Bayesian Analysis of Benchmark Examples for Data-Driven Site Characterization

Mavritsakis, Antonis; Schweckendiek, Timo; Teixeira, Ana; Smyrniou, Eleni; Nuttall, Jonathan

DOI
10.1061/AJRUA6.RUENG-975

Publication date
2023

Document Version
Final published version

Published in
ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering

Citation (APA)
Mavritsakis, A., Schweckendiek, T., Teixeira, A., Smyrniou, E., & Nuttall, J. (2023). Bayesian Analysis of Benchmark Examples for Data-Driven Site Characterization. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, *9*(2), Article 04023008. https://doi.org/10.1061/AJRUA6.RUENG-975

Bayesian Analysis of Benchmark Examples for Data-Driven Site Characterization

Antonis Mavritsakis[1]; Timo Schweckendiek, Ph.D.[2]; Ana Teixeira, Ph.D.[3]; Eleni Smyrniou[4]; and Jonathan Nuttall, Ph.D.[5]

Abstract: Data-driven site characterization (DDSC) aids geotechnical engineering by inferring and mapping soil parameters of the subsurface domain. In practice, the limited availability of site investigation data may hinder the performance of traditional machine learning methods and implies significant uncertainty in the predictions, which is typically not quantified. In this study, a framework for Bayesian site characterization (BaySiC) is applied on a benchmark example. Adopting Bayesian statistics enables the framework to deal with small training data sets and allows for coherent quantification of uncertainty, which is valuable to engineering practice for assessing the reliability and the determining characteristic values. BaySiC uses site investigation data to infer statistical estimators of cone penetration test (CPT) parameters and their dependence, as well as to learn spatial correlations. Consecutively, it generates a three-dimensional (3D) map of the subsurface by predicting the CPT parameter values and classifying the material type over the soil domain. For the benchmark example, the study formulated two models within the BaySiC framework and demonstrated their conduct in several cases of varying complexity. Eventually, the performance of the models was evaluated and compared in both deterministic and probabilistic terms. One of the models proved highly effective in predicting the material type at new locations of the subsurface domain, whereas the other provided accurate mapping of the CPT parameters even in complex stratigraphic cases. Also, investigating and comparing the results of the models led to insights regarding the effectiveness of their formulation. Moreover, the paper used hypothesis testing as a means of assessing the predictive power of the model independently from the validation data set. Stemming from the benchmark example, the paper draws conclusions that are meaningful to geotechnical engineering and decision-making. **DOI: 10.1061/AJRUA6.RUENG-975.**

Introduction

Data-driven site characterization (DDSC) comprises techniques that aim to characterize the subsoil of a project site solely by using measured data (Phoon et al. 2022a). With the rapid growth of the fields of artificial intelligence and machine learning over the last years, developing, reexamining, and improving DDSC methods becomes more relevant than ever. To that end, Phoon et al. (2022b) have provided a DDSC benchmark example as a means for development and comparison of techniques.

The use of cone penetration test (CPT) readings lies at the core of the benchmark example. Specifically, the example entailed training a model using CPT measurements at training locations and predicting CPT parameter values at specific validation locations in an artificial, multilayered subsoil domain. The CPT parameters of interest are the cone resistance (q_t) and sleeve friction (f_s). Moreover, the example required the prediction of the material type at the validation locations of the subsoil, as classified using CPT measurements. Eventually, the predictions of the model were evaluated using metrics defined by the benchmark exercise.

The setting in which the benchmark exercise took place is highly relevant to geotechnical practice. Site characterization is a fundamental step in all geotechnical projects. In addition, CPT soundings have traditionally been used in assessing geotechnical parameters of soils and deriving the stratigraphy of the subsoil. Apart from inferring CPT parameters, DDSC methods are perceptive of the spatial counterpart of CPT data, so they can address the spatial variability of CPT parameters. Thus, the benchmark exercise can stimulate fruitful developments that are meaningful to geotechnical practice.

This paper approaches the benchmark example from a Bayesian perspective by presenting two methods within a Bayesian site characterization framework (BaySiC), building further on the work of Mavritsakis et al. (2022). Opting for Bayesian statistics holds significant advantages. First, Bayesian inference techniques are effective in thoroughly exploring complex probability domains, leading to more effective quantification of uncertainty. Second, Bayesian statistics can support inference even with small training data sets, a feature that bears value as long as the associated uncertainty is properly quantified in the results. Also, Bayesian statistics aligns well with the geotechnical engineering mindset (Baecher 2017).

This study used random fields (RFs) for modelling the soil domain. RFs represent random variables per point of the domain (Vanmarcke 2010) and can be used to model spatially variable parameters. The RF can be conditioned to observations of the variables, a feature that enables inference and prediction using spatial data. Moreover, the joint distribution of a subset of the variables can be derived by marginalizing the RF over the rest of the variables.

[1]Deltares, Boussinesqweg 1, Delft 2629 HV, Netherlands (corresponding author). ORCID: https://orcid.org/0000-0001-6784-9867. Email: Antonis.Mavritsakis@deltares.nl

[2]Deltares, Boussinesqweg 1, Delft 2629 HV, Netherlands; Dept. of Hydraulic, Delft Univ. of Technology, Stevinweg 1, Delft 2628 CD, Netherlands. Email: Timo.Schweckendiek@deltares.nl

[3]Deltares, Boussinesqweg 1, Delft 2629 HV, Netherlands. Email: Ana.MartinsTeixeira@deltares.nll

[4]Deltares, Boussinesqweg 1, Delft 2629 HV, Netherlands. Email: Eleni.Smyrniou@deltares.nl

[5]Deltares, Boussinesqweg 1, Delft 2629 HV, Netherlands. Email: Jonathan.Nuttall@deltares.nl

Note. This manuscript was submitted on July 31, 2022; approved on November 17, 2022; published online on February 6, 2023. Discussion period open until July 6, 2023; separate discussions must be submitted for individual papers. This paper is part of the ***ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering***, © ASCE, ISSN 2376-7642.

Last, RFs for different parameters can be correlated by compiling their joint distribution, which is enabled by introducing cross-correlation hyperparameters (Zhu et al. 2017).

This paper aimed to provide a solution to the DDSC problem presented by the benchmark example by adopting Bayesian statistics and is structured as follows. The "Methodology" section formulates the Bayesian models for inference and prediction. Moreover, it provides the statistical formulations for (1) the evaluation of the metrics defined by the benchmark example; and (2) hypothesis testing the predictive power of the models. The "Implementation in the Benchmark Exercise" section starts by introducing the setting of the benchmark example. It then discusses the implementation of two different Bayesian models in one of the benchmark cases with the purpose of exploring the behavior and evaluating their performance. In the section "Results for All Benchmark Cases," the results for all benchmark cases are examined to determine the effect of problem complexity on the performance of the models. Last, "Conclusions" draws conclusions regarding the operation and performance of BaySiC.

Methodology

Random Field Modelling

In the context of the benchmark exercise, BaySiC adopts cross-correlated RF modelling in order to infer and predict the considered CPT parameters, q_t and f_s. The CPT parameter values per point, whose three-dimensional (3D) coordinates are indicated by the vector $\mathbf{z}$ ($z \in R^3_+$), are denoted by q_t (z) and f_s (z), respectively. The variables of the RFs are put together into the vector of the parameters per point, denoted by $Y(z)$ [Eq. (1)]. As suggested by Geyer et al. (2021), observations (Y_m) per measurement point (of coordinates z_m) can be described by the RF variable of the point plus a measurement error e, as indicated in Eq. (2). In this study, the measurement error was considered negligible. This means that the observation Y_{m_i} per domain point equals the point-specific distribution $Y(z_{m_i})$ of the RF variables

$$Y(z) = [q_t(z), f_s(z)]^T \tag{1}$$

$$Y_{m_i} = Y(z_{m_i}) + e \tag{2}$$

When jointly modelling RFs, the following hyperparameters are used: the mean, standard deviation, and autocorrelation of each RF, as well as the cross-correlation between them. In this case, the joint distribution is given by Eq. (3) (Ching and Phoon 2019), which assumes that the CPT parameters follow a multivariate normal distribution (Phoon et al. 2022b). For simplicity, both RFs are assumed to follow the same spatial variability pattern. In detail, μ is the vector containing the mean per parameter ($\mu = [\mu_{q_t}, \mu_{f_s}]$), Σ is the covariance matrix of the parameters given by Eq. (4), and ρ is the cross-correlation coefficient between the CPT parameters. The symbol 1_n simply represents an $nx1$ vector of ones, where n is the number of locations in the RF. The implementation takes advantage of the Kronecker product (indicated by the symbol $\otimes$) to combine the autocorrelation of the field and the cross-correlation of the parameter into the expression of the joint distribution for the cross-correlated RF. Last, the approach assumes stationarity, which practically means that RFs are modelled with constant means and the autocorrelation structure is the same for all points (Mariethoz and Caers 2014)

$$Y(z) \sim N(1_n \otimes \mu, C \otimes \Sigma) \tag{3}$$

$$\Sigma = \begin{bmatrix} \sigma_{q_t} & 0 \\ 0 & \sigma_{f_s} \end{bmatrix} \times \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \times \begin{bmatrix} \sigma_{q_t} & 0 \\ 0 & \sigma_{f_s} \end{bmatrix} \tag{4}$$

The autocorrelation matrix between a subset of the domain points is calculated through an autocorrelation function R and the autocorrelation lengths (θ) in each direction [Eq. (5)]. Also, q_t and f_s share the same autocorrelation structure, meaning that their RFs are described by the same autocorrelation lengths. Because the benchmark exercise considers horizontal spatial variability isotropic, θ has only a vertical and one horizontal component: $\theta = [\theta_h, \theta_v]^T$

$$C = R(\theta, z) \tag{5}$$

The implementation described previously suggests that the RF model is similar to a Gaussian process regression (GPR) (Rasmussen and Williams 2006) for two cross-correlated outputs, a concept that is akin to the regression method of co-Kriging (Helterbrand and Cressie 1994). These two concepts are not used later in the paper, even though the RF model strongly resembles them.

The RF model is described by a set of hyperparameters: the mean and standard deviation per CPT parameter, the cross-correlation between them, and the autocorrelation lengths. The hyperparameters are collected into the hyperparameter vector $\mathbf{X}$ [Eq. (6)]

$$\mathbf{X} = [\mu_{q_t}, \mu_{f_s}, \sigma_{q_t}, \sigma_{f_s}, \theta_v, \theta_h, \rho]^T \tag{6}$$

Bayesian Inference

The RF jointly models the parameter distributions $Y(z)$ per point of the domain, and at the same time, its behavior on the domain level is determined by a set of hyperparameters X. Bayesian inference aims at identifying X by conditioning the RF on the observations Y_m at the measurement points. Because the Bayesian model focuses on inferring the hyperparameter vector of the RF, $\mathbf{X}$ will be denoted as the model parameter vector in the continuation of the study.

Bayesian inference employs Bayes' theorem [Eq. (7)] in order to draw conclusions on the random variable (X) by conditioning on observations Y_m (Gelman et al. 2013). Bayesian inference entails the setup of a statistical model that distinguishes the sources of epistemic (reducible) and aleatory (irreducible) uncertainty present in the examined problem. Also, Bayesian inference aims to reduce epistemic uncertainty and provide a better description of aleatory uncertainty. Key components of every Bayesian model are the prior distribution $P(X)$ and the likelihood function $L(Y_m|X)$. The former expresses initial information on X, whereas the latter evaluates the accuracy of X in describing the observations. Bayes' theorem combines the two components into the posterior distribution $P(XY_m)$, which reflects the updated knowledge on X

$$P(X|Y_m) = \frac{L(Y_m|X)P(X)}{\int L(Y_m|X)P(X)dX} \tag{7}$$

According to Gelman et al. (2013), Bayesian inference is performed in three steps. First, a full probabilistic model is formulated, which jointly models all quantities of the problem. This step comprises the definition of the prior distribution and the likelihood function of the model. Second, the model is conditioned on the observations in order to derive the posterior distribution. Last, the predictive model is set up, allowing for prediction and evaluation of the fit achieved by inference. The following sections describe how

these steps are applied within BaySiC toward solving the benchmark exercise.

Formulation of the Bayesian Model

Prior Distributions

The prior distribution of the model parameters reflects initial knowledge on the problem. However, the notion of initial knowledge is not so relevant in an artificial case such as the benchmark example because the geotechnical background of the examined case is missing. Some prior knowledge was available for the model parameters. For example, it was known that all parameters but ρ should be nonnegative and that ρ was bound in the range $[-1, 1]$. Also, the alternative of adopting noninformative prior distributions was rejected because the use of specific prior distributions for computational convenience did not constitute a strong rationale, especially when performing Bayesian inference with modern Markov chain Monte Carlo (MCMC) methods. Therefore, the study opted for the use of weakly informative prior distributions.

In the context of this analysis, weakly informative prior distributions are designed to bound the model parameters in reasonable ranges of values without conveying significant information regarding the parameters themselves. Such boundaries were described at the beginning of this section. Also, BaySiC assumes that the model parameters are a priori uncorrelated. Thus, instead of using one prior distribution to model initial knowledge, a set of prior distributions can be employed. For the mean and standard deviation parameters, this is achieved by adopting half-normal distributions that are truncated at zero. The used variance of those priors is set to be orders of magnitude greater than the expected values of the model parameters. The prior distribution for ρ is set to be a beta distribution with parameters $a = 2$, $b = 2$, modified to support $\rho \in [-1, 1]$. Last, uniform distributions are assigned to the autocorrelation lengths, which means that no specific preference is attributed to any values of θ. The formulation of the prior distributions per model parameter is given by Table 1.

By adopting weakly informative prior distributions, the Bayesian model gives greater weight to the likelihood function and so inference is dominated by the data. This approach leads to a solution of the benchmark problem that minimizes the influence of the subjectiveness of the analyst in the end results. In a real-case application, informative priors should be adopted when possible.

Likelihood Function

So far, the components of the RF generation and the role of the elements of the model parameter vector $\mathbf{X}$ are defined. The latter is subject to inference by Bayesian updating, which means that the formulation of the likelihood function must describe each measurement Y_{m_i} using $\mathbf{X}$.

First, all measurements are collected in the measurement matrix Y_m per Eq. (8). As explained, the relationship between q_t and f_s of all domain points is described by a multivariate normal distribution. At this point, it is acknowledged that a distribution that disables negative values (such as the lognormal) would have been more suitable, but the multivariate normal is adopted as in Phoon et al. (2022b) for the sake of result comparability within the context of the benchmark. In order to account for both covariance between CPT parameters Σ and autocorrelation between points C_m, the covariance of said distribution is the Kronecker product of the two components Σ_{kron_m} [Eq. (9)]. Both components are functions of elements of $\mathbf{X}$, as given by Eqs. (4) and (5). Naturally, when deriving the formulation of the likelihood function, only measurement points of the RF are considered in the derivation of C_m. Last, the multivariate distribution is also parameterized by a mean vector. In this case, the mean vector μ_{kron_m} is derived by the Kronecker product of a vector of ones of size $mx1$ (1_m), where m is the number of training points) and the vector of means per CPT parameter [Eq. (15)]

$$Y_m = [Y_{m_1}, Y_{m_2}, \ldots, Y_{m_k}]^T \tag{8}$$

$$\Sigma_{\text{kron}_m} = C_m \otimes \Sigma \tag{9}$$

$$\mu_{\text{kron}_m} = 1_m \otimes [\mu_{q_t}, \mu_{f_s}] \tag{10}$$

The components defined here, which are all functions of $\mathbf{X}$, control the mean and uncertainty of the RF. The likelihood function takes the form of the probability density function (PDF) of the multivariate normal distribution (Gut 2009), but it is parameterized by the components defined previously. Ultimately, the likelihood function of $\mathbf{X}$ given the measurements Y_m is formulated as a multivariate distribution parameterized by said components, as given by Eq. (11)

$$L_{Y_m}(X) = \det(2\pi\Sigma_{\text{kron}_m})^{-\frac{1}{2}} \exp \times \left(-\frac{1}{2}(Y_m - \mu_{\text{kron}_m})^T \Sigma_{\text{kron}_m}^{-1}(Y_m - \mu_{\text{kron}_m}\right) \tag{11}$$

When defining the likelihood function, the matter of separability of variance and covariance is considered. According to Dutilleul (1999), inference explicitly addresses the covariance Σ_{kron_m}, which is a Kronecker product, but not its components, the individual autocorrelation and cross-correlation matrices, which are keys for predictions at the validation points. Essentially, inference will return combinations of C_m and Σ that lead to the same Σ_{kron_m}. The issue is also inflated by the lack of strong priors on the model parameters determining the components. Because Σ_{kron_m} is a separable covariance structure, the likelihood function can take the form of a matrix normal distribution (Gupta and Nagar 1999) [Eq. (12), modified for the examined situation]. In this case, the observations are put in an $kx2$ matrix denoted by Y'_m. Rows indicate different training points, whereas two columns represent the CPT parameters. Hence, the row correlation is C_m, and the column covariance is Σ. In this way, a distinction of the influence of the two components is achieved to some level. In this case, where the autocorrelation follows a pattern imposed by the autocorrelation function, opting for the matrix normal distribution does not lead to reduction of the model parameters. Ultimately, the matrix normal likelihood function is adopted in the analysis. The graph of the Bayesian model formulated in the section "Methodology" is illustrated in Fig. 1

$$L_{Y'_m}(X) = \frac{\exp\left(-\frac{1}{2}\text{tr}[\Sigma^{-1}(Y'_m - \mu)^T C_m^{-1}(Y'_m - \mu)]\right)}{(2\pi)^k |\Sigma|^{k/2} |C_m|} \tag{12}$$

Table 1. Prior distribution per model parameter

Model parameter	Prior distribution type	Distribution parameters
μ_{q_t}	Half normal	$\mu = 0$, $\sigma = 10$
μ_{f_s}	Half normal	$\mu = 0$, $\sigma = 100$
σ_{q_t}	Half normal	$\mu = 0$, $\sigma = \sqrt{10}$
σ_{f_s}	Half normal	$\mu = 0$, $\sigma = 10$
ρ	$\text{Beta}_{[-1,1]}$	$a = 2$, $b = 2$
θ_v	Uniform	$a = 0$, $b = 10$
θ_h	Uniform	$a = 0$, $b = 200$

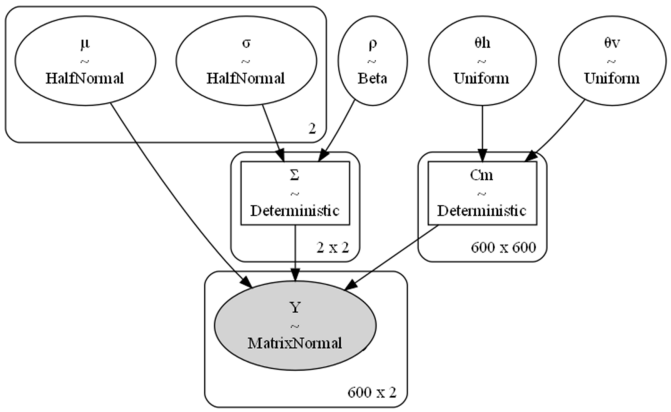

Fig. 1. Graph of the two-parameter Bayesian model.

Probabilistic Model Conditioning

The Bayesian inference problem is solved by conditioning the Bayesian model on the available observations. Although several techniques are available for applying Bayes' theorem and deriving the posterior distribution, many of them are unable to deal with a large number of random variables or Bayesian models of greater complexity. This study adopts the Markov chain Monte Carlo method for performing inference.

This study used the Hamiltonian Monte Carlo (HMC) algorithm, a member of the MCMC algorithmic family. HMC simulates sampling from the posterior with a physics analogue; the sampler is represented as a solid whose motion through the variable domain is described with Hamiltonian mechanics (Neal 2011). The gradient of the log-posterior to the random variables is central to the operation of HMC. Its calculation can be computationally exhausting in traditional programming paradigms but is undertaken by modern statistical packages. Aside from that, HMC is efficient in terms of sampling because it exhibits competency in approximating the typical set of the posterior distribution (Betancourt 2017). Furthermore, it is able to achieve greater acceptance rates than traditional MCMC algorithms (Wang et al. 2019) and reduces the correlation between successive samples, mitigating the MCMC issue of burn-in period (Meyn and Tweedie 1993). This study used the no U-turn sampler (NUTS) variation of HMC for its efficiency because it has been shown to typically require fewer evaluations of the likelihood function (Hoffman and Gelman 2014).

Probabilistic programming (PP) is the programming archetype that enhances the development of probabilistic models and enables automatic inference (van de Meent et al. 2018). The Python PP package PyMC3 (Salvatier et al. 2016) is adopted for probabilistic modelling. According to the PP paradigm, orchestrating a Bayesian model in PyMC3 requires only the definition of the prior distribution of the random variables and the formulation of the likelihood function. Moreover, PyMC3 allows for efficient HMC sampling through the use of automatic differentiation (AD). PyMC3 translates the operations of the Bayesian model in graph structure, breaking them down to simple mathematical operations. By application of the chain rule, AD is able to efficiently calculate the gradient of the log-posterior (Rall 1981), as required by the HMC sampler. Even though the use of the gradient by HMC posed as a computational obstacle, PyMC3 is able to overcome it and fully exploit the efficiency of HMC.

Predictive Model

After conditioning the Bayesian model to the data of the measurement points, the posterior distribution of the model parameter vector $\mathbf{X}$ is retrieved. Through a predictive model, the posterior can be used to make predictions of q_t and f_s at the prediction points of the soil domain. The multivariate normal distribution is used in the predictive model for its computational ease. The matrix normal distribution was adopted for inference because it displayed advantages over its multivariate normal counterpart. However, for prediction it is equivalent to a multivariate normal distribution, whose covariance is the Kronecker product of the autocovariance and cross-covariance.

This multivariate normal distribution is again parameterized by $\mathbf{X}$. Its autocorrelation matrix poses as the means of connecting the measurement and prediction points. Eq. (13) provides an update of the autocorrelation matrix that includes all points of the RF. Essentially, the greater autocorrelation matrix can be broken down to four submatrices: C_{mm} connects the measurement points, C_{pp} connects the prediction points and C_{mp} and C_{pm} establish a connection between measurement and prediction points. Developing further, the covariance matrix of the multivariate normal distribution between all points of the RF is the Kronecker product between the cross-covariance matrix S and the new autocorrelation matrix [Eq. (14)]. Inherited by C, similar submatrices exist for Σ_{kron}. Last, the mean of the multivariate normal μ_{kron} distribution is the Kronecker product of a vector of ones of size $m + p$ (1_{m+p}), where p is the number of the prediction points in the RF and the vector of global means [Eq. (15)]. The mean vector can be divided in a part of size $mx2$ for the measurement points and one of size $px2$ for the prediction points, even though the values of the mean per parameter are the same. Ultimately, the PDF of the multivariate normal distribution f for all variables Y of the RF is given by Eq. (16) and is a function of $\mathbf{X}$

$$C = R(\theta, z) = \begin{bmatrix} C_{\text{mm}} & C_{\text{mp}} \\ C_{\text{pm}} & C_{\text{pp}} \end{bmatrix} \tag{13}$$

$$\Sigma_{\text{kron}} = C \otimes \Sigma = \begin{bmatrix} \Sigma_{\text{kron}_{mm}} & \Sigma_{\text{kron}_{mp}} \\ \Sigma_{\text{kron}_{pm}} & \Sigma_{\text{kron}_{pp}} \end{bmatrix} \tag{14}$$

$$\mu_{\text{kron}} = 1_{m+p} \otimes [\mu_{q_t}, \mu_{f_s}] \tag{15}$$

$$f(Y|X) = \det(2\pi\Sigma_{\text{kron}})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(Y - \mu_{\text{kron}})^T \Sigma_{\text{kron}}^{-1}(Y - \mu_{\text{kron}}\right) \tag{16}$$

In order to make predictions of Y at the prediction points, the multivariate distribution needs to be conditioned to the measurements. According to Marriott and Eaton (1984), the conditional distribution of all $Y_{p_j} = Y(z_{p_j})$, $j = 1, 2, \ldots, p$ at the prediction points is again multivariate normal and depends on the values of Y_m at the measurement points. Its mean $\mu_{p|Y_m}$ and covariance $\Sigma_{p|Y_m}$ are estimated by Eqs. (17) and (18), respectively. The conditional distribution of the CPT parameters at the prediction points Y_p is a function of $\mathbf{X}$, and the PDF $f_{p|Y_m}(X)$ is given by Eq. (19)

$$\mu_{p|Y_m} = \mu_p + \Sigma_{\text{kron}_{\text{pm}}} \Sigma_{\text{kron}_{\text{mm}}}^{-1}(Y_m - \mu_m) \tag{17}$$

$$\Sigma_{p|Y_m} = \Sigma_{\text{kron}_{\text{pp}}} - \Sigma_{\text{kron}_{\text{pm}}} \Sigma_{\text{kron}_{\text{mm}}}^{-1} \Sigma_{\text{kron}_{\text{mp}}} \tag{18}$$

$$f_{p|Y_m}(Y_p|X) = \det(2\pi\Sigma_{p|Y_m})^{-\frac{1}{2}} \exp \times \left(-\frac{1}{2}(Y_p - \mu_{p|Y_m})^T \Sigma_{p|Y_m}^{-1}(Y_p - \mu_{p|Y_m}\right) \tag{19}$$

Following, Eq. (20) conditions the parametrization of the predictive model to the observations by marginalizing the conditional distribution $f_{p|Y_m}$ over the posterior distribution of X $[P(X|Y_m)]$. Additionally, because the posterior distribution is only known through the samples collected by the HMC sampler (of number N), the predictive model is also formulated in a discretized form by Eq. (21) (Krüger et al. 2016). Essentially, the predictive model can be employed for the generation of RFs that are conditional to the data at the measurement points

$$P(Y_p|Y_m) = \int f_{p|Y_m}(Y_p|X)P(X|Y_m)dX \tag{20}$$

$$P(Y_p|Y_m) = \frac{1}{N}\sum_{i=1}^{N} f_{p|Y_m}(Y_p|X_i) \tag{21}$$

Bayes Estimators

Representative point estimates of the posterior can be derived by minimizing Bayes risk (Murphy 2022). Bayes risk is defined as the expectation of a loss function (l) over the posterior [Eq. (22)]. The loss function quantifies a distance metric between the selected estimator $\hat{X}$ and any value of X. The Bayes estimator connected to a distance metric is the value of X that minimizes the Bayes risk [Eq. (23)]

$$\mathrm{BR}(X) = E_{P(X|Y_m)}[l(X,\hat{X})] = \int l(X,\hat{X})P(X|Y_m)dX \tag{22}$$

$$\hat{X} = \mathrm{argmin}(\mathrm{BR}) \tag{23}$$

Widely adopted loss functions are the mean square error (MSE) and the 0–1 loss function. In the first instance, the estimator that minimizes Bayes risk is the expectation of the posterior [Eq. (24)]. This estimator is expected to provide the minimum risk on average. For the case of the 0–1 loss function, the optimizer of Bayes risk is known as the maximum a posteriori (MAP), or simply the mode of the posterior [Eq. (25)], that is, the most likely M according to the posterior (Lehmann and Casella 1998)

$$\hat{X}_{\mathrm{MSE}} = E_{P(X|Y_m)}[X] \tag{24}$$

$$\hat{X}_{\mathrm{MAP}} = \mathrm{argmax}(P(X|Y_m)) \tag{25}$$

Prediction Evaluation Metrics

The benchmark exercise provides a set of q_t and f_s values for all points in the subsoil domain, which means that apart from the observations at the measurement points Y_m, a validation data set at the prediction points Y_p is given. The latter can be used for the validation of the efficacy of BaySiC by comparing Y_p with the predictions at the prediction points, much like a cross-validation scheme. Phoon et al. (2022b) arbitrated the effectiveness of their approach in terms of two loss-based metrics, and so it is convenient to adopt the same metrics for the sake of approach comparability.

The first metric is the root mean squared error (RMSE) of the q_t, which is a form of quadratic loss function. The RMSE is a means of assessing the average distance between the cone resistance of the validation data set Y_{p,q_t} and the prediction per location [Eq. (26)]. Because the RMSE does not consider the uncertainty of the prediction, an estimator $\hat{Y}_{p,q_t}$ of the q_t predictive distribution is used in the evaluation of the RMSE per prediction point

$$\mathrm{RMSE} = \sqrt{\frac{1}{p}\sum_{i=1}^{p}(Y_{p,q_t} - \hat{Y}_{p,q_t})} \tag{26}$$

The second metric is the identification rate (IR), which acts as a 0–1 loss function and estimates the average number of prediction points where the predicted soil type matches the one estimated by the validation data. The soil type is indicated by the soil behavior type (SBT), as introduced by Robertson (2016). The SBT is determined from I_c, which in turn is calculated based on the predictions of q_t and f_s [Eq. (27)]. Following, the IR is quantified by Eq. (28), with $1(z_{p_i})$ being the indicator function that returns 1 when the SBT prediction at the prediction point z_{p_i} matches the SBT of the validation data and returns 0 otherwise. As with the RMSE calculation, the I_c and SBT are estimated for an estimator $\hat{Y}$ of the predictive distribution per prediction point

$$I_c = I_c(q_t, f_s) = I_c(Y) \tag{27}$$

$$\mathrm{IR} = \frac{1}{k}\sum_{i=1}^{k} 1(z_{p_i}) \tag{28}$$

As mentioned, both the RMSE and IR are evaluated for the estimators $\hat{Y}$ derived from the predictive distribution per prediction point. The selected estimators are connected to the posterior estimators introduced in the "Methodology" section. Assuming that $\hat{X}$ represents either the MSE or the MAP estimator ($\hat{X}_{\mathrm{MSE}}$ or $\hat{X}_{\mathrm{MAP}}$, respectively), then a plug-in of the posterior distribution can be given by a Dirac delta function δ, as in Eq. (29) (Murphy 2022). The plug-in can be used in Eq. (20) to derive the predictive distribution for a specific value of the posterior. Ultimately, the predictive distribution after plug-in takes the form of Eq. (30). Even though the full posterior distribution appears not to contribute in prediction anymore, the derivation of $\hat{X}$ heavily depends on the posterior. Consequently, the expectation of the plug-in predictive distribution is selected as the estimator $\hat{Y}$ for the evaluation of the metrics [Eq. (31)]. $\hat{Y}$ is the vector of conditioned means as defined by Eq. (17), evaluated at $X = \hat{X}$. Because the predictive distribution is normal, $\hat{Y}$ is both the mean and mode of the prediction. A version of Y exists for each variation of $\hat{X}$, which are labeled as $\hat{Y}_{\mathrm{MSE}}$ and $\hat{Y}_{\mathrm{MAP}}$ for the MSE and MAP estimators, respectively. This method lets the fully Bayesian approach adopted by BaySiC degenerate to a point estimate in the sake of for comparability and simplicity

$$P_{\text{plug-in}}(X|Y_m) = \delta(X - \hat{X}) \tag{29}$$

$$P(Y_p|Y_m) = \int f_{p|Y_m}(Y_p|X)\delta(X - \hat{X})dX = f_{p|Y_m}(Y_p|\hat{X}) \tag{30}$$

$$\hat{Y} = E[Y|X = \hat{X}] = \int f_{p|Y_m}(Y|\hat{X})dY \tag{31}$$

Last, both metrics are expected to incorporate the impact of both observed variables in an explicit or even implicit fashion. For example, the RMSE is evaluated for the cone resistance, but cone resistance predictions are influenced by the sleeve friction via the correlation between the two.

Bayesian Hypothesis Testing

Even though estimation of metrics in a cross-validation setting is already applied in order to assess the accuracy of BaySiC, hypothesis testing is adopted because it can lead to generalized insights regarding the predictive power of the framework. Hypothesis

testing is the statistical procedure of examining whether the available data can reject a particular hypothesis with sufficient confidence (Wasserman 2004). In a simple setting, two competing and collectively exhaustive hypotheses are compared, with H_o traditionally indicating the status quo, whereas H_1 represents new insights gained by the data. Hypothesis testing can either reject H_o or fail to reject, with none of the possible outcomes indicating the credibility of H_1. Hypotheses are compared by the means of a test statistic Ψ. Frequentist statistics offers a variety of options for Ψ, but this study chose to construct it from a Bayesian perspective.

In order to dismiss the suggestion that the patterns identified though Bayesian inference are not credible, hypothesis testing is focused on random variables of X that control the predictive power of the models. Such variables are the cross-correlation coefficient ρ and the vertical and horizontal autocorrelation lengths (θ_v and θ_h), which define cross-correlation and autocorrelation, respectively. The status quo accepts that no correlation exists. In order to establish the cross-correlation between q_t and f_s, ρ should be nonzero and thus is tested by Eq. (32). Because autocorrelation is strictly nonnegative, θ_v and θ_h are tested to be greater than cut-off values ($\theta_v^{\min}$ and $\theta_h^{\min}$, respectively) that lead to almost zero autocorrelation between points [Eqs. (33) and (34)]. The cut-off values are defined based on the distances between points of the domain in the following section

$$H_0{:}\rho = 0, \quad H_1{:}\rho \neq 0 \tag{32}$$

$$H_0{:}\theta_v < \theta_v^{\min}, \quad H_1{:}\theta_v \geq \theta_v^{\min} \tag{33}$$

$$H_0{:}\theta_h < \theta_h^{\min}, \quad H_1{:}\theta_h \geq \theta_h^{\min} \tag{34}$$

The current goal of hypothesis testing is to constrain the probability of Type I error, that is, the probability that H_0 is true but gets rejected. Typically, a significance level of 5% is adopted as the constraint for the Type I error. Given that Ψ is a suitable test statistic, the constraint of Type I error probability is expressed by Eq. (35)

$$P(\text{TypeI}) = P(\Psi = 1|H_0) \leq 0.05 \tag{35}$$

As suggested by Kruschke (2013), Bayesian statistics can provide a powerful testing alternative to frequentist testing. The 95% highest density interval (HDI) is the neighborhood of the X posterior distribution that possesses two properties: first, every X in the HDI has greater probability density that any point outside, and second, the probability mass of the HDI is equal to 0.95. The suggested test consists of checking whether H_0 is included in the 95% HDI. Furthermore, Kruschke (2018) expands by defining the region of practical equivalence (ROPE), the neighborhood of H_0 that practically yields the same effect as H_0 on the predictive power of the model. Then, hypothesis testing consists of checking whether the ROPE and HDI overlap. In case of no overlap, H_0 can be rejected with sufficient confidence. It is assumed that no correlation is reflected with a correlation coefficient in the range of $(-0.3, +0.3)$, which constitutes the ROPE for ρ. Also, the ROPE for the autocorrelation lengths is of the form $\text{ROPE}_\theta = [0, \theta^{\min}]$. The hypothesis tests take the forms given by Eqs. (36)–(38), where 1 is the indicator function returning 1 if the expression in the parenthesis is true and 0 otherwise. If the tests return $\Psi = 1$ using the 95% HDI, then H_0 is rejected with confidence 0.95

$$\Psi_\rho = 1(\text{ROPE}_\rho \cap \text{HDI}_\rho = \emptyset) \tag{36}$$

$$\Psi_{\theta_v} = 1(\text{ROPE}_{\theta_v} \cap \text{HDI}_{\theta_v} = \emptyset) \tag{37}$$

$$\Psi_{\theta_h} = 1(\text{ROPE}_{\theta_h} \cap \text{HDI}_{\theta_h} = \emptyset) \tag{38}$$

whereas the selected hypothesis tests examine each component of covariance individually, hypothesis testing for the predictive power of the model is required on the final correlation matrix of Eq. (14), which is a product of the cross-correlation and autocorrelation. Clearly, rejecting the null hypothesis for either of the components does not guarantee rejection of the equivalent null hypothesis formulated for the final correlation matrix. Therefore, all tests should be performed simultaneously (in the sense of multiple hypothesis testing), and the total probability of Type I error should be below the adopted significance level. However, a different, tailor-made hypothesis testing approach was adopted in this study that aimed to check for the predictive power of the specific model.

In this instance, where data on both q_t and f_s are available at all training locations, the cross-correlation coefficient does not affect the predictive radius of the model, the maximum distance between training and validation points for which the model can establish sufficient correlation (greater than 0.3). The cross-correlation coefficient adjusts the relationship between predictions of q_t and f_s at validation points, a feature that can be of high importance for predicting the soil type and evaluating competent metrics. Hence, hypothesis testing for ρ can be performed independently from θ_v and θ_h. Also, isolating the cross-correlation coefficient for hypothesis testing can lead to conclusions on the relationship between q_t and f_s that bear geotechnical value. The predictive radius is determined by the autocorrelation matrix, and performing hypothesis testing on it is equivalent to testing for the combined contribution of θ_v and θ_h in the model. Such a test can be performed by constructing the posterior of the autocorrelation matrix using the posterior samples of the scales of fluctuation.

Implementation of a Single-Parameter Model

The "Methodology" section described how inference and prediction are achieved when using a two-parameter model for q_t and f_s. Here is a simpler version of the model for a single parameter, which is used later in the analysis.

The single-parameter model creates an RF for only one parameter. Thus, it is similar to the two-parameter model but uses a reduced version of the model parameter vector X' [Eq. (39)] by removing the mean and standard deviation of the second parameter, as well as the cross-correlation coefficient. No specific names are given to the hyperparameters because this model is intended for inference of any parameter

$$X' = [\mu, \sigma, \theta_v, \theta_h]^T \tag{39}$$

In the "Methodology" section, the matrix normal distribution was applied by making a distinction between the autocovariance and cross-covariance matrices. Because the latter is not present in this setup, the matrix normal distribution can be used as the likelihood function of the single-parameter model to distinguish between the vertical and horizontal autocorrelation matrices (C_{v_m} and C_{h_m}, respectively) [Eq. (40)]. These matrices are derived by applying the autocorrelation function R separately per dimension. Also, C_{h_m} is multiplied by σ^2 and is considered a covariance matrix. The observations are reorganized in columns per CPT, with rows being different depths, and thus the variable matrix Y'_m is of size

$n_{row} \times n_{col}$. The graph of the single-parameter model is presented in Fig. 2

$$L_{Y'_m}(X') = \frac{\exp\left(-\frac{1}{2}\mathrm{tr}[C_{h_m}^{-1}(Y'_m - \mu)^T C_{v_m}^{-1}(Y'_m - \mu)]\right)}{(2\pi)^{\frac{n_{row}n_{col}}{2}}|C_{h_m}|^{\frac{n_{row}}{2}}|C_{v_m}|^{\frac{n_{col}}{2}}} \quad (40)$$

Similarly to the two-parameter model, prediction can still be achieved using a multivariate distribution whose covariance is the Kronecker product $C_{v_m} \otimes C_{h_m}$. Adjusting the description from the "Methodology" section to the single-parameter model, conditioning of the multivariate normal distribution to the observations and the derivation of the predictive model is achieved.

The mathematical devices explained in the "Methodology" section can be modified for the single-parameter model. Thus, it is considered that the Bayes estimators, metric evaluation, and hypothesis test are already available for the single-parameter model.

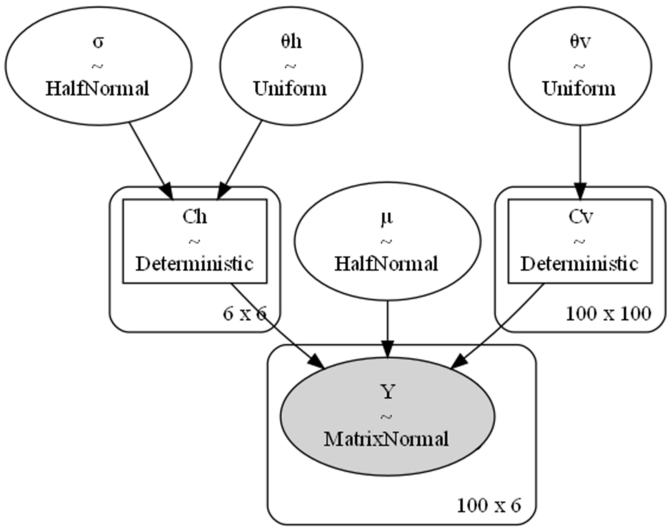

Fig. 2. Graph of the single parameter Bayesian model.

Implementation in the Benchmark Exercise

Introduction to the Benchmark Exercise

The design of the benchmark exercise was introduced in Phoon et al. (2022b). The exercise is divided in four different cases (S-VG1 to S-VG4), with each one having a different stratigraphy of the subsoil domain always composed of clay, sand, and silt layers, as shown in Fig. 2 of Phoon et al. (2022b). Fig. 3 illustrates the training and validation locations of the CPT soundings in the domain. The exercise allows inference using two sets of training CPT locations of different sizes; set T1 includes only three CPT soundings, whereas set T2 includes six. The validation set includes 12 CPT soundings. The training and validation locations of the CPTs are common between all stratigraphic cases. Combining the stratigraphic cases with the different training sets, the exercise provides eight cases for solving. Moreover, Table 2 presents the actual parameter used in generating the benchmark RFs.

In the current section, a first application of BaySiC is presented in order to showcase practical details about its operation and performance. Hence, it is sensible to use BaySiC in the simplest case among all combinations: stratigraphic case SVG-1, which is composed of horizontal soil layers without any depth trend, and training

Table 2. Parameter per layer type used in generating the data per case

Input parameter	Sand	Clay	Silt
μ_{q_t}(MPa)	9.00	2.00	55.00
μ_{f_s}(kPa)	110.00	75.00	85.00
σ_{q_t}(MPa)	2.16	0.40	1.32
σ_{f_s}(kPa)	26.40	15.00	20.40
θ_v(m)	1.00	1.20	1.00
θ_h(m)	10.00	15.00	10.00
ρ	0.70	0.70	0.70

Source: Data from Phoon et al. (2022b).

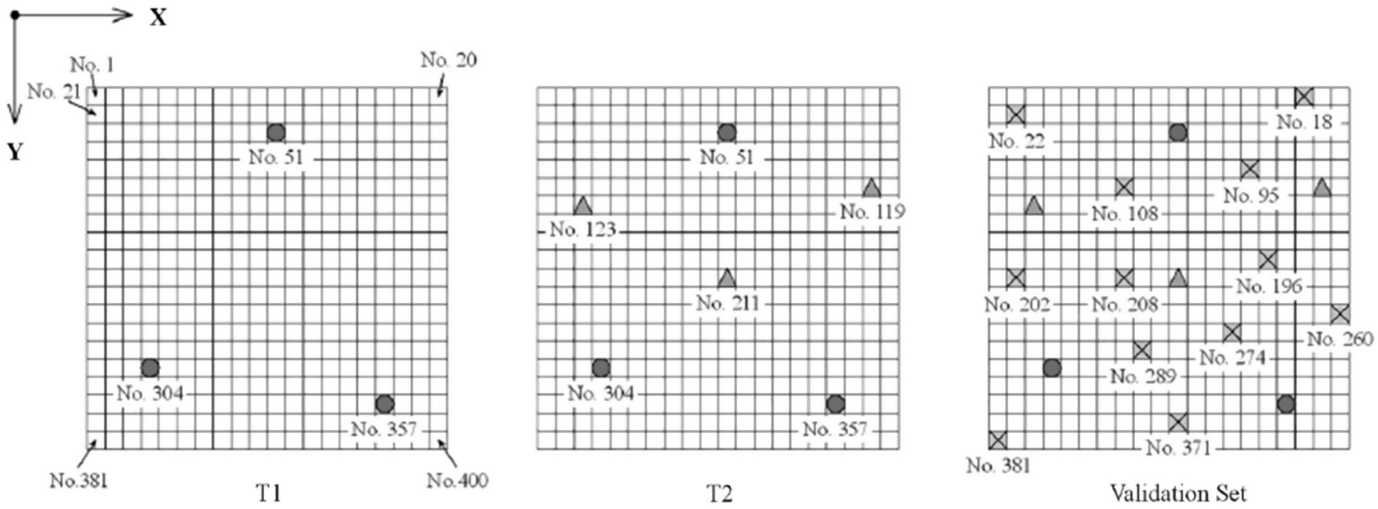

Fig. 3. Locations of the training and validation locations of the CPT soundings in the top view of the domain. ("Benchmark examples for data-driven site characterisation," K.-K. Phoon, T. Shuku, J. Ching, and I. Yoshida, *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, © 2022, reprinted by permission of Taylor & Francis Ltd., http://tandfonline.com.)

ASCE-ASME J. Risk Uncertainty Eng. Syst., Part A: Civ. Eng., 2023, 9(2): 04023008

set T2, which is larger and thus expected to convey more information (case: SVG-1-T2).

Strategy of Analysis

Aggregated Model

The aggregated model perceives a single soil material throughout the entire domain. In this approach, no distinction is made between layers, and BaySiC is applied as if a single soil layer exists. Although this model does not infer the CPT parameters per material, which means that the derived posterior distributions are not directly comparable to the distributions that were used in generating the data sets (Table 2), it is able to establish relationships between points of the domain and thus make predictions. Essentially, BaySiC chooses to infer and predict q_t and f_s as fundamental parameters that describe soil behavior and then derive I_c based on prediction of said parameters. This perspective aims to infer the correlation structures between q_t and f_s over the soil domain, which are expected to affect the determination of I_c.

The aggregated model was set up according to the "Methodology" section and aimed to infer all model parameters of X. Even though its fundamental assumption of a single layer was expected to increase model bias, the aggregated model managed to incorporate information of both q_t and f_s.

Split Model

The split model performs inference individually per soil material recognized at the training CPT locations. Thus, it infers the parameters of the Bayesian model for each soil formation separately and leads to results similar to the ones given by Table 2. However, this model suffers from a serious drawback; because the material type is unknown at the validation points, the framework is incapable of selecting the material whose parameters should be used, and so prediction is disabled. At this point, elements of the aggregated model can be used to avoid this pitfall.

The split model is adjusted to operate in two steps. In the first step, a single-parameter, aggregated model is used to infer and predict I_c and so the material type per validation point. Specifically, an aggregated model for I_c is set up using the single-parameter model from the "Methodology" section for performing inference and prediction. The I_c training data is determined by q_t and f_s data per training point. Through I_c, this step predicts the material type per validation point. In the second step of the split model, a two-parameter model is used to infer X per material type and predict q_t and f_s at every validation point. A two-parameter model is trained for each material recognized at the training points. Then, the posterior distribution of X for a material type is used to predict q_t and f_s at the validation locations where the first step of the split model predicts the same material.

The split model holds two conceptual benefits. First, it allows the derivation of posterior parameters per material, which makes sense for geotechnical engineering. Second, it uses the single-parameter model for soil type prediction, which is able to separate the contributions of the vertical and horizontal autocorrelations through the use of the matrix normal distribution as a likelihood function. The latter feature is expected to distinguish the influence of θ_v and θ_h more effectively, leading to enhanced prediction accuracy.

Last, another setup of the split layer model is explained. In case the single-parameter model can lead to greater prediction accuracy, it would be appropriate to use it in inferring q_t per layer in the second step of the split layer model. Such a device would solely aim to enhance prediction accuracy for q_t and achieve lower RMSE scores. However, the goal of this study is to present a robust framework with general applicability. This means that equal attention is paid to both q_t and f_s, and the results of inference are meaningful for geotechnical engineering. Therefore, this study decided not to adopt this approach, and the two-parameter model is used for the second step of the split layer model.

Analysis of SVG-1-T2

Model Tuning

The first step in analyzing case SVG-1-T2 is tuning the examined models by selecting an appropriate autocorrelation function. To that end, three kernel functions are tested: the Markov, Gaussian, and Matern 32 kernels, as given by Eqs. (41)–(43). The values of Δh and Δv are the distances between two points of the domain in the horizontal and vertical direction, respectively

$$R_{\text{Markov}}(\theta, z) = \exp\left(-\sqrt{\left(\frac{2\Delta v}{\theta_v}\right)^2 + \left(\frac{2\Delta h}{\theta_h}\right)^2}\right) \tag{41}$$

$$R_{\text{Gauss}}(\theta, z) = \exp\left(-\sqrt{\left(\frac{\Delta v}{2\theta_v}\right)^2 + \left(\frac{\Delta h}{2\theta_h}\right)^2}\right) \tag{42}$$

$$R_{\text{Matern32}}(\theta, z) = \left(1 + \frac{\sqrt{3\Delta v^2}}{\theta_v}\right)\left(1 + \frac{\sqrt{3\Delta h^2}}{\theta_h}\right) \times \exp\left(-\left(\frac{\sqrt{3\Delta v^2}}{\theta_v} + \frac{\sqrt{3\Delta h^2}}{\theta_h}\right)\right) \tag{43}$$

According to Murphy (2022), the most appropriate model $\hat{M}$ over all possible models M is the one that maximizes the likelihood of the model given the data [Eq. (44)]. Because all autocorrelation functions employ the same number of variables, using the likelihood as a comparison metric will yield the same result as adopting a metric that penalizes models for the number of their parameters. The metric is checked using the likelihood scores gathered during HMC sampling. Thus, Bayesian analysis is performed per autocorrelation function, and the achieved likelihoods of the associated $\hat{X}_{\text{MAP}}$ estimators are compared to identify the most fitting kernel

$$\hat{M} = \text{argmax}_{m \in M}[p(m|Y_m)] \tag{44}$$

The process described previously is performed for the aggregated and split models, and Table 3 presents the log-likelihood results per kernel function. The table shows the log-likelihood for the single-parameter I_c model used in the first step of the split model. Evidently, the Markovian kernel leads to the most descriptive aggregated model. For the split model, the Gaussian kernel marginally achieves the highest log-likelihood score. Even though the respective results are not presented here, the Markovian kernel is adopted for the second step of the split model.

Table 3. Log-likelihood score of the $\hat{X}_{\text{MAP}}$ for the two models model per autocorrelation function

Model	Autocorrelation function	Log-likelihood of the aggregated model	Log-likelihood of the split model
1	Markov	–3,183	162.9
2	Gauss	–3,268	163.5
3	Mattern 32	–3,711	163.1

Summary of the Results for SVG-1-T2

In the analysis of case SVG-1-T2, inference for both models is performed using four HMC chains, with each one retrieving 2,500 samples from the posterior distribution. Also, each chain draws 1,000 samples for warm-up that are discarded later. The runtimes for inference and prediction of the aggregated and split models are approximately 1,500 and 640 seconds, respectively. Clearly, BaySiC could easily handle a site characterization task of the same magnitude as case SVG-1-T2 in a real project setting. For reference, the analysis was performed on a laptop with an Intel i7 processor at 2.80 GHz and 32 GB of RAM.

Posterior Distribution Analysis

Fig. 4 provides a visualization of the posterior sample for the aggregated model. By examining the marginal histograms, located in the diagonal, it is evident that all variables follow a unimodal distribution. Moreover, the two Bayes factors, $\hat{X}_{\mathrm{MSE}}$ and $\hat{X}_{\mathrm{MAP}}$, lie close in the variable domain for all variables. In the cases of σ_{q_t}, σ_{f_s}, θ_v, and θ_h, more samples have been drawn at neighborhoods of higher values, leading to slight positive skewness. The same variables, which control the correlation between points of the domain, appear to be highly and positively correlated, as exhibited by their respective two-dimensional histograms. A strong positive correlation between σ_{q_t} and σ_{f_s} is present because of the cross-correlation coefficient ρ. Additionally, the standard deviation variables are strongly correlated to the scales of fluctuation. When the scales of fluctuation increase, the autocorrelation rises, and so likelihood scores that are accepted by the sampler are achieved even by samples of greater standard deviations. Furthermore, μ_{q_t} and μ_{f_s} are correlated because they are the means of a bivariate distribution between q_t and f_s, whose correlation is defined by ρ.

The posterior investigation was also performed for the split model. Fig. 5 visualizes the posterior of the first step, which infers model parameters for I_c. Again, θ_h is positively correlated to the standard deviation σ_{I_c}. Opposite the aggregated model, θ_h and θ_v are negatively correlated. Moreover, Fig. 6 presents the posterior histograms of θ_v, θ_h, and ρ per material derived by the second step. In this figure, $\hat{X}_{\mathrm{MSE}}$ is compared to the actual value that was used in generating the RF for the benchmark. The autocorrelation lengths are approximated sufficiently well by the posterior mean for all cases, except for θ_h for the silt material. Training points along the interface between the sand and clay layers are consistently classified as silt due to their SBT value, which complicates further the inference of the autocorrelation length for silt. Also, the cross-correlation coefficient is not estimated adequately for all materials. A possible cause for this can be limited data availability. For example, if inference is performed using a common cross-correlation coefficient for all materials (because all of them have the same actual ρ), then the posterior of approaches the actual value sufficiently, as validated by running BaySiC in the associated setting.

Hypothesis Testing of Inference Results

After examining the posterior distributions, hypothesis testing was performed in order to draw insights regarding the predictive power of the model.

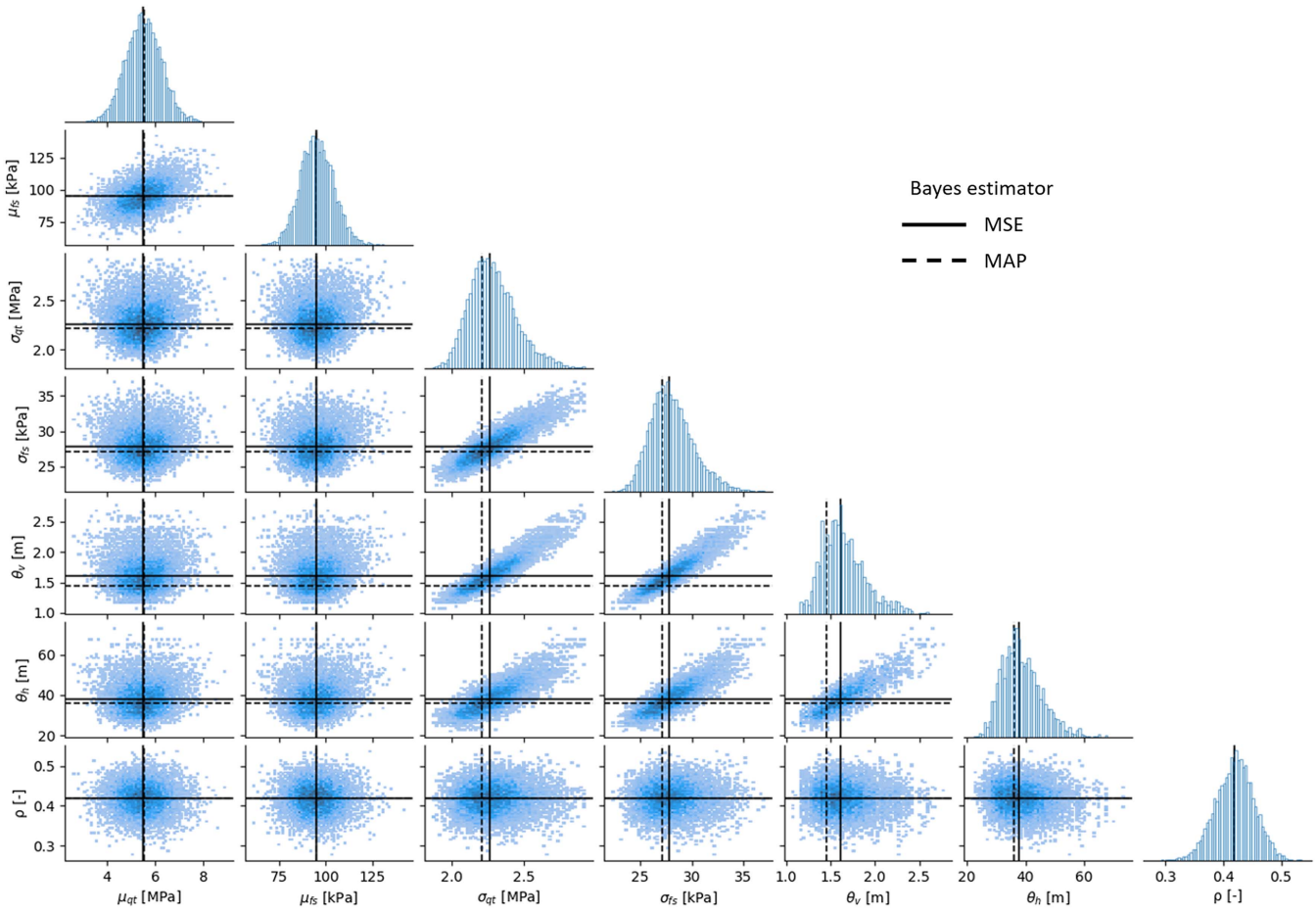

Fig. 4. Histogram of the posterior samples per variable and scatter plot of posterior samples per variable combination of the aggregated model with indication of $\hat{X}_{\mathrm{MSE}}$ and $\hat{X}_{\mathrm{MAP}}$.

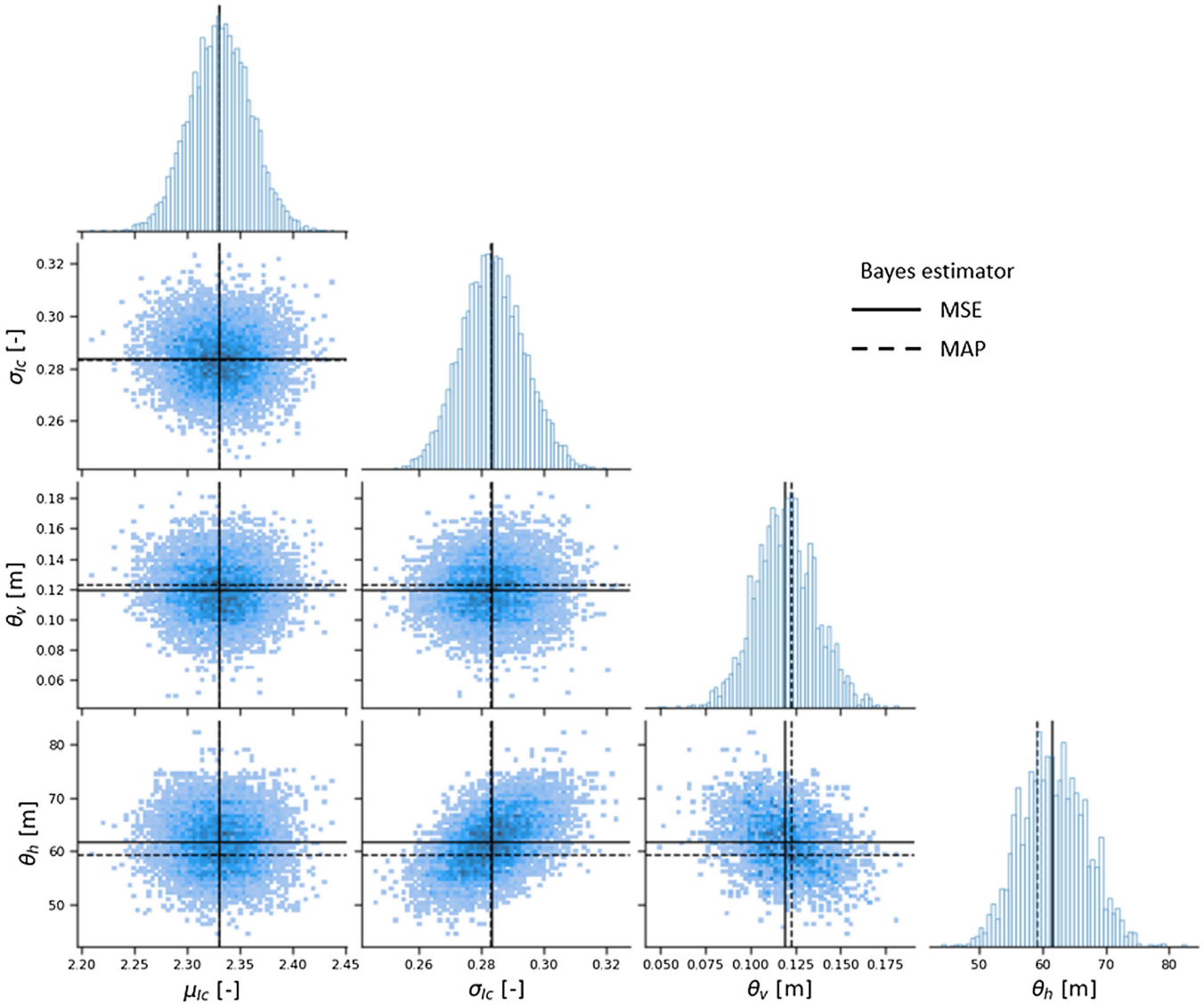

Fig. 5. Histogram of the posterior samples per variable and scatter plot of posterior samples per variable combination of the split model with indication of $\hat{X}_{\mathrm{MSE}}$ and $\hat{X}_{\mathrm{MAP}}$.

The autocorrelation lengths play a central role in prediction because they establish the relationship between the training and validation points of the domain. As discussed in the "Methodology" section, $\theta_v^{\min}$ and $\theta_h^{\min}$ are the minimum values of the respective lengths that lead to autocorrelation values greater than 0.3 and are defined based on the interpoint distances met in the domain. The critical validation point for prediction is the one with the greatest distance to its closest training point. This approach is acknowledged to be conservative, considering that the autocorrelation matrix can establish connections between validation and training points through other validation points. Moreover, because the points of the domain are aligned in vertical profiles over regular vertical distances, identifying the critical point can be performed individually per direction. A further implication of this is that θ_v does not affect the predictive radius of the model but strongly affects the predictive power by influencing the magnitude of autocorrelation. As a result, $\theta_v^{\min}$ and $\theta_h^{\min}$ can be estimated independently. For the horizontal direction, the critical point and distance are identified by examining the matrix of horizontal interdistances between points. Due to the vertical stratification of the points with regular intervals, the critical distance in the vertical direction is the interval distance. After identifying the critical distance for the vertical and horizontal directions, $\theta_v^{\min}$ and $\theta_h^{\min}$ can be evaluated as the values of θ_v and θ_h that lead to an autocorrelation of 0.3 when the autocorrelation function is evaluated individually per direction.

Table 4 lists the ROPE per variable for hypothesis testing in the aggregated model, including the values of $\theta_v^{\min}$ and $\theta_h^{\min}$ as estimated with the procedure described previously and the result of the respective hypothesis test. Fig. 7 visualizes the histogram per variable, along with the 95% HDI and ROPE. For all variables, the HDI does not overlap with the ROPE, which means that the null hypothesis can be rejected with sufficient confidence. An implication of rejecting the null hypothesis is that the scales of fluctuation can be inferred by the model on the specific training point grid. A typical problem with learning θ_h is that the horizontal distances between training points are too large to perform meaningful inference of the parameters.

Similarly, Table 5 provides the ROPE and hypothesis testing results for the split model. For the first step, $\theta_v^{\min}$ and $\theta_h^{\min}$ have been adjusted for the Gaussian kernel autocorrelation function. The model

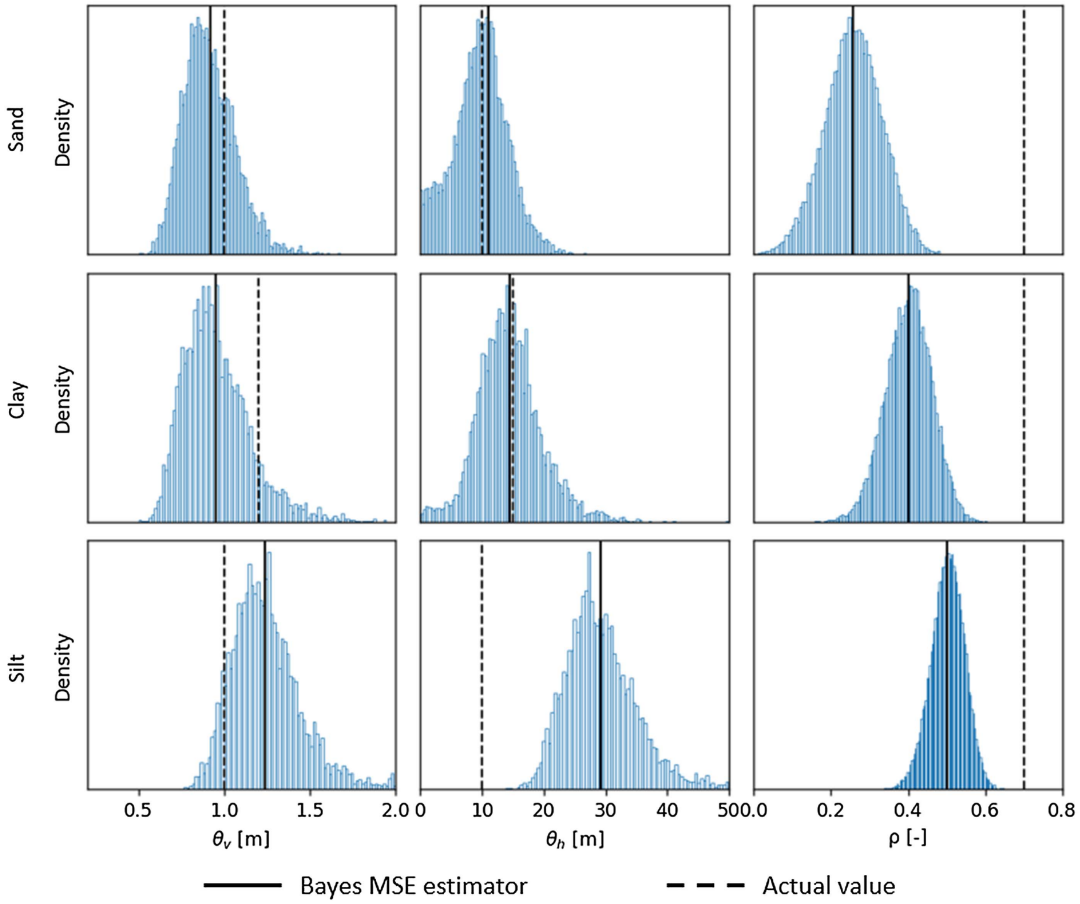

Fig. 6. Posterior histograms of θ_v, θ_h, and ρ per material type, with indication of $\hat{X}_{\text{MSE}}$ and the actual value used in generating the RF of the exercise.

Table 4. ROPE per examined variable and result of the associated hypothesis testing for the aggregated model

Variable	ROPE	Hypothesis test outcome
ρ	$(-0.30, 0.30)$	Reject H_0
θ_v(m)	$[0.00, 0.17]$	Reject H_0
θ_h(m)	$[0.00, 14.98]$	Reject H_0

rejects the null hypothesis for θ_h but fails to reject for θ_v, meaning that low autocorrelation is expected in the vertical direction. However, as explained previously, the role of the vertical autocorrelation is limited because CPT data are provided in columns, and missing data is not a problem in this exercise. Furthermore, the null hypothesis is not rejected for the cross-correlation coefficient of the sand and clay materials. As shown in Fig. 8, inference was not that effective for these parameters, and their posterior distributions are located within the ROPE. Moreover, the model fails to reject the null hypothesis for θ_h in sand and clay. However, this is not a shortcoming of inference; the model is approximating the actual values used in setting up the RF. This means that the setup of the benchmark exercise leads to validation locations that are outside the predictive radius of the model.

Hypothesis testing can lead to further conclusions regarding the prowess of the aggregated model by assessing its predictive radius and area of influence per training point. First, a two-dimensional mesh of points is constructed in the horizontal and vertical directions with the purpose of representing distances from point (0, 0). In this instance, point (0, 0) represents a training point, and prediction will be carried out at all points in the mesh. Because θ_h is taken to be isotropic, a single horizontal distance component suffices. Next, posterior samples of the autocorrelation matrix C for the mesh are formed by evaluating the autocorrelation function for the posterior samples of X. Hypothesis testing on all terms of C for ROPE $= [0, 0.3)$ and an HDI of 95% indicates whether sufficient autocorrelation can be established between the training point and the points in the mesh based on the posterior of X. Fig. 9 presents the result of this analysis for the aggregated model. The area of the domain where the null hypothesis is rejected resembles the predictive area of the model.

The analysis shows that the horizontal predictive radius of the aggregated model is 17.0 m, whereas the vertical one is 0.78 m. The shape of the predictive area is determined by the autocorrelation function, as well as the values of the of θ_v and θ_h posterior samples and the correlation between them. The same figure displays the validation locations as vertical profiles, which are placed from the training point at their respective critical distances. This illustration shows that all validation locations are well within the horizontal predictive radius, a point that was already validated by defining the critical point and distance in the derivation of $\theta_h^{\max}$. Also, because training points are placed over the entire depth of the domain, validation points of the profiles that lie outside the predictive area are covered by training points of greater depths. Hence, attention in the analysis is given to the horizontal direction.

Furthermore, Fig. 9 also visualizes the lower bound of the 95% HDI of C over the mesh in contours. This plot displays the autocorrelation coefficient that is expected to be surpassed with a confidence of 95% and thus indicates a minimum boundary for the

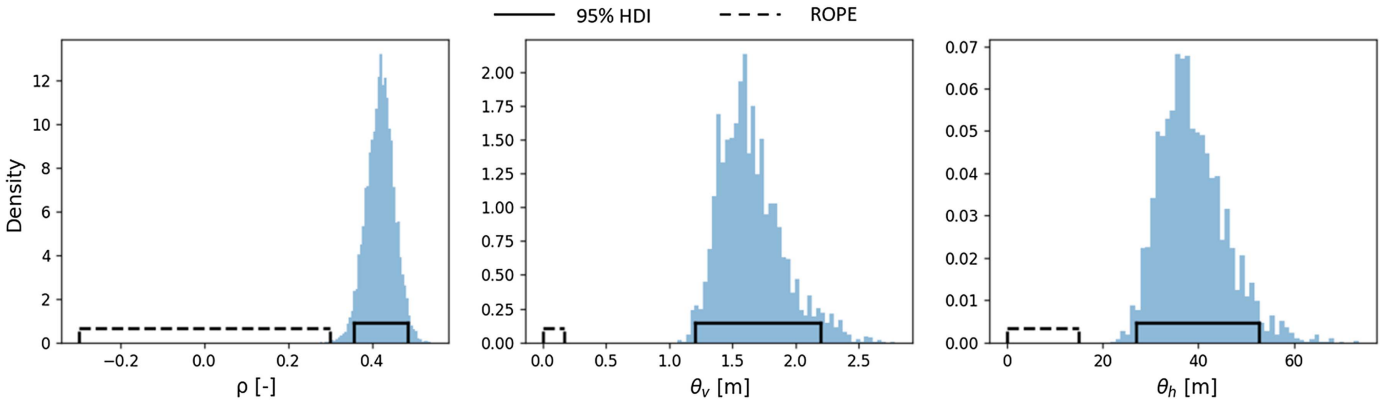

Fig. 7. Posterior histograms with 95% HDI and ROPE per parameter examined in hypothesis testing for the aggregated model.

Table 5. ROPE per examined variable and result of the associated hypothesis testing for the split model

	First step			Second step		
Variables	ROPE	I_c	ROPE	Sand	Clay	Silt
ρ	—	—	$(-0.30, 0.30)$	Fail to reject H_0	Fail to reject H_0	Reject H_0
θ_v(m)	$[0.00, 0.15)$	Fail to reject H_0	$[0.00, 0.17)$	Reject H_0	Reject H_0	Reject H_0
θ_h(m)	$[0.00, 1.13)$	Reject H_0	$[0.00, 14.98)$	Fail to reject H_0	Fail to reject H_0	Reject H_0

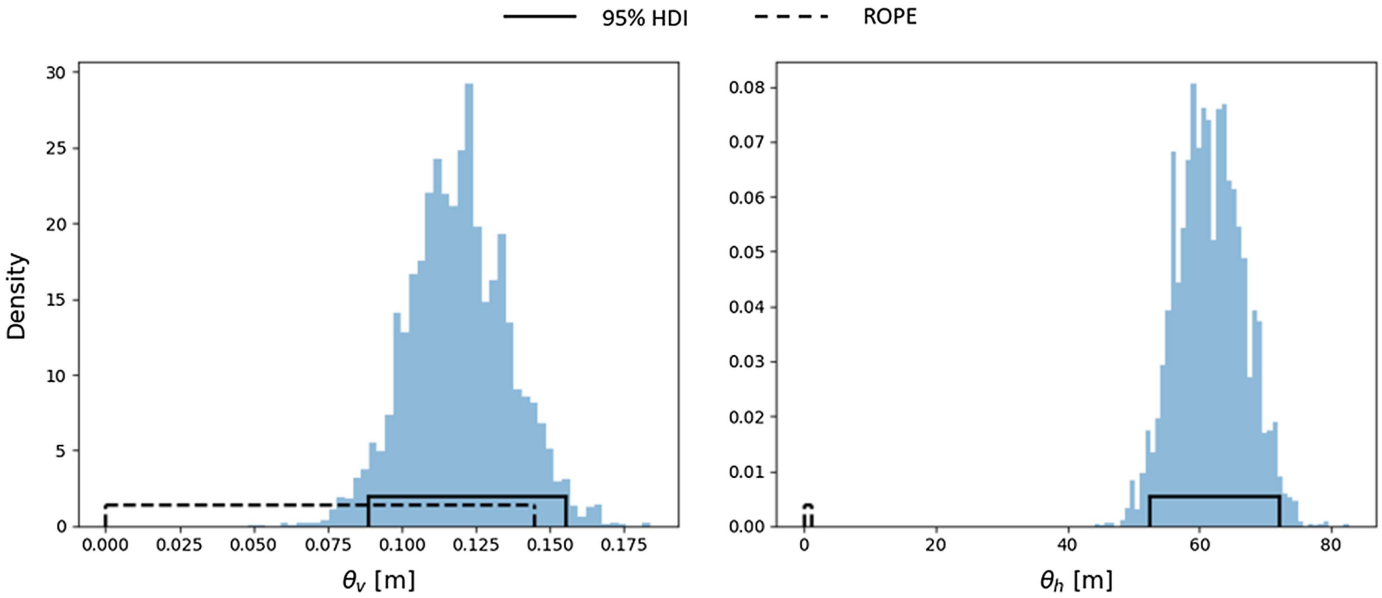

Fig. 8. Posterior histograms with 95% HDI and ROPE per parameter examined in hypothesis testing for the first step of the split model.

expected predictive power per validation location. These results only quantify the impact of the closest training location per validation location and do not consider additional influence arriving from other training points.

This extension of hypothesis testing leads to valuable conclusions regarding the predictive power of the model that cannot be derived through cross-validation. The insights gained by performing cross-validation with a set of validation locations are valuable, as long as the set is representative of the situations expected to be met in practice. Otherwise, cross-validation only offers a quantification of the predictive power of the model within a limited extent of the cases that could be actually met. On the other hand, the use of hypothesis testing allows BaySiC to answer application-oriented questions. First, testing for ρ checks the premise that q_t and f_s are correlated. Second, hypothesis testing for C enables the definition of the predictive radius of the model, which demonstrates the area of influence per training point, as well as the area of the site where credible predictions should be expected. Ultimately, hypothesis testing in BaySiC facilitates practice-oriented use of the framework, quantifies the impact of training data, and promotes advanced applications, such as prediction-driven site investigation.

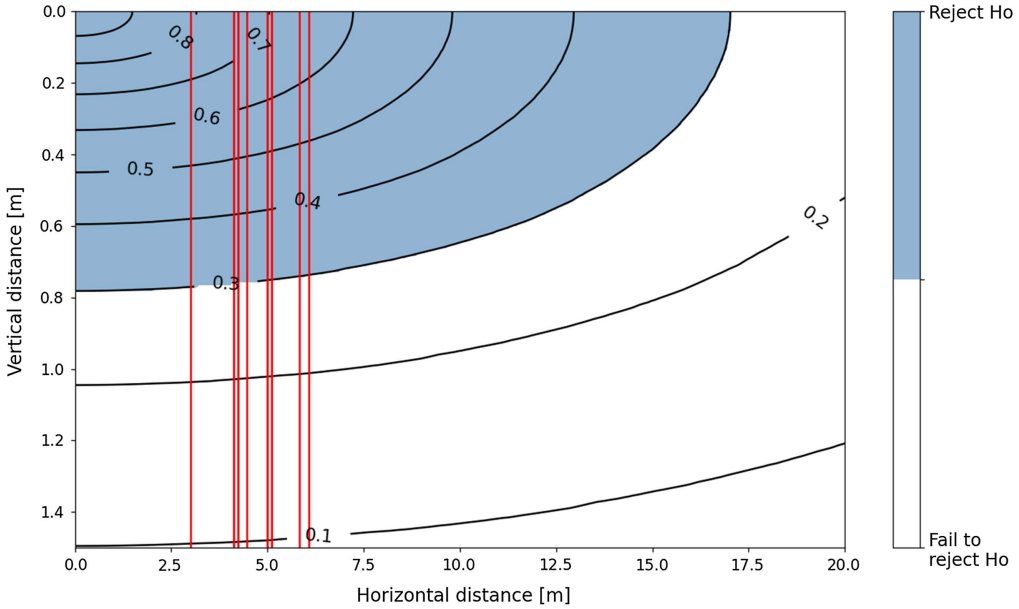

Fig. 9. Predictive area and lower bound of the 95% HDI of the autocorrelation for the aggregated model with validation profiles placed at critical horizontal distances.

Table 6. Estimators of RMSE and IR for $\hat{X}_{\mathrm{MSE}}$ and $\hat{X}_{\mathrm{MAP}}$ of the aggregated model

	Aggregated model				Split model			
Metric	$\hat{X}_{\mathrm{MSE}}$		$\hat{X}_{\mathrm{MAP}}$		$\hat{X}_{\mathrm{MSE}}$		$\hat{X}_{\mathrm{MAP}}$	
estimator	RMSE	IR	RMSE	IR	RMSE	IR	RMSE	IR
Average	1.23	0.81	1.23	0.81	1.13	0.92	1.14	0.92
Minimum	0.89	0.73	0.89	0.72	0.92	0.89	0.93	0.89
Maximum	1.71	0.89	1.73	0.89	1.58	0.96	1.58	0.96

Even though it is not presented here for conciseness, a similar analysis has been performed for the split model. Comparable conclusions regarding its predictive area are drawn.

Metric Evaluation

By examining the metrics of the two models, it is clear that the split model outperforms the aggregated model. Table 6 summarizes the results of the aggregated and split models per metric and estimator in terms of average, minimum, and maximum values. For the split model, the IR achieved by the first step is significantly higher than that of the aggregated model. Also, the IR score demonstrates that the split model is highly effective in predicting the material type per validation point. This prediction is the basis for the second step, which estimates a substantially lower RMSE than the aggregated model. In summary, whereas the aggregated model approximates the performance of the analysis shown in (Phoon et al. 2022b), the split model surpasses it and achieves better metrics. Also, the $\hat{X}_{\mathrm{MSE}}$ and $\hat{X}_{\mathrm{MAP}}$ lead to almost identical results, which was expected due to the proximity of the estimator values. Because of this, the paper will only report the results of $\hat{X}_{\mathrm{MSE}}$ from now on because it is considered a more convenient estimator.

Fig. 10 presents the mean prediction of the aggregated model for q_t as derived using $\hat{X}_{\mathrm{MSE}}$ against the validation data over a vertical profile per validation location. Although the prediction follows the general behavior of the data, it often misses localized fluctuations. This is a result of adopting the aggregated model, which leads to the estimation of scales of fluctuation that describe well the soil domain in general, but do not offer sufficient flexibility to capture layer-specific variations. Also because of the aggregate layer approach, the prediction is making smooth transitions between layers instead of sharp changes as seen in the data. The framework does not perceive the existence of layers and thus does not exhibit sharp changes on layer interfaces.

Moreover, Fig. 10 visualizes the prediction uncertainty in the form of the 95% prediction interval of the predictive distribution per validation point. The illustration of the 95% prediction interval aims to provide insight regarding prediction uncertainty because the presented metrics do not directly demonstrate its quantification, which is one of the main reasons for adopting Bayesian statistics. The greatest share of the validation data is covered by the envelope of the prediction interval. This happens even in cases where the mean prediction does not capture intense local fluctuations of the data, which implies that even if the accuracy of the prediction is lower, the prediction uncertainty can capture the variations of the data. Furthermore, the impact of adopting the aggregated model is demonstrated again. Although the prediction is precise in the sand and silt layers, as evidenced by the coefficient of variation (CoV) being close to 0.20, it is not so precise for the clay layer, where the CoV increases. The aggregated model tends to average out the standard deviations of all layers, and this leads to high CoV values for clay, where the mean q_t is significantly smaller than other layers.

Fig. 11 presents the mean prediction of the aggregated model for the SBT as derived using $\hat{X}_{\mathrm{MSE}}$ against the validation data over a vertical profile per location. The model can identify the sand and clay layers with limited competency but largely misclassifies the silt layer. For instance, a predictive anomaly consistently appears over the silt layer. A part of the layer is misclassified as sand (SBT = 6), and the anomaly region can change depths or sizes but persists at all locations. Because the full RF of the benchmark exercise has not been made available, this issue cannot be settled reliably. However, this behavior may be explained to some extent by the weak correlation between q_t and f_s. Because I_c and SBT are calculated using both CPT parameters and inference could not approximate well the strong correlation between the two, it might

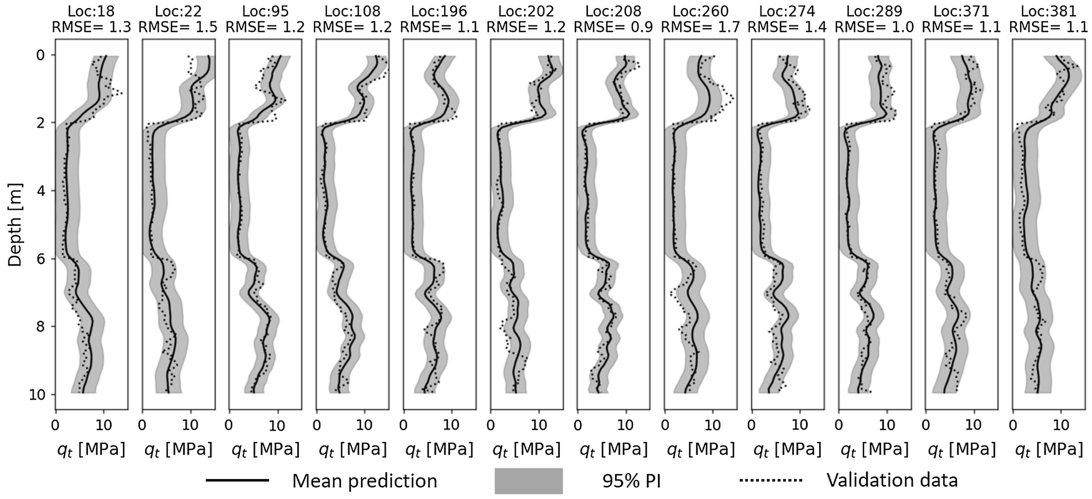

Fig. 10. Average prediction of the aggregated model for q_t 95% prediction interval according to $\hat{X}_{\text{MSE}}$, along with the validation data per validation location.

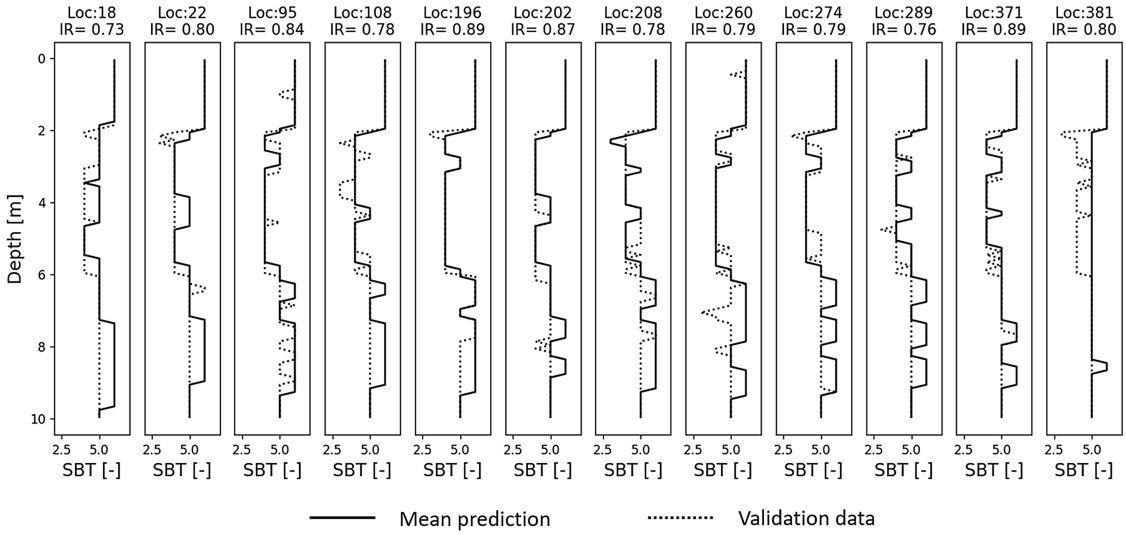

Fig. 11. SBT prediction of the aggregated model according to $\hat{X}_{\text{MSE}}$, along with the SBT validation data per validation location.

be that material prediction is hampered, even if the approximation of q_t validation data is satisfactory.

On the other hand, Fig. 12 shows that the split model is considerably more competent in classifying the material type at validation points. The SBT prediction of the split model is able to match well the SBT per layer. The validation SBT data exhibit several cases of pockets of foreign material within a layer. Although the model cannot find all of them, it is still potent in detecting a good amount, sometimes even if they appear as a single, localized anomaly. Moreover, the results show a low amount of misclassified points due to falsely predicted anomalies, as happened with the aggregated model. Ultimately, I_c acts as a strong predictor for the material type because it incorporates information of both q_t and f_s.

Additionally, Fig. 13 demonstrates the prediction of the split model for q_t. In this case, the prediction is able to make swift changes along layer interfaces because the predictions per material are independent. The mean prediction is successful in approximating the fluctuations of q_t for the clay and silt layers not only in terms of general trend but also when they appear as pockets within other layers. In such occasions, the mean prediction changes and the 95% CI expands sharply to account for localized uncertainty. For example, such an instance is met at a depth of 4.0 m in validation location 18. On the other hand, the prediction is less flexible for sand. The predictive distribution per point tends to be fixated to the mean of the material and compensates with a larger standard deviation instead of following the fluctuations of the q_t data. This can be a result of the low θ_h met during inference in the sand layer, which notably is also the actual θ_h used in generating the RF. This behavior validates the conclusions drawn by hypothesis testing regarding the predictive power of the model in sand. Specifically, because such a low θ_h is met for the sand layer, prediction in sand is limited by low autocorrelation and is destined to be close to the mean of the layer. In case more accurate prediction was required, CPT soundings should be performed in a denser grid.

Fig. 14 shows the RMSE and IR score of the split model per validation location over the map of the domain. In general, points that lie at the center of the domain tend to perform better than points at the perimeter. This is because they are expected to be closer to training points, gaining a better autocorrelation and receiving significant influence from multiple training soundings simultaneously.

ASCE-ASME J. Risk Uncertainty Eng. Syst., Part A: Civ. Eng., 2023, 9(2): 04023008

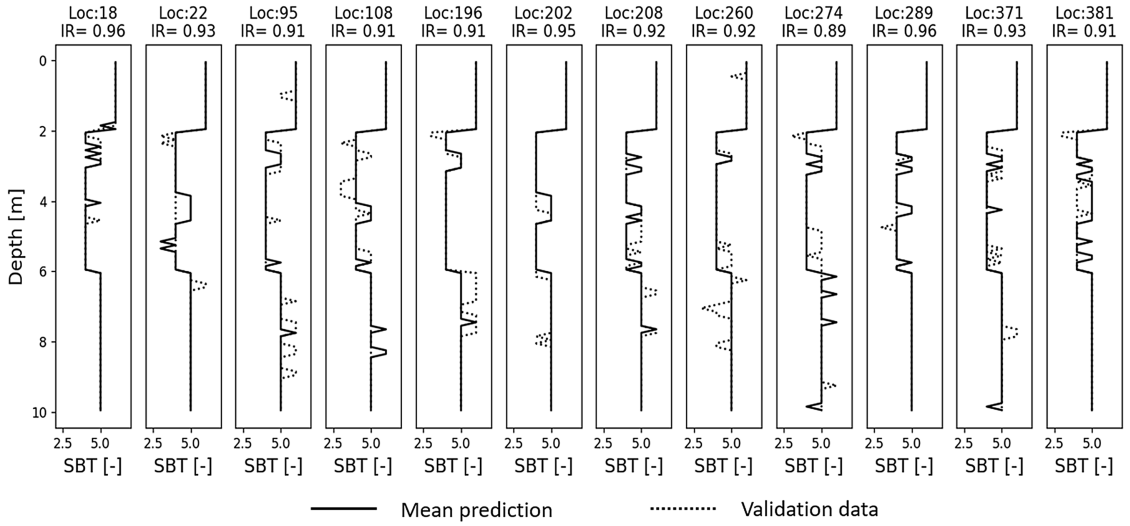

Fig. 12. SBT prediction of the split model according to $\hat{X}_{\mathrm{MSE}}$, along with the SBT validation data per validation location.

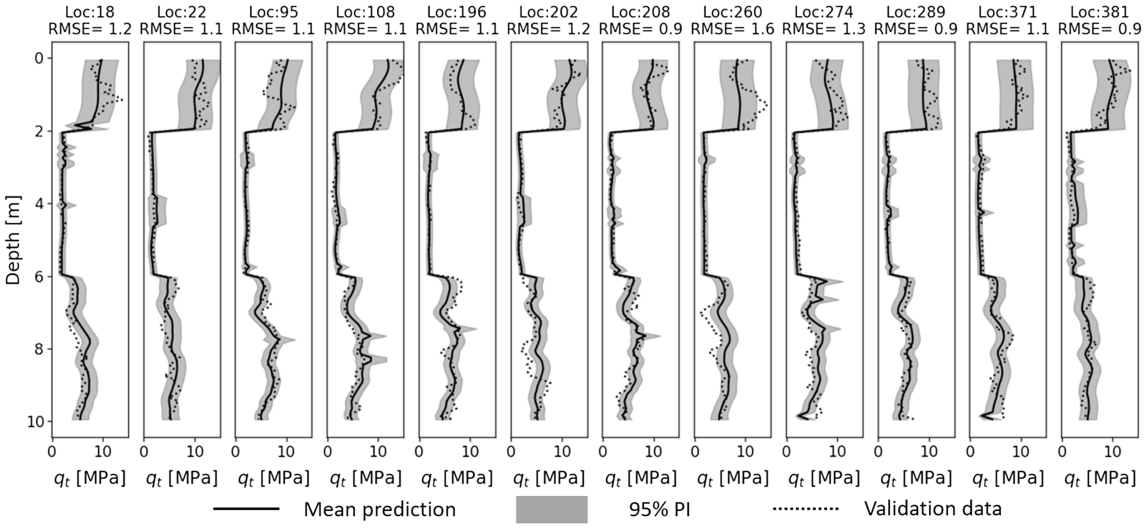

Fig. 13. Average prediction of the aggregated model for q_t 95% prediction interval according to $\hat{X}_{\mathrm{MSE}}$, along with the validation data per validation location.

Also, predictions at points lying in the perimeter of the domain are achieved by extrapolation of the predictive model, which is expected to have lower accuracy and precision. Soundings 208 and 289 are instances of validation locations that lie at the center of the domain and pose as the best performers in terms of RMSE and IR, respectively. Both soundings are surrounded by training CPTs, whose strong influence on prediction accumulates. At the same time, the distances to their respective closest training neighbors are the lowest among the validation locations, yielding the greatest estimates of expected autocorrelation. On the other hand, some perimetrical points outperform center points. For example, Point 381 exhibits considerably low RMSE values, whereas high RMSE values are expected, just as met in other perimetrical points, like 18 and 22. This pattern was encountered in the results of both models, as well as in the results of Phoon et al. (2022b). This behavior strengthens the notion that the accuracy of predictions is not only a function of the predictive power of the model but of the training and test data. This topic cannot be explored further dependably because the full RF of the benchmark exercise has not been made available.

Results for All Benchmark Cases

This section presents and compares the performance of the aggregated and split models in each benchmark case. The four stratigraphic cases reflect situations of increasing complexity. A distinction is also made for each set of training CPT soundings to showcase the effect of the size of the training data.

Fig. 15 summarizes the average of the performance metrics per stratigraphic case and training set size. Although the split model outperforms the aggregated model in all cases and training sets for the IR, the picture is not the same for the RMSE. The RMSE of both models increases with greater complexity, but this rise is more intense for the split model. Whereas the split model outperforms the aggregated model for SVG-1, this changes from SVG-2 and so on. Low data availability is the cause for this shift.

The split model consistently leads to better IR scores, meaning that the first step of the model is quite competent in identifying the material type. Thus, the steep RMSE increase is attributed to the second step, which maps q_t over the subsoil. In more complex

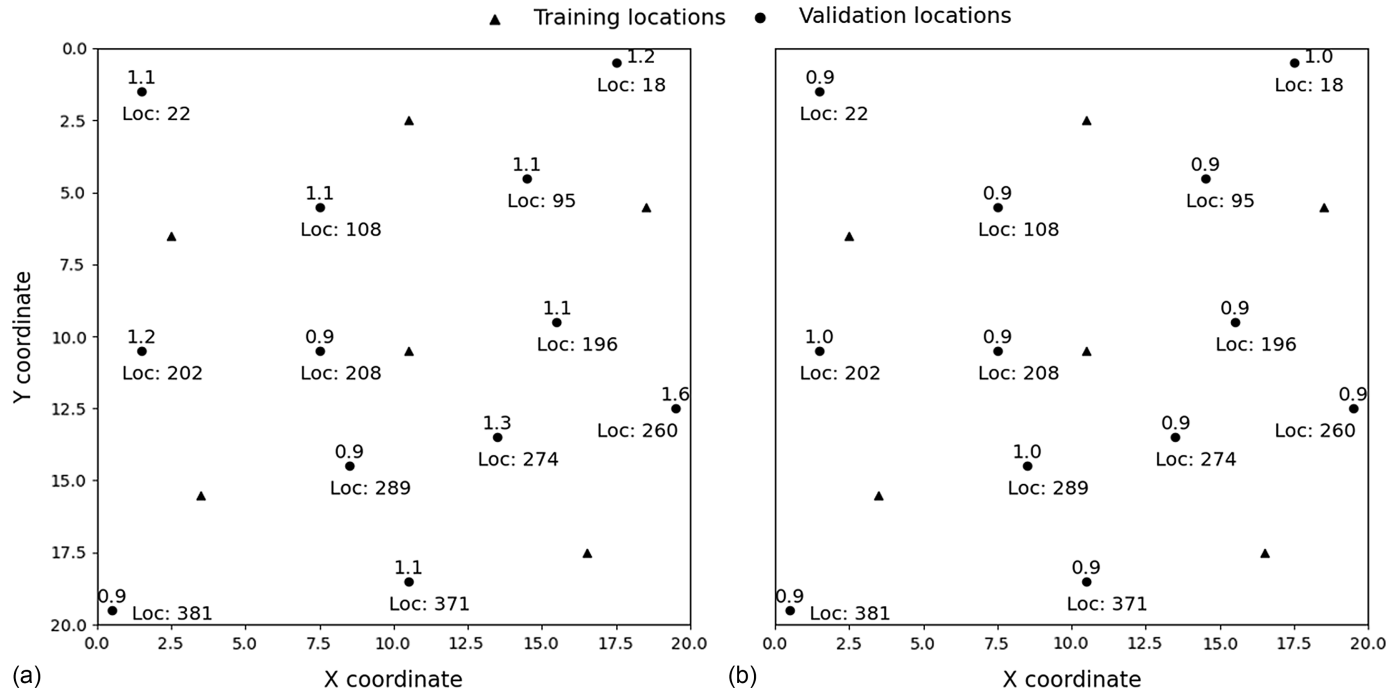

Fig. 14. (a) RMSE; and (b) IR scores per validation location for Bayes estimator $\hat{X}_{\mathrm{MSE}}$ of the split model.

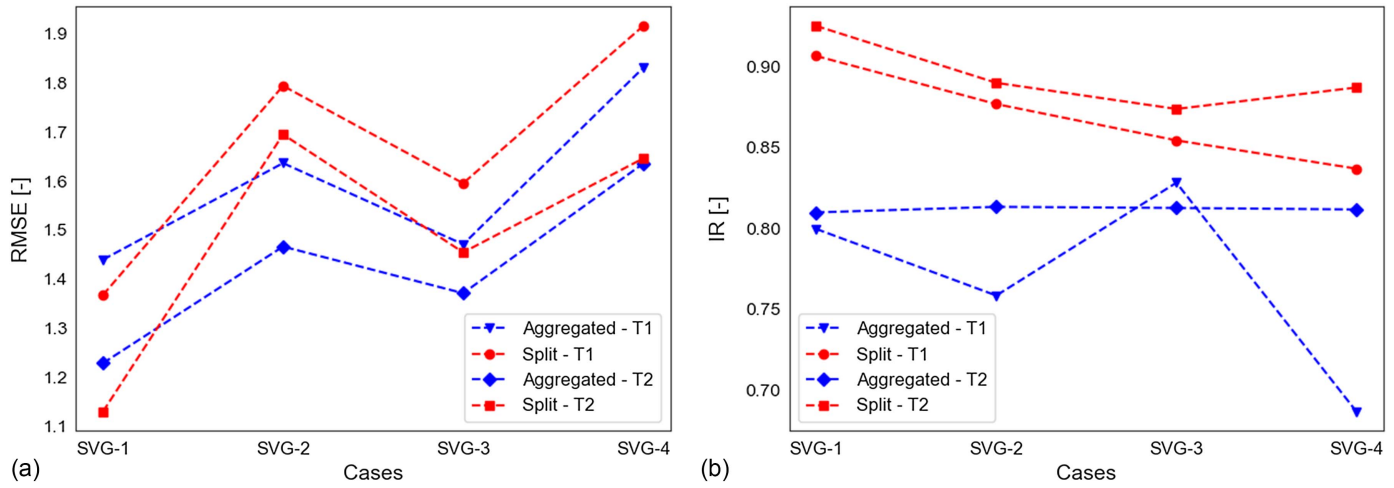

Fig. 15. Comparison of the average: (a) RMSE; and (b) IR for each case as estimated by $\hat{X}_{\mathrm{MSE}}$ per model and training set.

stratigraphic cases, training per individual layer can prove challenging because the ability of inference in capturing the spatial variability patterns can be hindered by the limited amount of data available per layer. On the other hand, the aggregated model concept has an advantage in such settings. At this point, the authors speculate that with larger training data sets, the split model would get an edge even in cases of greater complexity.

Last, the effect of training data set size is investigated. As expected, the larger training data set leads to better metrics. In all cases, models trained with set T2 outperform those trained with set T1. An exception is raised for the IR of the aggregated model in case SVG-3, which possibly occurs due to the complexity of the setting; in this case, the CPT parameters follow a linear trend over depth, and the model formulation has not been prepared accordingly. Moreover, the spread between the split models for T1 and T2 diverge for both RMSE and IR, with increasing case complexity. This point highlights that the amount of training data has a greater impact in more complex stratigraphic settings.

Conclusions

The paper demonstrates a Bayesian approach to the DDSC problem by using a framework for Bayesian site characterization in a benchmark example. The BaySiC models used in the paper achieve accurate predictions through uncertainty reduction. Random field modelling of the CPT parameters in the subsoil is the basis of the models in both inference and prediction. The article showcases the formulation of the aggregated and the split models and their implementation on the benchmark example.

In inference, BaySiC deals with separable covariance structures, so the matrix normal distribution is adopted as the likelihood function. Specifically, the aggregated model separates the spatial variability, which is represented by the autocorrelation matrix, and cross-correlation of the CPT parameters. The aggregated model describes all materials with a single set of parameters. Although this modelling approach is inherently biased, it is able to exploit the full training data set. On the other hand, the novelty of the split model lies in distinguishing the horizontal and vertical autocorrelations. The split model, which infers the parameters per material, leads to parameter posterior distributions that are meaningful to geotechnical engineering but in some cases appears to be hindered by low data availability per material.

Both strategies demonstrate their own advantages in prediction. The split model leads to consistently high identification rates of the material type. On the other hand, the aggregated model is more accurate in predicting the cone resistance in complex cases. Also, the precision of model predictions is quantified through the visualization of the 95% prediction interval, which can be used in reliability assessments or the derivation of characteristic profiles.

Moreover, the paper shows the value of Bayesian hypothesis testing as an alternative means for validating the predictive power of the model. Because the full RFs of the benchmark cases are not available, the capacity of the model for prediction can only be quantified by metrics at validation locations. Given the limited size of the validation data set, conclusions on model performance have to be treated with care because random effects can be relevant. On the other hand, hypothesis testing is independent of the validation data and can evaluate the predictive power of the model on a probabilistic level. In this study, hypothesis testing was used to assess the predictive radius of each training CPT sounding per model and the autocorrelation expected at each validation location. This information can be used further in geotechnical investigation, decision-making, and optimization.

Last, BaySiC has shown efficient performance in the benchmark exercise, which highlights the potential of the framework for application in real cases. The aggregated model exhibits runtimes of approximately 25 min for inference and prediction. For the split model, the respective figure lies close to 10 min. Such runtimes render the use of BaySiC feasible within a project setting and suggest it would remain appropriate with even larger data sets.

Data Availability Statement

All models and/or code that support the findings of this study are available from the corresponding author upon reasonable request. All data used during the study were provided by a third party. Direct requests for these materials may be made to the provider, as indicated in the Acknowledgments.

Acknowledgments

The authors would like to thank Kok-Kwang Phoon, Takayuki Shuku, Jianye Ching, and Ikumasa Yoshida for creating the benchmark example and driving developments in ML and statistical methods for DDSC. The data supporting the findings of this study is presented in the work of Phoon et al. (2022b), whose authors can be contacted to provide the benchmark data. The authors offer their special thanks to Jianye Ching for the fruitful discussions that have improved the quality of the paper.

References

Baecher, G. B. 2017. "Bayesian thinking in geotechnics." In *Geo-Risk 2017*, 1–18. Reston, VA: ASCE.

Betancourt, M. 2017. "A conceptual introduction to Hamiltonian Monte Carlo." Preprint, submitted January 10, 2017. https://doi.org/10.48550/arXiv.1701.02434.

Ching, J., and K.-K. Phoon. 2019. "Constructing site-specific multivariate probability distribution model using Bayesian machine learning." *J. Eng. Mech.* 145 (1): 04018126. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001537.

Dutilleul, P. 1999. "The MLE algorithm for the matrix normal distribution." *J. Stat. Comput. Simul.* 64 (2): 105–123. https://doi.org/10.1080/00949659908811970.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. *Bayesian data analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Geyer, S., I. Papaioannou, and D. Straub. 2021. "Bayesian analysis of hierarchical random fields for material modeling." *Probab. Eng. Mech.* 66 (Oct): 103167. https://doi.org/10.1016/j.probengmech.2021.103167.

Gupta, A. K., and D. K. Nagar. 1999. *Matrix variate distributions*. Boca Raton, FL: Chapman & Hall.

Gut, A. 2009. *An intermediate course in probability*. New York: Springer.

Helterbrand, J. D., and N. Cressie. 1994. "Universal cokriging under intrinsic coregionalization." *Math. Geol.* 26 (2): 205–226. https://doi.org/10.1007/BF02082764.

Hoffman, M. D., and A. Gelman. 2014. "The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.* 15 (1): 1593–1623. https://doi.org/10.5555/2627435.2638586.

Krüger, F., S. Lerch, T. L. Thorarinsdottir, and T. Gneiting. 2016. "Predictive inference based on Markov chain Monte Carlo output." *Int. Stat. Rev.* 89 (2): 274–301. https://doi.org/10.1111/insr.12405.

Kruschke, J. K. 2013. "Bayesian estimation supersedes the *t* test." *J. Exp. Psychol. Gen.* 142 (2): 573–603. https://doi.org/10.1037/a0029146.

Kruschke, J. K. 2018. "Rejecting or accepting parameter values in Bayesian estimation." *Adv. Methods Pract. Psychol. Sci.* 1 (2): 270–280. https://doi.org/10.1177/2515245918771304.

Lehmann, E. L., and G. Casella. 1998. *Theory of point estimation*. New York: Springer.

Mariethoz, G., and J. Caers. 2014. *Multiple-point geostatistics*. New York: Wiley.

Marriott, F. H. C., and M. L. Eaton. 1984. "Multivariate statistics: A vector space approach." *Appl. Stat.* 33 (3): 319. https://doi.org/10.2307/2347710.

Mavritsakis, A., T. Schweckendiek, A. Teixeira, and E. Smyrniou. 2022. "Bayesian subsurface mapping using CPT data." In *Proc., 8th Int. Symp. for Geotechnical Safety & Risk, ISGSR 2022*. Adelaide, Australia: Engineers Australia.

Meyn, S. P., and R. L. Tweedie. 1993. *Markov chains and stochastic stability*. London: Springer.

Murphy, K. P. 2022. *Probabilistic machine learning: An introduction*. Cambridge, MA: MIT Press.

Neal, R. M. 2011. *MCMC using Hamiltonian dynamics*. Boca Raton, FL: Chapman and Hall/CRC.

Phoon, K.-K., J. Ching, and T. Shuku. 2022a. "Challenges in data-driven site characterization." *Georisk: Assess. Manage. Risk Eng. Syst. Geohazards* 16 (1): 114–126. https://doi.org/10.1080/17499518.2021.1896005.

Phoon, K.-K., T. Shuku, J. Ching, and I. Yoshida. 2022b. "Benchmark examples for data-driven site characterization." *Georisk: Assess. Manage. Risk Eng. Syst. Geohazards* 16 (4): 1–23. https://doi.org/10.1080/17499518.2022.2025541.

Rall, L. B. 1981. Vol. 120 of *Automatic differentiation: Techniques and applications*. Berlin: Springer.

Rasmussen, C. E., and C. K. I. Williams. 2006. *Gaussian processes for machine learning*. Cambridge, UK: MIT Press.

Robertson, P. 2016. "Cone penetration test (CPT)-based soil behaviour type (SBT) classification system—An update." *Can. Geotech. J.* 53 (12): 1910–1927. https://doi.org/10.1139/cgj-2016-0044.

Salvatier, J., T. V. Wiecki, and C. Fonnesbeck. 2016. "Probabilistic programming in Python using pymc3." *PeerJ Comput. Sci.* 2: e55. https://doi.org/10.7717/peerj-cs.55.

van de Meent, J.-W., B. Paige, H. Yang, and F. Wood. 2018. "An introduction to probabilistic programming." *Found. Trends Mach. Learn*. Preprint, submitted September 27, 2018. https://arxiv.org/abs/1809.10756.

Vanmarcke, E. 2010. *Random fields: Analysis and synthesis*. Singapore: World Scientific.

Wang, Z., M. Broccardo, and J. Song. 2019. "Hamiltonian Monte Carlo methods for subset simulation in reliability analysis." *Struct. Saf.* 76 (Jan): 51–67. https://doi.org/10.1016/j.strusafe.2018.05.005.

Wasserman, L. 2004. *All of statistics*. New York: Springer.

Zhu, H., L. Zhang, T. Xiao, and X. Li. 2017. "Generation of multivariate cross-correlated geotechnical random fields." *Comput. Geotech.* 86 (Jun): 95–107. https://doi.org/10.1016/j.compgeo.2017.01.006.