the gas pressure distribution, given specific values (*i.e.*, μ = 278 kPa, σ = 97 kPa, and λ = 50 m) of random field parameters. As outlined in Section 7.2, titled "Expected state-identification probability given different investigation schemes", the optimal scheme determined by the proposed method is $\mathbf{S}_{20}^{*}$, with a corresponding optimal number of boreholes of 20. Utilizing the mean value and standard deviation of gas pressure predicted with simulated data corresponding to the optimal scheme $\mathbf{S}_{20}^{*}$, the coefficient of variation (COV) for each unknown location is obtained, as illustrated in Figure 9.

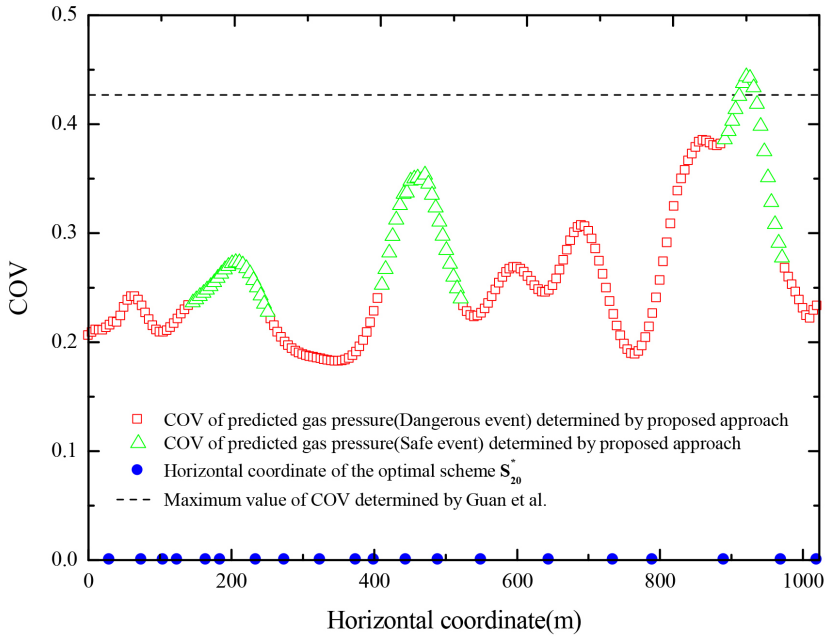

Figure 9. COV of location $L_{N\text{-}n}$ along the horizontal direction corresponding to optimal scheme $\mathbf{S}_{20}^{*}$.

In **Figure 9**, it can be seen that the maximum COV among all unknown locations (*i.e.*, $\mathbf{L_{N\text{-}n}}$) determined by the proposed method in this study is 44.37%. This value is close to the maximum COV (42.69%) obtained by Guan *et al.*'s approach when the number of boreholes is 20. It is important to note that, in this study, the mean values μ, standard deviations σ, and scale of fluctuation λ are defined as uniform random variables within their respective typical ranges (*i.e.*, $\mu \in$ (0 kPa, 300 kPa], $\sigma \in$ (0 kPa, 125 kPa], and $\lambda \in$ (0 m, 100 m]) rather than specific values (*i.e.*, μ = 278 kPa, σ = 97 kPa, and λ = 50 m) as in Guan *et al.*'s approach. This choice introduces larger uncertainty and relatively less informative prior knowledge on random field parameters. Therefore, the result that the maximum COV (44.37%) determined by the proposed method, given the same number of boreholes (n = 20), is relatively larger than that of Guan *et al.*'s approach is reasonable. This finding substantiates the effectiveness of the method proposed in this study.

6.4. Effect of the Range of Prior Knowledge

Employing the proposed approach, the determination of the optimal scheme relies on the prior knowledge concerning the random field parameters of gas pressure. In the previous discussion, the prior knowledge has been defined as $\mu \in$ (0 kPa, 300 kPa], $\sigma \in$ (0 kPa, 125 kPa], and $\lambda \in$ (0 m, 100 m]), referred to as Priori I in this research. To discuss the impact of varying prior knowledge, this subsection explores a new set of parameters (*i.e.*, $\mu \in$ (0 kPa, 150 kPa], $\sigma \in$ (0 kPa, 62.5 kPa], and $\lambda \in$ (0 m, 50 m]]), denoted as Priori II, in the determination of the optimal scheme using the proposed method.

Figure 10 demonstrates the $E\left(p_a\left(y_{\min}\right)\middle|\mathbf{S_n}\right)$ values of optimal schemes with

varying numbers (n) of measuring points, determined using Priori I and II. The results are depicted by lines with squares and circles, respectively. For a specific number of measuring points, the $E\left(p_a\left(y_{\min}\right)|\mathbf{S_n}\right)$ associated with Priori II surpasses that of Priori I. This indicates that, with the same number of measuring points, gas pressure exhibits lower uncertainty when considering Priori II compared to Priori I. This discrepancy arises from the relatively higher informativeness of Priori II, leading to a more substantial reduction in uncertainty concerning gas pressure variability given an equivalent amount of measurement data. While the proposed approach was developed based on the squared exponential correlation function, it can be adapted to accommodate other correlation functions. During the optimization phase of the scheme, the choice of correlation function can be made based on existing knowledge of gassy soil. If multiple correlation functions are considered, the uncertainty associated with model selection should be integrated into the optimization process. Bayesian model selection methods can be utilized to quantify this uncertainty. Incorporating the proposed method to address model selection uncertainty will be a focus of future research endeavors.

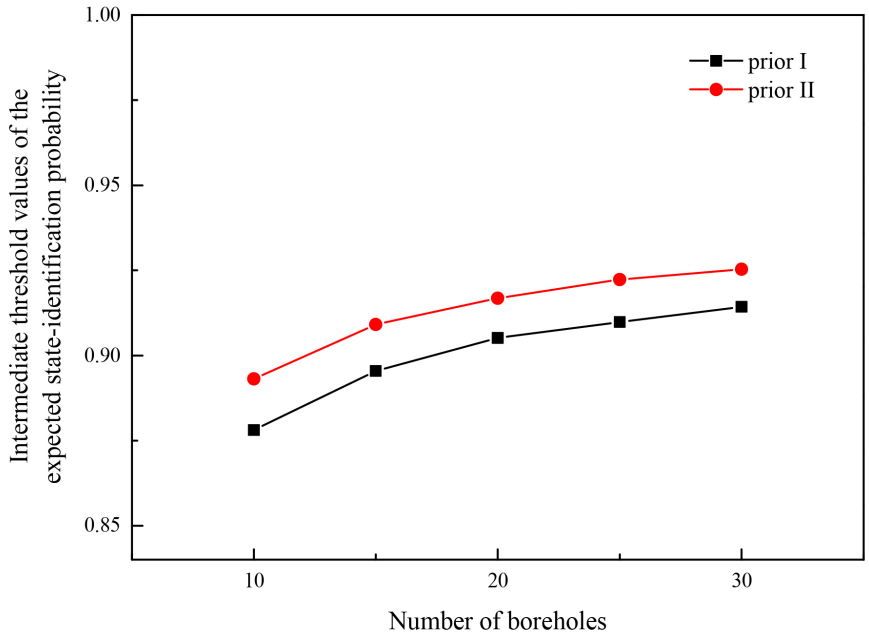

Figure 10. Comparison of the expected state-identification probability of optimal schemes obtained using different prior knowledge for different numbers of boreholes.

7. Summary and Conclusions

The study has devised a probabilistic method for optimizing site investigation, aiming to determine the most effective investigation scheme while considering the statistical uncertainty associated with random field parameters. This approach allows for the accurate identification of the risk state and simultaneous reduction of the corresponding uncertainty. The key findings are summarized as follows:

1) The space of potential site investigation schemes is established through

discretization along the horizontal dimension of gassy soil areas. The expected state-identification probability, quantifying the risk and uncertainty of gassy soils, is computed using simulated data based on GP. An optimization process is employed to identify the site investigation scheme with the maximum expected state-identification probability at the minimum value location. This scheme is considered optimal if its probability value surpasses a predetermined threshold.

2) The proposed approach is applied and validated using a site investigation example from the literature concerning gassy soils. Results demonstrate that, for a given number of measuring points (n), the maximal expected state-identification probability at the minimum value location increases progressively with an elevated number of simulation levels. This value ultimately converges to a maximum. The determined optimal scheme corresponds to $n = 20$ measuring points, identified by the SSO, as it exceeds the specified threshold probability value (0.9).

3) The effectiveness of the proposed method is verified by comparing it with an alternative approach for planning site investigation schemes. Using the optimal number of measuring points determined by our method ($n = 20$), the maximum COV of gas pressure among all unknown locations is found to be 44.37%, surpassing the 42.69% obtained through the alternative approach. The advantage of our proposed site investigation method lies in its consideration of the prior knowledge of parameters, defined as uniform random variables within typical ranges, aligning more closely with actual engineering conditions. This results in larger uncertainty but provides a more informative and realistic representation of prior knowledge on random field parameters. Notably, the method allows for the identification of gas pressure risk states during the site investigation stage, a facet overlooked in previous studies.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Au, S. and Wang, Y. (2014) Engineering Risk Assessment with Subset Simulation. Wiley. https://doi.org/10.1002/9781118398050

[2] Deng, Z., Jiang, S., Niu, J., Pan, M. and Liu, L. (2020) Stratigraphic Uncertainty Characterization Using Generalized Coupled Markov Chain. *Bulletin of Engineering Geology and the Environment*, **79**, 5061-5078. https://doi.org/10.1007/s10064-020-01883-y

[3] Deng, Z., Li, D., Qi, X., Cao, Z. and Phoon, K. (2017) Reliability Evaluation of Slope Considering Geological Uncertainty and Inherent Variability of Soil Parameters. *Computers and Geotechnics*, **92**, 121-131. https://doi.org/10.1016/j.compgeo.2017.07.020

[4] Deng, Z., Pan, M., Niu, J. and Jiang, S. (2022) Full Probability Design of Soil Slopes Considering Both Stratigraphic Uncertainty and Spatial Variability of Soil Properties.

Bulletin of Engineering Geology and the Environment, **81**, 1-13. https://doi.org/10.1007/s10064-022-02702-2

[5] Deng, Z., Pan, M., Niu, J., Jiang, S. and Qian, W. (2021) Slope Reliability Analysis in Spatially Variable Soils Using Sliced Inverse Regression-Based Multivariate Adaptive Regression Spline. *Bulletin of Engineering Geology and the Environment*, **80**, 7213-7226. https://doi.org/10.1007/s10064-021-02353-9

[6] Ding, S., Li, D., Cao, Z. and Du, W. (2022) Two-Stage Bayesian Experimental Design Optimization for Measuring Soil-Water Characteristic Curve. *Bulletin of Engineering Geology and the Environment*, **81**, Article No. 142. https://doi.org/10.1007/s10064-022-02598-y

[7] Guan, Z., Wang, Y., Cao, Z. and Hong, Y. (2020) Smart Sampling Strategy for Investigating Spatial Distribution of Subsurface Shallow Gas Pressure in Hangzhou Bay Area of China. *Engineering Geology*, **274**, Article ID: 105711. https://doi.org/10.1016/j.enggeo.2020.105711

[8] Huang, F., Zhang, J., Zhou, C., Wang, Y., Huang, J. and Zhu, L. (2019) A Deep Learning Algorithm Using a Fully Connected Sparse Autoencoder Neural Network for Landslide Susceptibility Prediction. *Landslides*, **17**, 217-229. https://doi.org/10.1007/s10346-019-01274-9

[9] Huang, F., Xiong, H., Yao, C., Catani, F., Zhou, C. and Huang, J. (2023) Uncertainties of Landslide Susceptibility Prediction Considering Different Landslide Types. *Journal of Rock Mechanics and Geotechnical Engineering*, **15**, 2954-2972. https://doi.org/10.1016/j.jrmge.2023.03.001

[10] Huang, S.P., Quek, S.T. and Phoon, K.K. (2001) Convergence Study of the Truncated Karhunen-Loeve Expansion for Simulation of Stochastic Processes. *International Journal for Numerical Methods in Engineering*, **52**, 1029-1043. https://doi.org/10.1002/nme.255

[11] Jiang, W., Ye, Z., Zheng, H., Yong, Z. (1997) Quaternary Shallow Gas Characteristics in Hangzhou Bay and Exploration Method. *Natural Gas Industry*, **17**, 20-23. (In Chinese)

[12] Li, H. and Cao, Z. (2016) Matlab Codes of Subset Simulation for Reliability Analysis and Structural Optimization. *Structural and Multidisciplinary Optimization*, **54**, 391-410. https://doi.org/10.1007/s00158-016-1414-5

[13] Li, H. and Ma, Y. (2015) Discrete Optimum Design for Truss Structures by Subset Simulation Algorithm. *Journal of Aerospace Engineering*, **28**, Article ID: 04014091. https://doi.org/10.1061/(asce)as.1943-5525.0000411

[14] Li, L., Zhao, Y. and Yu, L. (2009) Exploration for Quaternary Shallow Biogenic Gas by Sealed Core Drilling and Modified CPT. *Coal Geology and Exploration*, **37**, 72-76. (In Chinese)

[15] Li, Y. and Lin, C. (2010) Exploration Methods for Late Quaternary Shallow Biogenic Gas Reservoirs in the Hangzhou Bay Area, Eastern China. *AAPG Bulletin*, **94**, 1741-1759. https://doi.org/10.1306/06301009184

[16] Lin, C.M., Gu, L.X., Li, G.Y., Zhao, Y.Y. and Jiang, W.S. (2004) Geology and Formation Mechanism of Late Quaternary Shallow Biogenic Gas Reservoirs in the Hangzhou Bay Area, Eastern China. *AAPG Bulletin*, **88**, 613-625. https://doi.org/10.1306/01070403038

[17] Phoon, K.K., Huang, S.P. and Quek, S.T. (2002) Implementation of Karhunen-Loeve Expansion for Simulation Using a Wavelet-Galerkin Scheme. *Probabilistic Engineering Mechanics*, **17**, 293-303.

https://doi.org/10.1016/s0266-8920(02)00013-9

[18] Rasmussen, C.E. and Nickisch, H. (2010) Gaussian Processes for Machine Learning (GPML) Toolbox. *The Journal of Machine Learning Research*, **11**, 3011-3015.

[19] Vick, S.G. (2002) Degrees of Belief: Subjective Probability and Engineering Judgment. ASCE Press.