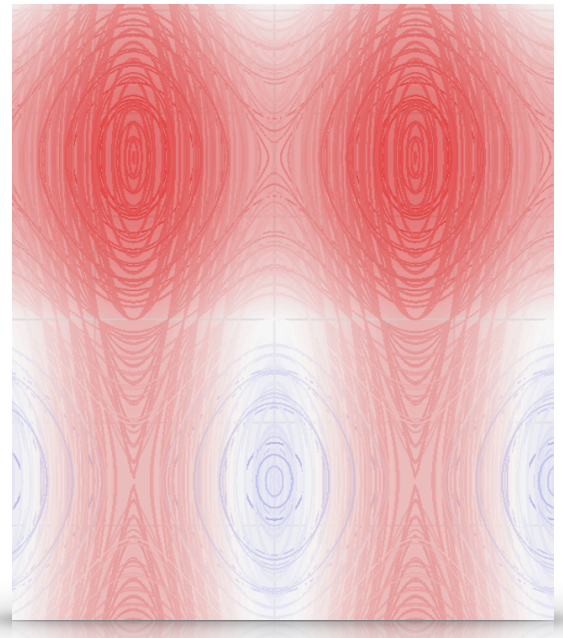
A novel active learning-based Gaussian process metamodelling strategy for estimating the full probability distribution in forward UQ analysis



Z. Wang
M. Broccardo

Structural Safety
10.1016/j.strusafe.2020.101937

This page is intentionally left blank.

A novel active learning-based Gaussian process metamodelling strategy for estimating the full probability distribution in forward UQ analysis

Ziqi Wang[a], Marco Broccardo[b,c]

[a]*Earthquake Engineering Research and Test Center, Guangzhou University, China*
[b]*Swiss Seismological Service, SED, ETH Zürich, Switzerland*
[c]*Department of Civil Engineering and Industrial Design, University of Liverpool, United Kingdom*

Abstract

This paper proposes an active learning-based Gaussian process (AL-GP) metamodelling method to estimate the cumulative as well as complementary cumulative distribution function (CDF/CCDF) for forward uncertainty quantification (UQ) problems. Within the field of UQ, previous studies focused on developing AL-GP approaches for reliability (rare event probability) analysis of expensive black-box solvers. A naive iteration of these algorithms with respect to different CDF/CCDF threshold values would yield a discretized CDF/CCDF. However, this approach inevitably leads to a trade off between accuracy and computational efficiency since both depend (in opposite way) on the selected discretization. In this study, a specialized error measure and a learning function are developed such that the resulting AL-GP method is able to efficiently estimate the CDF/CCDF for a specified range of interest without an explicit dependency on discretization. Particularly, the proposed AL-GP method is able to simultaneously provide accurate CDF and CCDF estimation in their median-low probability regions. Three numerical examples are introduced to test and verify the proposed method.

Keywords: Active learning, distribution function, Gaussian process model, rare event simulation, uncertainty quantification

1. Introduction

In a broad sense, uncertainty quantification (UQ) refers to the theory and practice to obtain quantitative understanding on the influences of uncertainties present within computational or real physical models. An incomplete list of possible UQ analysis includes the prediction of probability [1], the interpolation/extrapolation for the most likely outcome [2], the validation/calibration/correction of computational model [3], etc. There are intrinsic connections between various branches of UQ, and attempts are made to develop unified UQ frameworks [4][5]. This study focuses on a central problem in forward UQ problems, the estimation of distribution function, i.e. cumulative and complementary cumulative distribution function (CDF/CCDF). Specifically, consider a system with a finite set of basic random variables $\boldsymbol{X} = [X_1, X_2, ..., X_n]$ representing the source of randomness. Given the joint probability distribution of $\boldsymbol{X}$, we are interested in the probability distribution of a system output, Y, propagated from $\boldsymbol{X}$. The output Y can be any variables selected to describe the system performance or behavior of interest. The deterministic mapping from $\boldsymbol{X}$ to Y is written as

$$Y = \mathcal{M}(\boldsymbol{X}) \ . \tag{1}$$

In a more general setting Y can also be a vector of random variables, yet this paper focuses on the case where Y is unidimensional*. Although Eq.(1) looks trivial, the model function $\mathcal{M}(\cdot)$ may involve computationally expensive models, e.g. finite element algorithms.

[1]Corresponding author: Ziqi Wang; ziqidwang@yahoo.com
[2]Corresponding author: Marco Broccardo; bromarco@ethz.ch
*Section 5 provides more details on extending to a multidimensional output $\boldsymbol{Y}$

Using Eq.(1), the CDF of Y, denoted by $F_Y(y)$, can be expressed as

$$F_Y(y) \equiv \mathbb{P}\left(Y \leq y\right) = \mathbb{P}\left(\mathcal{M}\left(\boldsymbol{X}\right) \leq y\right) = \int_{\mathcal{M}(\boldsymbol{x}) \leq y} f_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x}\,, \tag{2}$$

where $\mathbb{P}(\cdot)$ denotes probability, $f_{\boldsymbol{X}}(\boldsymbol{x})$ denotes the joint probability density function (PDF) of the basic random variables $\boldsymbol{X}$. The CCDF of Y, denoted by $\bar{F}_Y(y)$ can be expressed as

$$\bar{F}_Y(y) \equiv \mathbb{P}\left(Y > y\right) = \int_{\mathcal{M}(\boldsymbol{x}) > y} f_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x} = 1 - F_Y(y)\,. \tag{3}$$

In terms of the trivial connection between Eq.(2) and Eq.(3), it seems given the CDF $F_Y(y)$ one could compute the CCDF simply by $\bar{F}_Y(y) = 1 - F_Y(y)$. However, to obtain satisfactory accuracy in the low probability region of the CCDF, instead of using $1 - F_Y(y)$, one typically needs to solve the integration in Eq.(3)*. Note that by definition the low probability region of the CDF is in the left tail of $f_Y(y)$, while the low probability region of the CCDF is in the right tail. This suggests that in the space of $\boldsymbol{X}$ the critical domains that contribute to the tail probability of the CDF and the CCDF are different. Therefore, one typically needs to separately run numerical integration algorithms to capture these critical domains so that the low probability regions of the CDF and CCDF can be accurately estimated.

The major challenges of solving Eq.(2) or Eq.(3) include: i) $\mathcal{M}(\cdot)$ is computationally expensive, especially when it involves complex computational codes which are difficult to be manipulated intrusively; ii) the dimensionalilty of $\boldsymbol{X}$ is large; iii) the tail accuracy of the CDF/CCDF is of particular interest. This paper focuses on addressing the i) and iii) challenges, while the possible extension to high-dimensional problems is discussed in Section 5. Moreover, the paper illustrates CDF/CCDF estimations down to the tail region with probability 10^{-4}, for lower probability problems one may need to introduce variance-reduction Monte Carlo techniques (see Section 5 for a discussion).

In UQ practice, non-intrusive numerical methods, e.g. deterministic cubature [6], metamodelling [7] and Monte Carlo simulation techniques [8], are becoming increasingly popular in solving Eq.(2) or Eq.(3), given the vast legacy of complex computational codes for model function $\mathcal{M}(\cdot)$. Intrusive approaches such as the stochastic spectral methods [9] can be attractive when $\mathcal{M}(\cdot)$ is not given in a black box fashion and can be customized for the given UQ problem. In this paper we focus on non-intrusive approaches where $\mathcal{M}(\cdot)$ is assumed to be a black-box solver. Among various non-intrusive UQ methods, Gaussian process (also called Kriging) and polynomial chaos expansion (PCE) based matamodelliing approaches are being widely used to estimate the CDF/CCDF and other response statistics [10][11][12][13], due to their flexibility and efficiency in approximating complex model functions.

To further improve the accuracy and efficiency of metamodelling approaches, instead of performing a global design of experiment, active learning techniques can be introduced into the metamodelling framework. In particular, for reliability analysis an active learning-based Gaussian process (AL-GP) metamodelling strategy [14] was recently introduced to estimate rare event probabilities. The approach has proven a remarkable success in solving low-medium dimensional reliability problems. In the context of Eq.(2), the AL-GP approach for reliability analysis can be interpreted as to train a Gaussian process model to perform interpolation/extrapolation within the domain defined by $G(\boldsymbol{x}|y) = \mathcal{M}\left(\boldsymbol{x}\right) - y \leq 0$, so that the probability $\mathbb{P}(G(\boldsymbol{x}|y) \leq 0)$ can be estimated by the metamodel. Clearly, $\mathbb{P}(G(\boldsymbol{x}|y) \leq 0)$ as a function of y is by definition the CDF of Y. Therefore, the AL-GP approach can be directly used to estimate the CDF for a specified threshold y. Similarly, by manipulating $G(\boldsymbol{x}|y)$ the AL-GP approach can also be used to estimate the CCDF. However, if one is interested in the CDF/CCDF corresponding to a relatively wide range of y, a naive iteration of the conventional AL-GP method for a sequence of y values (i.e., a discrete mesh) would lead to a trade-off between accuracy and computational efficiency. Within this research gap, this study introduces a novel "mesh free" global AL-GP strategy to estimate the CDF/CCDF. The proposed AL-GP

*Observe that we implicitly assume non-symmetric distributions or distributions for which are *not* known a-priory whether they are symmetric or not

strategy possesses an attractive property such that the median-low probability regions of both CDF and CCDF can be simultaneously estimated.

It is important to note that the Gaussian process metamodel is equipped with an error/uncertainty measure that can be directly used in an active learning strategy, therefore the combination of Gaussian process metamodel and active learning is straightforward. The use of PCE in the active learning context is not as straightforward as the Gaussian process case, but the bootstrap technique [15] is recently proposed to explore this combination. The possible extension of the proposed AL-GP approach to PCE metamodels will be investigated in future studies, while the PCE estimates on CDF/CCDF using a global experiment design will be studied and compared with the proposed approach in Section 4.

Finally, it did not escape to our attention that for specific UQ applications specialized methods are developed to estimate the distribution function (PDF/CDF) and other statistical measures. In structural and mechanical engineering applications, approaches such as the Smoluchowski/FokkerPlanck equation, perturbation, moment closure, stochastic finite element, stochastic linearization, probability density evolution, etc (see [9][16][17][18] for a review), are developed to estimate the response statistics of mechanical systems. The synergy between these approaches and the proposed method for structural/mechanical problems is worthy of further investigations. Moreover, there are recent advances in combining advanced Monte Carlo simulation with active learning based metamodelling method [19][20][21], and we leave this possible extension to future studies.

The paper is organized as follows. Section 2 provides a basic framework of the global AL-GP method to estimate the distribution function. Section 3 provides theoretical and technical details for the proposed method. Section 4: first, it introduces a toy example with semi-analytical CDF/CCDF solution to gain a deep understanding of the proposed method; second, it shows an example with highly nonlinear analytical model function; and third, it presents a structural dynamics example with hysteretic force-deformation behavior. Section 5 discusses various practical issues, limitations and future research topics regarding the proposed method. Finally, Section 6 presents the conclusion.

2. The basic framework of the global AL-GP metamodelling method

2.1. Gaussian process model

To make this paper self-contained, the essential concept of Gaussian process model will be briefly reviewed. In Gaussian process metamodelling, the metamodel of $\mathcal{M}(\boldsymbol{x})$, denoted by $\hat{\mathcal{M}}(\boldsymbol{x})$, is considered as a realization of a Gaussian process [22], i.e.

$$\hat{\mathcal{M}}(\boldsymbol{x}) \sim GP\left(\mu(\boldsymbol{x}), \kappa(\boldsymbol{x}, \boldsymbol{x}')|\boldsymbol{\theta}\right) , \tag{4}$$

where

$$\mu(\boldsymbol{x}) = \mathrm{E}\left[\hat{\mathcal{M}}(\boldsymbol{x})\right] \tag{5}$$

is the mean/trend function, and

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{E}\left[\left(\hat{\mathcal{M}}(\boldsymbol{x}) - \mu(\boldsymbol{x})\right)\left(\hat{\mathcal{M}}(\boldsymbol{x}') - \mu(\boldsymbol{x}')\right)\right] \tag{6}$$

is a positive definitive kernel/covariance function, and $\boldsymbol{\theta}$ is a set of parameters describing both the mean and the kernel functions.

Given a training set $\{\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}\} = \{(\boldsymbol{x}_i, \mathcal{M}(\boldsymbol{x}_i)), i = 1, 2, ..., d\}$, the parameters $\boldsymbol{\theta}$ are typically estimated by generalized least-squares solution [22]. Provided a test set $\{\boldsymbol{\mathcal{X}}_*, \boldsymbol{\mathcal{Y}}_*\} = \left\{(\boldsymbol{x}_i, \hat{\mathcal{M}}(\boldsymbol{x}_i)), i = 1, 2, ..., s\right\}$ for which the predictions $\boldsymbol{\mathcal{Y}}_*$ are desired, by definition of the Gaussian process the following holds,

$$\begin{pmatrix} \boldsymbol{\mathcal{Y}} \\ \boldsymbol{\mathcal{Y}}_* \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}_* \end{pmatrix}, \begin{pmatrix} \boldsymbol{K} & \boldsymbol{K}_* \\ \boldsymbol{K}_*^T & \boldsymbol{K}_{**} \end{pmatrix}\right] , \tag{7}$$

where $\mathcal{N}$ denotes joint Gaussian distribution, $\boldsymbol{\mu} = \mu(\boldsymbol{\mathcal{X}})$, $\boldsymbol{\mu}_* = \mu(\boldsymbol{\mathcal{X}}_*)$, $\boldsymbol{K} = \kappa(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}})$, $\boldsymbol{K}_* = \kappa(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}}_*)$, and $\boldsymbol{K}_{**} = \kappa(\boldsymbol{\mathcal{X}}_*, \boldsymbol{\mathcal{X}}_*)$. Using Eq.(7), the conditional distribution (or posterior distribution) $f_{\boldsymbol{\mathcal{Y}}_*}(\boldsymbol{\mathcal{Y}}_*|\boldsymbol{\mathcal{Y}})$ can be obtained as

$$\begin{aligned} & f_{\boldsymbol{\mathcal{Y}}_*}(\boldsymbol{\mathcal{Y}}_*|\boldsymbol{\mathcal{Y}}) = \mathcal{N}\left(\boldsymbol{\mathcal{Y}}_*|\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}_*}, \boldsymbol{K}_{\boldsymbol{\mathcal{Y}}_*}\right) \\ & \boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}_*} = \boldsymbol{\mu}_* + \boldsymbol{K}_*^T \boldsymbol{K}_*^{-1}\left(\boldsymbol{\mathcal{Y}} - \boldsymbol{\mu}\right) \\ & \boldsymbol{K}_{\boldsymbol{\mathcal{Y}}_*} = \boldsymbol{K}_{**} - \boldsymbol{K}_*^T \boldsymbol{K}_*^{-1} \boldsymbol{K}_* \end{aligned} \tag{8}$$

There are several open source toolboxes for training Gaussian process models, e.g. DACE [23], ooDACE [24], GPML [25], etc. In this study, the ooDACE package is used to produce the results in Section 4.

2.2. Model selection for Gaussian process

In Gaussian process metamodelling practice, the trend function $\mu(\boldsymbol{x})$ as well as the covariance function $\kappa(\boldsymbol{x}, \boldsymbol{x}')$ are parameterized by prespecified basis functions. The trend function $\mu(\boldsymbol{x})$ is usually in the form of a constant, linear function or quadratic polynomials. The covariance $\kappa(\boldsymbol{x}, \boldsymbol{x}')$ typically takes the form of a *stationary correlation* function:

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\sum_{i=1}^{n} \alpha_i |x_i - x_i'|^p\right), \tag{9}$$

where n is the dimension of $\boldsymbol{x}$, $p = 2$ corresponds to the Gaussian correlation, and $p = 1$ corresponds to the exponential correlation. Besides the stationary correlation function, the linear, cubic, spline and Matérn [26] functions are also widely used in the literature. In general, for smooth model functions the smooth covariance models (e.g., Gaussian) may be preferable, and less smooth covariance models (e.g., Matérn) are more suitable for coarse model functions. In fact, theoretical results on the selection of covariance models are fairly rare. A most recent theoretical result [27] suggests that a less smooth covariance model can be more robust in prediction, while a smoother correlation function can achieve a higher rate of convergence. In this paper, since a black-box model function assumption is made (i.e., no prior knowledge on the smoothness), for all numerical examples the covariance functions are uniquely set to Gaussian*.

To select the trend function, we employ the "blind Kriging" approach [28] such that $\mu(\boldsymbol{x})$ takes the form

$$\mu(\boldsymbol{x}) = \sum_{i=1}^{m} \mu_i v_i(\boldsymbol{x}), \tag{10}$$

where $v_i(\boldsymbol{x})$ are a set of unknown basis functions, and μ_i are the corresponding weights. In blind Kriging $v_i(\boldsymbol{x})$ and μ_i are identified using the Bayesian variable selection technique. Note that the blind Kriging approach is implemented in the ooDACE package.

2.3. Procedures of the global AL-GP metamodelling method

The basic procedures of the AL-GP approach for distribution function estimation are listed as follows.

Algorithm 1 Procedures of global AL-GP metamodelling to estimate the CDF $F_Y(y)$

Step 1: Initialization

- Generate a large set of training candidates $\boldsymbol{\mathcal{X}}_c = \{\boldsymbol{x}_i, i = 1, 2, ..., N\}$.
- Generate an initial training set $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}_i, i = 1, 2, ..., d\}$, $d << N$.
- Evaluate the true model function for $\boldsymbol{\mathcal{X}}$ to obtain $\boldsymbol{\mathcal{Y}} = \{\mathcal{M}(\boldsymbol{\mathcal{X}})\}$.

*Note that perhaps by trying different covariance models, one may achieve better results than those reported in Section 4. However, the main contribution of this paper is to propose a global AL-GP strategy for CDF/CCDF estimation, and the "optimal" setting of a Gaussian process model is not the critical concern in this context.

Step 2: Train the Gaussian process model

- Using $\{\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}\}$, train a Gaussian process metamodel $\hat{\mathcal{M}}(\boldsymbol{x})$.

Step 3: Monte Carlo simulation on the metamodel

- Perform Monte Carlo simulation with $\hat{\mathcal{M}}(\boldsymbol{x})$ and $\boldsymbol{y}$ to obtain a three-fold estimate of $F_Y(y)$, denoted by $\hat{F}_Y^+(\boldsymbol{y})$, $\hat{F}_Y^0(\boldsymbol{y})$, $\hat{F}_Y^-(\boldsymbol{y})$, and $\forall y_i \in \boldsymbol{y}$, $\hat{F}_Y^+(y_i) \geq \hat{F}_Y^0(y_i) \geq \hat{F}_Y^-(y_i)$.

Step 4: Stopping criterion check

- Evaluate an error measure using $\hat{F}_Y^+(\boldsymbol{y})$, $\hat{F}_Y^0(\boldsymbol{y})$, and $\hat{F}_Y^-(\boldsymbol{y})$.
- If a specified stopping criterion is met, terminate the algorithm; else, proceed to **Step 5**.

Step 5: Update the training data set

- Search in $\boldsymbol{\mathcal{X}}_c$ for the optimal sample that maximizes a specified learning function.
- Add the optimal sample to the training set $\boldsymbol{\mathcal{X}}$.
- Evaluate the true model function for the optimal sample and update $\boldsymbol{\mathcal{Y}}$.
- Generate a new set of training candidates $\boldsymbol{\mathcal{X}}_c$.
- Return to **Step 2**.

To obtain the CCDF, we simply use $\bar{F}_Y(y) = 1 - F_Y(y)$. The error measure and learning function will be specially designed such that $1 - F_Y(y)$ will be accurate in the low probability region of $\bar{F}_Y(y)$.

In this study we develop on Algorithm 1 by providing a novel AL-GP strategy such that: a) the CDF as well as the CCDF estimation for a relatively wide range of y is accurate, and b) the efficiency of the method is not sensitive to the CDF/CCDF discretization, i.e. it is mesh free. To achieve these goals, in the following section the crucial ingredients of Algorithm 1 will be developed. Note that active learning based metamodellings all share a similar set of procedures as described in Algorithm 1. The novelty of this study lies in the global learning strategy to obtain the distribution function (CDF as well as CCDF). In particular, the strategy is based on a novel error measure combined together with a two step learning function, which allows a mesh free estimate of the CDF/CCDF.

3. A detailed development of the global AL-GP metamodelling method

3.1. Initialization of training samples (Step 1 of Algorithm 1)

The training candidates $\boldsymbol{\mathcal{X}}_c$ can be generated via Monte Carlo simulation using the PDF $f_{\boldsymbol{X}}(\boldsymbol{x})$, and the sample size N is typically of the order of 10^6 (or larger). The initial training set $\boldsymbol{\mathcal{X}}$ can be generated by quasi Monte Carlo or low-discrepancy sequence techniques such as the Latin hypercube sampling and the Sobol sequence, and the sample size d is typically of the order of 10^1.

3.2. Monte Carlo simulation for the metamodel (Step 3 of Algorithm 1)

In Gaussian process metamodelling, the $\mathcal{M}(\boldsymbol{x})$ is considered as a realization of a Gaussian process at location $\boldsymbol{x}$, which is completely defined by the mean $\mu_{\hat{\mathcal{M}}}(\boldsymbol{x})$ and variance $\sigma^2_{\hat{\mathcal{M}}}(\boldsymbol{x})$ (this mean and variance correspond to the posterior distribution described in Eq.(8)). It follows that a Gaussian process model describes a family of models expressed by

$$\hat{\mathcal{M}}(\boldsymbol{x}|k) = \mu_{\hat{\mathcal{M}}}(\boldsymbol{x}) + k\sigma_{\hat{\mathcal{M}}}(\boldsymbol{x})\,. \tag{11}$$

For $k = -\bar{k}, 0, \bar{k}$, one obtains three metamodels, i.e.

$$\begin{aligned} \hat{\mathcal{M}}^{+}(\boldsymbol{x}) &= \mu_{\hat{\mathcal{M}}}(\boldsymbol{x}) - \bar{k}\sigma_{\hat{\mathcal{M}}}(\boldsymbol{x}) \\ \hat{\mathcal{M}}^{0}(\boldsymbol{x}) &= \mu_{\hat{\mathcal{M}}}(\boldsymbol{x}) \\ \hat{\mathcal{M}}^{-}(\boldsymbol{x}) &= \mu_{\hat{\mathcal{M}}}(\boldsymbol{x}) + \bar{k}\sigma_{\hat{\mathcal{M}}}(\boldsymbol{x}) \end{aligned} \tag{12}$$

where $\bar{k}$ is typically fixed around 2. Using Eq.(2) and Eq.(12), one obtains a three-fold CDF estimate of Y,

$$\hat{F}_Y^a(y) = \int_{\hat{\mathcal{M}}^a(\boldsymbol{x}) \leq y} f_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x}\,, \tag{13}$$

where $a = +, 0, -$. If a crude Monte Carlo simulation is used to solve Eq.(13), to make the estimates consistent, the same set of random samples should be used in estimating $\hat{F}_Y^+(y)$, $\hat{F}_Y^0(y)$, and $\hat{F}_Y^-(y)$*.

3.3. The error measure (Step 4 of Algorithm 1)

The error measure is the critical ingredient for meeting the goal of simultaneously estimating the CDF and CCDF. To satisfy this goal, the error measure needs to be defined such that the CDF and the CCDF errors are measured *symmetrically*. Formally, consider a functional $D(\boldsymbol{F})$, where $\boldsymbol{F}$ denotes a set of CDFs or CCDFs with their discrepancy being measured by $D(\cdot)$. In this study, $\boldsymbol{F}$ may include $\hat{F}_Y^+(y)$, $\hat{F}_Y^0(y)$, and $\hat{F}_Y^-(y)$. Ideally, $D(\boldsymbol{F})$ should satisfy the following symmetry,

$$D(\boldsymbol{F}) = D(\boldsymbol{1} - \boldsymbol{F})\,. \tag{14}$$

The symmetry indicates that the error measure should be invariant under the transformation $\boldsymbol{F} \leftarrow \boldsymbol{1} - \boldsymbol{F}$. In other words, the error measure $D(\cdot)$ cannot tell if it is the CDF or the CCDF being measured. Clearly, in Algorithm 1, the computations should stop if $\hat{F}_Y^+(y)$ is sufficiently close to $\hat{F}_Y^-(y)$. To measure the discrepancy between CDFs $\hat{F}_Y^+(y)$ and $\hat{F}_Y^-(y)$, a natural metric that satisfies Eq.(14) is the Wasserstein (Kantorovich-Rubinstein) distance [29], expressed by

$$W(\hat{F}_Y^+, \hat{F}_Y^-) = \int_{-\infty}^{+\infty} \left| \hat{F}_Y^+(y) - \hat{F}_Y^-(y) \right| dy\,. \tag{15}$$

Eq.(15) is not the definition of the Wasserstein metric, but for the unidimensional case one can show that Eq.(15) is equivalent to the Wasserstein metric [29]. The Wasserstein metric corresponds to the minimum "cost" of turning one distribution into another. In this context, the cost is defined as the amount of probability mass needed to be transported integrated over the "transportation distance." The absolute function in Eq.(15) seems redundant because by definition $\hat{F}_Y^+(y)$ must be larger or equal to $\hat{F}_Y^-(y)$. However, the absolute function cannot be deleted because otherwise the symmetric condition (Eq.(14)) would be violated. Specifically, the absolute function is reserved for the scenario such that one replaces $\hat{F}_Y^+(y)$ and $\hat{F}_Y^-(y)$ with $1 - \hat{F}_Y^+(y)$ and $1 - \hat{F}_Y^-(y)$, respectively, and in this scenario Eq.(15) should provide the identical answer.

The problem of using Eq.(15) is that the contribution from the distribution tail is negligible. To highlight the tail contribution, we introduce the following symmetric measure

$$W^*(\hat{F}_Y^+, \hat{F}_Y^0, \hat{F}_Y^-) = \int_{-\infty}^{+\infty} w^*(y) dy. \tag{16}$$

where

$$w^*(y) \equiv \frac{\left| \hat{F}_Y^+(y) - \hat{F}_Y^-(y) \right|}{\min\left[\hat{F}_Y^0(y), 1 - \hat{F}_Y^0(y) \right]}\,. \tag{17}$$

*Due to statistical errors of Monte Carlo solutions, if different set of random samples are used, the property $\hat{F}_Y^+(y) \geq \hat{F}_Y^0(y) \geq \hat{F}_Y^-(y)$ may not hold. This is what we meant by "consistent".

It is easy to verify that Eq.(16) satisfies Eq.(14). In practice, we replace the infinite integral bound to $[y_{\min}, y_{\max}]$, i.e. the range of practical interest, and thus the potential zero denominator issue in Eq.(16) can be avoided. With Eq.(16), the stopping criterion can be specified as

$$W^*(\hat{F}_Y^+, \hat{F}_Y^0, \hat{F}_Y^-) < \epsilon \,. \tag{18}$$

The threshold ϵ can be set to

$$\epsilon = \bar{\epsilon}(y_{\max} - y_{\min}) \,, \tag{19}$$

where $\bar{\epsilon}$ is a specified tolerance. The tolerance $\bar{\epsilon}$ indicates that the integrand of Eq.(16), $w^*(y)$, on average (with respect to y) should be smaller than $\bar{\epsilon}$.

3.4. The learning function (Step 5 of Algorithm 1)

The learning function predicts the value of information gained by adding a new sample to the training set. In the AL-GP approach for reliability analysis, the learning function is defined as the probability of misclassification (failure/safe domain) [14][30], i.e.

$$\boldsymbol{L}(\boldsymbol{x}|y) = \Phi\left[-\frac{|y - \hat{\mathcal{M}}^0(\boldsymbol{x})|}{\sigma_{\hat{\mathcal{M}}}(\boldsymbol{x})}\right] \,, \tag{20}$$

where $\Phi[\cdot]$ denotes the CDF function of the univariate standard Gaussian distribution. The learning function Eq.(20) implies that one should select the optimal training sample, $\boldsymbol{x}^*$, such that: a) $\boldsymbol{x}^*$ is close to the limit-state surface $y - \hat{\mathcal{M}}^0(\boldsymbol{x}) = 0$, and b) the metamodelling uncertainty, $\sigma_{\hat{\mathcal{M}}}$, at $\boldsymbol{x}^*$ is large.

The learning function Eq.(20) cannot be used in the global AL-GP method because for CDF/CCDF estimation y is not fixed, in other words, for any candidate $\boldsymbol{x}$ there will be a corresponding y to exactly have $y - \hat{\mathcal{M}}^0(\boldsymbol{x}) = 0$. A simple remedy to this issue is to use the maximum of variance (MoV) learning criterion, i.e.

$$\boldsymbol{L}_{MoV}(\boldsymbol{x}) = \sigma_{\hat{\mathcal{M}}}(\boldsymbol{x}) \,. \tag{21}$$

The MoV learning function may not be most effective because it only uses a small portion of the available information and it is not directly related to the error measure Eq.(16). In the following, we will develop an alternative learning function such that: a) it makes full use of the available information, and b) it is directly related to the error measure. We first apply a kernel to Eq.(16) to obtain a localized error measure for the distribution function, i.e.

$$W_L^*(y') = \int_{-\infty}^{+\infty} w^*(y)\psi(y|y', \boldsymbol{\nu})\, dy = \int_{-\infty}^{+\infty} \frac{\left|\hat{F}_Y^+(y) - \hat{F}_Y^-(y)\right|}{\min\left[\hat{F}_Y^0(y), 1 - \hat{F}_Y^0(y)\right]} \psi(y|y', \boldsymbol{\nu})\, dy \,, \tag{22}$$

where $\psi(y|y', \boldsymbol{\nu})$, $\psi(y|y', \boldsymbol{\nu}) \geq 0$, $\int_{-\infty}^{+\infty} \psi(y|y', \boldsymbol{\nu})\, dy = 1$, denotes the kernel function centered at y' and parametrized by $\boldsymbol{\nu}$. Eq.(22) measures the localized contribution to the CDF/CCDF error in the neighbourhood of y'. The learning function is used to predict the specific "location" to assign the training sample, therefore it must contain some "localized" information on the error measure. At the same time, the learning function should be provided with some "global" knowledge, since the ultimate goal of learning is to reduce a global error measure on the distribution function. Clearly, Eq.(22) bridges the gap between global and local error measures on the distribution function, and therefore being an ideal medium for constructing a learning function. Given that, we introduce a two-step learning criterion. Specifically, first, we select the candidate threshold, y^*, which is located at the region with maximum localized distribution function error; second, we select the training sample, $\boldsymbol{x}^*$, where the model uncertainty is large. Formally, the two-step learning function is defined as

$$\boldsymbol{L}(\boldsymbol{x}) = \Phi\left[-\frac{|y^* - \hat{\mathcal{M}}^0(\boldsymbol{x})|}{\sigma_{\hat{\mathcal{M}}}(\boldsymbol{x})}\right] \,, \tag{23}$$

where

$$y^* = \arg\max_{y'} \left\{ W_L^*(y') | y' \in [y_{\min}, y_{\max}] \right\} . \tag{24}$$

Note that the constraint in Eq.(24) is applied because in practice we are only interested in the CDF/CCDF within the range $[y_{\min}, y_{\max}]$. With the learning function, the optimal training sample is selected as

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x} \in \boldsymbol{\mathcal{X}}_c} \left\{ \boldsymbol{L}(\boldsymbol{x}) \middle| \hat{\mathcal{M}}^0(\boldsymbol{x}) \in \left[y_{\min} - \bar{k}\sigma_{\hat{\mathcal{M}}}(\boldsymbol{x}), y_{\max} + \bar{k}\sigma_{\hat{\mathcal{M}}}(\boldsymbol{x}) \right] \right\} . \tag{25}$$

With this strategy, for a given training candidate set $\mathcal{X}_c$ $\boldsymbol{x}^*$ is the point associated with the maximum localized distribution function error and—at the same time—with the largest model uncertainty. Similar to Eq.(24), the constraint in Eq.(25) is applied because $[y_{\min}, y_{\max}]$ is of interest, and the term $\bar{k}\sigma_{\hat{\mathcal{M}}}(\boldsymbol{x})$ is used to consider the model uncertainty of $\hat{\mathcal{M}}(\boldsymbol{x})$.

3.4.1. Kernel selection: Dirac kernel

A special case of Eq.(22) is the use of Dirac kernel $\delta(y - y')$, and Eq.(22) reduces to

$$W_L^*(y') = \int_{-\infty}^{+\infty} w^*(y)\delta(y - y')\, dy = w^*(y) . \tag{26}$$

The use of Dirac Kernel corresponds to a greedy approach, and for this case the optimization described by Eq.(24) is trivial.

3.4.2. Kernel selection: Gaussian kernel

If a Gaussian kernel is used, Eq.(22) can be written as

$$W_L^*(y') = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} w^*(y) e^{-\frac{(y-y')^2}{2\sigma^2}}\, dy . \tag{27}$$

The standard deviation σ controls the width of the kernel. The limiting case $\sigma \to 0$ corresponds to the Dirac kernel, and $\sigma \to +\infty$ corresponds to a uniform distribution. Setting $\sigma \to +\infty$ is clearly not desirable, because this implies an almost constant y^* solution for Eq.(24). Setting $\sigma \to 0$, i.e. the Dirac kernel, may also be undesirable, because it does not consider the possible contribution to the error measure W^* from the neighborhood of y^*. Figure 1 illustrates this concept. In terms of Figure 1, if σ is properly selected, the next training point would be put on the ideal location which properly accounts for the possible contribution from the vicinity of y^*, thus leading to an effective route to reduce the global error measure W^*.

With the insight of how σ influences the training, we note that σ is related to the "effective bandwidth" of applying a training sample at location y'. The information of the effective bandwidth can be extracted from Eq.(12). It is seen from Eq.(12) that applying a training sample at $y' = \hat{\mathcal{M}}^0(\boldsymbol{x})$ directly influences probability estimate of $y' \pm \bar{k}\sigma_{\hat{\mathcal{M}}}(\boldsymbol{x})$*; thus, one can simply set the effective bandwidth $b_e = \bar{k}\sigma_{\hat{\mathcal{M}}}(\boldsymbol{x})$. Since $\bar{k}$ is commonly fixed to 2, and assuming the Gaussian kernel is "effective" in the $\pm 2\sigma$ regions around the mean, we can set $\sigma = \sigma_{\hat{\mathcal{M}}}(\boldsymbol{x})$. However, $\sigma_{\hat{\mathcal{M}}}(\boldsymbol{x})$ is a function of $\boldsymbol{x}$, while, in the current context, it is ideal to let σ be solely determined by y'. In principle this is not possible because the effective bandwidth b_e must be a function of $\boldsymbol{x}$. However, observe that in practical implementation we have a finite training candidate set $\mathcal{X}_c$. Therefore, in $\mathcal{X}_c$ we can select the sample in the closest neighborhood of y' and assign the corresponding $\sigma_{\hat{\mathcal{M}}}$ as the σ for the Gaussian kernel.

Formally, in the Gaussian kernel approach we solve the following optimization.

$$y^* = \arg\max_{y'} \left\{ \frac{1}{Z} \int_{y_{\min}}^{y_{\max}} w^*(y) e^{-\frac{(y-y')^2}{2\sigma^2(y')}}\, dy \middle| y' \in [y_{\min}, y_{\max}] \right\} , \tag{28}$$

*Note that $\mathbb{P}\left(\hat{\mathcal{M}}^{\pm} \leq y'\right) = \mathbb{P}\left(\hat{\mathcal{M}}^0 \leq y' \pm \bar{k}\sigma_{\hat{\mathcal{M}}}(\boldsymbol{x})\right)$.

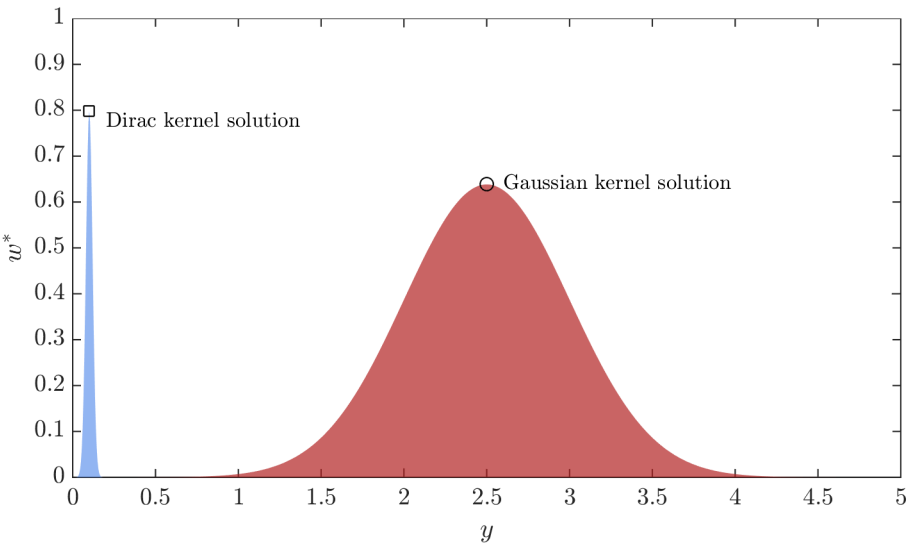


Figure 1: **Illustration of the effects of kernel function**. *The figure shows the integrand of Eq.(16), $w^*(y)$, as a function of y. If a Dirac kernel is used, the next training candidate will be put on the left peak (the maximum). (Ignoring, for the moment, the influence of model uncertainty $\sigma_{\hat{\mathcal{M}}}$.) This threshold candidate may not be as effective as putting on the middle peak to reduce the global error measure W^*, since putting a training candidate on the middle peak will have the benefit to suppress the widely distributed error in the neighborhood regions.*

where the normalizing constant Z is expressed as

$$Z = \sqrt{2\pi}\sigma(y')\left(\Phi\left(\frac{y_{\max} - y'}{\sigma(y')}\right) - \Phi\left(\frac{y_{\min} - y'}{\sigma(y')}\right)\right) , \tag{29}$$

and $\sigma(y')$ is solved from

$$\sigma(y') = \left\{\sigma_{\hat{\mathcal{M}}}(\boldsymbol{x}')\middle|\boldsymbol{x}' = \arg\min_{\boldsymbol{x}\in\mathcal{X}_c}\left|y' - \hat{\mathcal{M}}^0(\boldsymbol{x})\right|\right\} \tag{30}$$

Note that Eq.(30) is adopted for simplicity, yet one could develop other approaches via interpolation or the k-Nearest Neighbor. Also note that the introduction of a correction in the normalizing constant of Gaussian distribution is because the Gaussian kernel is truncated by the range $[y_{\min}, y_{\max}]$.

In the global AL-GP approach with Gaussian kernel there is no parameter to tune. Although analytical solution of Eq.(28) cannot be developed, the one dimensional constrained optimization problem can be effectively solved by various global optimization algorithms (e.g., simulated annealing [31] and differential evolution [32]). We finally remark that the error measure as well as the learning procedure are constructed based on a global description of the CDF/CCDF, thus naturally the proposed AL-GP method is mesh free (no discretization of CDF or CCDF are introduced a-priori). Moreover, the proposed error measure is equipped with the important CDF-CCDF symmetry; as a consequence, the proposed AL-GP strategy is designed to accurately estimate the CDF and CCDF at the same time.

4. Numerical examples

4.1. A toy example

Consider an analytical two dimensional model function $\mathcal{M}(\boldsymbol{X})$ expressed by

$$Y = \mathcal{M}(\boldsymbol{X}) = \min\left[X_1 - X_2, X_1 + X_2\right] , \tag{31}$$

where X_1 and X_2 are independent standard Gaussian random variables. The CDF of Y can be expressed by

$$F_Y(y) = \Phi\left(\frac{y}{\sqrt{2}}\right)\left(2 - \Phi\left(\frac{y}{\sqrt{2}}\right)\right) . \tag{32}$$

Although Eq.(32) is not analytical, most scientific computing codes provide highly accurate built-in univariate standard Gaussian CDF $\Phi(\cdot)$, and we denote the corresponding solution as "exact." Then, the proposed global AL-GP approaches with the Dirac and Gaussian kernels in Eq.(22), and the approach using the maximum of variance (MoV) learning criterion* are used to estimate the distribution function. The conventional AL-GP approach† and PCE with a global design of experiment ‡ are also considered for a comparison. The CDF/CCDF range of interest, $[y_{\min}, y_{\max}]$, is set to $[-5, 3]$. The range $[y_{\min}, y_{\max}]$ is uniformly discretized into 100 intervals. We use $\epsilon = \bar{\epsilon}(y_{\max} - y_{\min})$, $\bar{\epsilon} = 0.2$, for the stopping criterion. The number of samples in the training candidate set $\mathcal{X}_c$ is 10^6. In the following examples except the range $[y_{\min}, y_{\max}]$, other settings of the AL-GP method will be the same. Table 1 reports the performance of the global AL-GP approaches, the conventional AL-GP approach and the global PCE approach in estimating the distribution function averaged over 50 independent runs.

Table 1: **Distribution function estimations averaged over** 50 **independent runs**. *In the table,* $\mathrm{E}[\epsilon_e]$ *denotes expectation of the error measure* ϵ_e*;* $\sigma(\epsilon_e)$ *denotes standard deviation of* ϵ_e*;* $\mathrm{E}[N_{\mathcal{M}}]$ *denotes expectation of the number of (true) model function evaluations, and it is composed by the initial training set plus the adaptively added training set. The crude number of model function calls for the global PCE to achieve a similar accuracy of the active learning approaches is also reported in the "[]" brackets.*

Methods	$\mathrm{E}[\epsilon_e]$	$\sigma(\epsilon_e)$	$\mathrm{E}[N_{\mathcal{M}}]$
Gaussian	0.018	0.010	12+29.06
Dirac	0.021	0.022	12+28.74
MoV	0.028	0.043	12+27.34
Conventional	0.013	0.011	12+86.22
Global PCE	0.439	0.025	50 [5000]

In the table, the error ϵ_e is defined as

$$\epsilon_e = \frac{1}{y_{\max} - y_{\min}} \int_{y_{\min}}^{y_{\max}} \frac{\left|\hat{F}_Y^0(y) - F_Y(y)\right|}{\min\left[F_Y(y), 1 - F_Y(y)\right]} dy , \tag{33}$$

where $F_Y(y)$ is the "exact" CDF obtained from Eq.(32). This error measure also satisfies the CDF-CCDF symmetry condition (Eq.(14)), and it has a clear interpretation. For example, $\epsilon_e = 0.02$ indicates, within a specified range of interest, the average (with respect to y) relative CDF/CCDF error is 2 percent. From the table, it can be seen that the global AL-GP approaches are noticeably more efficient than the conventional AL-GP and global PCE approaches. The global AL-GP approaches with Gaussian and Dirac kernels are more effective than that with the MoV learning function. Also it is seen from the table that the Global PCE takes around **5000** model function calls to achieve a similar accuracy of the active learning approaches. Figure 2 illustrates a typical Gaussian process metamodel for the model function Eq.(31). Figure 3 illustrates

*The approach using the MoV learning function is also viewed as a global AL-GP method, because it also uses the global error measure Eq.(16).

†In the conventional AL-GP approach, we start from estimating the probability for a fixed threshold with learning function (20) and stopping criterion $w^* < \bar{\epsilon}$, and then we iteratively reuse the current metamodel and perform AL-GP to estimate the probability for the next nearby threshold.

‡The Latin hypercube sampling is used for global experiment design of PCE, and for a comparison the number of training samples is set to be similar to that of the proposed AL-GP approach. The coefficients of PCE are obtained using least-square minimization. The degrees of polynomials are adaptively selected to minimize the leave-one-out cross validation error. The UQLab [33] is used in training the PCE model for this paper.

the iterative process for the CDF and CCDF estimation from the global AL-GP approach using the Gaussian kernel.

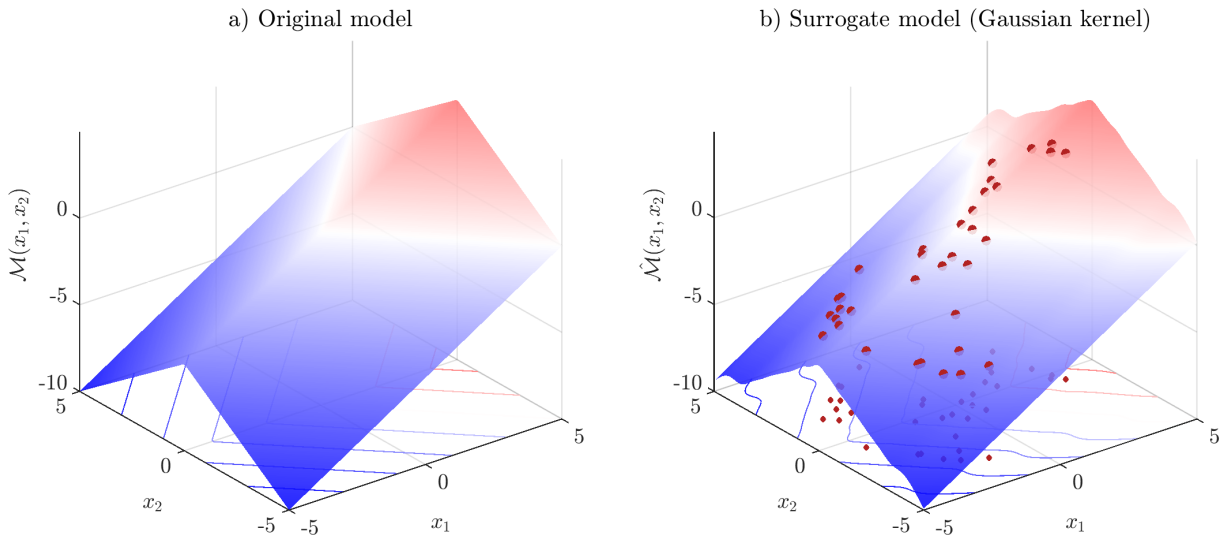


Figure 2: **A typical Gaussian process metamodel for Eq.**(31). a) Original model, b) metamodel based on the red samples (obtained via global AL-GP with Gaussian kernel)

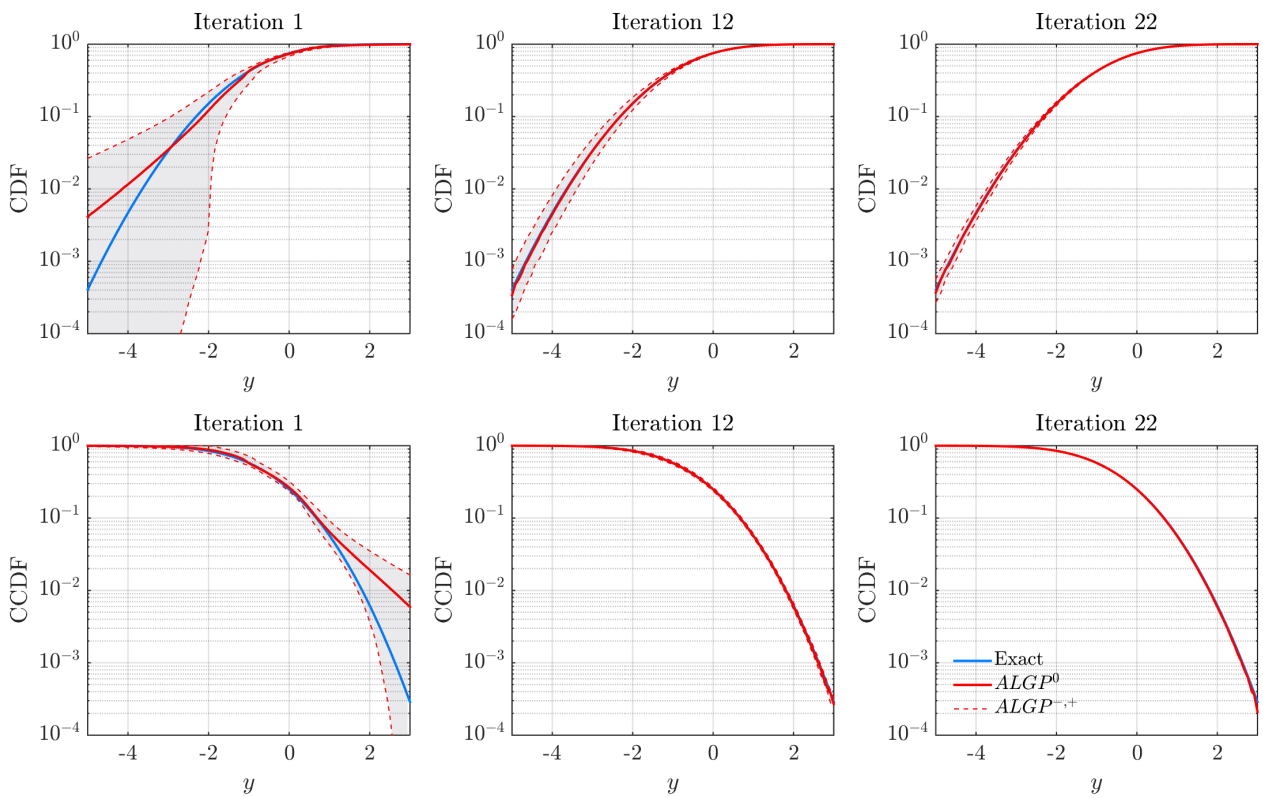


Figure 3: **Iterative process in distribution function estimation**.

Figure 4 shows a typical iterative process for the first four statistical moments of the estimated distribution from the global AL-GP approach with Gaussian kernel. The "exact" moments are obtained from the following equation,

$$\mathrm{E}(Y^j) = j \int_0^{+\infty} y^{j-1}(1 - F_Y(y) + (-1)^j F_Y(-y))\, dy\,. \tag{34}$$

Note that Eq.(34) is in terms of the CDF instead of the PDF, thus being especially useful in this study. Table 2 shows the estimated mean, standard deviation, skewness and kurtosis, averaged over 50 independent runs. The absolute coefficient of variation (c.o.v.) of these estimates are also shown in the table. Note that the number of model function evaluations is already reported in Table 1. Also note that the aforementioned statistical measures can be computed from the first four moments, thus their "exact" solutions are obtained via Eq.(34). It can be seen from the table that the global AL-GP method is able to accurately estimate various statistical measures on the distribution function. More importantly, this accuracy is achieved at the cost of only a few (less than 40) model function evaluations.

Table 2: **Global measures of the estimated distribution function averaged over** 50 **independent runs**.

Methods	mean/$\lvert c.o.v. \rvert$	standard deviation/$\lvert c.o.v. \rvert$	skewness/$\lvert c.o.v. \rvert$	kurtosis/$\lvert c.o.v. \rvert$
Gaussian	-0.7946/0.0115	1.1668/0.0048	-0.1391/0.0728	3.0619/0.0057
Dirac	-0.7960/0.0113	1.1691/0.0079	-0.1386/0.1335	3.0576/0.0084
MoV	-0.7996/0.0092	1.1700/0.0073	-0.1312/0.2358	3.0538/0.0040
Conventional	-0.7989/0.0074	1.1664/0.0037	-0.1370/0.0743	3.0663/0.0088
Global PCE	-0.8204/0.0044	1.1798/0.0025	-0.3737/0.0569	4.0867/0.0378
Exact	-0.7979	1.1676	-0.1369	3.0617

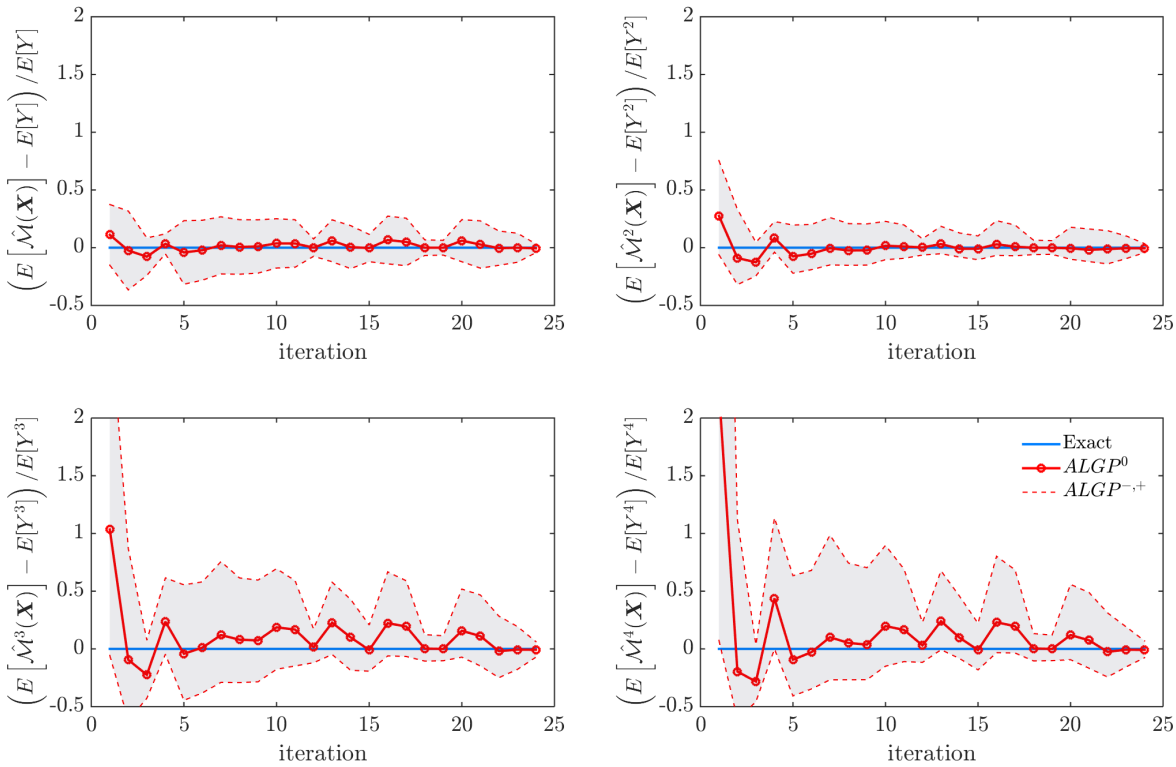


Figure 4: **Iterative process in moment estimation**.

Figure 5 shows locations of the training samples and contours of the metamodel. Figure 6 shows histograms on the y locations of the training samples, i.e. for each training sample $\boldsymbol{x}$, location y is obtained by $y = \mathcal{M}(\boldsymbol{x})$. Note that Figure 5 and Figure 6 illustrate all the adaptively added training samples obtained from 50 independent runs, thus they depict the spatial distribution of the training samples.

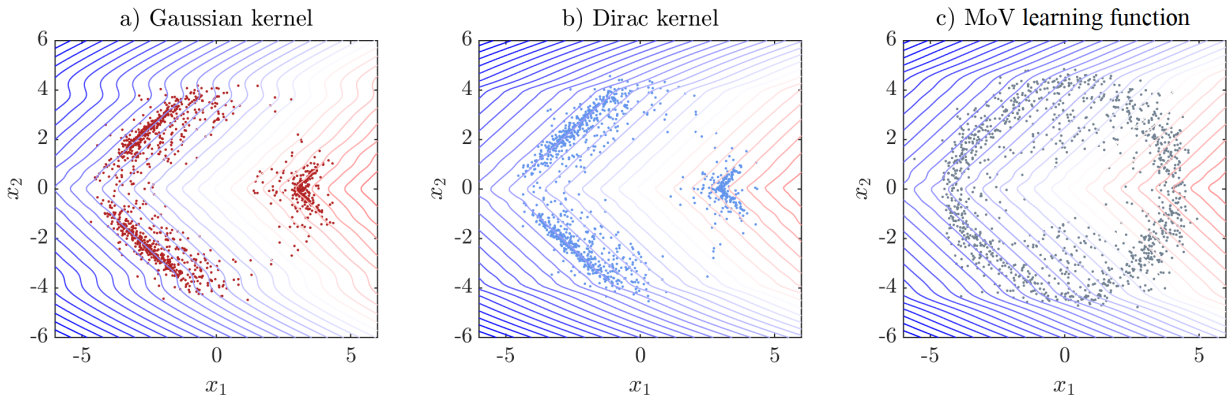


Figure 5: **Locations of the training samples**.

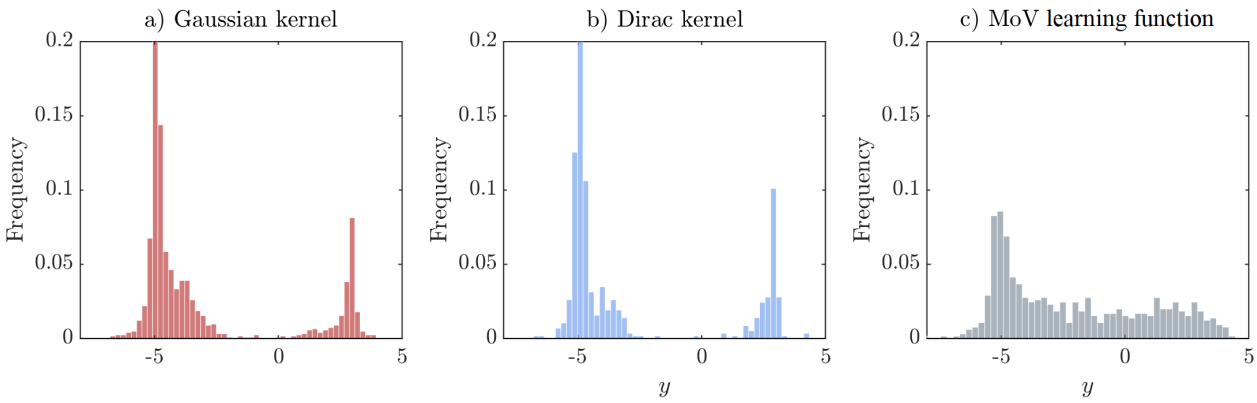


Figure 6: **Histograms on the y locations of the training samples**.

It can be seen from Figure 6 that for the global AL-GP approaches with Dirac or Gaussian kernels, most training samples come from the tails; while for the global AL-GP approach using MoV learning function, training samples are more dispersively distributed. Figure 6 also suggests that the y^* solution (Eq.(24)) from the Gaussian kernel, in most cases, is close to the solution from the Dirac kernel. Therefore, the situation discussed in Figure 1 may rarely happen.

4.2. Ishigami function

Consider the Ishigami function [34] expressed by

$$Y = \mathcal{M}(\boldsymbol{X}) = \sin(X_1) + a\sin^2(X_2) + bX_3^4 \sin(X_1)\,, \tag{35}$$

where $a = 7$, $b = 0.1$, and X_1, X_2 and X_3 are independent uniform random variables within $[-\pi, \pi]$. The CDF/CCDF range of interest, $[y_{\min}, y_{\max}]$, is set to $[-10, 15]$. The performance of the global AL-GP approaches are reported in Table 3, and the estimations for the distribution mean, standard deviation, skewness and kurtosis are reported in Table 4. The distribution function estimated via a crude Monte Carlo simulation with 10^7 samples is used as the reference solution. It is seen from the tables that except the skewness (which is zero theoretically), the global AL-GP approaches provide accurate estimation on the distribution function and various statistical measures. Figure 7 and Figure 8 show typical iterative processes for the distribution function and moment estimations from the global AL-GP approach with Gaussian kernel. Observe that for even moments the three fold monotonic property of the CDF is not necessarily preserved (this can be clearly seen from Eq.(34)).

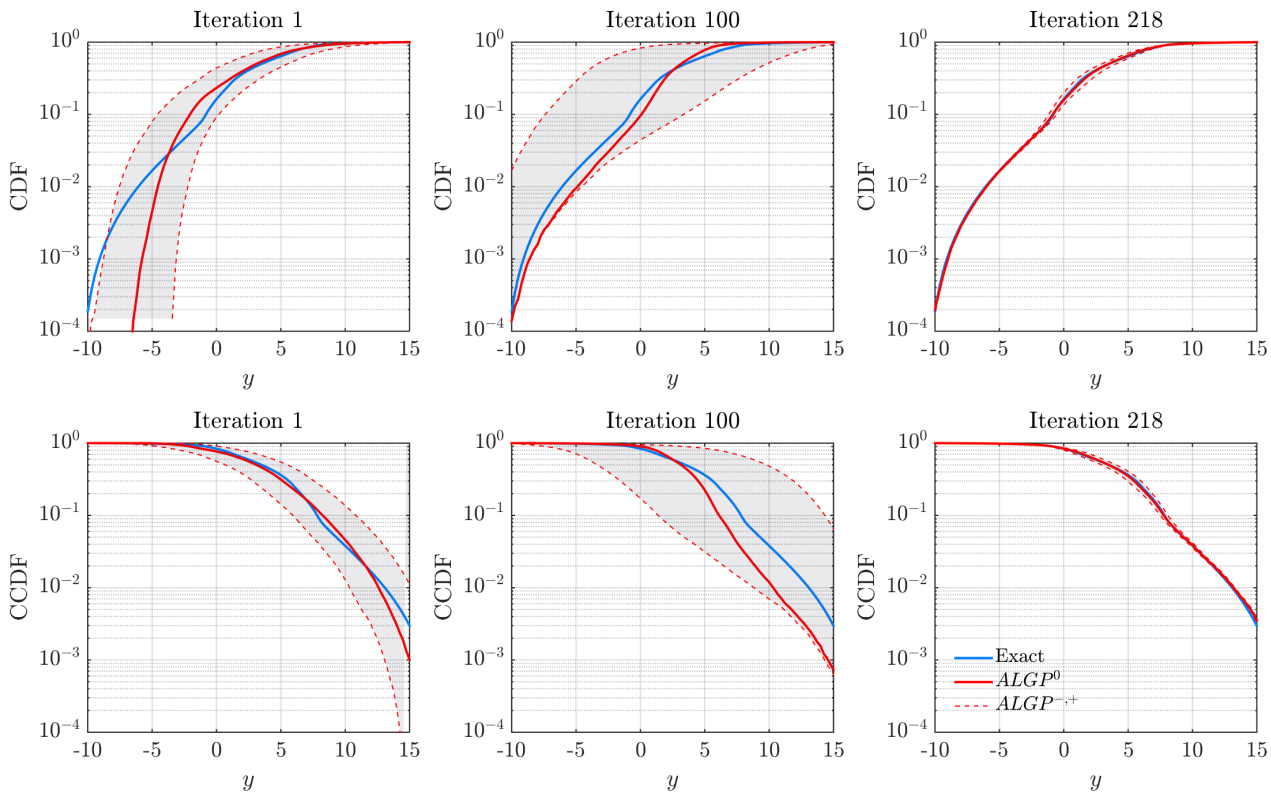


Figure 7: **Iterative process in distribution function estimation**.

Table 3: **Distribution function estimations averaged over 50 independent runs**.

Methods	$\mathrm{E}[\epsilon_e]$	$\sigma(\epsilon_e)$	$\mathrm{E}[N_{\mathcal{M}}]$
Gaussian	0.028	0.0074	12+230.20
Dirac	0.027	0.0074	12+225.60
MoV	0.060	0.0133	12+212.20
Conventional	0.019	0.0035	12+366.80
Global PCE	0.071	0.0069	250 [500]

Table 4: **Global measures of the estimated distribution function**.

Methods	mean/$\lvert c.o.v. \rvert$	standard deviation/$\lvert c.o.v. \rvert$	skewness/$\lvert c.o.v. \rvert$	kurtosis/$\lvert c.o.v. \rvert$
Gaussian	3.5023/0.0088	3.6844/0.0027	0.0113/1.6956	3.5948/0.0090
Dirac	3.5204/0.0101	3.7062/0.0043	-0.0036/5.1806	3.5603/0.0133
MoV	3.4933/0.0046	3.7789/0.0045	-0.0010/11.5284	3.5087/0.0125
Conventional	3.5002/0.0026	3.7236/0.0018	0.0047/2.3023	3.5066/0.0054
Global PCE	3.4384/0.0037	3.7534/0.0025	0.0406/0.2444	3.3849/0.0052
Monte Carlo	3.5018	3.7204	0 (exact)	3.5106

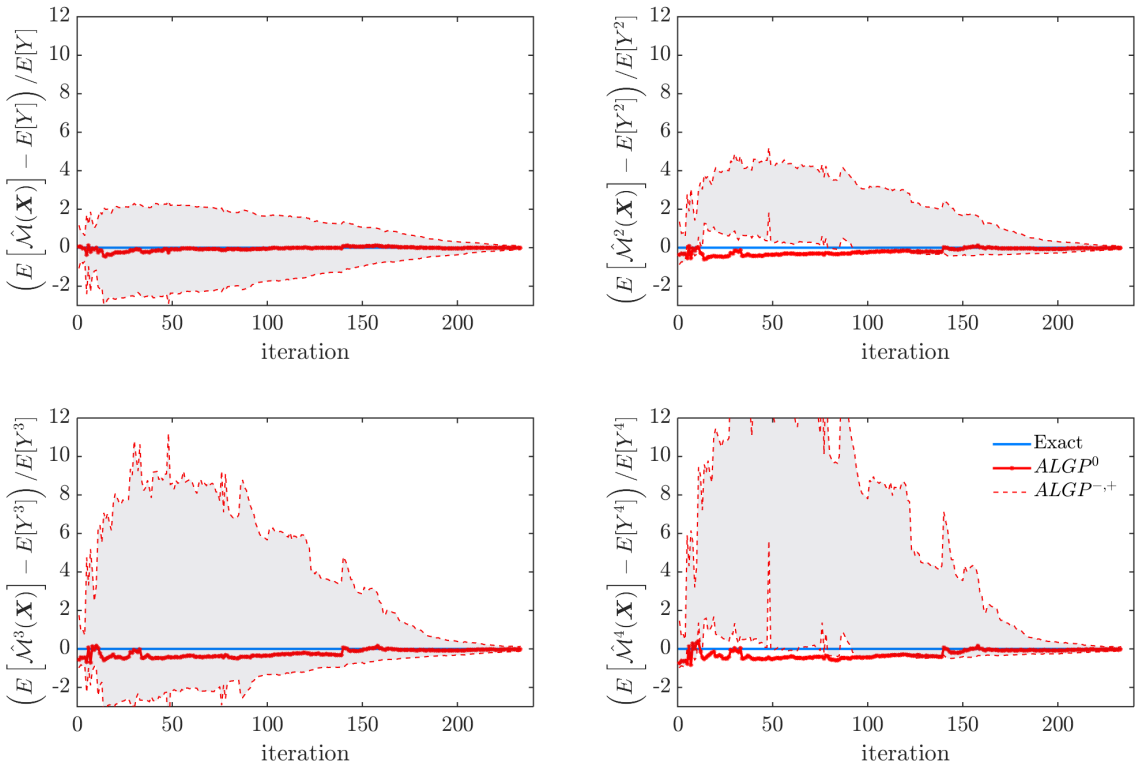


Figure 8: **Iterative process in moment estimation**.

Figure 9 shows projection of the original model and the Gaussian process model in various planes. Figure 10 shows projection of the training samples in the (x_1, x_2) plane. Figure 11 shows histograms of the y locations of the training samples. Similar to the previous example, the results obtained from multiple independent runs are used to produce Figure 10 and Figure 11, and only the adaptively added training samples are shown in the figures. Due to the oscillating behavior of the Ishigami function, it is seen from Figure 11 that the training samples for Gaussian and Dirac kernels concentrate around multiple modes.

4.3. Structural dynamics analysis of a shear-frame structure

Consider a dynamics analysis of the three stories shear-frame structure shown in Figure 12 (a similar structural model is studied in [35]). The interstory mechanical behavior is inelastic with a force-interstory-drift relationship described by the Bouc-Wen hysteretic model [36][37]:

$$\begin{aligned} &k(\alpha u_1(t) + (1-\alpha)u_2(t)) = f(t) \\ &\dot{u}_2(t) = -\gamma\,|\dot{u}_1(t)|\,|u_2(t)|^{\bar{n}-1}\,u_2(t) - \eta\,|u_2(t)|^{\bar{n}}\,\dot{u}_1(t) + A\dot{u}_1(t) \end{aligned} \tag{36}$$

where $u_1(t)$ represents the linear response component, and $u_2(t)$ represents the hysteretic response component. The parameters of the Bouc-Wen model are set as: $\alpha = 0.1$, $\bar{n} = 5$, $A = 1$, and $\gamma = \eta = 1/(2u_y^{\bar{n}})$, in which $u_y = 0.04$ [m] is the yielding displacement. The typical hysteretic force-deformation behavior of the Bouc-Wen model is illustrated in Figure 13. The initial inter-story stiffness, $\boldsymbol{k} = [k_1, k_2, k_3]$, and mass, $\boldsymbol{m} = [m_1, m_2, m_3]$, values are reported in Table 5. The Rayleigh damping with 5 percent damping ratio for the first and second mode is adopted.

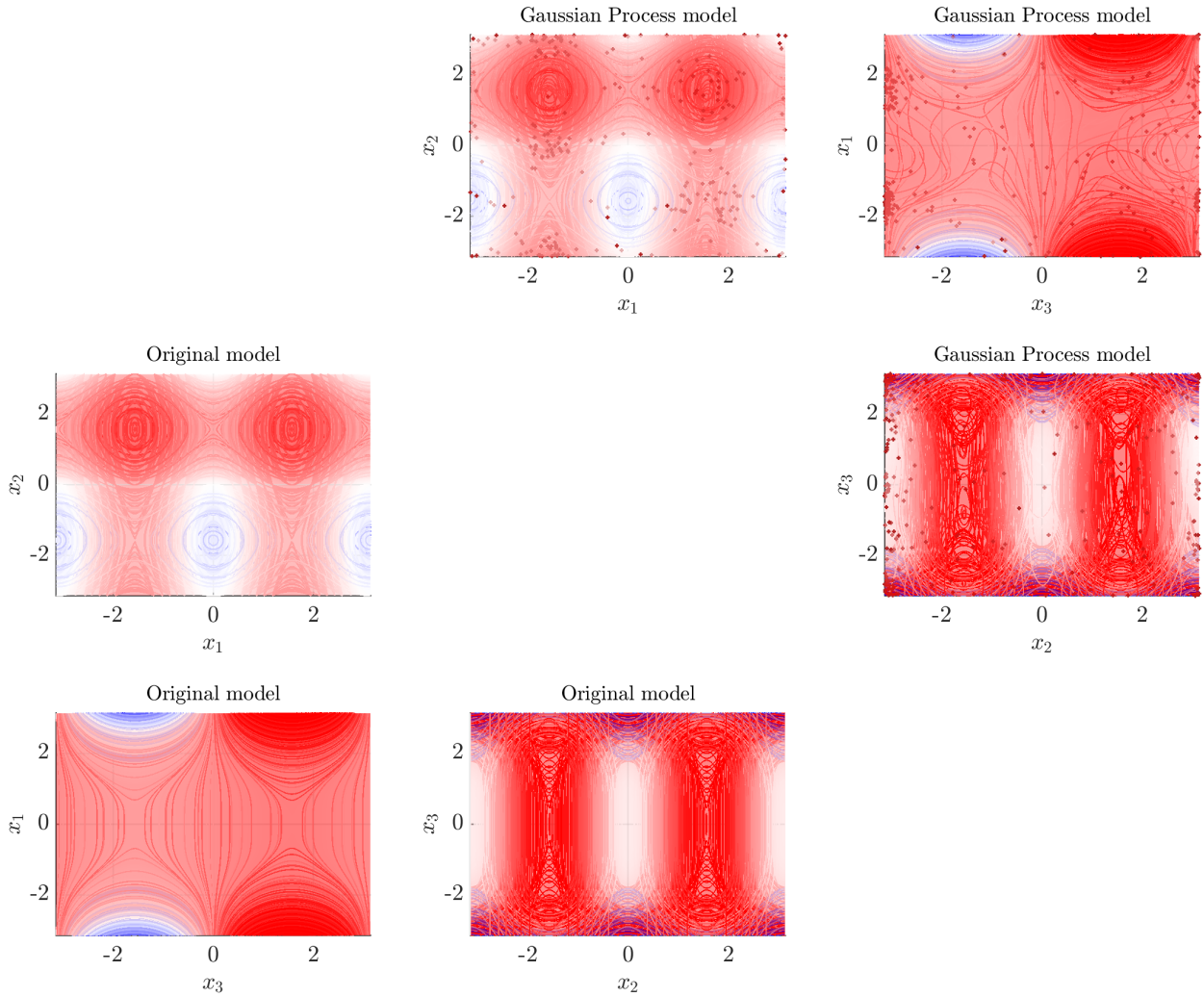


Figure 9: **Projection of the original model and Gaussian process model in various planes**. *The figure shows the projection of level sets* $\{\boldsymbol{x}|m - \mathcal{M}(\boldsymbol{x}) = 0\}$ *and* $\left\{\boldsymbol{x}|m - \hat{\mathcal{M}}(\boldsymbol{x}) = 0\right\}$ *in various planes. The training samples used to generate the metamodel is also shown in the figure.*

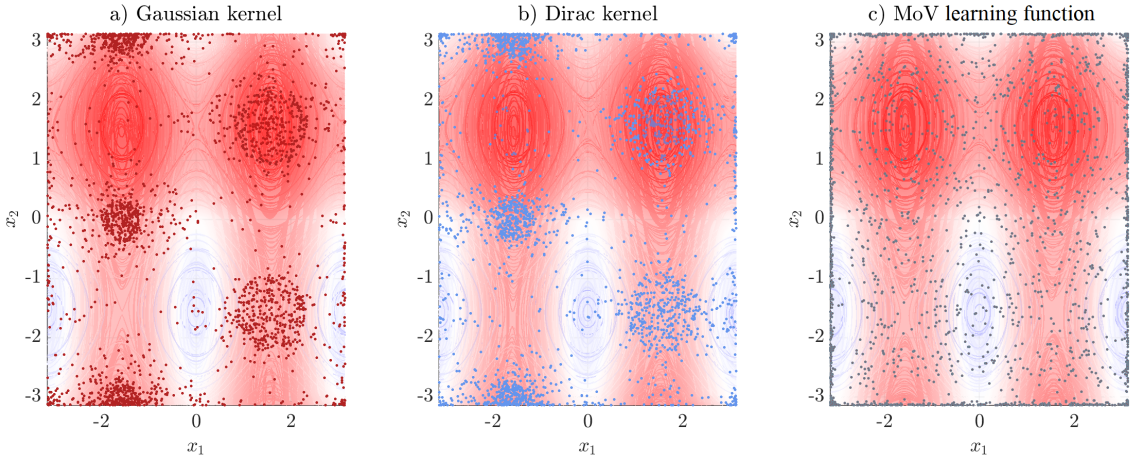


Figure 10: **Projection of the training samples in the (x_1, x_2) plane.**

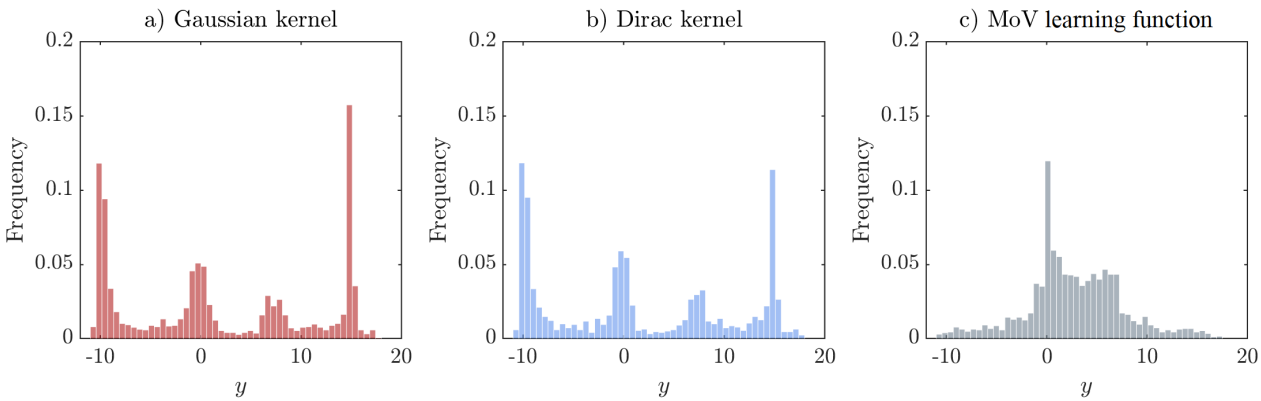


Figure 11: **Histograms on the y locations of the training samples.**

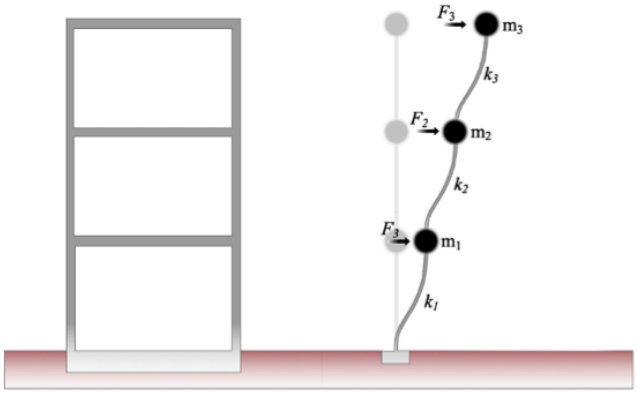


Figure 12: **Structural archetype**.

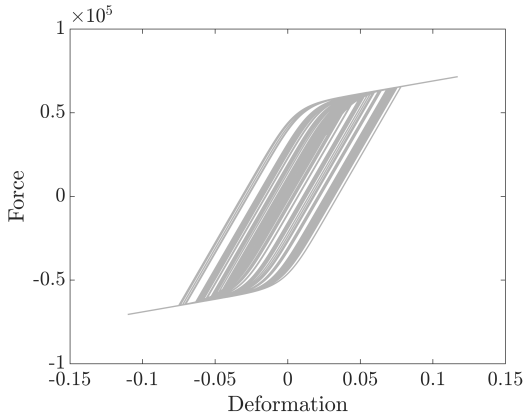


Figure 13: **Typical hysteretic loops of the Bouc-Wen model**.

Table 5: **Structural properties**.

	k_i[N/m]	m_i[kg]
Story 1	3.0×10^8	1×10^6
Story 2	2.8×10^8	1×10^6
Story 3	1.5×10^8	1×10^6

The horizontal forces for each story, f_i, $i = 1, 2, 3$, are considered as a combination of harmonic waves with random amplitudes:

$$f_i(\boldsymbol{X}, t) = \frac{1}{6} m_i \left(X_1 \sin(2\pi t) + X_2 \sin(4\pi t) + X_3 \cos(8\pi t) + X_4 \sin(16\pi t) \right) , \tag{37}$$

where m_i is the mass of the i-th story, and $\boldsymbol{X} = [X_1, X_2, X_3, X_4]$ are independent standard Gaussian random variables. The duration of the excitation is 10 seconds. The model function describing the maximum absolute interstory drift is defined as

$$y = \mathcal{M}(\boldsymbol{X}) = \max_{i=1,2,3} \max_{t \in [0,10]} |v_i(\boldsymbol{X}, t)| , \tag{38}$$

where v_1, v_2 and v_3 are the first, second and third interstory drift, respectively. The CCDF of the maximum absolute drift is of practical interest, since it is an indicator of structural reliability. Therefore, with a

trivial modification on the error measure (see Section 5), the global AL-GP method is used to estimate the CCDF. The CCDF range of interest, $[y_{\min}, y_{\max}]$, is set to $[0, 0.12]$. The performance of the global and conventional AL-GP approaches is reported in Table 6*, and the estimations for the distribution mean, standard deviation, skewness and kurtosis are reported in Table 7. The CCDF estimated via a crude Monte Carlo simulation with 10^6 samples is used as the reference solution.

Table 6: **CCDF estimations averaged over 50 independent runs**.

Methods	$\mathrm{E}[\epsilon_e]$	$\sigma(\epsilon_e)$	$\mathrm{E}[N_{\mathcal{M}}]$
Gaussian	0.039	0.0080	12+124.30
Dirac	0.038	0.0072	12+115.60
MoV	0.044	0.0067	12+265.70
Conventional	0.033	0.0069	12+288.40
Global PCE	0.465	0.0063	1000 [$\gg 10^4$]†

Table 7: **Global measures of the estimated distribution function**.

Methods	mean/$\lvert c.o.v. \rvert$	standard deviation/$\lvert c.o.v. \rvert$	skewness/$\lvert c.o.v. \rvert$	kurtosis/$\lvert c.o.v. \rvert$
Gaussian	0.0219/0.0122	0.0160/0.0149	1.2520/0.0403	5.5601/0.0423
Dirac	0.0221/0.0201	0.0159/0.0223	1.2773/0.0519	5.6365/0.0401
MoV	0.0218/0.0150	0.0164/0.0168	1.2002/0.0446	5.1714/0.0446
Conventional	0.0222/0.0055	0.0159/0.0069	1.2603/0.0334	5.5753/0.0394
Global PCE	0.0218/0.0027	0.0156/0.0043	1.1644/0.0997	5.7548/0.3577
Monte Carlo	0.0221	0.0160	1.2256	5.4799

Figure 14 and Figure 15 show typical iterative processes for the CCDF and moment estimations from the global AL-GP method with Gaussian kernel. Figure 16 shows projection of the original model and the Gaussian process model in various planes. Figure 17 shows projection of the training samples in the (x_3, x_4) plane. Figure 18 shows histograms of the y locations of the training samples.

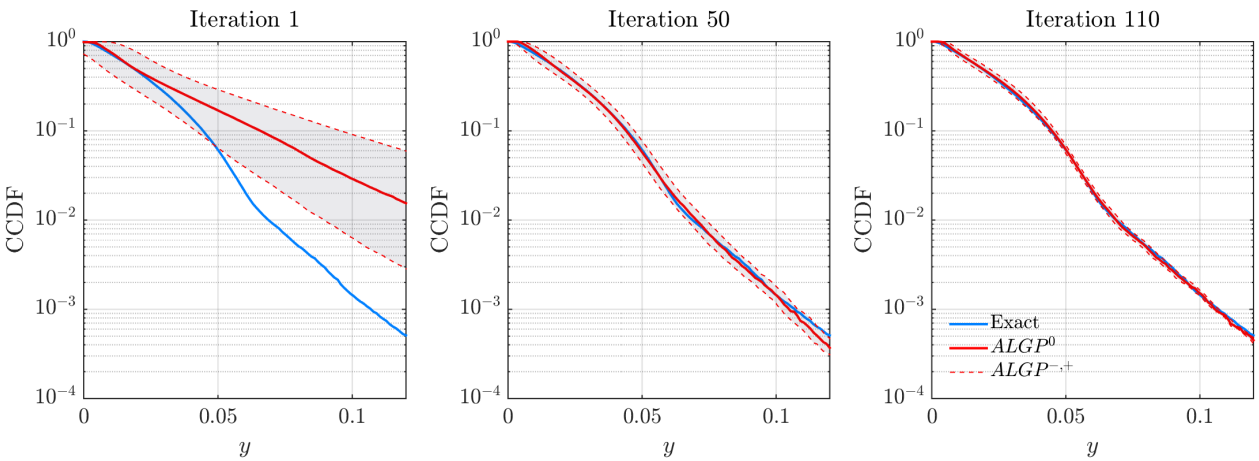


Figure 14: **Iterative process in CCDF estimation**.

*Note that in Table 6 the ϵ_e is slightly different from that in the previous examples, specifically, in Eq.(33) the denominator is replaced by $1 - F_Y(y)$.

†By using 10^4 samples, we obtain an error of 0.43.

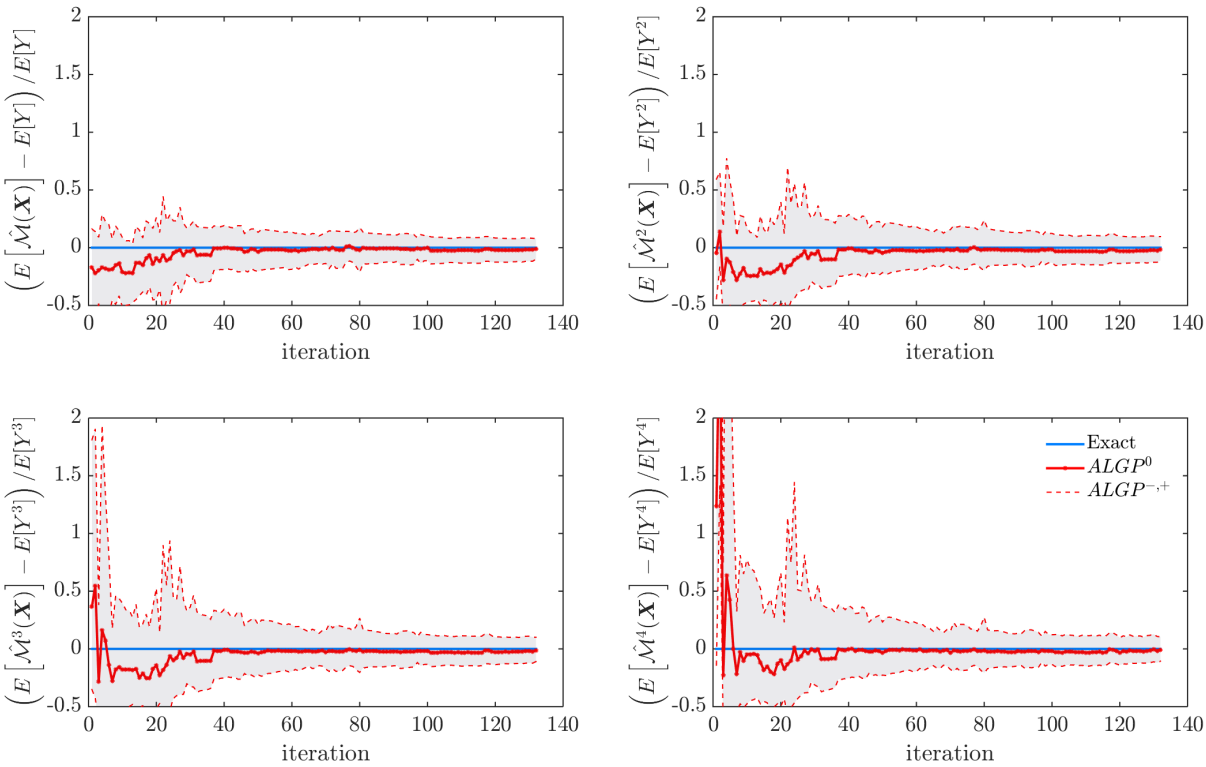


Figure 15: **Iterative process in moment estimation**.

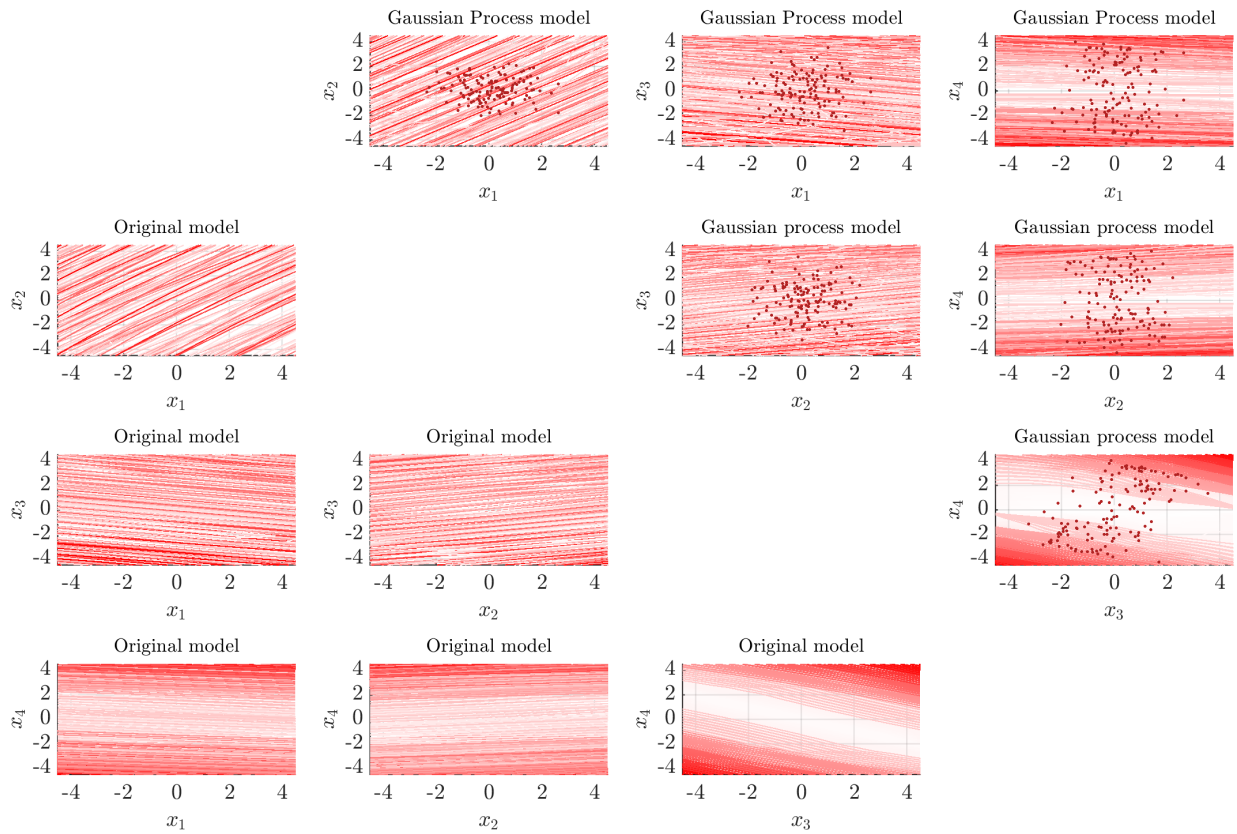


Figure 16: **Projection of the original model and the Gaussian process model in various planes**. *The figure shows the projection of level sets* $\{\boldsymbol{x}|m - \mathcal{M}(\boldsymbol{x}) = 0\}$ *and* $\left\{\boldsymbol{x}|m - \hat{\mathcal{M}}(\boldsymbol{x}) = 0\right\}$ *in various planes. The training samples used to generate the metamodel is also shown in the figure.*

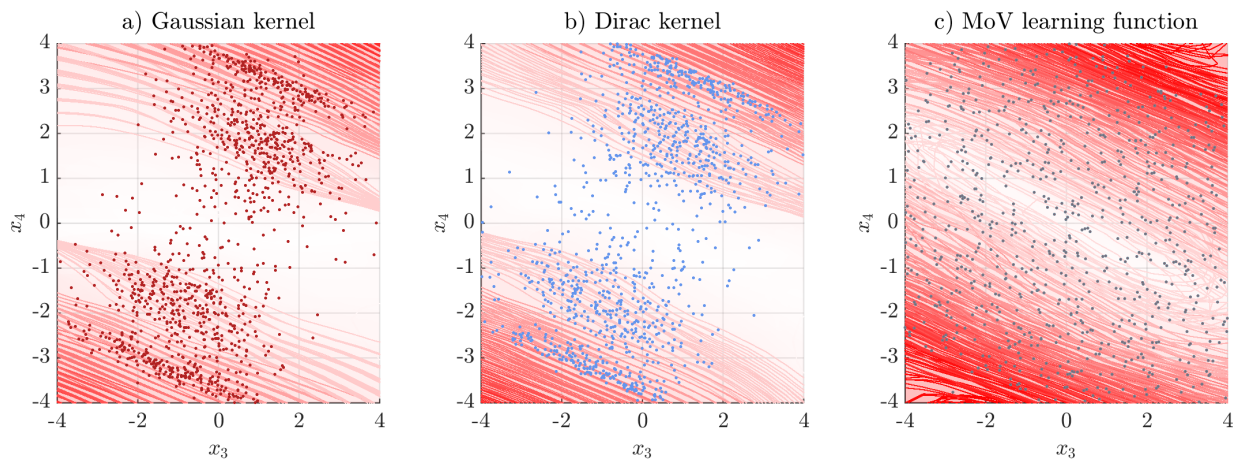


Figure 17: **Projection of the training samples in the** (x_3, x_4) **plane**.

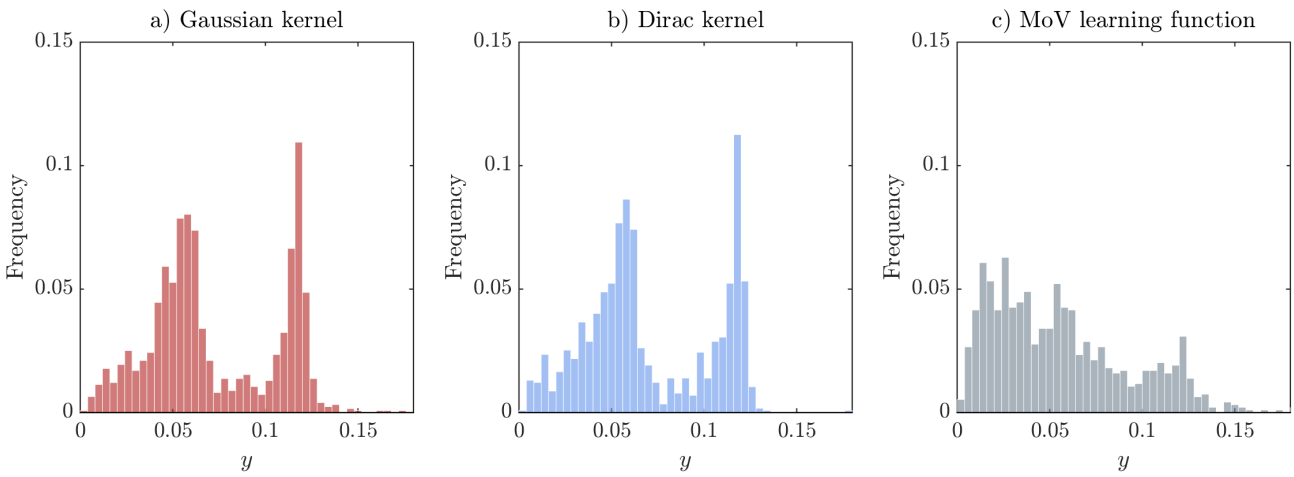


Figure 18: **Histograms on the y locations of the training samples**.

It is seen from Figure 18 that the training samples are concentrated around the tail and the transition region where the stiffness degrades significantly.

From all the examples studied, it can be observed that both the proposed global (mesh-free) AL-GP and the conventional (discretized) AL-GP methods are fairly accurate in the estimation of CDF/CCDF, and the errors in terms of Eq.(33) are typically less than 4%. However, the proposed global AL-GP method with the two-step learning strategy only requires (approximately) half of the (true) model function calls of the conventional AL-GP method. Also it is noted that the global PCE approach is not preferable over the active learning approaches, but this is expected from its non-adaptive nature.

5. Practical issues, limitations, and future directions

5.1. The selection of CDF/CCDF range

One practical issue deserves attention is the selection of the CDF/CCDF range $[y_{\min}, y_{\max}]$. For an UQ problem without any prior knowledge, an arbitrarily selected CDF/CCDF range may include zero or extremely small probabilities, and consequently the integrand of Eq.(16), $w^*(y)$, may include infinity or/and NaN (Not a Number) outputs. A practical solution to this issue is to set infinity and NaN outputs to zero, and in this way the algorithm will automatically skip the corresponding infeasible region and perform CDF/CCDF estimation within the capability of the specified Monte Carlo simulation technique. Moreover, due to similar reasons, the $w^*(y)$ values corresponding to unreliable probability estimates should also be set to zero. For example, in Step 3 of Algorithm 1, if a crude Monte Carlo simulation with 10^6 samples is used, the region with probability estimates smaller than 10^{-5} (this corresponds to a coefficient of variation around $1/\sqrt{10}$, i.e. unreliable) should be skipped. Clearly, another solution for the aforementioned issue is to adaptively increase the sample size of Monte Carlo simulation so that the probability estimates within $[y_{\min}, y_{\max}]$ are always reliable. However, we do not recommend this approach since the sample size can become infeasibly large.

5.2. The stopping criterion

Another practical issue that deserves attention is the stopping criterion Eq.(18). In particular, it should be compatible with the capability of the specified Monte Carlo simulation technique. For example, if a crude Monte Carlo simulation with 10^6 samples is used in Step 3 of Algorithm 1, and we intend to estimate a CDF/CCDF value as low as 10^{-4}, then the coefficient of variation of using 10^6 samples is circa 10^{-1}. Consequently, the $\bar{\epsilon}$ in Eq.(19) should be specified larger than 0.1, otherwise the tolerance is unnecessarily tight (which requires unnecessarily more model function evaluations).

5.3. Only the CDF or the CCDF is of interest

A practical, albeit trivial, issue is that for some applications only the CDF *or* the CCDF (e.g., Example 3) needs to be accurately estimated. For this case, one can simply replace the $\min[\hat{F}_Y^0(y), 1-\hat{F}_Y^0(y)]$ term in error measures W^* and W_L^* with $\hat{F}_Y^0(y)$ or $1-\hat{F}_Y^0(y)$, and the rest part of the algorithm remains intact. The same holds for outputs with symmetric probability distribution (which is known a-priori). Observe that in Example 2, we compute both tails because we assumed no a-priori knowledge on the shape of $f_Y(y)$.

5.4. Singularities in the model function

We mention an important issue for future studies: the existence of bifurcation/singularity in the model function. Suppose the model function has singularity at $y_s = \mathcal{M}(\boldsymbol{x})$, i.e. the behavior of $M(\cdot)$ for $y < y_s$ is fundamentally different from the behavior for $y > y_s$, and the transition is abrupt. If the three-fold metamodel approach (see Eq.(12)) is used to estimate the probability around y_s, the $\hat{\mathcal{M}}^+$ and $\hat{\mathcal{M}}^-$ model may span different sides of y_s, and consequently an error measure using the discrepancy between $\hat{\mathcal{M}}^+$ and $\hat{\mathcal{M}}^-$ could be extremely large regardless of numerous training samples may already be applied to the singularity region. In the context of distribution function estimation, if $y_s \in [y_{\min}, y_{\max}]$, the algorithm would keep applying training samples around y_s yet it could hardly converge. A possible solution to this issue is to use transformation techniques such that in the projected feature space the singularity may be smoothed out [38][39].

5.5. High dimensional problems

Another important topic for future studies is the long-standing challenge of Gaussian process model in handling high-dimensional input. As the dimensionality of input $\boldsymbol{X}$ increases, the parametric space of the covariance function would expand dramatically. Intrusive approaches to handle this issue include sparsity constraint [40] or built-in dimensionality reduction [41], while non-intrusive approaches include advanced nonlinear dimensionality reduction and manifold embedding techniques [42][43]. However, note that all these methods would work under the condition that $y = \mathcal{M}(\boldsymbol{x})$ is equipped with a low-dimensional structure. This condition can be met in many UQ problems involving physical processes. Therefore, the further development of the proposed AL-GP approach in the physics-informed high dimensional UQ analysis context would be particularly meaningful.

5.6. Low probability problems

In the manuscript we estimated the CDF/CCDF down to 10^{-4}, for smaller probabilities the method would require more (true) model function evaluations and more Monte Carlo simulation samples for the metamodel. If an extremely small probability (e.g., 10^{-8}) is of interest, performing the crude Monte Carlo simulation even for the metamodel would be inefficient, and thus variance-reduction Monte Carlo approaches would be preferable. There are recent advances in this line of research [19][20][21], and the synergy between these methods and the proposed AL-GP method for CDF/CCDF estimation is worthy of future studies.

5.7. Multivariate output problems

The typical application of Gaussian process model is for multivariate-input-single-output problems. However, extensions of conventional Gaussian process model to multivariate output are developed in the literature [44][45][46]. For multivariate-output problems, one may encounter the following three scenarios: a) the outputs are statistically independent; b) the outputs are correlated and the correlation structure is known; c) the outputs are correlated and the correlation structure is unknown. The scenarios a) and b) are relatively trivial, the scenario c) can be addressed by methods such as *coKriging* [44], *twin Gaussian processes* [46], *Convolved Gaussian processes* [45], etc. The application of these approaches in the active learning context deserves further studies.

5.8. Metamodels other than Gaussian process

It is also important to note that the proposed active learning strategy for distribution function estimation is not restricted to Gaussian process model. In fact, the approach can be used with any metamodels as long as the model error/uncertainty can be evaluated. This idea instantly generates meaningful research topics for future studies, e.g. the use of PCE (with bootstrap) [15] or PCE-Kriging [30] in distribution function estimation.

6. Conclusions

This study proposes a global active learning-based Gaussian process metamodelling strategy for estimating the probability distribution function in forward uncertainty quantification analysis. The strategy is mesh free in the sense that a-priory discretization (mesh) of the distribution function is not required. A novel error measure is developed such that it satisfies a symmetric condition between cumulative and complementary cumulative distribution functions. As a result of this symmetry, the proposed method is able to simultaneously provide accurate CDF and CCDF estimation in their median-low probability regions. Moreover, a two-step learning function is proposed such that it makes full use of the available information and it is compatible with the error measure. The proposed metamodelling strategy has been tested for three benchmark examples showing both high accuracy and efficiency.

Acknowledgement

Dr. Ziqi Wang was supported by the National Science and Technology Major Project of the Ministry of Science and Technology of China (Grant No. 2016YFB0200605), National Natural Science Foundation of China (Grant No.1808149), and the Natural Science Foundation of Guangdong Province (Grant No.2018A030310067). Dr. Marco Broccardo was supported by DESTESS, a projected which has received funding from the European Union's Horizon 2020 research and innovation programme under Grant No.691728. This support is gratefully acknowledged. Any opinions, findings, and conclusions expressed in this paper are those of the authors, and do not necessarily reflect the views of the sponsors.

References

[1] O. Ditlevsen and H.O Madsen. *Structural reliability methods*. J. Wiley and Sons, Chichester, 1996.
[2] James L. Beck. Bayesian system identification based on probability logic. *Structural Control and Health Monitoring*, 17(7):825–847, 2010.
[3] L.J. Lucas, H. Owhadi, and M. Ortiz. Rigorous verification, validation, uncertainty quantification and certification through concentration-of-measure inequalities. *Computer Methods in Applied Mechanics and Engineering*, 197(51):4591 – 4609, 2008.
[4] R. E. Barlow and Proschan F. *Mathematical Theory of Reliability*. SIAM, Philadelphia, 1996.
[5] H. Owhadi, C. Scovel, T. Sullivan, M. McKerns, and M. Ortiz. Optimal uncertainty quantification. *SIAM Review*, 55(2):271–345, 2013.
[6] Michael Evans and Tim Swartz. Methods for approximating integrals in statistics with special emphasis on bayesian integration problems. *Statistical Science*, 10(3):254–272, 1995.
[7] M. S. Eldred. Recent advances in non-intrusive polynomial chaos and stochastic collocation methods for uncertainty analysis and design. In *50th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, 2009.
[8] R. Y. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo Method*. Wiley, December 2007.
[9] Roger G Ghanem and Pol D Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer, New York, 1991.
[10] Amandine Marrel, Bertrand Iooss, Franois Van Dorpe, and Elena Volkova. An efficient methodology for modeling complex computer codes with gaussian processes. *Computational Statistics & Data Analysis*, 52(10):4731 – 4744, 2008.
[11] Dongbin Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press, 2010.
[12] Roland Schöbi. *Surrogate models for uncertainty quantification in the context of imprecise probability modelling*. PhD thesis, ETH Zurich, Zürich, 2017.
[13] Bruno Sudret. Meta-models for structural reliability and uncertainty quantification. *arXiv:1203.2062*, 2012.
[14] B Echard, N Gayton, and M Lemaire. Ak-mcs: an active learning reliability method combining kriging and monte carlo simulation. *Structural Safety*, 33(2):145–154, 2011.

[15] Stefano Marelli and Bruno Sudret. An active-learning algorithm that combines sparse polynomial chaos expansions and bootstrap for structural reliability analysis. *Structural Safety*, 75:67 – 74, 2018.
[16] T. T. Soong and Mircea Grigoriu. *Random vibration of mechanical and structural systems*. Prentice Hall, 1993.
[17] J. B. Roberts and P. D. Spanos. *Random vibration and statistical linearization*. Dover, 2003.
[18] Jie Li and Jianbing Chen. *Stochastic dynamics of structures*. Wiley, 2009.
[19] B. Echard, N. Gayton, M. Lemaire, and N. Relun. A combined importance sampling and kriging reliability method for small failure probabilities with time-demanding numerical models. *Reliability Engineering & System Safety*, 111:232 – 240, 2013.
[20] Clment Walter. Moving particles: A parallel optimal multilevel splitting method with application in quantiles estimation and meta-model based algorithms. *Structural Safety*, 55:10 – 25, 2015.
[21] Xufang Zhang, Lei Wang, and John Dalsgaard Srensen. Akois: An adaptive kriging oriented importance sampling method for structural system reliability analysis. *Structural Safety*, 82:101876, 2020.
[22] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.
[23] H. B. Nielsen, S. N. Lophaven, and J. Søndergaard. DACE - a matlab kriging toolbox, 2002.
[24] Ivo Couckuyt, Tom Dhaene, and Piet Demeester. oodace toolbox: A flexible object-oriented kriging implementation. *Journal of Machine Learning Research*, 15:3183–3186, 2014.
[25] C. Rasmussen and H Nickisch. Gaussian processes for machine learning (gpml) toolbox. *Journal of Machine Learning Research*, 11:3011–3015, 2010.
[26] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.
[27] Wenjia Wang, Rui Tuo, and CF Jeff Wu. On prediction properties of kriging: Uniform error bounds and robustness. *Journal of the American Statistical Association*, pages 1–27, 2019.
[28] V. Roshan Joseph, Ying Hung, and Agus Sudjianto. Blind Kriging: A New Method for Developing Metamodels. *Journal of Mechanical Design*, 130(3), 02 2008. 031102.
[29] S. Vallender. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974.
[30] R Schöbi, Bruno Sudret, and Stefano Marelli. Rare event estimation using polynomial-chaos kriging. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 3(2):D4016002, 2016.
[31] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.
[32] Rainer Storn and Kenneth Price. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, Dec 1997.
[33] S. Marelli and B. Sudret. UQLab user manual – Polynomial chaos expansions. Technical report, Chair of Risk, Safety and Uncertainty Quantification, ETH Zurich, Switzerland, 2019. Report # UQLab-V1.3-104.
[34] T. Ishigami and T. Homma. An importance quantification technique in uncertainty analysis for computer models. In *[1990] Proceedings. First International Symposium on Uncertainty Modeling and Analysis*, pages 398–403, Dec 1990.
[35] Z. Wang, M. Broccardo, and J. Song. Hamiltonian monte carlo methods for subset simulation in reliability analysis. *Structural Safety*, 76:51 – 67, 2019.
[36] R. Bouc. Forced vibration of mechanical systems with hysteresis. In *Proceedings of 4th Conference Nonlinear Oscillation*, page 315. Prague, Czechoslovakia, 1967.
[37] Y. K. Wen. Equivalent Linearization for Hysteretic Systems Under Random Excitation. *Journal of Applied Mechanics*, 47:150, 1980.
[38] R. Calandra, J. Peters, C. E. Rasmussen, and M. P. Deisenroth. Manifold gaussian processes for regression. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3338–3345, July 2016.
[39] Edward Snelson, Carl Edward Rasmussen, and Zoubin Ghahramani. Warped gaussian processes. In *IN ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NIPS)*, page 2003. MIT Press, 2003.
[40] Josip Djolonga, Andreas Krause, and Volkan Cevher. High-dimensional gaussian process bandits. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1025–1033. Curran Associates, Inc., 2013.
[41] Rohit Tripathy, Ilias Bilionis, and Marcial Gonzalez. Gaussian processes with built-in dimensionality reduction: Applications to high-dimensional uncertainty propagation. *Journal of Computational Physics*, 321:191 – 223, 2016.
[42] Ronald R. Coifman and Stphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5 – 30, 2006. Special Issue: Diffusion Maps and Wavelets.
[43] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.
[44] Jay M. Ver Hoef and Ronald Paul Barry. Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference*, 69(2):275 – 294, 1998.
[45] Mauricio A. lvarez and Neil D. Lawrence. Computationally efficient convolved multiple output gaussian processes. *Journal of Machine Learning Research*, 12:1425, 00 2011.
[46] Liefeng Bo and Cristian Sminchisescu. Twin gaussian processes for structured prediction. *International Journal of Computer Vision*, 87(1):28, Feb 2009.