University of London
Imperial College of Science, Technology and Medicine
Department of Civil and Environmental Engineering

Data-driven uncertainty quantification and predictive digital twin for offshore piles

Ningxin Yang

Submitted on December 1, 2023

Abstract

Proper estimation of offshore piles is vital to the life and permanence of the foundation in question. However, due to the uncertainties of soil parameters in the field, the offshore piles are greatly affected. The source of soil uncertainties may come from various reasons, such as lack of uniformity between in-situ test and laboratory experiment; spatial variability of soil profile and rationality of the constitutive model, etc. Traditional statistical analysis which is based on Monte Carlo, is time-consuming and laborious. Though Bayesian theorem provides ways to understand and update the uncertainties, the amount of the inference analysis is still computationally heavy, thus bringing big challenges for the pile design. Additionally, to mimic the geotechnical structures and bearing behaviors as accurate as possible, digital twin is seeping into all kinds of engineering problems, enabling to evolve over time to persistently represent a unique physical asset and achieving data-driven decision making process. However, state-of-the-art digital twins are still relying on considerable expertise and deployment resources, leading to an only one-off implementation and remaining limitation on providing adaptive digital models on unique offshore piles.

In this thesis, we hope to introduce probabilistic graphical model (PGM) involving in Bayesian inverse analysis to speed up the calculation and provide reasonable posterior results for the soil parameters. Besides, as a mathematical and rigorous foundation, partially observed PGM is proposed to support the transition from custom defined model towards accessible digital twins at scale. Based on such flexible asset-specific models, the entire loading life-cycle can be incorporated into a digital twin forming a unified and accessible foundation for a wide range of offshore piles. Combined with monitored data, the proposed dynamic updated digital twin provides rapid analysis results for reliable soil parameters and enables intelligent decision making on the pile bearing behaviors.

Contents

List of Tables

List of Figures

Nomenclature

$\boldsymbol{x}$	Vector of input parameter
$\boldsymbol{y}$	Model response
ϵ	Gaussian discrepancy
$\mathcal{X}$	Experimental design
$\mathcal{D}_{\boldsymbol{X}}$	Input parameters space
$\mathcal{M}$	Computational model
$\mathcal{M}_d$	Deterministic simulator
$\mathcal{M}_d(\boldsymbol{x})$	Deterministic output
$\mathcal{M}_s$	Stochastic simulator
$\mathcal{M}_s(\boldsymbol{x})$	Stochastic output
$\tilde{\mathcal{M}}_d$	Surrogate model
ε	Generalization error

Acronyms

PDF Probability Density Function. 8, 10

MAP maximum a posterior. 10, 11

MCMC Markov Chain Monte Carlo. 17, 18

Chapter 1

Introduction

1.1 Background

Offshore monopiles are increasingly favored in wind farm installations due to their advantages in clean energy generation and easy deployment. These cylindrical steel structures are driven into the seabed to provide a stable foundation for wind turbines. While offshore wind energy presents a promising source of clean and sustainable power, the utilization of offshore monopiles has introduced certain engineering challenges. One of the key concerns associated with offshore monopiles is the need to address the potential issues related to excessive pile displacements induced during their installation and operation(Byrne & Houlsby, 2003; Randolph et al., 2005). Excessive pile movements can lead to significant displacements and rotations in supporting structures, which, in turn, may result in damage or structural instability. Consequently, it becomes crucial to accurately predict and manage pile deformations when designing and analyzing support systems for offshore monopile installations.

When soil properties, design load and pile dimensions are acquired from a technical report, estimating the pile response is typically done through empirical solutions in guidelines or numerical simulations. Several design methods have been provided in the design codes (API, 2011; Bhattacharya, 2019) to predict offshore pile $p-y$ curve. However, it is challenging to incorporate all influential factors, such as pile length, soil layer, soil prop-

erties and loading conditions into a simplified empirical model for predicting monopile displacement. More recently, the rapid advancement of computational techniques has led to the increased application of numerical models (Randolph & Gourvenec, 2017; Taborda et al., 2020; Zdravković et al., 2020; Royston et al., 2022). While numerical modeling serves as a potent analytical tool, it demands a substantial number of simulation runs based on observed data. This presents a significant challenge in the context of monopile analysis and prediction, especially when attempting manual back-calculation of soil properties. Because the observed data is acquired incrementally during construction stages, as opposed to simultaneous data collection.

In recent research endeavors, Bayesian probability frameworks have garnered increasing recognition as an efficacious approach for inverse parameter estimation and response prediction (Finno & Calvello, 2005; Nakamura et al., 2011; Hsein Juang et al., 2013; Nguyen & Nestorović, 2016; Wagner et al., 2020; Jin et al., 2021; Tao et al., 2021; Buckley et al., 2023; Tang et al., 2023). In contrast to conventional back-analysis methodologies, which primarily focus on the determination of fixed input variable values, a probabilistic framework takes an approach where the parameters of interest are considered stochastic variables. Subsequently, the updated parameters are expressed in terms of posterior distributions. In such circumstances, the Bayesian framework emerges as a powerful tool within the probabilistic context, facilitating parameter learning and informed decision-making. Based on this, digital twin (DT) can be constructed and enable the real time data exchange between the digital and physical twins. In offshore engineering, in particular, it has shown substantial potential in various domains, including health monitoring, pile penetration and long-term bearing capacities (Wang et al., 2021; Zhao et al., 2023; Stuyts et al., 2023).

In practice, through adaptive Bayesian updating soil parameters and constructing digital twin on offshore piles, a field engineer would benefit from: (1) properly accounting the uncertainties of input variables (2) real-time monitoring and adaptively predicting the pile response in probabilistic setting (3) providing an efficient tool for data-driven decision making on pile operation and design.

1.2 Problem statement

Modern pile installation and proper estimation is becoming increasingly complex and vital to the reliability and permanence of the foundation in question. However, in the construction process, uncertainties and insufficient information about the soil parameters lead to inaccurate predictions of pile-soil response and bearing capacities. The source of soil uncertainties may come from various reasons, such as lack of uniformity between in-situ test and laboratory experiment; spatial variability of soil profile and rationality of the constitutive model, etc. Dealing with different uncertainties sources is a challenging task. One typical soil profile can be illustrated in Figure 1.1, which shows uncertainties sources:

- Fluctuating curve indicates the spatial variability
- Non-uniformity exists between in-situ test and laboratory experiment

Furthermore, geotechnical engineering problems inherently belong to the high-dimensional realm with substantial uncertainties. The quantity of unknown distribution parameters may become excessively large to be inferred accurately from the limited sample size within the available dataset, resulting in an underdetermined problem. Although some well-established approaches for fitting pile deformations are proposed to infer the underlying soil parameters and reduce the uncertainties, this task becomes nontrivial when the number of input variables is large (i.e., $\mathcal{O}(10^2 - 10^4)$) (Lataniotis, 2019). Even if an adequate probabilistic input model can be obtained, performing the inference analysis through Monte Carlo simulation is still expensive. This poses challenges in understanding uncertainties and providing timely predictions for pile design. In such cases, the underlying model is substituted by a surrogate. In high dimension, however, the performance of surrogate models decreases, while the cost of computing and storing them increases. This is a well-known issue known as the curse of dimensionality (Verleysen & François, 2005). The surrogate computation may even be intractable when the number of input parameters is large (Lataniotis, 2019).

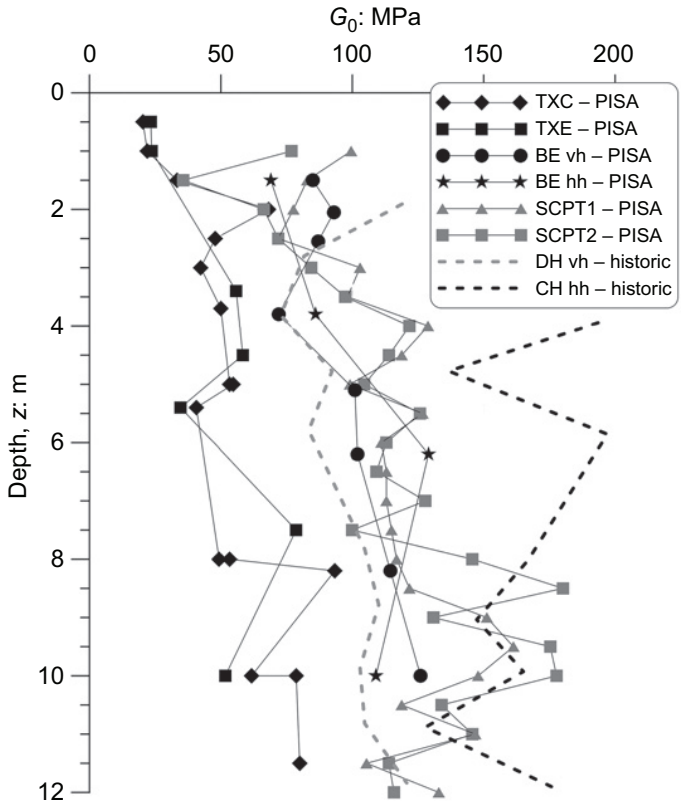

Figure 1.1: Stiffness characteristics at Cowden from Zdravković et al. (2020)

To address these challenges in real time, digital twin (DT) has gained popularity in handling abundant data and predict the pile's response in a more organized and accurate manner (Wang et al., 2021). DT makes full use of data such as physical models, sensor updates and operating history, and integrates simulation processes to real-time reproduce the dynamics of a physical system in the virtual space. More importantly, the DT model cannot only describe the current state of the physical entity, but also predict the future state. However, state-of-the-art digital twins are still relying on considerable expertise and deployment resources (Kapteyn et al., 2021), leading to an only one-off implementation and remaining limitation on providing adaptive digital models on unique offshore piles. Thus, unified and scalable models should be developed and incorporated into digital twin to enable a intelligent decision making. Details about the need for unified and scalable digital twin for piles are outlined next.

1.3 Urgent need for a unified and scalable digital twins for piles

The demand for efficiency, reliability, and safety continues to grow in offshore wind foundation constructions. Computational models are an invaluable tool for understanding complex pile behaviors for new designs, operating conditions and control strategies. This reduces the need for costly experiments or field tests. However, insights gained from a computational model are contingent on the model being an accurate reflection of the underlying soil parameters. Moreover, real-world pile foundations are constantly changing and evolving throughout their lifecycle. Using a single static computational model that ignores these differences fundamentally limits the specificity, and thus the accuracy, of the model and any insights gained through its use.

While the value proposition of digital twin has become widely appreciated in geotechnical engineering, the pile design process remains in a custom production phase. Current digital twin for offshore piles are still bespoken, relying on highly specialized implementations and thus requiring considerable resources and expertise to deploy and maintain. Therefore, it is necessary to move toward digital twins at scale by developing a rigorous and unified mathematical foundation. Based on a robust computational approach and probabilistic graphical model (Kapteyn et al., 2021), this unified mathematical foundation enables a promising application at scale in the offshore pile bearing response.

1.4 Objectives and outline

The underlying objective of this thesis is to develop a robust and scalable digital twin model for offshore piles. This endeavor will leverage cutting-edge methodologies in surrogate modeling and uncertainty quantification, all geared towards facilitating extensive predictive digital twin capabilities. In particular, the specific goals of this thesis are:

- Develop a surrogate model suitable for offshore piles characterized by high input dimensions.

- Accelerate Bayesian inversion calculations for soil parameters to reduce the uncertainties, and providing real-time pile response predictions through adaptive enrichment of observed monitoring data.

- Develop a unifying mathematical foundation for predictive digital twins for offshore piles in the form of a probabilistic graphical model.

Chapter 2

Bayesian probabilistic theory

In probabilistic theory, two main interpretations prevail: frequentist and Bayesian. The frequentist perspective views probabilities as the long-term frequencies observed in infinite trials. For example, in this context, the statement implies that, over many coin flips, heads are expected roughly half the time.

On the other hand, the Bayesian interpretation associates probability with uncertainty and information, rather than repeated trials. From the Bayesian viewpoint, the statement suggests an equal likelihood of the coin landing heads or tails in the next toss.

Depending on the amount of available data, which may range from zero to infinite, various techniques may be used:

- when no data is available to characterize the input parameters, a probabilistic model may be prescribed purely by expert judgment;
- when a large amount of data is available, the tools of statistical inference may be fully applied, like the method of moments (Wagner et al., 2020);
- when both expert judgment and very limited observations are available, Bayesian inference may be resorted to.

One big advantage of the Bayesian interpretation is that it can be used to model our events that do not have long term frequencies. Take, for example, the assessment

of the probability of structural damage to a high-rise building, the collapse of a tunnel, or the occurrence of irreversible deformation in bridge piers.This event is anticipated to occur only a limited number of times over the structure's lifetime and is not expected to happen repeatedly. Nevertheless, we ought to be able to quantify our uncertainty about this event and take appropriate actions (see chapter 3 and chapter 4).

Since data collection is inherently constrained during the progression of most engineering projects, Bayesian theory stands out as a highly effective method. Therefore, this thesis exclusively explores Bayesian methods next, while detailed information on frequentist approaches can be found in Murphy (2012).

2.1 Bayesian inference

When dealing with a limited number of data points, direct statistical estimation becomes unreliable due to substantial statistical uncertainty in the sample estimates. In this context, *Bayesian inference* provides a solution by integrating prior knowledge on parameters with a small set of observed data points. Operating in this fully probabilistic setting, all unknowns are treated as random vectors. Distribution parameters can be denoted by $\boldsymbol{x}$ as realisations of the random vector $\boldsymbol{X} : \Omega \to \mathcal{D}_{\boldsymbol{X}}$. The joint probability distribution of the combined random vector $(\boldsymbol{X}, \boldsymbol{Y}) : \Omega \to \mathcal{D}_{\boldsymbol{X}} \times D_{\boldsymbol{Y}}$ is represented by $\pi(\boldsymbol{x}; \boldsymbol{y})$. Leveraging the fundamental *sum rule* and *product rule* in probabilistic theory, the Probability Density Function (PDF) of the parameters and the data can be expressed as

$$\pi(\boldsymbol{x}|\boldsymbol{y}) = \frac{\mathcal{L}(\boldsymbol{x}; \boldsymbol{y}) \cdot \pi(\boldsymbol{x})}{\pi(\boldsymbol{y})} \tag{2.1}$$

which is also known as *Bayes' theorem* or *Bayes' rule*. In Bayesian terminology, this distribution $\pi(\boldsymbol{x}|\boldsymbol{y})$ is called the posterior distribution and it is calculated by prior $\pi(\boldsymbol{x})$, likelihood $\mathcal{L}(\boldsymbol{x}; \boldsymbol{y}) \stackrel{\text{def}}{=} \pi(\boldsymbol{y}|\boldsymbol{x})$ and the evidence $\pi(\boldsymbol{y})$. These terms in Equation (2.1) have practical significance that we will briefly summarise next.

- Prior $\pi(\boldsymbol{x})$: In the Bayesian paradigm, before considering the data the parameters $\boldsymbol{x}$ are treated as realisations from a random vector $\boldsymbol{X}$ which is assumed to follow the so-called prior distribution.

- Likelihood function $\mathcal{L}(\boldsymbol{x};\boldsymbol{y})$: The likelihood function is a measure of how well the prescribed parametric distribution $\pi(\boldsymbol{y}|\boldsymbol{x})$ describes the data. To evaluate the likelihood $\mathcal{L}(\boldsymbol{x};\boldsymbol{y})$, some ingredients are needed: a computational forward model $\mathcal{M}$, a set of input parameters $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{X}}$ that need to be inferred, and a set of experimental data $\boldsymbol{y}$. The forward model $\boldsymbol{x} \rightarrow \boldsymbol{M}(\boldsymbol{x})$ is a mathematical representation of the system under consideration. All models are always simplifications of the real world. Thus, to connect model predictions to the observations $\boldsymbol{y}$, a *discrepancy term* $\boldsymbol{\varepsilon}$ shall be introduced. We consider the following well-established format:

$$\boldsymbol{y} = \mathcal{M}(\boldsymbol{x}) + \boldsymbol{\varepsilon} \tag{2.2}$$

where $\epsilon \in \mathbb{R}^{N_{\text{out}}}$ is the term that describes the discrepancy between an experimental observation $\boldsymbol{y}$ and the model prediction. For the sake of simplicity, we consider it as an additive *Gaussian discrepancy* with zero mean and a covariance matrix $\boldsymbol{\Sigma}$ in this introduction:

$$\boldsymbol{\varepsilon} \in \mathcal{N}(\boldsymbol{\varepsilon}|\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{2.3}$$

It is noted that simple Gaussian discrepancy assumption is only one out of many possible models. In a more general setting, other distributions for the discrepancy are used as well (Wagner et al., 2022). Due to the widespread used of the additive Gaussian models in engineering disciplines, the thesis is limited to Gaussian type. If N independent measurement $\boldsymbol{y_i}$ are available and gathered in the data set $\boldsymbol{y} \overset{\text{def}}{=} \{\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(N)}\}$, the likelihood can thus be written as:

$$\begin{aligned} \mathcal{L}(\boldsymbol{x};\boldsymbol{y}) &= \prod_{i=1}^{N} N(\boldsymbol{y_i}|\mathcal{M}(\boldsymbol{x}), \boldsymbol{\Sigma}) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{(2\pi)^{N_{\text{out}}} \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}\left(\boldsymbol{y_i} - \mathcal{M}(\boldsymbol{x})\right)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{y_i} - \mathcal{M}(\boldsymbol{x})\right)\right) \end{aligned} \tag{2.4}$$

- Evidence $\pi(\boldsymbol{y})$: In Bayesian inference, $\pi(\boldsymbol{y})$ is often seen as a normalizing factor that

ensures that posterior PDF integrates to one:

$$\pi(\boldsymbol{y}) \overset{\text{def}}{=} \int_{\mathcal{D}_{\boldsymbol{X}}} \mathcal{L}(\boldsymbol{x};\boldsymbol{y})\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{2.5}$$

A schematic Bayesian inference in two dimensional space is displayed in Figure 2.1. The plots show the various elements of the Bayesian inference procedure in the parameter and data spaces. In the parameter space, with new experimental data comes in, the posterior is more concentrated than the prior distribution.

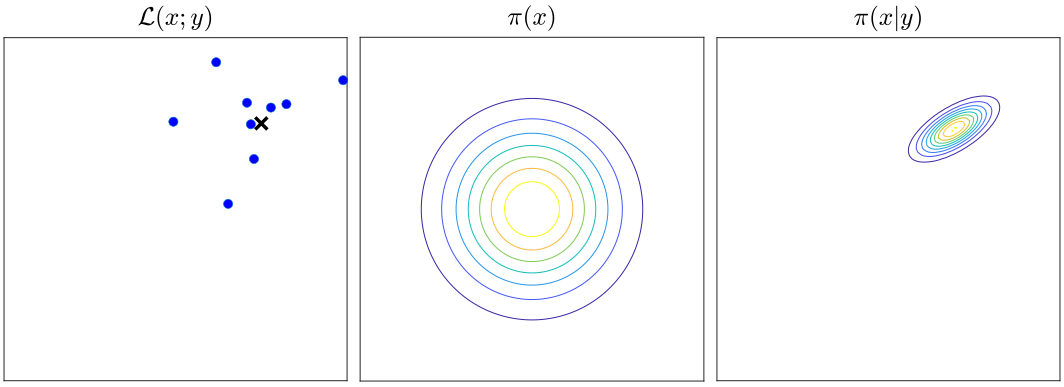

Figure 2.1: Bayesian inference in 2D space

2.2 Posterior quantities of interest

Under the Bayesian paradigm the posterior distribution $\pi(\boldsymbol{x}|\boldsymbol{y})$ is the solution of the inverse problem. However, in practice it can also serve as an intermediate result that is further processed for interpretation or prediction purpose. Furthermore, the full distribution can contain too much information to allow statements about the inferred parameters. Therefore, it is common to process the posterior and extract certain quantities of interest that summarize the inversion results more concisely.

In many applications, one is only interested in a single parameter $\hat{\boldsymbol{x}}$, i.e., the one that characterise the inversion most suitably. The two most common *point estimation* methods are the posterior mean and maximum a posterior (MAP). The posterior mean is given as:

$$x^{\text{mean}} = \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}] = \int_{\mathcal{D}_{\boldsymbol{X}|\boldsymbol{y}}} x\pi(\boldsymbol{x}|\boldsymbol{y})\mathrm{d}\boldsymbol{x} \tag{2.6}$$

It reflects what we expect the parameter value to be after the inference. The MAP parameter, as the mode of the posterior distribution on the other hand, is the one maximises the posterior

The practical computation of posterior distributions is not trivial, since computing evidence $\pi(\boldsymbol{y})$ is usually not a tractable problem.

To make the calculation more feasible, we usually choose the *conjugate prior* (Gelman et al., 1995) to the likelihood, so the integral can be represented analytically. However, in the general cases, sampling methods shall be used.

2.3 Exact inference

2.3.1 Variant elimination

2.3.2 Belief propagation

2.4 Approximation inference

2.4.1 Expectation maximization algorithm

2.4.2 Ensemble Kalman filter

2.4.3 Sequential Monte Carlo

2.4.4 Markov Chain Monte Carlo

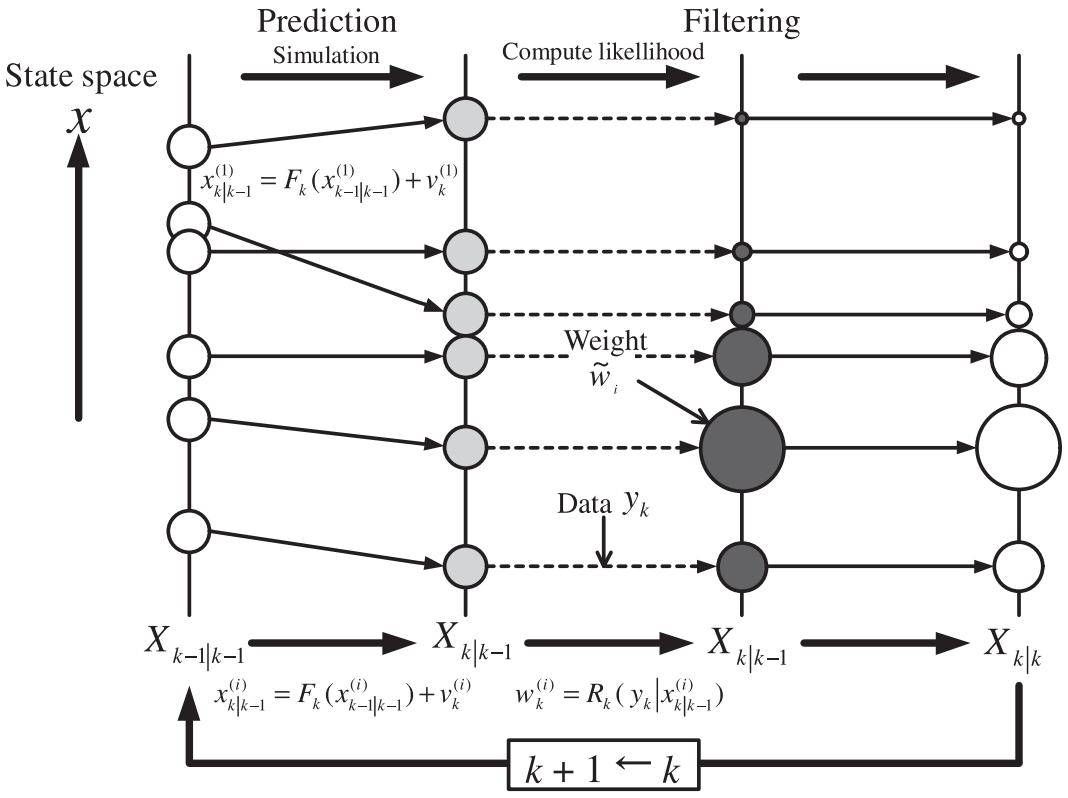

Figure 2.2: Bayesian inference in 2D space

Chapter 3

Uncertainty quantification in parameter identification

3.1 Inverse problem

3.1.1 Forward problem

3.1.2 High dimensional problem

3.2 Surrogate model-Spectral method

Computer modelling is used in nearly every field of science and engineering. Often, these computer codes model complex phenomena, have many input parameters, and are expensive to evaluate. In order to explore the behavior of the model under uncertainty (e.g., uncertainty propagation, parameter calibration from data or sensitivity analysis), many model runs are required. However, if the model is costly, only a few model evaluations can be afforded, which often do not suffice for thorough uncertainty quantification. In engineering and applied sciences, a popular work-around in this situation is to construct a reduced-order surrogate model. A reduced-order surrogate model is a cheap-to-evaluate proxy of the original model, which typically can be constructed from a relatively small number of model evaluations and approximates the input-output relation of the original

model well. Since the surrogate model is cheap to evaluate, uncertainty quantification can be performed at a low cost by using the surrogate model instead of the original model. Therefore, surrogate modelling aims at constructing a metamodel that provides an accurate approximation to the original model while requiring as few model evaluations as possible for its construction. A surrogate model $\tilde{\mathcal{M}}$ can be expressed as:

$$\tilde{\mathcal{M}}(\boldsymbol{X}) \overset{\text{def}}{=} \mathcal{M}(\boldsymbol{X}) - \mathcal{R}(\boldsymbol{X}) \tag{3.1}$$

$$\tilde{\mathcal{M}}(\boldsymbol{X}) \overset{\text{def}}{=} \mathcal{M}(\boldsymbol{X}) - \mathcal{R}(\boldsymbol{X}) \tag{3.2}$$

where $\mathcal{R}$ is the residual between the original model and the surrogate.

Why not neural networks?

Model reduction lets us create approximate models that are fast to solve, and — importantly — it provides us a rigorous mathematical basis on which to establish strong guarantees of accuracy of the low-dimensional model. This is in contrast with black-box machine learning methods (ANN), where we just have to hope that our training data was rich enough to yield a sufficiently accurate surrogate model. This is especially problematic for engineering applications where we often need to issue extrapolatory predictions.

3.2.1 Polynomial chaos expansion

Metamodelling (or surrogate modelling) attempts to offset the increased costs of stochastic modelling by substituting the expensive-to-evaluate computational models (e.g. finite element models, FEM) with inexpensive-to-evaluate surrogates. Polynomial chaos expansions (PCE) are a powerful metamodelling technique that aims at providing a functional approxmation of a computational model through its spectral representation on a suitably built basis of polynomial functions.

Different from other surrogate modelling approaches such as support vector machine, Gaussian process regression, or neural networks, the mathematical theory underlying reduced-order models can lead to more reliable and robust predictive capability (Frangos et al., 2010; Kapteyn et al., 2021).

Different from Monte Carlo simulation (MCS) which is based on point-to-point exploring the output space, PCE assumes a generic structure, which better exploits the available runs of the FE realizations.

3.3 Dimensionality reduction

Geotechnical problems inherently involve high dimensionality, posing challenges for learning methods like surrogate modeling. Technical constraints impact the storage and processing of such huge amount of data. Furthermore, as input and output data expand, independent scalar surrogate models show inadequate in accurately capturing the covariance matrix of the original data, leading to less reliable predictions. Consequently, in high dimensional space, the need for dimensionality reduction technique (DR) becomes more critical.

3.3.1 Linear DR technique

3.3.2 Nonlinear DR technique

3.4 DR-based surrogate modelling

3.5 Inverse problems in UQ

When faced with large-scale forward models characteristic of many engineering and science applications, high computational cost arises from: (1) In the large-scale setting, performing thousands or millions of forward simulations is often computationally intractable; (2) the dimension of the input space are complex; (3) sampling may be complicated by the large dimensionality of the input space. Thus, to reduce the computational cost of solving of a statistical inverse problem, methods can be broadly in three groups: (1) Surrogate models to accelerate a forward simulation; (2) Reduce the dimension of the input space,i.e., sensitive analysis; (3) Efficient sampling method to posterior,i.e., MCMC.

3.5.1 Inverse problems

3.5.2 Bayesian inference

3.5.3 Bayesian calibration

3.5.4 Sampling methods

Using the CDF

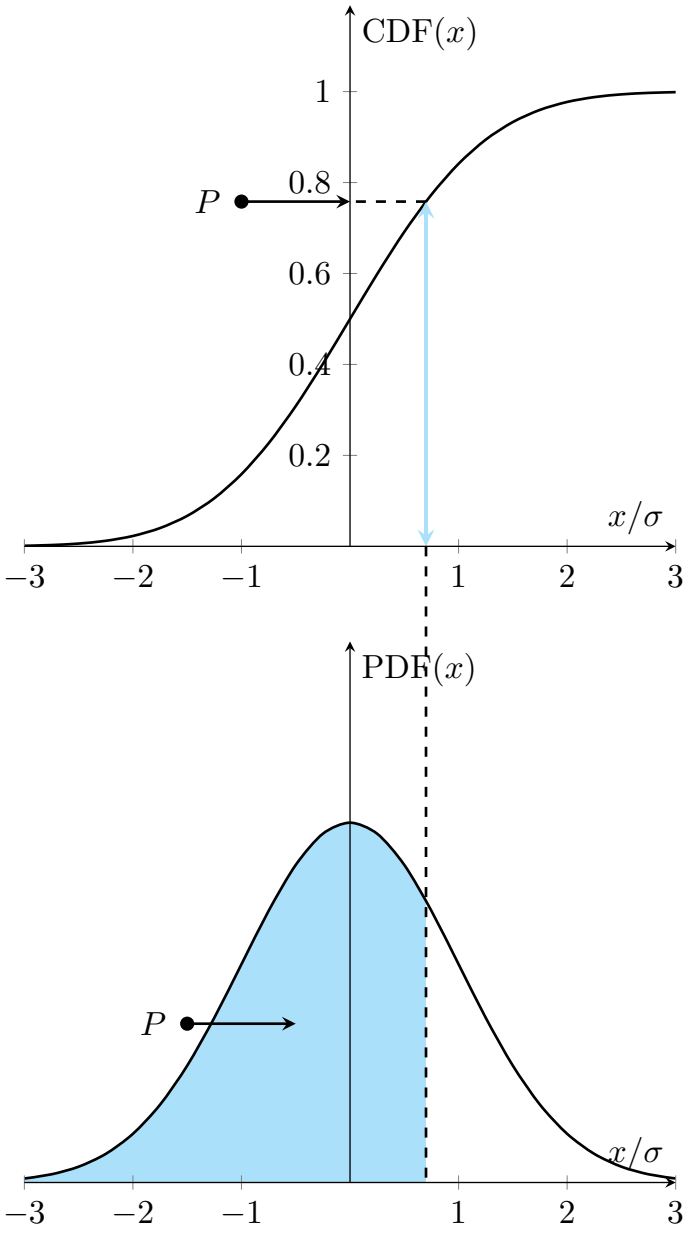

Figure 3.1: Sampling using an inverse CDF

The simplest method for sampling from a univariate distribution is based on the inverse probability transform. As shown in Figure 3.1, if we can get the cumulative probability density function, then we can easily generate samples by computing $x = \text{CDF}(\mathcal{U})$. $\mathcal{U}$ follows uniform distribution $\mathcal{U} \sim U(0, 1)$.

Rejection sampling

When the inverse cdf method cannot be used, one simple alternative is to use rejection sampling. In rejection sampling, we create a proposal distribution $q(x)$ which satisfies $Mq(x) \geq \tilde{p}(x)$, for some constant M, where $\tilde{p}(x)$ is an unnormalized version of $p(x)$ (i.e., $p(x) = \tilde{p}(x)/Z$ for some unknown constant Z). The function $Mq(x)$ provides an upper envelope for $\tilde{p}$. We then sample $x \sim q(x)$, which corresponds to picking a random x location, and then we sample $u \sim U(0, 1)$ which corresponds to picking a random height (y location) under the envelope. If $u \geq \frac{\tilde{p}(x)}{Mq(x)}$, we reject the sample,otherwise we accept it. See Figure 3.2, where acceptance region is shown shaded, and the rejection region is the white region between the shaded zone and the upper envelope. But, large-dimensional

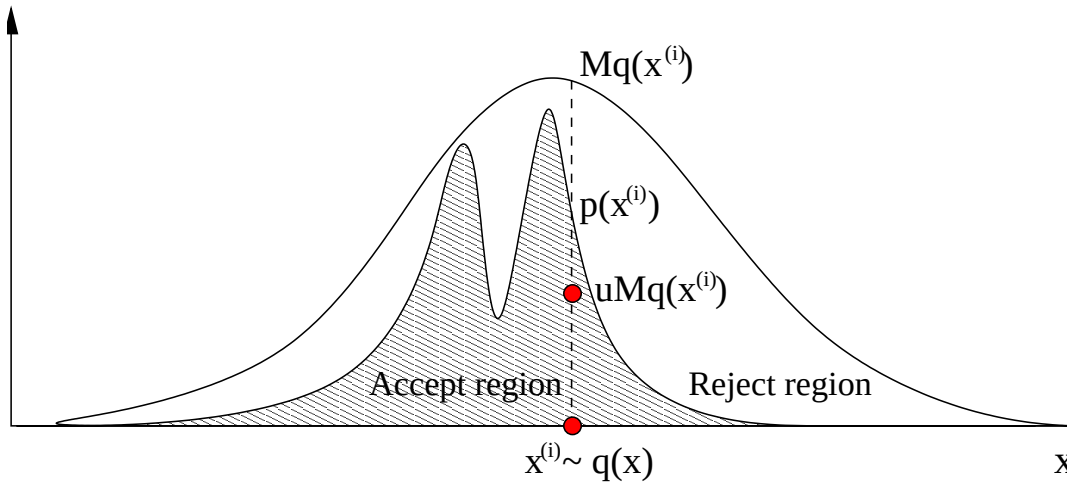

Figure 3.2: Schematic illustration of rejection sampling from Andrieu et al. (2003)

spaces tend to be very empty, and the chances that this method accepts a point may be dramatically low when working with multidimensional spaces. MCMC

Importance sampling

Often, in practical Bayesian models, it is not possible to obtain samples directly from p.x j y1WT / due to its complicated functional form.

Sequential Monte Carlo

Markov chain Monte Carlo

MCMC Some Monte Carlo methods, including rejection sampling, importance sampling and particle filtering. The trouble with these methods is that they do not work well in high dimensional spaces. The most popular method for sampling from high-dimensional distributions is Markov chain Monte Carlo or MCMC. In a survey by *SIAM News* , MCMC was placed in the top 10 most important algorithms of the 20th century (Murphy, 2012).

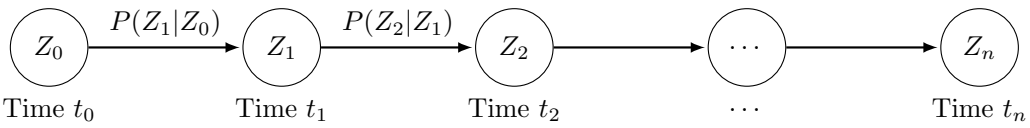

Figure 3.3: Markov Chain process

3.5.5 Choice of sampling method

3.5.6 Seqential Bayesian inference

3.6 Sequential enrichment for surrogate model

Traditional large-scale physics-based models are intractable to solve real-time, many-query context problem.

Instead of sampling the whole experimental design at once, it has been proposed to use sequential enrichment. Starting with a small experimental design, additional points are chosen based on the last computed sparse solution. In the context of machine learning, sequential sampling is also known as active learning. In all cases, numerical examples show that the sequential strategy generally leads to solutions with a smaller validation error compared to non-sequential strategies

3.7 Sensitivity analysis

Chapter 4

Predictive digital twins at scale for piles

This chapter develops a mathematical and computational foundation for digital twins of piles.

While the value proposition of digital twins has become widely appreciated, the technology itself remains in a custom production phase.

4.1 State space model

State space model, also called dynamic model, usually represents a class of directed probabilistic graphical model that describes the dependence between the hidden variables $\boldsymbol{Z} = (\boldsymbol{Z_0}, \boldsymbol{Z_1}, \boldsymbol{Z_2}, ..., \boldsymbol{Z_t})$ and the observed variables $\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, ..., \boldsymbol{X_t})$. This model also enables a dynamical updating for state estimation and control because they allow for recursive analysis of the systems in the time domain. As shown in Figure 4.1, this sequential updating feature is very suitable to geotechnical problems, because in our real life the observations typically appear in time series.

Based on the different assumptions for the model, state space model can be categorized as: Hidden Markov model, linear Gaussian state space model and Nonlinear non-Gaussian state space model.

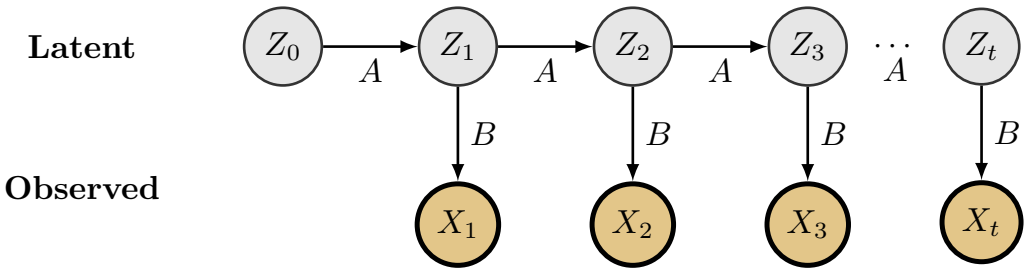

Figure 4.1: State space model

where gray circles $\boldsymbol{Z}$ are latent variable; bold outlines $\boldsymbol{X}$ are observed quantities; A is the transition equation; B is the observation equation. A and B are not time-varying.

4.1.1 Hidden Markov model

If the latent variables $\boldsymbol{Z} = (\boldsymbol{Z_0}, \boldsymbol{Z_1}, \boldsymbol{Z_2}, ..., \boldsymbol{Z_t})$ in **??** are discrete, we can treat the state space model as a hidden Markov model. A hidden Markov model (HMM) allows us to talk about both observed events $\boldsymbol{X}$ and hidden discrete events $\boldsymbol{Z}$. A first-order hidden Markov model instantiates two simplifying assumptions as:

- First, as with a first-order Markov chain, the probability of a particular state depends only on the previous state:

$$P(Z_t|Z_0, Z_1, ..., Z_{t-1}) = P(Z_t|Z_{t-1}) \tag{4.1}$$

- Second, the probability of an observation X_t depends only on the state that produced the observation Z_t and not on any other states or any other observations:

$$P(Z_t|Z_0, Z_1, ..., Z_{t-1}) = P(Z_t|Z_{t-1}) \tag{4.2}$$

In geotechnical area, filtering problem is more a quantity of interest. That is to say, given the current belief state $p(Z_{t-1}|X_{1:t-1})$, the primary concern next is how to calculate the belief state $p(Z_t|X_{1:t-1})$. A prominent application for Hidden Markov model is addressing soil classification challenges with time series data. This is due to the discrete nature of the labeling predicament inherent in this context. However, the limitation for the HMM

is also obvious. For the continuous state changing with time (e.g., soil stress, loading force or cumulative plastic strain), HMM is not suitable for such problem because infinite transition matrix A does not exist. To solve this, we will discuss linear Gaussian model and nonlinear non-Gaussian model next.

4.1.2 Linear Gaussian state space model

Linear Gaussian state space model is an exact Bayesian filtering solution, also called Kalman filter. Kalman filter describe a very specific setting, i.e., linear transition/observation equations and Gaussian noises. Since everything is Gaussian, we can perform the prediction and update steps in closed form, as we explain below:

$$\begin{aligned} &\boldsymbol{z_t} = \boldsymbol{A z_{t-1}} + \boldsymbol{w_t} && \boldsymbol{w_t} \sim \mathcal{N}(0, \boldsymbol{Q}) \\ &p(\boldsymbol{z_t}|\boldsymbol{z_{t-1}}) = \mathcal{N}(\boldsymbol{A z_{t-1}}, \boldsymbol{Q}) \\ &\boldsymbol{x_t} = \boldsymbol{B z_t} + \boldsymbol{v_t} && \boldsymbol{v_t} \sim \mathcal{N}(0, \boldsymbol{R}) \\ &p(\boldsymbol{x_t}|\boldsymbol{z_t}) = \mathcal{N}(\boldsymbol{B z_t}, \boldsymbol{R}) \end{aligned} \tag{4.3}$$

which stands for the prediction step and correction step, respectively.

The detailed principle of Kalman filter is not extensively emphasized here, because in geotechnical engineering, the transition and observation equations always show high nonlinearity. Once we add the SSM parameters to the state space model, the model is generally no longer linear Gaussian. Consequently we must use some of the approximate online inference methods to be discussed below.

4.1.3 Nonlinear non-Gaussian state space model

Desipte the fact that some methods like Ensemble Kalman filter (EnKF), Unscented Kalman filter (UKF) or Extended kalman filter (EKF) can relax the linearity at some degree. These methods are still designed only for Gaussian posterior, which is not suitable for geotechnical problems with high nonlinearity. To handle the arbitrary posteriors, particle filtering which is based on Monte Carlo, can approximate the posterior with the increasing sampling points.

4.2 Probabilistic graphical model: Control theory

4.3 Partially observable Markov decision process

As shown in Figure 4.2, based on Markov Chain, with introducing Rewards and Actions, it can form the basis of Partially observed Markov decision process.

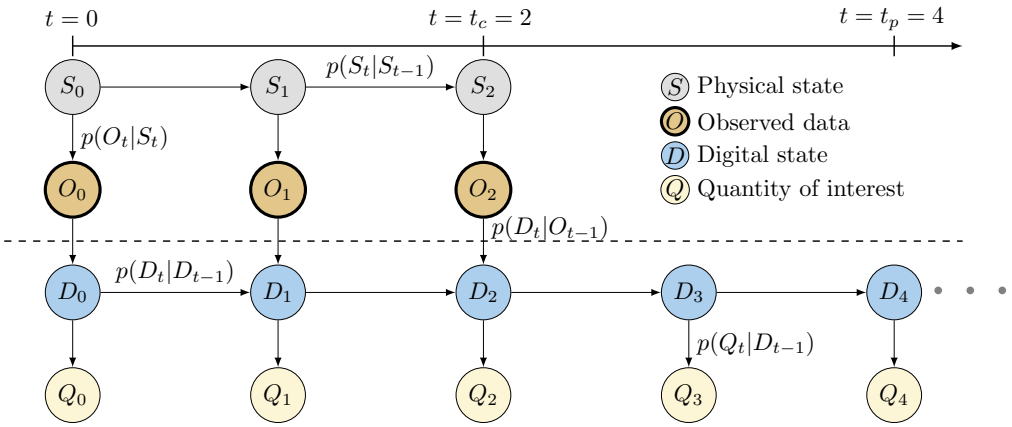

Figure 4.2: Digital twin

Generally, Digital Twin can be divided into two main parts, including (1) calibration and assimilation (2) Prediction, as shown in Equation (4.4) and Equation (4.5).

$$\begin{aligned} &p(D_0, ..., D_{t_c}, Q_0, ..., Q_{t_c}, R_0, ..., R_{t_c} | o_0, ..., o_{t_c}, u_0, ..., u_{t_c}) \\ &= \prod_{t=0}^{t_c} [\phi_t^{update} \phi_t^{QoI} \phi_t^{evaluation}] \end{aligned} \tag{4.4}$$

$$\begin{aligned} &p(D_0, ..., D_{t_p}, Q_0, ..., Q_{t_p}, R_0, ..., R_{t_p}, U_{t_c+1}, ..., U_{t_p} | o_0, ..., o_{t_c}, u_0, ..., u_{t_c}) \\ &\propto \prod_{t=0}^{t_p} [\phi_t^{dynamics} \phi_t^{QoI} \phi_t^{evaluation}] \prod_{t=0}^{t_c} \phi_t^{assimilation} \prod_{t=t_c+1}^{t_p} \phi_t^{control} \end{aligned} \tag{4.5}$$

4.4 Computational model-ICFEP

4.5 Planning and prediction via digital twin

Chapter 5

Work Plan

5.1 Stage 1

Calibrate the models in Figure 5.1 with partially observed Markov decision process.

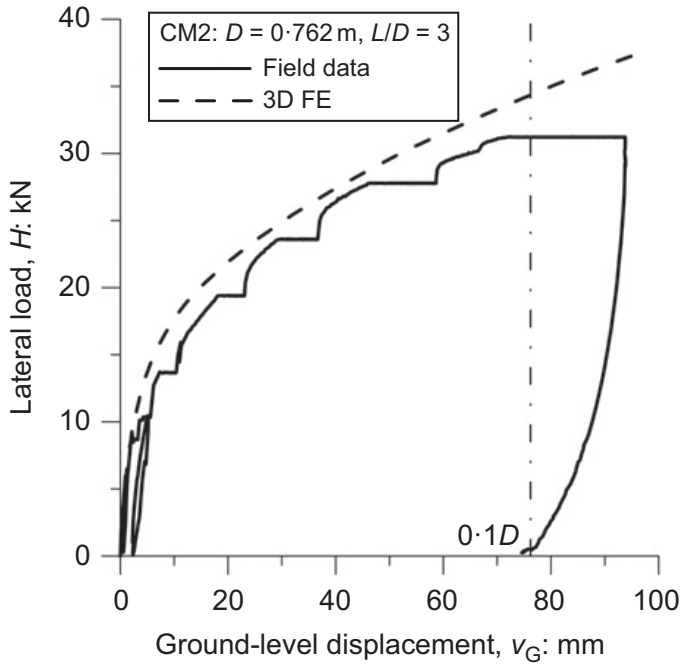

Figure 5.1: CM2 pile load displacement from Zdravković et al. (2020)

- Software: ICFEP-Likelihoods and observed data.
- Constitutive model: clay in Zdravković et al. (2020) and sand in Taborda et al. (2020).

- Consider the soil profile variance-Create the random field (scale of fluctuation in ICFEP).

Objective: Ensure the soil parameters in digital model can reveal unique characteristics of piles.

5.2 Stage 2

In operational Phase, Based on Partially observed Markov decision process method, continue the assimilation process: extend the digital twin capability to capture the piles response during loading.

5.3 Stage 3

Extension to Prediction

5.4 Time plan

Table 5.1: PhD timeline

month	0	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48
Literature review	✓	✓	✓														
Numerical modelling (Data collection)		✓	✓	✓	✓	✓	✓										
Statistics Methods learning		✓	✓	✓	✓	✓	✓	✓	✓								
Statistics analysis calibration			✓	✓	✓												
Statistics analysis assimilation						✓	✓	✓	✓	✓	✓						
Statistics analysis prediction												✓	✓	✓			
Thesis writing															✓	✓	✓
Journal/Conference								✓				✓					✓

References

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to mcmc for machine learning. *Machine learning*, *50*, 5–43.

API. (2011). *Geotechnical and foundation design considerations.* API Washington, DC.

Bhattacharya, S. (2019). *Design of foundations for offshore wind turbines.* John Wiley & Sons.

Buckley, R., Chen, Y. M., Sheil, B., Suryasentana, S., Xu, D., Doherty, J., & Randolph, M. (2023). Bayesian optimization for cpt-based prediction of impact pile drivability. *Journal of Geotechnical and Geoenvironmental Engineering*, *149*(11), 04023100.

Byrne, B. W., & Houlsby, G. T. (2003). Foundations for offshore wind turbines. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *361*(1813), 2909–2930.

Finno, R. J., & Calvello, M. (2005). Supported excavations: observational method and inverse modeling. *Journal of geotechnical and geoenvironmental engineering*, *131*(7), 826–836.

Frangos, M., Marzouk, Y., Willcox, K., & van Bloemen Waanders, B. (2010). Surrogate and reduced-order modeling: a comparison of approaches for large-scale statistical inverse problems. *Large-Scale Inverse Problems and Quantification of Uncertainty*, 123–149.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* Chapman and Hall/CRC.

Hsein Juang, C., Luo, Z., Atamturktur, S., & Huang, H. (2013). Bayesian updating of soil parameters for braced excavations using field observations. *Journal of Geotechnical and Geoenvironmental Engineering*, *139*(3), 395–406.

Jin, Y., Biscontin, G., & Gardoni, P. (2021). Adaptive prediction of wall movement during excavation using bayesian inference. *Computers and Geotechnics*, *137*, 104249.

Kapteyn, M. G., Pretorius, J. V., & Willcox, K. E. (2021). A probabilistic graphical model foundation for enabling predictive digital twins at scale. *Nature Computational Science*, *1*(5), 337–347.

Lataniotis, C. (2019). *Data-driven uncertainty quantification for high-dimensional engineering problems* (Unpublished doctoral dissertation). ETH Zurich.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Nakamura, K., Yamamoto, S., & Honda, M. (2011). Sequential data assimilation in geotechnical engineering and its application to seepage analysis. In *14th international conference on information fusion* (pp. 1–6).

Nguyen, L. T., & Nestorović, T. (2016). Nonlinear kalman filters for model calibration of soil parameters for geomechanical modeling in mechanized tunneling. *Journal of Computing in Civil Engineering*, *30*(2), 04015025.

Randolph, M., Cassidy, M., Gourvenec, S., & Erbrich, C. (2005). Challenges of offshore geotechnical engineering. In *Proceedings of the international conference on soil mechanics and geotechnical engineering* (Vol. 16, p. 123).

Randolph, M., & Gourvenec, S. (2017). *Offshore geotechnical engineering*. CRC press.

Royston, R., Sheil, B. B., & Byrne, B. W. (2022). Undrained bearing capacity of the cutting face for an open caisson. *Géotechnique*, *72*(7), 632–641.

Stuyts, B., Weijtjens, W., & Devriendt, C. (2023). Development of a semi-structured database for back-analysis of the foundation stiffness of offshore wind monopiles. *Acta Geotechnica*, *18*(1), 379–393.

Taborda, D. M., Zdravković, L., Potts, D. M., Burd, H. J., Byrne, B. W., Gavin, K. G., … others (2020). Finite-element modelling of laterally loaded piles in a dense marine sand at dunkirk. *Géotechnique*, *70*(11), 1014–1029.

Tang, C., Cao, Z.-J., Hong, Y., & Li, W. (2023). State space model of undrained triaxial test data for bayesian identification of constitutive model parameters. *Géotechnique*, 1–15.

Tao, Y.-q., Sun, H.-l., & Cai, Y.-q. (2021). Bayesian inference of spatially varying parameters in soil constitutive models by using deformation observation data. *International Journal for Numerical and Analytical Methods in Geomechanics*, *45*(11), 1647–1663.

Verleysen, M., & François, D. (2005). The curse of dimensionality in data mining and time series prediction. In *International work-conference on artificial neural networks* (pp. 758–770).

Wagner, P.-R., Fahrni, R., Klippel, M., Frangi, A., & Sudret, B. (2020). Bayesian calibration and sensitivity analysis of heat transfer models for fire insulation panels. *Engineering structures*, *205*, 110063.

Wagner, P.-R., Nagel, J., Marelli, S., & Sudret, B. (2022). *UQLab user manual – Bayesian inversion for model calibration and validation* (Tech. Rep.). Chair of Risk, Safety and Uncertainty Quantification, ETH Zurich, Switzerland. (Report UQLab-V2.0-113)

Wang, M., Wang, C., Hnydiuk-Stefan, A., Feng, S., Atilla, I., & Li, Z. (2021). Recent progress on reliability analysis of offshore wind turbine support structures considering digital twin solutions. *Ocean Engineering*, *232*, 109168.

Zdravković, L., Jardine, R. J., Taborda, D. M., Abadias, D., Burd, H. J., Byrne, B. W., … others (2020). Ground characterisation for pisa pile testing and analysis. *Géotechnique*, *70*(11), 945–960.

Zhao, X., Dao, M. H., & Le, Q. T. (2023). Digital twining of an offshore wind turbine on a monopile using reduced-order modelling approach. *Renewable Energy*, *206*, 531–551.