Structural Safety 96 (2022) 102174

Contents lists available at ScienceDirect

Structural Safety

journal homepage: www.elsevier.com/locate/strusafe

Active learning for structural reliability: Survey, general framework and benchmark

Maliki Moustapha *, Stefano Marelli, Bruno Sudret

Chair of Risk, Safety and Uncertainty Quantification, ETH Zurich, Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland

ARTICLE INFO

Keywords:
Structural reliability
Active learning
Surrogate models
Benchmark
Gaussian process (Kriging)
Polynomial chaos expansions

ABSTRACT

Active learning methods have recently surged in the literature due to their ability to solve complex structural reliability problems within an affordable computational cost. These methods are designed by adaptively building an inexpensive surrogate of the original limit-state function. Examples of such surrogates include Gaussian process models which have been adopted in many contributions, the most popular ones being the efficient global reliability analysis (EGRA) and the active Kriging Monte Carlo simulation (AK-MCS), two milestone contributions in the field. In this paper, we first conduct a survey of the recent literature, showing that most of the proposed methods actually span from modifying one or more aspects of the two aforementioned methods. We then propose a generalized modular framework to build on-the-fly efficient active learning strategies by combining the following four ingredients or modules: surrogate model, reliability estimation algorithm, learning function and stopping criterion. Using this framework, we devise 39 strategies for the solution of 20 reliability benchmark problems. The results of this extensive benchmark (more than 12,000 reliability problems solved) are analyzed under various criteria leading to a synthesized set of recommendations for practitioners. These may be refined with *a priori* knowledge about the feature of the problem to solve, *i.e.* dimensionality and magnitude of the failure probability. This benchmark has eventually highlighted the importance of using surrogates in conjunction with sophisticated reliability estimation algorithms as a way to enhance the efficiency of the latter.

1. Introduction

Structural reliability analysis is a central tool for the design and assessment of complex engineering systems. Such systems are affected by uncertainties, which may arise from natural variability in their physical properties (*e.g.*, material strength, manufacturing tolerances), operating conditions (*e.g.*, variable loads, environmental conditions) or simply because of an incomplete or lack of knowledge (*e.g.*, in the non-destructive assessment of existing structures). Structural reliability analysis aims at assessing the effects of such uncertainties, by estimating the associated failure probability with respect to some relevant limit states. In this paper, we consider a probabilistic setting, in which the uncertainties are represented through a set of random parameters $\boldsymbol{X} \in \mathcal{D}_{\boldsymbol{X}} \subset \mathbb{R}^M$ completely defined by their joint probability distribution function (PDF) $f_{\boldsymbol{X}}$. These parameters represent the state of the system, which can be evaluated through a so-called performance function (a.k.a. limit-state function), herein denoted by $g(\boldsymbol{X})$. By convention, the system is assumed to be in a failure (resp. safe) state when $g(\boldsymbol{x}) \leq 0$ (resp. $g(\boldsymbol{x}) > 0$). The probability of failure of the system can then be defined as

$$P_f = \mathbb{P}(g(\boldsymbol{X}) \leq 0) = \int_{\mathcal{D}_f} f_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{1}$$

This integration over an implicitly defined domain $\mathcal{D}_f = \{\boldsymbol{x} : g(\boldsymbol{x}) \leq 0\}$ is not straightforward to solve and has motivated the development of a rich variety of techniques [1–3]. These techniques can be broadly grouped in several classes. These include *approximation methods*, where the limit-state function is linearized (or otherwise approximated) around a so-called design point, *e.g.*, the most probable failure point (MPFP) in a suitably transformed probabilistic input space. This step allows one to then derive (semi-)analytically an approximation of the failure probability. This class includes the well-known first-order and second-order reliability methods (FORM and SORM) [4, 5]. This family, however, is known to suffer severe limitations when the limit-state function is strongly non linear, or in the presence of multiple failure modes. A second class of methods, namely that of *simulation techniques*, is widely used for the solution of Eq. (1). Monte Carlo simulation is certainly among the most widely-used methods in

* Corresponding author.
E-mail address: moustapha@ibk.baug.ethz.ch (M. Moustapha).

https://doi.org/10.1016/j.strusafe.2021.102174
Received 29 May 2021; Received in revised form 6 November 2021; Accepted 29 November 2021
Available online 16 January 2022

this category. It is known to be robust and unbiased, yet its convergence rate is extremely slow, especially when the target failure probability is small. This is problematic when the computational model used in the evaluation of the limit-state function is costly, a common occurrence when *e.g.*, finite element analysis is involved. More advanced methods, based on variance-reduction techniques, are constantly being developed. A non-exhaustive list of the latter include importance sampling [6], subset simulation [7], directional simulation [8], line sampling [9] and asymptotic sampling [10]. Numerous variants of these methods have been introduced in the recent literature in an attempt to further accelerate their convergence rates, *e.g.*, [11–13]. However, the computational cost remains unaffordably high (*i.e.* $\mathcal{O}(10^{3-4})$ model runs) when considering time-consuming computational models.

In the past decade, a different avenue that offers substantial savings in the computational budget, while retaining the favorable properties of simulation methods, has been explored in the reliability analysis literature: *surrogate-model aided methods*. Surrogate models are inexpensive approximations of the original computational model, which have consistently shown superior performance when combined with the traditional simulation methods introduced earlier. Originally, simple polynomial response surface models (RSM) were built using a set of carefully designed computer experiments [14,15] These RSM were then used *in lieu* of the original computational model to approximately solve Eq. (1). Borrowing from the machine learning community, this process has evolved into a more sophisticated methodology known as *active learning* [16,17]. In active learning, the surrogate model is not used as a mere proxy of the original computational model, but as a tool to help explore the random input variable space efficiently. The idea is to start with an initial small set of model evaluations, known as the *experimental design*, which is then sequentially enriched following a so-called *learning function*. The latter aims at finding which model evaluation would bring the most useful information for the purpose of accurately assessing the failure probability of the system under consideration. This starts from the premise that in Eq. (1), only the sign of the limit-state is required to characterize the failure domain $\mathcal{D}_f$ in simulation-based reliability algorithms. The goal is then to approximate the limit-state surface as parsimoniously as possible (*i.e.*, using the least number of model evaluations) to achieve the best possible accuracy for the estimated failure probability. In the past few years, an increasingly large number of contributions have been proposed in the field of active learning for reliability analysis. The most popular approaches are based on Kriging, a.k.a. Gaussian process modeling, owing to its built-in error measure. The most prominent examples are the efficient global reliability analysis (EGRA) proposed by Bichon et al. [16] and the active Kriging Monte Carlo simulation (AK-MCS) developed by Echard et al. [17]. The latter is a cornerstone of various methods derived incrementally by modifying one or another aspect of the AK-MCS algorithm [18] and commonly referred to as *AK methods*. For instance, replacing the Monte Carlo simulation part of the algorithm with importance sampling or subset simulation leads respectively to AK-IS [19] or AK-SS [20]. Similarly, other contributions have targeted the surrogate model type, introducing for instance support vector machines [21–24] or polynomial chaos expansions [25]. A comprehensive overview of recent developments in active-learning based reliability analysis can be found in [26].

This paper aims at achieving two goals. The first is to provide an in-depth characterization of the current trends in active-learning-based reliability analysis through a comprehensive survey, following the footsteps of Teixeira et al. [26]. In doing so, however we focus on highlighting common aspects in the numerous literature contributions. More specifically, we classify the methods with respect to the specific *novelty* put forward in each contribution.

We then propose a generalized framework that summarizes the survey and puts the entire reviewed literature under a consistent formal umbrella. This framework is built by combining non-intrusively four identified common ingredients of active learning-based methods: i. a surrogate model, ii. a reliability estimation algorithm, iii. a learning function and iv. a stopping criterion. In the second part of the paper, we then conduct the first-ever extensive benchmark of active learning methods considering, on the one hand, a collection of 20 problems of diverse characteristics and on the other hand, a total of 39 active learning schemes built by combining selected methods in each of the four aforementioned components.

This benchmark is mainly aimed at illustrating how easily the proposed framework can be configured to reproduce a wide class of recently published studies, and hence only a limited number of methods per component was considered. The methods were selected for their maturity and ease of deployment, *i.e.*, they do not require extensive tuning by the user and are relatively fast to run on standard workstations. For instance, only Kriging, polynomial chaos expansions and PC-Kriging surrogate models are considered in this study due to their prevalence in the reliability analysis literature and their off-the-shelf availability. Similarly the scope of the problems solved within the benchmark is limited to the ones typically considered in the reviewed active learning papers. More specifically, we do not consider time-variant (such as in [27]), dynamic or extremely high-dimensional problems (*i.e.*, in the order of hundreds) as they would require special treatment. For the former, dimensionality reduction techniques are often combined with surrogate modeling as in manifold learning [28], active subspace method [29,30] or in a more general setting as in [31]. Even though some of these methods may be used for mildly high dimensional problems, they are not considered in the survey or benchmark carried out in this paper.

The large batch of analyses resulting from the selected methods is repeated 15 times, to obtain statistically significant estimates on the stability of each method. This results in a set of over 12,000 reliability analyses which allows us to validate, repeat and assess most of the methods introduced in the recent literature, and at the same time to explore a large portion of new methods and combinations that have not been published yet. The results of this benchmark are used to give recommendations as to which type of methods performs the best generally or at least to be preferred given features of the reliability problem at hand.

The remainder of the paper is organized as follows. Section 2 presents a literature review of the current state-of-the art in surrogate-modeling based reliability analysis. Section 3 introduces a generalized active learning reliability framework inferred from the literature review. In Section 4, an extensive comparative benchmark study is carried out on a wide class of methods and benchmark problems. Finally, recommendations and conclusions are given in Sections 5 and 6.

2. A short overview of recent literature

2.1. Common rationale

At the core of active-learning reliability lies the idea of reducing the cost of simulation algorithms by introducing a surrogate model as an inexpensive approximation of the expensive-to-evaluate limit-state function. Surrogate models were first introduced in a static scheme to globally replace computer codes mainly for the purpose of visualization or optimization. Active learning pushes this concept further by aiming to an efficient allocation of resources, *i.e.*, computer simulations are performed sparingly and only when most needed. Active learning reliability (ALR) algorithms are practically devised using the general framework illustrated in Fig. 1. In the initialization step, a so-called *experimental design* $\mathcal{E}^{(0)} = \left\{ \left(\boldsymbol{\mathcal{X}}^{(i)}, \mathcal{Y}^{(i)}\right) : \mathcal{Y}^{(i)} = g\left(\boldsymbol{\mathcal{X}}^{(i)}\right) \in \mathbb{R}, \boldsymbol{\mathcal{X}}^{(i)} \in \mathbb{X} \subset \mathbb{R}^M, i = 1, \dots, m_0 \right\}$ is initially generated. The input sample set $\left\{ \boldsymbol{\mathcal{X}}^{(i)}, i = 1, \dots, m_0 \right\}$ is often drawn using space-filling methods such as Latin hypercube sampling (LHS, [32]) or randomized low-discrepancy sequences [33]. Typically m_0 is chosen small, *i.e.*, in the order of tens of samples. Following initialization, the algorithm enters in a four-step loop where:

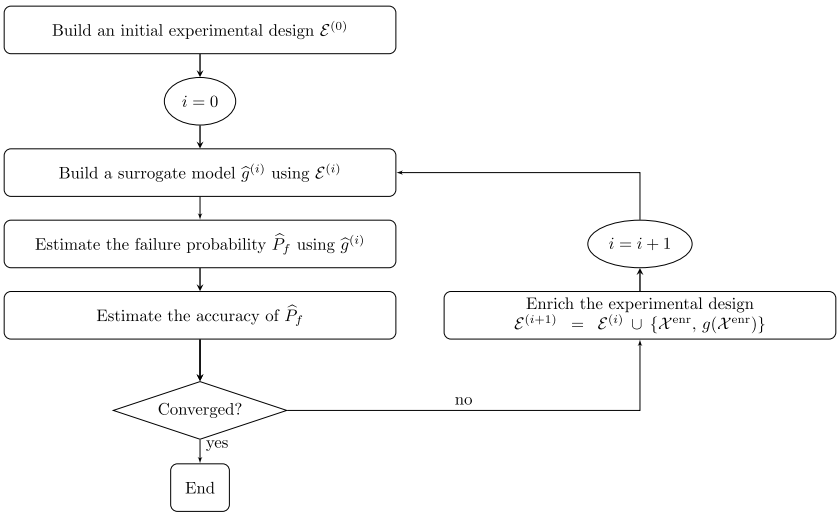


Fig. 1. General flowchart of active learning reliability.

1. A **surrogate model** is built using the current experimental design;
2. The failure probability is estimated using the current surrogate model and an appropriate **reliability estimation** algorithm;
3. The **convergence** of the algorithm is assessed;
4. An **enrichment of the experimental design** is carried out by appropriately selecting at least one pair of sample points $\left\{\boldsymbol{\mathcal{X}}^{(\text{enr})}, g\left(\boldsymbol{\mathcal{X}}^{(\text{enr})}\right)\right\}$, when convergence is not achieved. This is often achieved by evaluating a so-called *learning function* which gives information as to which points are most likely to increase the accuracy of the surrogate (and subsequently of the estimated failure probability) when added to the experimental design.

One of the first implementations of this flowchart was proposed by Bichon et al. [16] in their efficient global reliability analysis method (EGRA). In this work, they used Gaussian process regression as a surrogate model, Monte Carlo simulation as reliability estimation algorithm and the so-called expected feasibility function (EFF) as a means to find points to enrich the experimental design. The latter is actually an adaptation to contour estimation of the well-known expected improvement (EI) function [34] widely used in Bayesian optimization as first introduced in [35]. A noticeable improvement of EGRA was introduced by Echard et al. [17] in the widely known active Kriging-Monte Carlo simulation (AK-MCS) method. Contrary to EGRA, where reliability estimation is carried out only after the enrichment stage is completed, AK-MCS couples enrichment and reliability estimation. Furthermore, it introduces a new learning function, the so-called deviation number U, which is optimized with respect to a pre-defined sample set. This highly reduces the computational cost and the complexity of the active learning procedure.

While this algorithm is at the time of writing already ten years old, it has aged surprisingly well, with a number of recent methods proposing only minor variations to one or more of the steps just reported. A comprehensive survey on recent developments on this topic was recently proposed by Teixeira et al. [26]. In the following sections we identify and describe in more detail four key ingredients that are common to all of the aforementioned active-learning based methods.

2.2. Surrogate models in structural reliability

Various surrogate models were already used in adaptive schemes for the solution of reliability problems even before the emergence of the AK methods. Polynomial response surface models were arguably the first type of surrogates used in the context of structural reliability analysis. Faravelli [14] uses a second-order polynomial while [36] introduced a two-stage approach where a first quadratic response surface is used to locate the MPFP. A second response surface is then built close to that MPFP to refine the knowledge of the limit-state surface. A direct improvement of this approach which introduces an iterative procedure was proposed by Rajashekhar and Ellingwood [37]. Leonel et al. [38] considered various schemes using response surfaces for reliability analysis in crack propagation and concluded that a direct coupling (*i.e.*, building the surrogate only once the MPFP has been located) is more efficient. More recently, Roussouly et al. [39] proposed an iterative scheme that combines trust regions, sparse response surface and bootstrap for the identification of regions where enrichment is necessary. Radial basis functions (RBF), which have been popular in static surrogate-assisted reliability analysis, were introduced in a sequential approach as well. Li et al. [40] presented an MCS-based approach where an RBF is sequentially updated through a constrained min–max optimization problem which aims at finding points close to the limit-state surface while keeping a minimum distance to the existing ED points. Shi et al. [41] proposed two other learning schemes based on RBF considering either an ensemble of surrogates or cross-validation. In the former case, the interquartile range of the predictions using an ensemble of surrogates is considered as a measure of uncertainty to derive a learning function similar to the U-function of Echard et al. [17]. Another learning function similar to U was developed by Marelli and Sudret [25] using bootstrap and polynomial chaos expansions (PCE). More recently, sparse Bayesian PCE was used by Cheng and Lu [42] where a new learning function relying on the Gaussian process variance was proposed. Pan et al. [43] also used Bayesian regression PCE combined with the deviation number U.

Popular methods from the machine learning community, such as support vector machines or neural networks, have also been steadily introduced in structural reliability. Support vector machines for classification was first introduced by Hurtado [21]. Basudhar and Missoum [44] proposed an adaptive scheme combining SVM classification and Monte Carlo simulation. The enrichment scheme is based on finding the point belonging to the limit-state surface approximation that is the furthest from the existing training points. This is a maximin problem solved using a general-purpose optimization algorithm. Lacaze and Missoum [45] proposed an improvement of this maximin scheme by including a weight which accounts for the random variables joint PDF. Another improvement aiming at avoiding the optimization problem and relying on a candidate pool for enrichment has been proposed by Pan and Dias [46]. Combining SVM and subset simulation, Bourinet et al. [23] proposed a learning scheme where a classifier is built in each iteration of the SuS algorithm. SVM has also been widely used

in its regression form (SVR) for reliability analysis [24]. Bourinet [47] proposed an SVR scheme with three novelties: i. the sample set size is kept constant, meaning some samples are withdrawn from the training set as the algorithm is proceeding, ii. intermediate thresholds are used to approximate the limit-state functions and iii. the surrogate models built in each stage are combined in a weighted ensemble to keep information of all training points without increasing the computation time. Another popular machine learning method widely used in structural reliability analysis is neural networks. Even though most of the contributions are in a static scheme, the most recent ones consider adaptivity [48,49]. Sundar and Shields [50] proposed a two-stage algorithm where an artificial neural network (ANN) is first used together with parallel Markov Chains to identify (possibly disjoint) failure regions. The ANN is then enriched to accurately represent the limit-state surface. Finally, Gomes [51] introduced an active scheme combining artificial neural networks and Monte Carlo simulation using the bootstrap-based learning function introduced in [25].

Various other surrogate model types have been used to propose new active learning reliability algorithms, following similar schemes as introduced earlier, *e.g.*, polynomial chaos-Kriging [52], high-dimensional model reduction (HDMR) [53], deep neural networks [54] or stochastic spectral embedding [55], among others.

2.3. Reliability estimation algorithm

An immediate alternative strategy to AK-MCS can be devised by focusing on the reliability estimation algorithm. The benefits of replacing Monte Carlo simulation are two-fold. First, more sophisticated algorithms have been developed to reduce the variance of the failure probability estimate, and introducing them in active learning allows overcoming the pitfalls of MCS, *i.e.*, its slow convergence rate. Second, choosing another reliability estimation algorithm also allows one to modify the way sample candidates to enrichment are generated. In fact, for problems with low failure probability, the initial candidate set for enrichment may not contain any sample point at all in the actual failure domain. This can seriously reduce the chances of convergence of the active learning scheme. In contrast, more advanced reliability estimation algorithms may allow one to reach more easily areas associated with small probability densities, as well as disconnected failure regions.

Basically, almost all well-established simulation-based reliability estimation methods have been used together with active learning in the literature. A direct adaptation of AK-MCS, simply coined AK-IS, has been proposed by Echard et al. [19] using importance sampling [3] in lieu of Monte Carlo simulation. In this contribution, they first find the design point using FORM and the original model. They then build an importance density sample set around this point which is used both for computing the failure probability and as candidate pool for enrichment. Gaspar et al. [56] proposed to use a surrogate model even for the location of the design point, hence further reducing the computational cost. Zhao et al. [57] did not rely on the design point but rather uses Monte Carlo Markov Chain (MCMC) to generate points in the failure domains. Importance sampling is then performed around those points together with enrichment. This allows overcoming a major shortcoming of importance sampling related to the presence of multiple design points. Another line of research involving Kriging combined with IS includes the meta-IS algorithm where [58] proposed to use the Kriging model to approximate the optimal importance density in an iterative scheme. Cadini et al. [59] combined the work of Dubourg et al. [58] and Echard et al. [19] in a two-stage algorithm called metaAK-IS^2. Other sequential importance sampling methods have been adapted in an active Kriging strategy. For instance, Balesdent et al. [60] sequentially built and enriched Kriging models in intermediate steps of a cross-entropy and non-parametric adaptive importance sampling algorithm. Other contributions using adaptive importance sampling include [43,56,61–65].

Another popular reliability estimation algorithm that has been used in an active learning scheme is subset simulation [66]. Huang et al. [20] introduced AK-SS which, as its name suggests, is a declination of AK-MCS with the use of subset simulation for the computation of the failure probability. All other aspects are those of the original AK-MCS algorithm, including the candidate pool for enrichment which is obtained by an initial large Monte Carlo sample set. The obvious limitation here is that it may be difficult to find points in the failure region for problems where failure is an extremely rare event. Zhang et al. [67] then proposed an improvement where the first and last levels of subset simulation are used as candidate pool for enrichment. The first level being a global Monte Carlo and the last one leading to points closest to the limit-state surface, this method allows both exploration and exploitation of the random input space. Ling et al. [68] proposed an intermediate approach where a local Kriging model is built at each stage of subset simulation. Other similar methods include Bayesian subset simulation [69,70] which combines subset simulation, sequential Monte Carlo and Kriging and AK-SSIS [71] which combines subset simulation and importance sampling in an active Kriging strategy.

Finally, we shall note that even though importance sampling and subset simulation have been widely exploited in active learning methods, the use of other variance-reduction simulation methods has been explored. Examples include algorithms such as directional importance sampling [72], radial basis importance sampling [73] or line sampling [74].

2.4. Enrichment of the experimental design

A core feature of active-learning-based reliability methods is that the accuracy of the failure probability estimate is gradually increased by enriching the experimental design. A key component in this respect is the learning function (LF), which plays the central role of providing a measure of the information value of any experimental design enrichment candidates. Many authors have come up with new learning functions that can increase the efficiency of otherwise comparable methods. A direct improvement of the deviation number U was given for instance by Peijuan et al. [75], where a line search step is added to get even closer to the limit-state surface once the best next point with respect to U is found. Arguing that errors due to regions with small density would be negligible in the final estimate of the failure probability, Wen et al. [76] also used the random variables joint PDF to constrain the EFF learning function. Similarly, Sun et al. [77] proposed the least improvement function (LIF) which weights the probability of misclassification $\Phi(-U(\boldsymbol{x}))$ with the joint probability density of the samples $f_{\boldsymbol{X}}(\boldsymbol{x})$, an idea already used in [58]. Tong et al. [78] followed this idea and, adding more terms related to global/local uncertainty, they created a new learning function.

From another perspective, Lv et al. [74] introduced a new learning function based on the information theory with an analytical expression similar to EFF. Hu and Mahadevan [79] introduced a method which relies on computing the sensitivities of the failure probability to add new points to the experimental design. A new learning function using K-fold cross validation generalizable to any type of surrogate model was introduced by Xiao et al. [80]. More recently, Jiang et al. [81] proposed an approach which is based on splitting the space using Voronoi cells and finding out those points with the largest sensitivities to the estimated failure probability. The use of Voronoi cells allows the authors to both reduce the computational cost and spread the sample points as much as possible. The latter goal is also achieved through a pre-processing step by Zhang et al. [82] who then introduced a new LF inspired from the expected improvement.

2.5. Stopping criterion

The stopping criterion is an often overlooked yet crucial part of any active learning reliability algorithm. Three types of criteria have been proposed in the literature to halt the iterative enrichment scheme. The first one is directly based on the learning function. For instance, Bichon et al. [16] proposed to stop the enrichment scheme when the value of the expected feasibility function is lower than 10^{-3}. Similarly, Echard et al. [17] stops AK-MCS iterations when $U > 2$ for all candidate points. This actually means that the probability of misclassifying any point from the sample set used to evaluate the failure probability is below 2.28%. This criterion has shown to be extremely conservative, leading to unnecessarily added points. Some authors have tried softening it either by considering the whole candidate set through an average, for instance [18,77,83], or by considering convergence when only a small proportion of the candidate set does not comply with $U > 2$ [84,85].

The second family of convergence criteria are those based on the accuracy of the failure probability. Using the Kriging variance, Dubourg et al. [86] proposed a bound on the estimate $\hat{P}_f$ which accounts for the Kriging epistemic uncertainty. Similarly, Sun et al. [77] and Jian et al. [83] proposed an upper bound on $\left|\hat{P}_f - P_f\right|$ using the probability of misclassification $\Phi(-U)$. For surrogate models which do not possess a built-in error measure, similar bounds have been derived considering either cross-validation [41] or bootstrap replicates [25].

Finally, the third family of stopping criteria has been built using the stabilization of either the limit-state surface or the failure probability estimates within enrichment iterations. Basudhar and Missoum [44] tracked the fraction of some predefined convergence points that changed sign within two updates of an SVM model and assumed convergence when this fraction was relatively small. This criteria, often with slight adjustments, has been used in numbers of SVM-based active learning schemes [24]. As for the failure probability, the obvious approach is to track its variation within iterations. Stabilization criteria may often lead to premature convergence when the initial surrogate model is extremely inaccurate. A workaround consists in tracking the convergence over several iterations, on average 2 to 3 and in some contributions and up to 10 iterations [47]. An alternative is to smooth out the convergence criterion as in [44] by using an exponential curve fitted to the convergence criterion.

3. A generalized active learning reliability framework

3.1. Motivation

As anticipated in the previous section, the state-of-the-art in active learning reliability can essentially be summarized into four basic components or modules. These modules are namely the surrogate model, the reliability estimation algorithm, the learning function and the stopping criterion. Most of the contributions in the recent literature can be reconstructed by combining various methods within each module. In a few cases, elaborate ad-hoc techniques are devised by taking advantage of some highly specific combination of such methods. The modules are in these cases not independent anymore but intrusively linked to each other. However, the advantages brought by such configurations are most often only marginal and not systematically justified by benchmarks.

In this section, we present a modular framework for active learning whose first aim is to unify and present the plethora of active learning reliability methods from a single and consistent viewpoint. The interest of such an approach can be seen through the prism of the no-free lunch principle. Despite their claimed advantages, methods proposed in the literature perform best under certain conditions and do not generalize so well as to provide consistently superior performance in the wide spectrum of structural reliability problems. By framing active learning reliability under a modular framework, we can then take advantage of each method to solve a wide class of reliability problems, possibly identifying guidelines based on limited prior information, such as the problem dimensionality.

A second advantage of the framework we propose is that it decouples the four modules. This independence means that there is no need to alter or adapt a given method/module to fit in the overall workflow. Methods can be used solely through their input/output structures, as "black boxes" even allowing for a seamless interconnection to third-party software.

The core idea of the framework is depicted in Fig. 2. The four modules are shown in columns with examples of popular methods. Most of the contributions surveyed in the previous section can be retrieved by appropriately combining the methods. For instance, combining all methods on the first row, *i.e.*, Kriging, Monte Carlo simulation, deviation number U and the LF-based stopping criterion, leads to the well-known AK-MCS. In principle, all methods within each module block can be combined with any from the other blocks. The only exceptions are from the methods that specifically rely on the surrogate built-in error, *e.g.*, the Kriging variance. It should be noted however that alternatives have been proposed in the literature to estimate comparable local error measures when not directly provided by the surrogate model itself.

3.1.1. Surrogate models

Surrogate models lie at the core of any active learning reliability algorithm. A relevant aspect that needs to be stressed here is that they are merely used as a tool to explore the random input space in conjunction with the original computational model. They are not meant to replace the latter *per se*. Many surrogates have been adopted in the literature for active learning and can be further classified on the basis of different properties. One way is to consider interpolation *vs.* regression approaches. The former are often preferred in active learning schemes as they allow to precisely approximate the limit-state function in the vicinity of any point that belongs to the experimental design. One may also consider a classification with regards to the availability of a built-in error measure. Built-in errors have been instrumental in the proliferation of active learning schemes, because most enrichment schemes rely on them. Kriging for instance, with its local variance estimator, is arguably the most adopted surrogate in recent active learning contributions. However, as shown in the literature review above, alternatives also exist and have even shown to be quite efficient. Such alternatives may include a similar error measure (either through statistical methods such as bootstrap and cross-validation or plug-in methods) or the use of alternative learning functions, as will be explained shortly.

3.1.2. Reliability estimation algorithm

Despite approximation methods have been sometimes used in early surrogate-assisted reliability methods [36,37], most of the recent contributions rely on simulation-based methods. Monte Carlo simulation, thanks to its generality, is naturally one of the most commonly-used methods. The use of variance-reduction techniques such as subset simulation or importance sampling has also been widely explored. The latter can reduce the error due to the random nature of the sampling algorithm while keeping the number of model evaluations as low as possible, especially when the probability of failure is low.

In this contribution, we advocate for going even one step further by using an "overkill setting" of the reliability estimation algorithm, further capitalizing on the negligible computational costs associated with the use of surrogate models. By "overkill setting", we mean that the parameters of the algorithms are tuned to drastically reduce the coefficient of variation of the resulting failure probability estimate. This set of parameters depends on the reliability estimation algorithm. When using subset simulation for instance, the batch sample size and conditional failure probability are made larger than in the settings classically found in the literature: we choose here 10^5 samples per

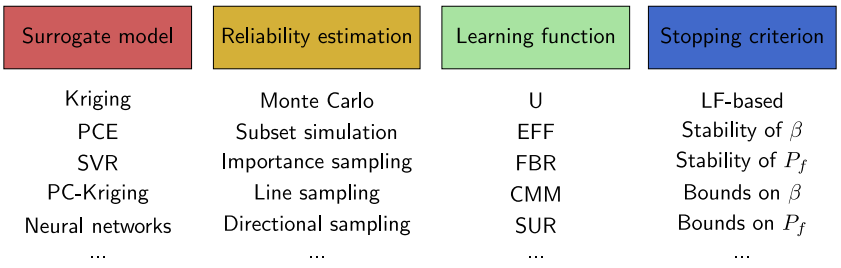


Fig. 2. Active learning reliability framework with example of methods.

simulation step and an intermediate probability of 0.25 instead of the usual 0.10 value.

This approach has a two-fold benefit. First, the stochastic error due to the reliability estimation algorithm is reduced as much as possible, hence leaving only the surrogate-induced error to dominate the global estimation uncertainty on the failure probability. It should be noted that the overall computational time is of course increased, especially when Kriging is used. However, this overhead is expected to be marginal when compared to that of an actual computational model, *e.g.*, a finite element analysis. Second, by over-calibrating the reliability algorithm, we allow for the random space to be even more thoroughly explored, hence increasing the likelihood of finding sample points in the failure regions when the latter is considerably small. Examples of such settings will be shown in the benchmark section and compared to more traditional settings.

3.1.3. Learning function

The learning function is used as a driver to add new points in the experimental design. It is often intrinsically linked to both the surrogate model and the reliability estimation algorithm. Indeed, its very definition often draws from the characteristics of the surrogate model, *e.g.*, variance or built-in error measure. This does not need to be the case systematically, as the same features can be replaced by statistical methods that provide comparable error metrics, such as bootstrap and cross-validation or even mere distance measures to the existing experimental design points.

Generally, new candidate enrichment points are obtained by minimizing (or maximizing) the learning function over the input domain. The optimization problem is most often simplified into a discrete approximation where the enrichment samples are chosen from a finite candidate pool. In most of the literature (starting from the original AK-MCS algorithm [17] and in most subsequent variants), the candidate pool is defined prior to the analysis using Monte Carlo sampling. However this is not an optimal approach. For instance, it is difficult to find points in the failure domain when the real failure probability is very small (*i.e.*, $P_f < 10^{-6}$) as this would require an extremely large candidate pool. An alternative and more efficient approach proposed here is to define the candidate pool as the set of samples generated by the chosen reliability estimation algorithm, which can in principle be updated throughout the ALR iterations. This allows us to fully exploit the benefits of more advanced reliability estimation algorithms that are more likely to locate the (multiple) failure domains and sample more points where needed most. To further avoid being intrusive, we consider as candidate pool for enrichment all samples that were used to estimate the failure probability in the previous iteration of the algorithm. To accelerate the procedure, it is possible to statistically reduce the size of the candidate pool by simple down-sampling or clustering. The latter approach can also serve as a way of simultaneously identifying multiple enrichment points so as to take advantage of any available parallelization capability.

Table 1
Methods selected in each module to create the 39 solution strategies used in the benchmark. Further details about each method are given in Sections 4.2 and in the supplementary materials (Appendix B).

Reliability	Metamodel	Learning function	Stopping criterion
Monte Carlo simulation Subset simulation Importance sampling	Kriging PC-Kriging	U EFF	Beta bounds Beta stability Combined
Monte Carlo simulation Subset simulation Importance sampling	PCE	FBR	Beta stability

3.1.4. Stopping criterion

The stopping criterion is an important part of the active learning scheme as the efficiency of the algorithm is ultimately and largely driven by its robustness. Too loose a stopping criterion can lead to premature convergence, while a too strict one would cause the unnecessarily addition of costly experimental design points. The criteria proposed in the literature can be classified into two groups. First are those based on the learning function value, *e.g.*, [16,17]. These have shown to often be extremely conservative. The second family is derived by directly monitoring the accuracy of the estimated failure probability. Confidence bounds on the latter can be derived [52,58], and convergence is assumed when such bounds are small enough. Alternatively, one may monitor their evolution and assume convergence when a certain degree of stability is observed. Finally, increased robustness may be achieved by either combining different stopping criteria and/or considering convergence only when the criteria are satisfied consistently within a given number of consecutive iterations.

4. Comparative study

4.1. Benchmark set-up

The ingredients shown in the previous section can be assembled non-intrusively to build active learning schemes. In this section we perform an extensive comparison of several framework configurations on a set of benchmark reliability problems representative of a wide range of real case applications. We selected such configurations by considering some of the most widely-used methods in each module. Table 1 shows the different algorithms considered for the benchmark in this paper. The first part of the table deals with methods that use the built-in surrogate model variance, while the second is based on a regression method and bootstrap error estimation. All possible combinations resulting from the tensor product of each compatible ingredient are considered. This amounts in a total of 39 strategies (36 for the first surrogate class and 3 for the second). UQLab [87], a MATLAB framework for uncertainty quantification was used to run these analyses.

Using these 39 strategies, a collection of 20 reliability problems are solved. 11 of these problems were collected from the TNO reliability benchmark repository [88]. The remaining were chosen from the literature with the aim of ensuring a large variety both in terms of limit-state

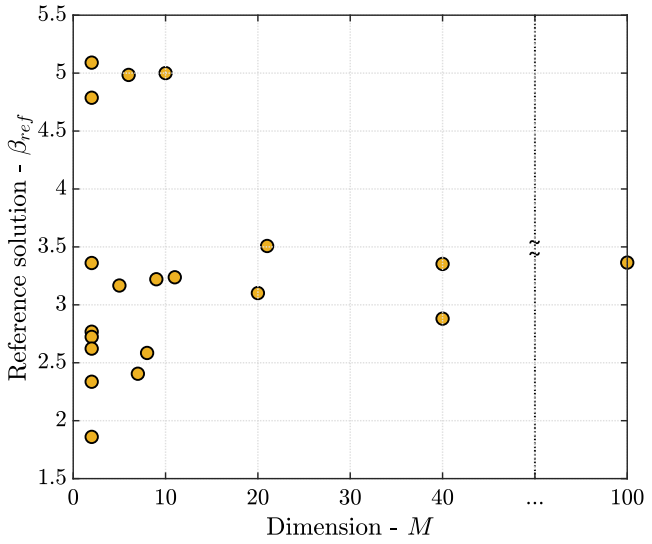


Fig. 3. Collection of problems selected for the benchmark in terms of dimensions M and reference reliability indices β_{ref}.

function dimensionality and reference failure probability/reliability index. Most of the limit-state functions are analytical, except for two which are based on a truss finite element model. A list of the problems together with some references is given in the supplementary materials (Appendix B). Fig. 3 summarizes these problems in terms of dimension against reference reliability index. They range from input dimension $M = 2$ to $M = 100$ and have a reliability index (resp. failure probability) that ranges from $\beta_{\text{ref}} \approx 1.86$ (resp. $P_{f,\text{ref}} \approx 3.14 \cdot 10^{-2}$) to $\beta_{\text{ref}} \approx 5.15$ (resp. $P_{f,\text{ref}} \approx 1.32 \cdot 10^{-7}$). The reference solutions are calculated using the original models and a large Monte Carlo set whose size is set adaptively until a coefficient of variation of 1% is reached. For the four problems that have a failure probability smaller than 10^{-7}, an overcalibrated subset simulation is used instead.

Throughout the benchmark, each analysis is repeated 15 times. The only exception being benchmark #8 of dimension 100, which is repeated only 10 times. It should be noted that within different strategies, the same initial random conditions/seeds are used for each of the repetitions. Hence, a total of 11,700 reliability analyses (39 strategies × 20 problems × 15 repetitions) are carried out for this benchmark.

4.2. Algorithmic settings

In this section, we will briefly review the algorithm settings used for each of the methods selected in Table 1. The methods selected for the first three components, *i.e.*, surrogate models, reliability estimation algorithm and learning functions are detailed in the supplementary materials (Appendix B). A summary of the most important settings for the surrogate models and reliability estimation algorithms is given in Fig. 4. The learning functions however do not possess any special setting and are used exactly as described in the supplementary materials.

This section therefore focuses on the three stopping criteria mentioned in Table 1:

Beta bounds:. This stopping criterion is based on the Kriging variance and reads:

$$\frac{\left|\widehat{\beta}^{+} - \widehat{\beta}^{-}\right|}{\widehat{\beta}} \leq \bar{\varepsilon}_{BB} \tag{2}$$

where $\widehat{\beta}^{+}$ and $\widehat{\beta}^{-}$ are the reliability indices respectively obtained using the limit-state functions $\mu_{\hat{g}}(\boldsymbol{x}) - 2\sigma_{\hat{g}}(\boldsymbol{x})$ and $\mu_{\hat{g}}(\boldsymbol{x}) + 2\sigma_{\hat{g}}(\boldsymbol{x})$, while $\widehat{\beta}$ is the reliability index obtained using the limit-state $\mu_{\hat{g}}(\boldsymbol{x})$.

The threshold $\bar{\varepsilon}_{BB}$ is set to 0.01 which is arguably a relatively large value. However, convergence is assumed only when this criterion is respected three times in a row, hence ensuring some degree of robustness.

Beta stability. This convergence criterion ensures the stability of the failure probability estimate assuming that convergence is achieved when adding new points do not noticeably modify the estimate. Using the reliability index, it reads:

$$\frac{\left|\widehat{\beta}^{(i)} - \widehat{\beta}^{(i-1)}\right|}{\widehat{\beta}^{(i)}} \leq \bar{\varepsilon}_{BS}, \tag{3}$$

where $\widehat{\beta}^{(i)}$ represents the estimated reliability index at the ith iteration.

The threshold is set to $\bar{\varepsilon}_{BS} = 0.005$ and convergence is considered only when this criterion is respected within three consecutive iterations.

Combined stopping criterion:. This stopping criterion is simply a combination of the previous two. Convergence is assumed when the two criteria in Eqs. (2) and (3) are met within two consecutive iterations.

4.2.1. Other common settings

Beside the method-specific settings introduced in the previous paragraph, others, which are common to all methods, need to be defined. These are mainly related to the initial experimental design which is drawn using the Latin hypercube sampling (LHS) method [32]. The number of initial ED points is set to $\max(10, 2M)$, where M is the problem dimensionality. This allows one to ensure a minimum of 10 points for low-dimensional problem while at the same time making sure that there are enough sample points w.r.t. the dimension when the latter increases. Similarly at the other end of the spectrum, the number of sample points is limited and only a maximum of $100+10M$ points can be added during the enrichment process. This number appears realistic for classical costly simulators used in engineering.

4.3. Criteria for the evaluation of the strategies

To properly compare different reliability analysis strategies, a performance measure needs to be defined. Due to the inherent complexity of the problem, we compare our benchmark results in terms of several different measures that take into account different performance metrics. Perhaps the most straightforward measures are i. how close the reliability estimate is to the reference and ii. how many points are needed to reach it. Focusing on the reliability index rather than the failure probability, a first criterion can be simply computed using the following relative reliability error estimator:

$$\epsilon^{(k)}_{\beta_{i,j}} = \left| \frac{\widehat{\beta}^{(k)}_{i,j} - \beta_{\text{ref},j}}{\beta_{\text{ref},j}} \right|, \tag{4}$$

where $\widehat{\beta}^{(k)}_{i,j}$ denotes the reliability index resulting from the kth replication of the ith strategy applied to the jth problem and $\beta_{\text{ref},j}$ is the reference solution for the jth problem. As a reminder, this benchmark comprises a set of 20 problems solved using 39 strategies, each repeated 15 times.

The criterion presented in Eq. (4) measures the accuracy of the resulting reliability index estimate. Additionally, we need to also assess the efficiency of the method, which simply relates to the number of model evaluations N_{eval} necessary to converge. The lower N_{eval}, the better the strategy. However N_{eval} alone is not a sufficient measure of the strategy efficiency, as premature convergence may occur. To avoid this, we will consider only solutions whose relative error, as computed in Eq. (4), is below a threshold arbitrarily set at 0.05 when ranking w.r.t. N_{eval}. Those with larger error will be automatically ranked in the last position, regardless of the number of model evaluations needed to converge.

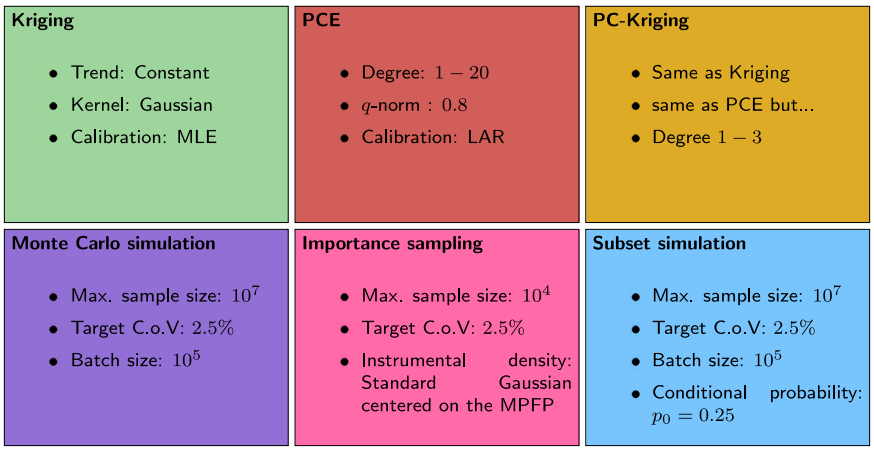


Fig. 4. A summary of the most important settings for the surrogate models and reliability estimation algorithms considered in this paper. The meaning of each of these parameters can be found in the supplementary materials (Appendix B).

Ideally, both criteria should be as low as possible, but they are by construction conflicting. Finding the best approach would therefore mean finding a good trade-off between relative error and computational cost. We therefore propose here a third criterion that combines these two criteria in one, making the ranking easier:

$$\Delta_{i,j}^{(k)} = \varepsilon_{\beta_{i,j}}^{(k)} \frac{N_{\text{eval},i,j}^{(k)}}{N_{\text{med},j}}, \tag{5}$$

where $N_{\text{med},j}$ is the median number of model evaluations considering all strategies, repetitions included, to solve the jth problem (in total, there are $15 \times 39 = 585$ runs for each problem).

All the 39 strategies are compared with each other and a ranking is established based on the three criteria defined above. The main goal of such a ranking is to find out if one strategy is consistently better than the others. If no such strategy were to be found, next is to find whether there are methods within each module that are consistently better than the others. Finally, the ranking will be used to assess whether given methods are better when applied to a specific feature of the problem at hand, *i.e.*, dimensionality and failure probability magnitude.

4.4. Methods ranking over different problems

4.4.1. Ranking of the strategies

At first, we compare different strategies considering all the criteria previously defined, to which we add as reference two plain simulation methods, *i.e.*, importance sampling and subset simulation (the total number of reliability analysis runs becomes then 12,300). They provide us with reference results without surrogates and serve as a benchmark baseline. For importance sampling, the MPFP is found using FORM and the MCS sample set is of size 10^3. For subset simulation, the subset sample set is of size 10^3 while p_0 is set to 0.1.

Additionally to the ranking, the robustness of each strategy is assessed. For each problem and replication, we observe whether a given strategy is within a certain distance from the best solution w.r.t. a chosen criterion. For the criterion "number of model evaluations", the distance is set to $\{2, 3, 5\} \times N_{\text{eval}}^*$ where N_{eval}^* is the smallest number of model evaluations among strategies whose relative error is below the threshold of 0.05. For the "relative error" (resp. Δ-criterion), the distance is measured as $\{5, 10, 20\} \times \varepsilon_\beta^*$ (resp. $\{5, 10, 20\} \times \Delta^*$) where ε_β^* (resp. Δ^*) is the smallest relative error (resp. Δ value) among all strategies. This count is aggregated over all problems and replications (in total, $20 \times 15 = 300$ analyses) and given in terms of percentage as illustrated by the bars in Figs. 5, 6 and 7. The mid-distance (*i.e.*, $3\, N_{\text{eval}}^*$, $10\, \varepsilon_\beta^*$ and $10\, \Delta^*$) is used to rank the methods in these figures (the best solutions are in the upper positions).

In general, the larger the bars, the more robust and accurate the associated method is. For instance looking at the first line of Fig. 5, the second bar shows that for the combination PCK + SuS + EFF + β-stability criterion and considering all the problems, the number of model evaluations required to converge in 69% of the repetitions is below 2 times the smallest number of model evaluations achieved for each given problem. The largest bar shows that this ratio increases to 88% when considering a threshold within 5 times the best achieved number of model evaluations. Finally, in each of these figures, the smallest and darkest bar represents the percentage of times a strategy was ranked first for a given experimental design.

Ranking with respect to the number of model evaluations. The first criterion we consider is the number of model evaluations, whose results are shown in Fig. 5. As expected, when considering only the number of model evaluations, the direct solutions (*i.e.*, subset simulation and importance sampling without the use of surrogates) rank last. The most robust and efficient solution w.r.t. the number of model evaluations is the combination of PCK with subset simulation, EFF and β-stability stopping criterion. Overall, considering the 10 best solutions, PCE or PC-Kriging as surrogates, subset simulation as reliability estimation algorithm and β-stability as stopping criterion seem to dominate. Regarding the learning function, there is no clear top performer as they all appear at least twice in the first ten positions.

Ranking with respect to the relative error. The next criterion we consider is the relative error as illustrated in Fig. 6. Here the direct solutions (without surrogates) are better ranked than with the previous criterion but they still rank worse than more than half of the methods considered. The better performance of surrogate-based methods is due to the "overkill" setup of the reliability solvers used in conjunction with the surrogates. As explained in Section 3.1.2, the computational efficiency of the surrogates allows one to use reliability solver configurations that maximize accuracy and minimize the stochastic uncertainty in the reliability estimator, without the traditional trade-offs associated. This shows that the use of surrogate models as an instrument for the exploration of the random input space can lead to results at least as equally accurate as a direct solution, *i.e.*, without the use of surrogates.

Regarding the best strategy, we can observe a few differences with the previous ranking. The overall most robust and efficient strategy is the combination of PC-Kriging with subset simulation, deviation number U and the combined stopping criterion. PCE is not so well classified when focus is put solely on accuracy. PC-Kriging is dominating the top ranking with a few occurrences of Kriging. As far as reliability estimation algorithm and learning function are concerned, subset simulation and the deviation number U are preferred. Finally, regarding the stopping criterion, β-stability (Eq. (3)) seems not to favor accuracy, in sharp contrast to the combined criterion, which now appears in the top performing combinations.

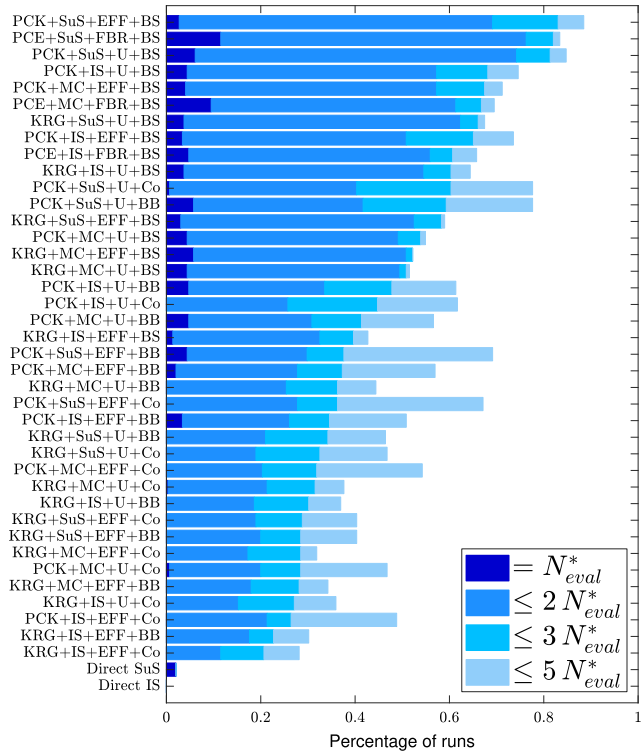


Fig. 5. Ranking of the strategies w.r.t. N_{eval}. The overall ranking is based on the number of times the method performs within $3\,N^*_{\text{eval}}$.

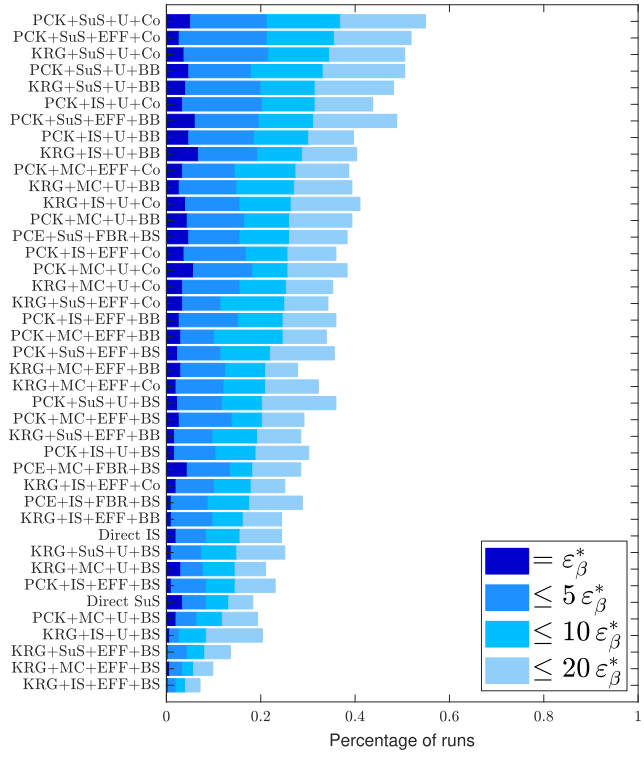


Fig. 6. Ranking of the strategies w.r.t. ε_β. The overall ranking is based on the number of times the method performs within $10\,\varepsilon^*_\beta$ (one order of magnitude).

Ranking with respect to the Δ-criterion. The last criterion considered is Δ (Fig. 7, Eq. (5)), which as expected results in a combination of

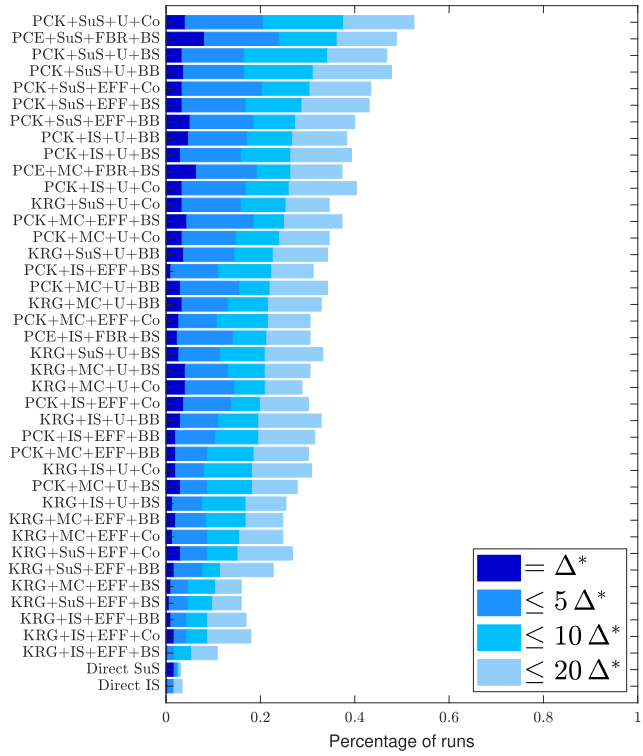


Fig. 7. Ranking of the strategies w.r.t. Δ. The overall ranking is based on the number of times the method performs within $10\,\Delta^*$ (one order of magnitude).

the two previous rankings. First the direct solutions are penalized by their relatively large number of model evaluations and rank again in the last two positions. Overall, this criterion favors solution accuracy, because the relative error in Eq. (4) can vary orders of magnitudes, while the range of variation of the number of model evaluations is not equally large (remember that the allowed number of model evaluations is limited to $100 + 10M$). Therefore the latter can only help make a difference within strategies that already lead to roughly the same accuracy.

As evidenced by Figs. 5, 6 and 7, there is not a single strategy that outperforms all others in every benchmark. However, some trends are emerging from these results. The next section dives deeper into the specific methods selected for each module.

4.4.2. Ranking of the methods within each module

In the previous section, we analyzed the strategies as a block, now we split them into their four components and perform the same statistical analysis. The strategies are once again ranked for each problem and replication and we count the number of occurrences of each method in a given ranking. The results are summarized in Figs. 8 and 9, where the percentage of times a given method is the best is shown in the last group of bars of each panel. To assess the variability in the ranking we also count the number of times a given method is within the first 5, 10 or 20 positions. Despite some minor variations in the share of each method in the top positions, the ranking remains unchanged if considering either the Δ-criterion or the relative error. In terms of surrogate models, PC-Kriging is the best performing choice, as it accounts for roughly half the occurrence in the best rankings. The reliability module is the most balanced, even though subset simulation shows a slight margin over the two others. In terms of learning function, the deviation number U outperforms both the expected feasibility function (EFF) and its PCE counterpart (FBR).

The convergence criterion is the only module, the results of which differ depending on the ranking criterion considered. The best method

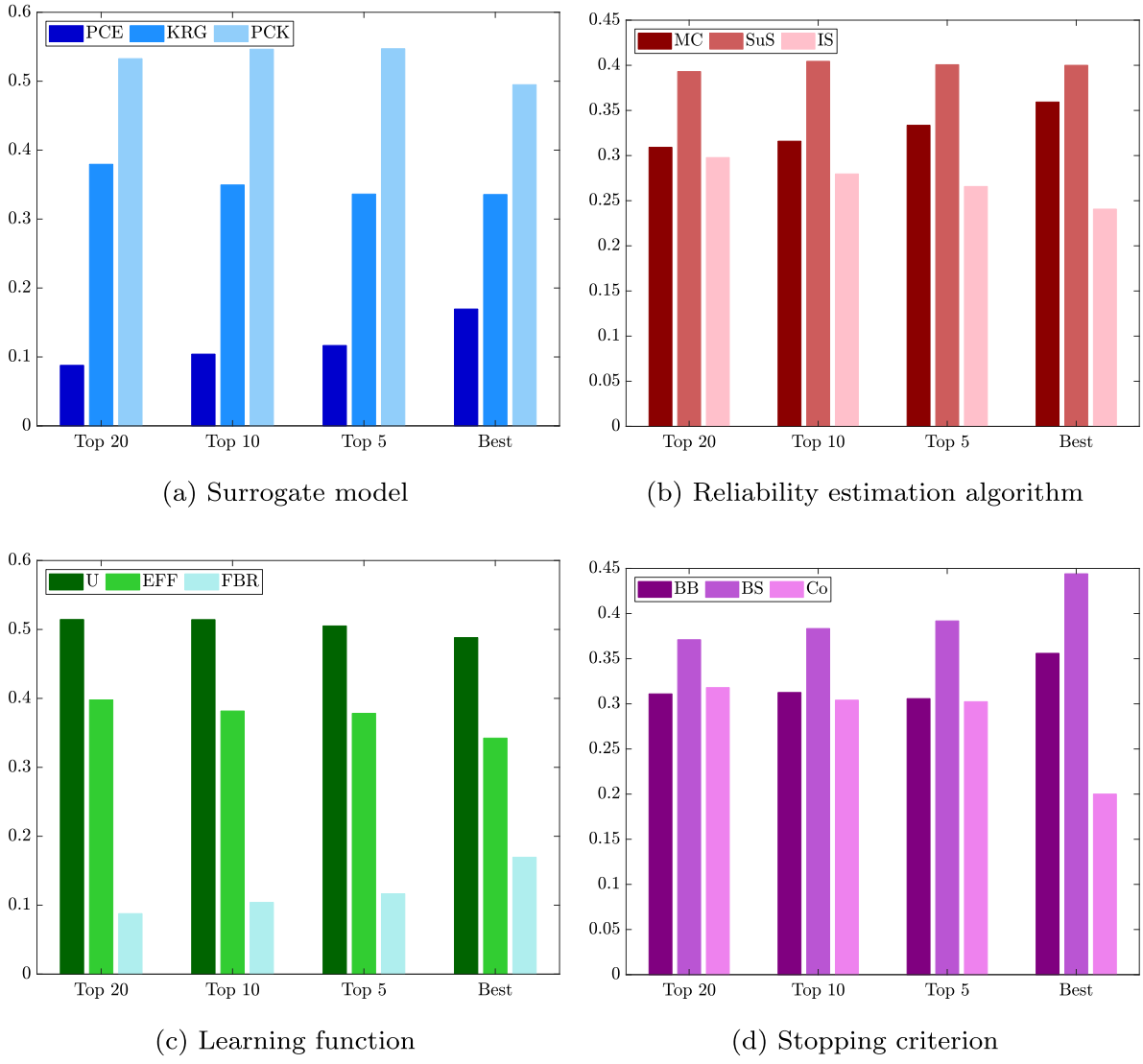


Fig. 8. Relative number of times a given method is among the top 20, top 10, top 5 or is the best. Ranking is made w.r.t. to the Δ-criterion.

seems to be β-stability when it comes to the Δ-criterion. However this turns to β-bounds when considering the relative error. The explanation is simply that β-stability converges faster than β-bounds. Hence, when accuracy is the prior concern, the second criterion is more suitable. However, when the computational budget is limited, β-stability is a more appropriate convergence criterion.

4.5. Aggregation of the results w.r.t. the problem features

Using the ranking from the previous two sections, it is clear now which methods lead on average to the best performance in terms of accuracy and efficiency. In this section, we go a step further and try to determine if these methods behave in the same way as a function of two selected features of the problem, namely input dimensionality and magnitude of the probability of failure, or if their performances are intrinsically linked to the type of problem at hand. We split the benchmark into low- ($M < 20$) and high-dimensional ($M \geq 20$) problems, as well as small ($\beta_{\text{ref}} < 3.5$) and high ($\beta_{\text{ref}} \geq 3.5$) reliability indices.

4.5.1. Performance with respect to dimensionality

Figs. 10 and 11 respectively show the Δ-criterion values and relative errors aggregated over all problems and then split for each method. The horizontal dotted black line represents the median over all problems. In each panel, the boxplots represent the aggregated median results (over all 15 replications) considering all (blue), low- to medium- (magenta) and high-dimensional problems (cyan). Starting with the surrogate models and looking at the Δ-criterion (Fig. 10a), we observe that PC-Kriging does not seem to be strongly affected by the problem dimensionality, as in all cases the conditional median remains slightly below the overall median, while PCE seems to improve its relative performance in higher dimensional problems. Kriging on the other hand performs noticeably poorly. The same trend is observed with the accuracy criterion ϵ_{β} (Fig. 11a).

Next are the reliability estimation algorithms (Fig. 10b) and as expected Monte Carlo simulation is essentially insensitive to the dimension. Subset simulation performs slightly worse in high-dimension but not as much as importance sampling. The latter becomes slightly worse when considering the purely accuracy-oriented criterion (Fig. 11b).

Regarding the learning functions (Figs. 10c and 11c), the deviation number (U) also does not seem to be noticeably affected by the dimension, contrary to the expected feasibility function (EFF) which gets considerably worse as the dimension increases. The fraction of bootstrap replicates (FBR) mirrors the behavior of PCE as there is a one-to-one mapping between the two. Their performance gets somehow better for high-dimensional problems when considering either of the criteria.

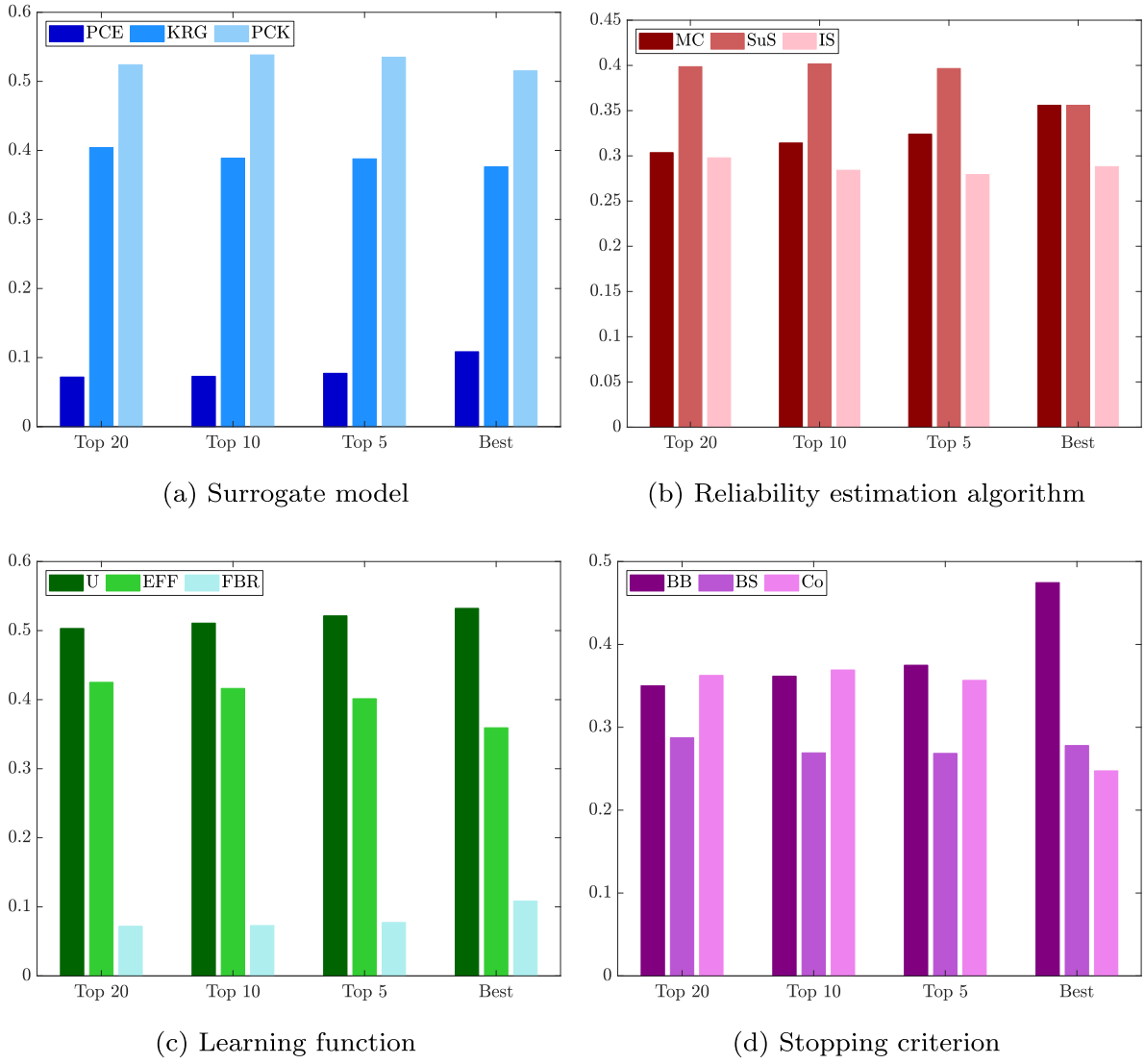


Fig. 9. Relative number of times a given method is among the top 20, top 10, top 5 or is the best. Ranking is made w.r.t. the relative error ϵ_β.

Finally, as already observed earlier, contradicting conclusions are obtained when considering either the Δ-criterion value (Fig. 10d) or the relative error (Fig. 11d) for the stopping criterion. In the former case, β-stability seems to be the best option when dimension increases whereas in the latter β-bounds, or even the combined criterion, seem to be the best option. This again can be explained by the fact that β-bounds is a stricter convergence criterion, especially in high-dimensional cases where the Kriging variance hardly shrinks as the experimental design is enriched.

4.5.2. Performance with respect to the magnitude of the failure probability

In contrast to the dimensionality cases where some methods were not particularly affected by the increase in dimensions, the level of the reliability index seems to affect pretty much all the methods. More specifically, the extremely low failure probability cases result on average in systematically poorer performances both in terms of relative error and Δ (See Figs. 12 and 13). As expected, Monte Carlo simulation performs considerably worse than its alternatives. The reason is that even in its "overkill" setting, the maximum number of model evaluations was set to 10^7, hence limiting the statistical estimator variance for extremely low failure probability problems.

4.6. Results with strategies aggregated over different problems

Now that we have a clear picture of the performances of each strategy as a whole and more particularly of different methods w.r.t. different problems and their features, let us have a look at the overall performance of all methods for each problem. The aggregated relative error and Δ values for all 20 problems (ordered in increasing dimensions) are shown in Fig. 14 as violin plots, *i.e.*, boxplots with an additional indication of the probability density of the data. The three problems with the overall worst median performances are highlighted in red. The first observation from Fig. 14 is that, regardless of the problem, there are still at least a few strategies that lead to good performances. As a matter of fact, the median values of both criteria are for most problems below the level arbitrarily set at 10^{-2}.

To end this benchmark, the methods which were ranked most often best are compared with the overall pool of strategies. This is shown in Fig. 15. More specifically, Figs. 15a and 15b show comparison on the Δ-criterion and relative error respectively, with respect to the combination of PC-Kriging with subset simulation, deviation number U and Combined criterion. This is the approach that was consistently best both in the strategies and methods ranking. In both cases, we can observe that this choice of methods improves the performance on most problems and also reduce the scatter in the results. This graph also

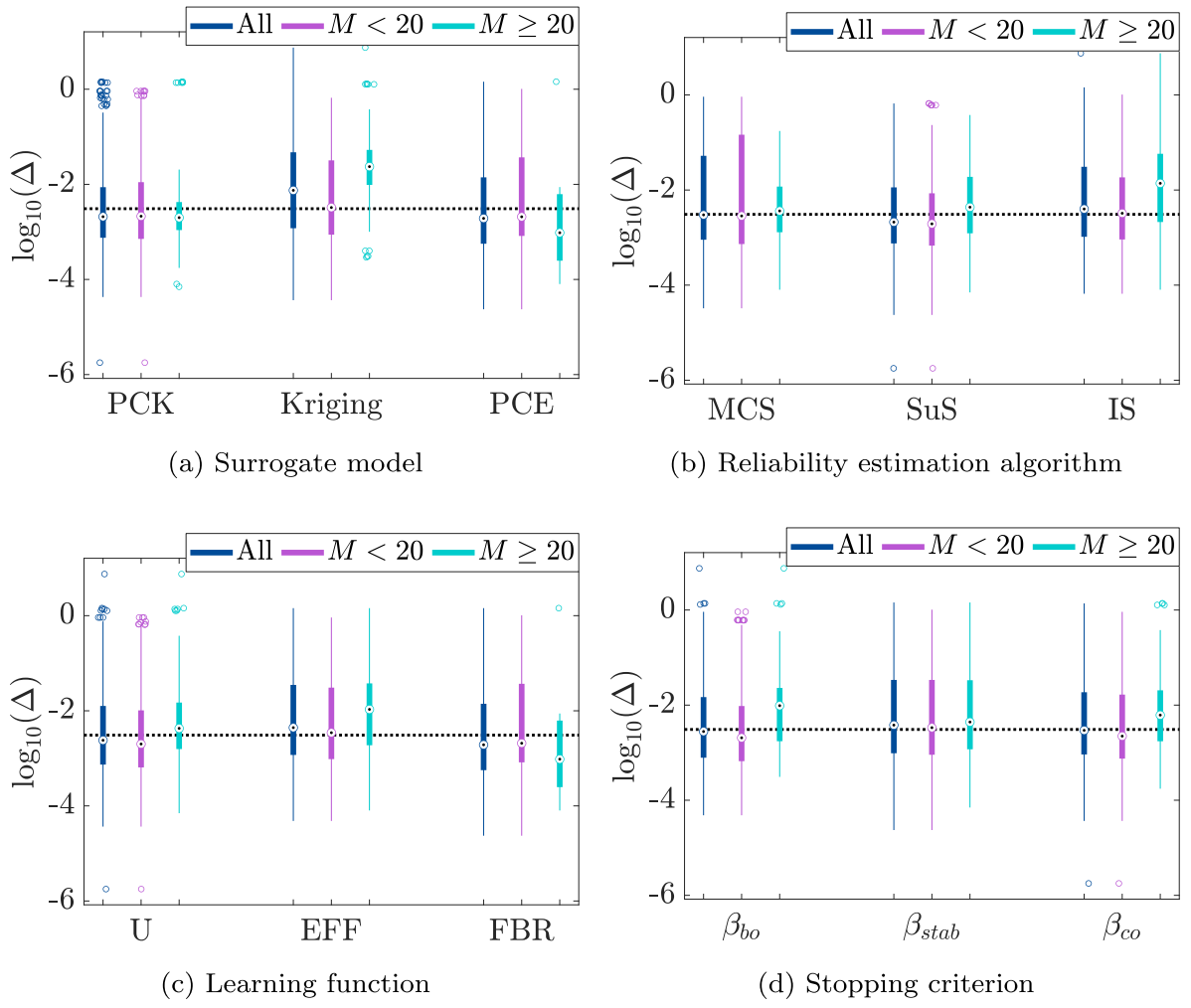


Fig. 10. Different methods compared w.r.t. the combined criterion Δ with problems split in two: low- ($M < 20$) and high-dimensional ($M \geq 20$). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

confirms the no-free-lunch principle exhibited by this benchmark as the overall best strategy is not necessarily the best for each problem.

4.7. Investigation of the most difficult problems

In this section, we closely look at the three problems that were most difficult to solve using the proposed ALR methods. By doing so, we note that some of these problems could not have been solved correctly even when considering the direct reliability estimation algorithms (without surrogates). Problem #11 contains four disjoint failure regions, which makes it impossible to solve using the standard importance sampling configuration considered in this benchmark. This also applies to the problem #8, which is spherical along the variables $\{X_2, X_3, \dots, X_{100}\}$. Furthermore, problem #11 cannot be solved with direct Monte Carlo simulation due to the low reference failure probability $p_f = 7.83 \cdot 10^{-7}$, which would require $N_{\text{MCS}} \approx 10^{9-10}$ samples to achieve sufficient accuracy.

These observations can be confirmed by Fig. 16 which shows the results corresponding to the solution of the benchmark problems using the three reliability estimation algorithms, with their overkill settings and without the aid of surrogate models. The red lines correspond to problems whose relative error is larger than 1. Using the threshold for acceptable accuracy at 10^{-2} as in the previous sections, we can see that Monte Carlo simulation fails to solve problems #10, #11, and #15 while importance sampling fail with problems #8, #11, #13, #18. This illustrates an important result of the benchmark, namely that the surrogate models were never the main cause of failure of the ALR strategies.

5. Research questions and recommendations

Our extensive benchmarking exercise allows us to showcase the framework introduced in this paper. We compared strategies built using this generalized active learning reliability framework with respect to various metrics. In this section, we summarize the findings from this benchmark and set up some guidelines as to the choice of the methods within the modules of the framework.

The first question that we set out to answer in this benchmark was whether there was one strategy that would consistently outperform the others with respect to all metrics and throughout all problems. The answer is clearly that there is no such strategy (no-free-lunch principle), yet using a surrogate model is always beneficial. For each analysis, the best strategy would vary according to the metric of interest and the type of problems. A natural follow-up question is whether any pattern could be uncovered by the benchmark. Without prior knowledge about the problem and using a non-intrusive combination of different methods, the conclusion is sharp: the best results are most likely obtained by combining PC-Kriging, subset simulation, deviation number U and β-bounds or combined stopping criterion.

This conclusion is obviously only limited to the methods selected for the benchmark. However, we may in some cases extrapolate to general guidelines by considering the characteristics of the different methods in

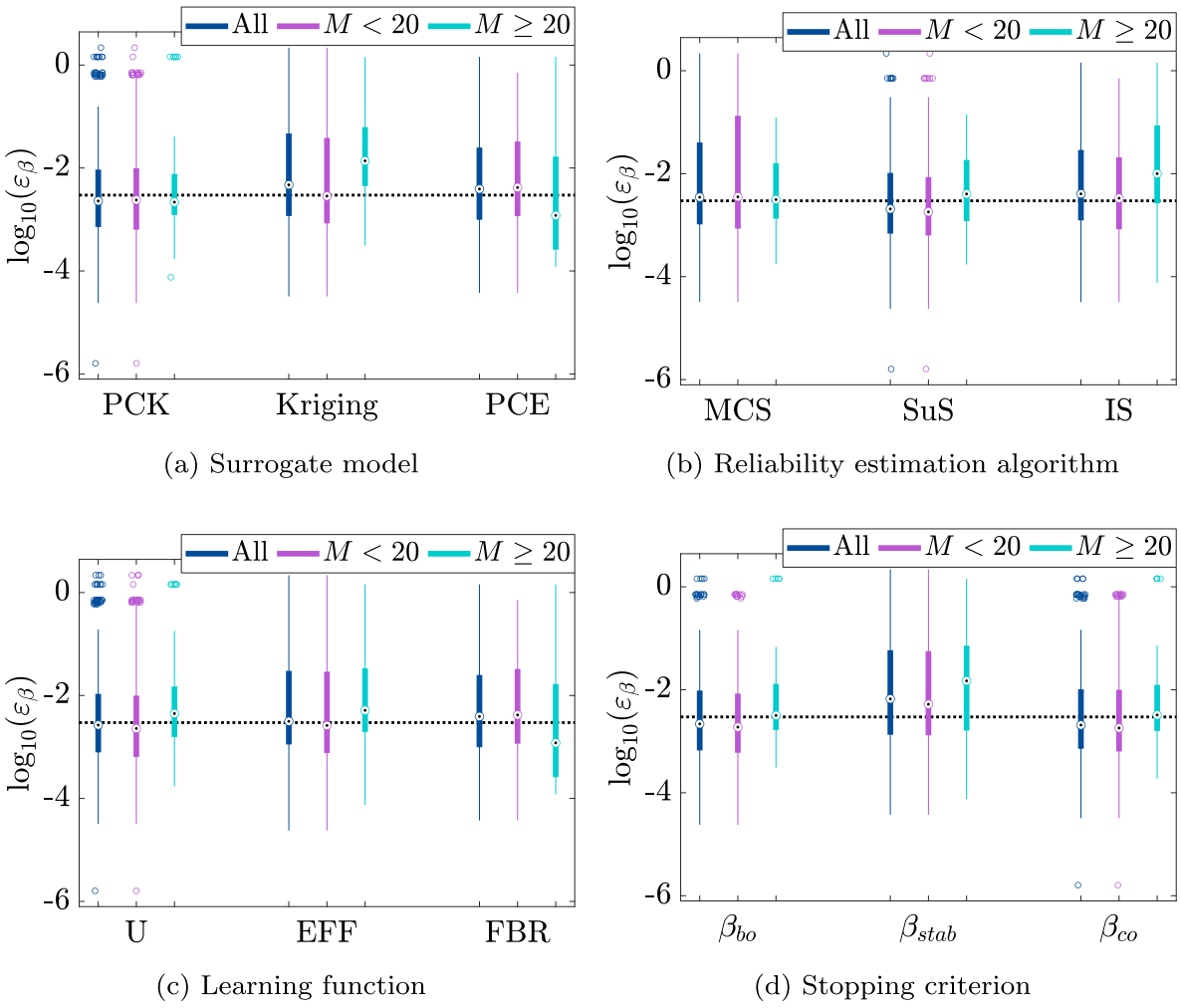


Fig. 11. Different methods compared w.r.t. the relative error ε_β with problems split in two: low- ($M < 20$) and high-dimensional ($M \geq 20$). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

relation to various features of the problems solved in the benchmark. We made the following observations regarding each module of the framework:

- **Surrogate model:** From this benchmark, it is clear that Kriging, which is the most used method in the literature, is not necessarily the best choice for active learning reliability. The main reason is that it performs poorly in fairly high-dimensional problems. PC-Kriging, which combines the global and local approximations capabilities of PCE and Kriging respectively, has shown consistently high performance throughout this benchmark. PCE has shown to perform better than Kriging for high-dimensional problems and this also benefits PC-Kriging to some extent. Finally, PC-Kriging possesses the same built-in error measure as Kriging, which makes it compatible with the various learning functions that take advantage of the Kriging variance.
- **Reliability estimation algorithm:** From this benchmark, it is clear that the active learning scheme inherits the pros and cons of the reliability estimation algorithm it uses, although this is somewhat mitigated by the ability to use "overkill" configurations. There is a direct correlation between the ability of a reliability estimation method to solve a given type of problems and the performance of the associated active learning scheme. Once again, Monte Carlo simulation, which is widely used in the literature, has proven not to be the best choice of algorithm according to the benchmark results. The introduction of surrogates does not eliminate the weakness of Monte Carlo simulation indeed when it comes to problems with extremely small failure probabilities. This is true also for the other algorithms, as importance sampling still showed limitations with problems where multiple failure regions exist. Note that other importance sampling densities may be used in practice when such a multiple-failure feature is known in advance (this was not considered in the benchmark). Regardless of the three algorithms used in the benchmark, it seems that the best course of action is to choose whatever algorithm the analyst thinks is best given his *a priori* knowledge on the problem. Finally, an important result highlighted by the benchmark is that over-calibrating the reliability estimation algorithm is beneficial and can lead to even better results than the non-surrogate equivalent using conventional settings. Therefore, surrogate models should be used to fully harness the benefits of the most sophisticated reliability estimation methods.
- **Learning function:** Considering this benchmark, the deviation number U seems to outperform EFF. The fraction of bootstrap replicates (FBR) is better than U when it comes to high-dimensional problems but loses its advantage when considering problems with small failure probabilities. All the methods selected for the benchmark are however very similar. It would have been interesting in this specific case to explore other type of learning functions, *e.g.*, those that also include the PDF of the input random variables.

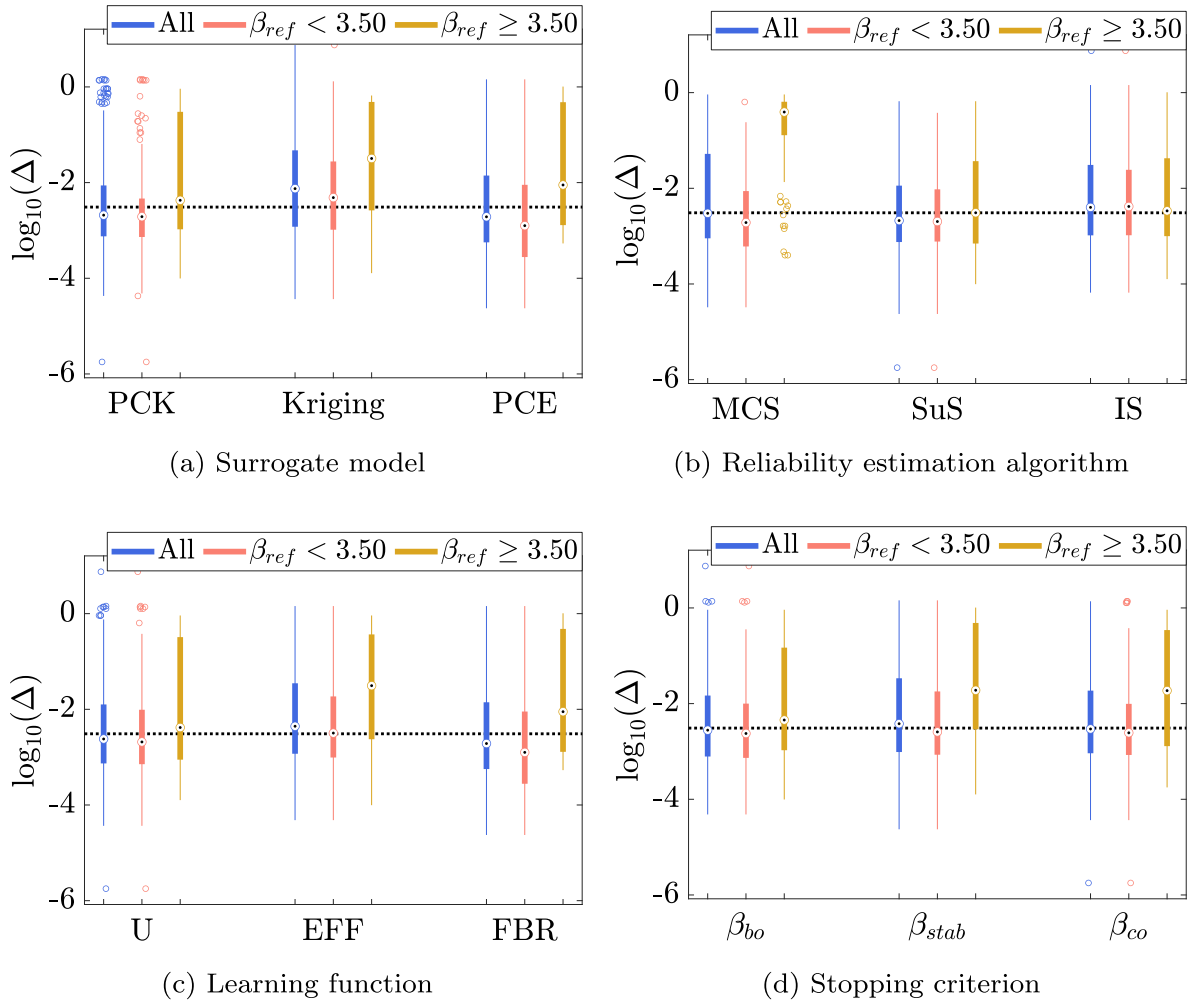


Fig. 12. Different methods compared w.r.t. the combined criterion Δ with problems split in two: small ($\beta_{\text{ref}} < 3.5$) and large ($\beta_{\text{ref}} \geq 3.5$) reliability indices.

- **Stopping criterion:** The stopping criteria selected for the benchmark are based on the accuracy of the P_f (or β) estimate rather than on the learning function. This finding is consistent with those identified in previous contributions [52]. This benchmark has however highlighted the difference between stopping criteria based on the local accuracy of the surrogate with those based on the stability of the limit-state surface. The former are shown to provide good results overall but to perform poorly in high-dimensional problems. This can be explained by the difficulty to sufficiently reduce the Kriging variance for high-dimensional problems. In contrast, stability criteria are somewhat more efficient for high-dimensional problems. However, they are prone to premature convergence. Combining these two criteria did not lead to a noticeable increase in performance.

To further summarize the observations made by analyzing the results of the benchmark, Table 2 gives a few recommendations on the basis of the benchmark.

6. Conclusions

This paper investigated the use of active learning strategies for the solution of structural reliability problems. We first conducted a literature survey and identified an underlying and recurring scheme. This scheme was used to propose a global framework for active learning reliability which is made of four components non-intrusively linked to each other. The four modules of this framework are surrogate model, reliability analysis, learning function and stopping criterion. We then showed that it is possible to combine various methods from each of the modules to build a wide array of viable solution strategies. On this basis, 39 solution strategies were built to solve a set of 20 selected problems. The results of this benchmark allowed us to identify patterns regarding the generalization capability of each method.

The first observation is that there is no strategy that consistently outperforms the others. We could however identify clear patterns to give recommendations on which types of methods should be preferred with regard to a given feature of the problem at hand. The flexibility of the presented framework is in this regard of great value as it allows the analyst to build tailored active learning reliability schemes.

The second observation is that there is essentially no drawbacks in using surrogate models. The latter allows one indeed to better exploit the reliability estimation algorithms through "overkill" settings.

Even though we ran an extensive benchmark a few aspects still need to be investigated. For instance, we did not explore the effect of the thresholds in the stopping criterion, nor did we explore more advanced learning functions. This includes techniques that weigh in the input PDF or allow for multiple points enrichment. By design, the scope of the analysis was limited to problems with single limit-state functions. More aspects need to be taken into account when considering multiple limit-state functions. Finally, extremely high-dimensional problems (M in the other of hundreds or even thousands) were not considered as they also would require special treatment. Finally the proposed optimal

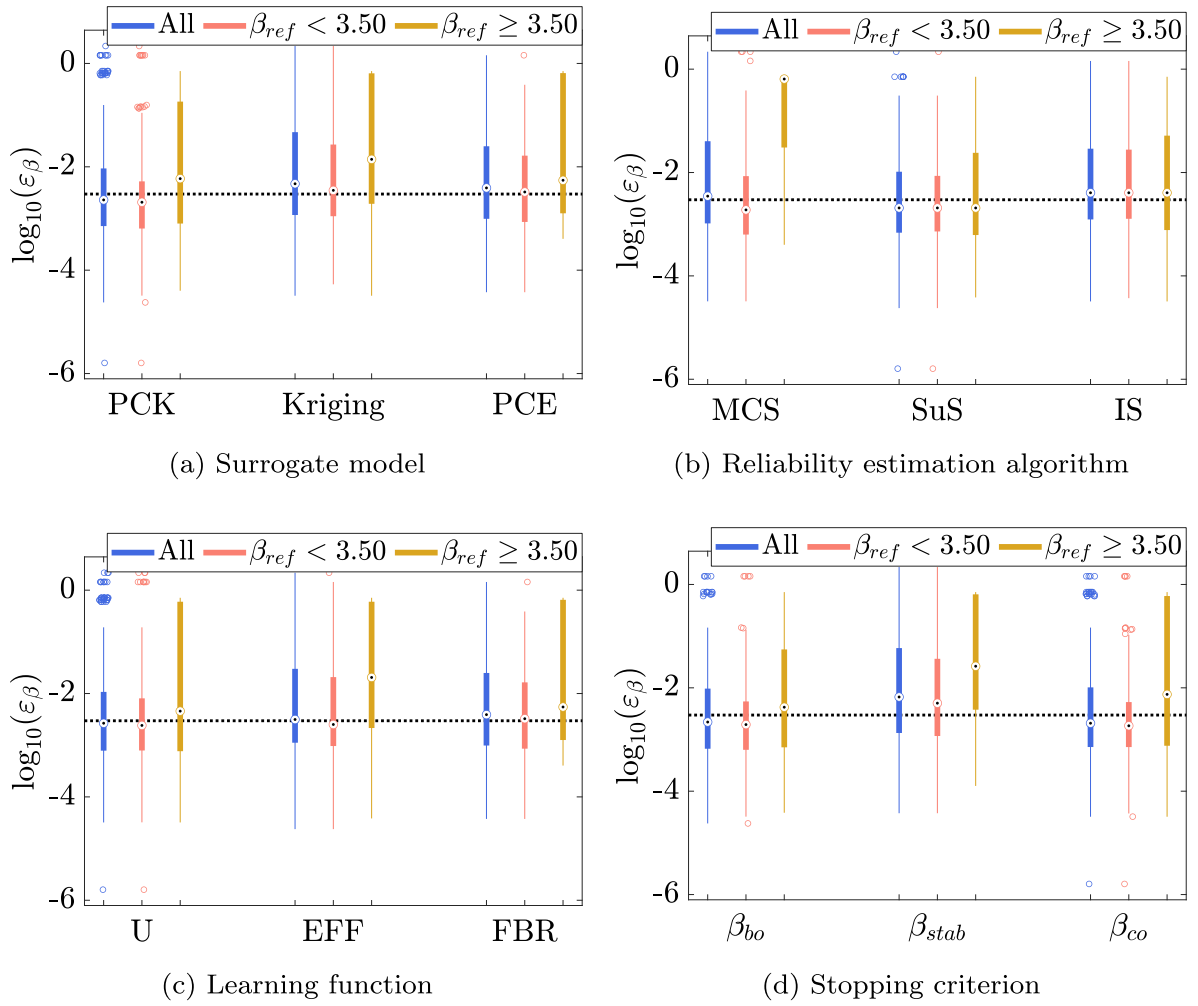


Fig. 13. Different methods compared w.r.t. the relative error ϵ_β with problems split in two: small ($\beta_{\text{ref}} < 3.5$) and large ($\beta_{\text{ref}} \geq 3.5$) reliability indices.

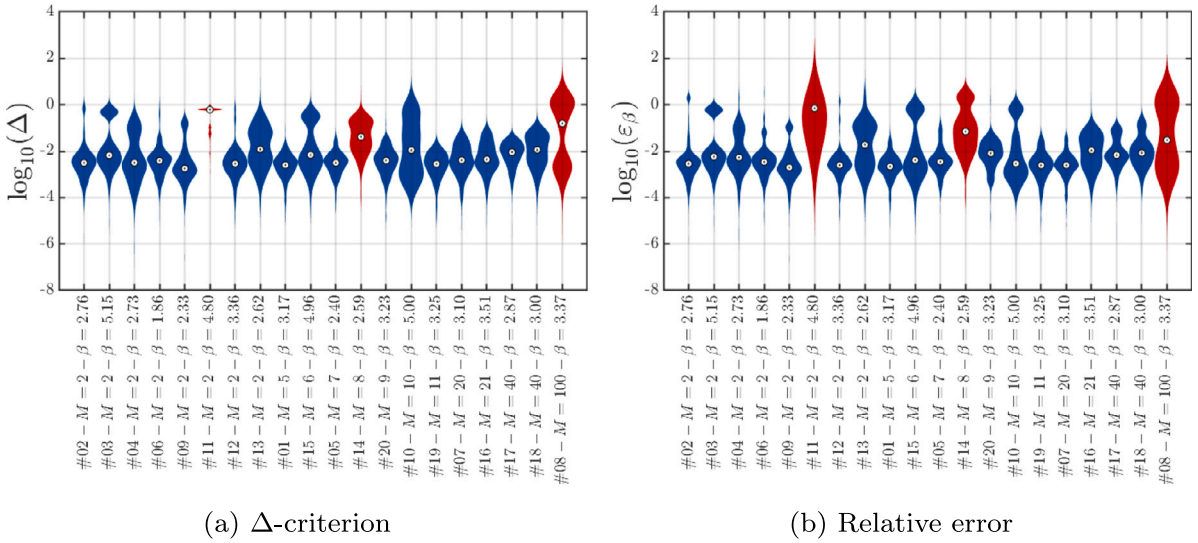
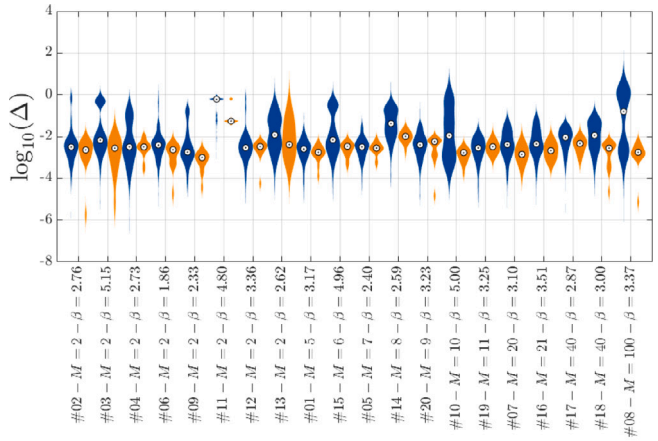


Fig. 14. Results aggregated for all solutions strategies and shown for each problem considering the combined criterion and the relative error. The three problems with the worst median results are highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

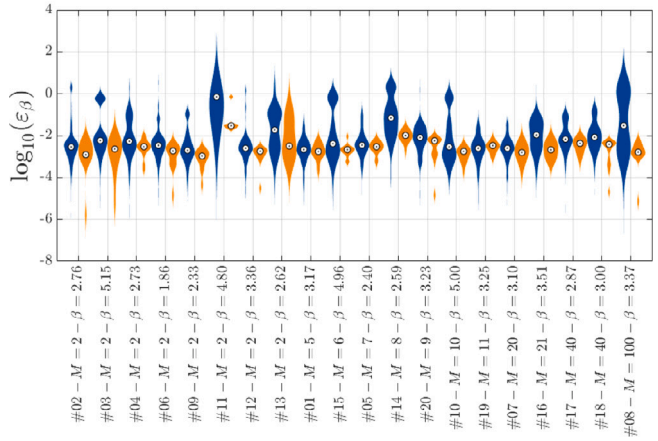
Table 2
General recommendations on the basis of the benchmark carried out in this paper. β_{bo} stands for β-bounds convergence, β_{stab} stands for β-stability convergence and β_{co} is the combined criterion.

Module	Dimensionality		Failure probability magnitude	
	$M < 20$	$20 \leq M \leq 100$	$\beta \leq 3.5$	$\beta \geq 3.5$
Surrogate model	PCK	PCE	PCE/PCK	PCK
Reliability estimation algorithm	SuS	SuS	SuS	SuS
Learning function	U	FBR	FBR/U	U
Stopping criterion	β_{bo}, β_{co}	β_{bo}, β_{co}[a]/β_{stab}[b]	β_{bo}, β_{co}	β_{bo}

[a]When considering accuracy only ϵ_β.
[b]When factoring in efficiency Δ.



(a) Δ with PCK+SuS+U+Co



(b) ε_β with PCK+SuS+U+Co

Fig. 15. Results aggregated for all solutions strategies (blue) compared to those corresponding to the overall best strategy (orange) w.r.t. to Δ and ϵ_β. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

framework has been applied blindly to a structural reliability context organized by TNO [88], whose results are available at https://rprepo.readthedocs.io/en/latest/results.html (last accessed on October 4th, 2021). Out of 16 component- and 11 system-reliability problems, our approach was the most efficient for 24 problems. A short summary of the results is provided in Appendix A for the sake of completeness.

It is worth mentioning that the codes and the set of examples used are made publicly available through the UQLab release R1.4, such that it will be easy to extend the results presented here to new problems.

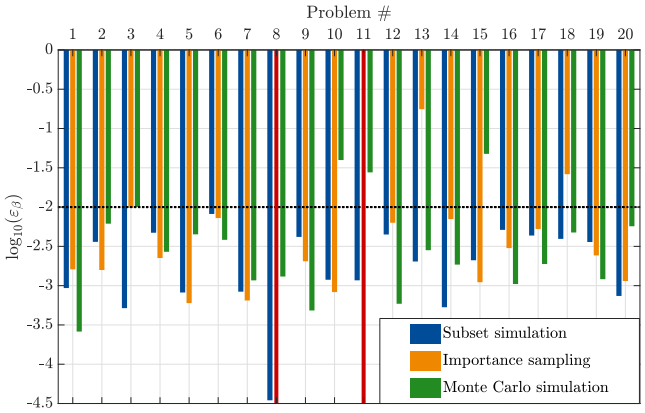


Fig. 16. Solution of the 20 problems using the three "overkill" reliability settings without surrogate models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Appendix A. TNO benchmark

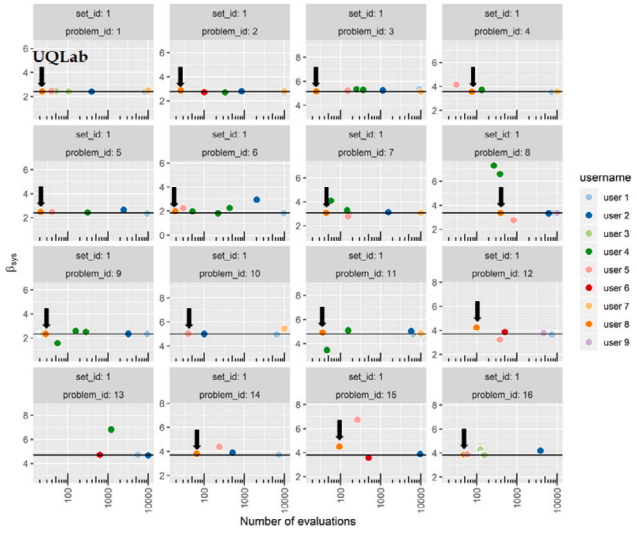
The TNO benchmark is a truly black-box benchmark of structural reliability analysis methods organized by TNO (Netherlands) in 2019 [88]. It is a two-part challenge which consists in a set of 16 component- and 11 system-reliability problems. It aims at assessing the efficiency and accuracy of various structural reliability methods. The limit-state functions were not known to the participants and were only accessible via an anonymous server API, *i.e.*, the participants could only submit a set of sample points and the server would return the corresponding model evaluations.

The methods highlighted by the benchmark in this paper, *i.e.*, a combination of PC-Kriging, subset simulation with overkill settings, deviation number U and combined stopping criterion, were used to participate in the challenge. The results were disclosed in terms of accuracy and efficiency.

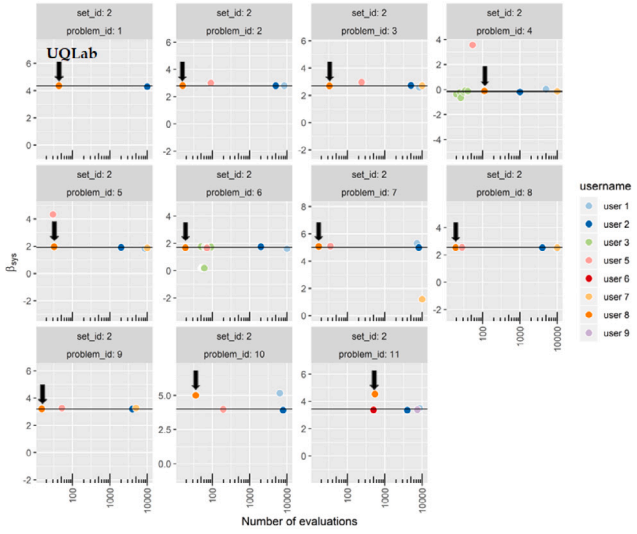
Fig. A.17 shows the results submitted anonymously by the nine research groups that took part in the challenge. The black arrows point to the results submitted using our approach. Our approach turned out to be both the most accurate and efficient in 24 out of 27 problems. This confirms the potential of such a flexible framework for the solution of a wide variety of structural reliability problems.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.strusafe.2021.102174.



(a) Part 1: Component-reliability problems



(b) Part 2: System-reliability problems

Fig. A.17. Results of the black-box reliability challenge as disclosed by Rozsas and Slobbe [88]. The results submitted using the approach highlighted in this paper are marked by the black arrow.

References

[1] Ditlevsen O, Madsen H. Structural reliability methods. Chichester: J. Wiley and Sons; 1996.

[2] Lemaire M. Structural reliability. Wiley; 2009.

[3] Melchers AT. Structural reliability analysis and prediction. John Wiley & Sons; 2018.

[4] Hasofer A-M, Lind N-C. Exact and invariant second moment code format. J Eng Mech 1974;100(1):111–21.

[5] Rackwitz R, Fiessler B. Structural reliability under combined load sequences. Comput Struct 1978;9:489–94.

[6] Melchers RE. Importance sampling in structural systems. Struct Saf 1989;6:3–10.

[7] Au SK, Beck JL. Estimation of small failure probabilities in high dimensions by subset simulation. Probl Eng Mech 2001;16(4):263–77.

[8] Ditlevsen O, Melchers RE, Gluver H. General multi-dimensional probability integration by directional simulation. Comput Struct 1990;36(2):355–68.

[9] Koutsourelakis PS, Pradlwarter HJ, Schuëller GI. Reliability of structures in high dimensions, part I: algorithms and applications. Probl Eng Mech 2004;19:409–17.

[10] Bucher C. Asymptotic sampling for high-dimensional reliability analysis. Probl Eng Mech 2009;24:504–10.

[11] Papaioannou I, Papadimitriou C, Straub D. Sequential importance sampling for structural reliability analysis. Struct Saf 2016;62:66–75.

[12] Wang Z, Broccardo M, Song J. Hamiltonian Monte Carlo methods for subset simulation in reliability analysis. Struct Saf 2019;76:51–67.

[13] Geyer S, Papaioannou I, Straub D. Cross entropy-based importance sampling using Gaussian densities revisited. Struct Saf 2019;76:15–27.

[14] Faravelli L. Response surface approach for reliability analysis. J Eng Mech 1989;115(12):2763–81.

[15] Lemaire M. Finite element and reliability : combined methods by response surfaces. In: Frantziskonis G, editor. Probamat-21st century, probabilities and materials : tests, models and applications for the 21st century. Kluwer Academic Publishers; 1998, p. 317–31.

[16] Bichon BJ, Eldred MS, Swiler L, Mahadevan S, McFarland J. Efficient global reliability analysis for nonlinear implicit performance functions. AIAA J 2008;46(10):2459–68.

[17] Echard B, Gayton N, Lemaire M. AK-MCS: an active learning reliability method combining kriging and Monte Carlo simulation. Struct Saf 2011;33(2):145–54.

[18] Lelièvre N, Beaurepaire P, Mattrand C, Gayton N. AK-MCSI: A kriging-based method to deal with small failure probabilities and time-consuming models. Struct Saf 2018;73:1–11.

[19] Echard B, Gayton N, Lemaire M, Relun N. A combined importance sampling and kriging reliability method for small failure probabilities with time-demanding numerical models. Reliab Eng Syst Saf 2013;111:232–40.

[20] Huang X, Chen J, Zhu H. Assessing small failure probabilities by AK–SS: An active learning method combining kriging and subset simulation. Struct Saf 2016;59:86–95.

[21] Hurtado JE. An examination of methods for approximating implicit limit state functions from the viewpoint of statistical learning theory. Struct Saf 2004;26:271–93.

[22] Deheeger F, Lemaire M. Support vector machines for efficient subset simulations: ^{2}SMART method. In *Proc. 10th int. conf. on applications of stat. and prob. in civil engineering (ICASP10)*, Tokyo, Japan, 2007.

[23] Bourinet J-M, Deheeger F, Lemaire M. Assessing small failure probabilities by combined subset simulation and support vector machines. Struct Saf 2011;33(6):343–53.

[24] Bourinet J-M. Reliability analysis and optimal design under uncertainty - Focus on adaptive surrogate-based approaches. Clermont-Ferrand, France: Université Blaise Pascal; 2018, p. 243, Habilitation à diriger des recherches.

[25] Marelli S, Sudret B. An active-learning algorithm that combines sparse polynomial chaos expansions and bootstrap for structural reliability analysis. Struct Saf 2018;75:67–74.

[26] Teixeira R, Nogal M, O'Connor A. Adaptive approaches in metamodel-based reliability analysis: A review. Struct Saf 2021;89:102019.

[27] Kroetz HM, Moustapha M, Beck AT, Sudret B. A two-level kriging-based approach with active learning for solving time-variant risk optimization problems. Reliab Eng Syst Saf 2020;203:107033.

[28] Zheng N, Xue J. Manifold learning. In: Statistical learning and pattern analysis for image and video processing. advances in pattern recognition. London: Springer; p. 87–119.

[29] Constantine PG, Dow E, Wang Q. Active subspace methods in theory and practice: Applications to k riging surfaces. SIAM J Sci Comput 2014;36(4):A1500–24.

[30] Zhou T, Peng Y. Structural reliability analysis via dimension reduction, adaptive sampling, and Monte Carlo simulation. Struct Multidiscip Optim 2020;62:2629–51.

[31] Lataniotis C, Marelli S, Sudret B. Extending classical surrogate modeling to high dimensions through supervised dimensionality reduction: a data-driven approach. Int. J. Uncertain Quantif 2020;10(1):55–82.

[32] McKay MD, Beckman RJ, Conover WJ. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 1979;2:239–45.

[33] Sobol' IM. Distribution of points in a cube and approximate evaluation of integrals. USSR Comput Math Math Phys 1967;7:86–112.

[34] Ranjan P, Bungham D, Michailidis G. Sequential experiment design for contour estimation from complex computer codes. Technometrics 2008;50:527–41.

[35] Jones DR, Schonlau M, Welch WJ. Efficient global optimization of expensive black-box functions. J Global Optim 1998;13(4):455–92.

[36] Bucher C, Bourgund U. A fast and efficient response surface approach for structural reliability problems. Struct Saf 1990;7:57–66.

[37] Rajashekhar M-R, Ellingwood B-R. A new look at the response surface approach for reliability analysis. Struct Saf 1993;12:205–20.

[38] Leonel ED, Beck AT, Venturini WS. On the performance of response surface and direct coupling approachesin solution of random crack propagation problems. Struct Saf 2011;33:261–74.

[39] Roussouly N, Petitjean F, Salaun M. A new adaptive response surface method for reliability analysis. Probl Eng Mech 2013;32:103–15.

[40] Li X, Gong C, Gu L, Gao W, Jing Z, Su H. A sequential surrogate method for reliability analysis based on radial basis function. Struct Saf 2018;73:42–53.

[41] Shi L, B. S, Ibrahim DS. An active learning reliability method with multiple kernel functions based on radial basis function. Struct Multidiscip Optim 2019;60:211–29.

[42] Cheng K, Lu Z. Active learning polynomial chaos expansion for reliability analysis by maximizing expected indicator function predictor error. Struct Saf 2020;82:1–13.

[43] Pan Q, Qu X, Liu L, Dias D. A sequential sparse polynomial chaos expansion using Bayesian regression for geotechnical reliability estimations. Int J Numer Anal Methods Geomech 2020;44:874–89.

[44] Basudhar A, Missoum S. An improved adaptive sampling scheme for the construction of explicit boundaries. Struct Multidiscip Optim 2008;42(4):517–29.

[45] Lacaze S, Missoum S. A generalized "max-min" sample for surrogate update. Struct Multidiscip Optim 2014;49:683–7.

[46] Pan Q, Dias D. An efficient reliability method combining adaptive support vector machines and Monte Carlo simulation. Struct Saf 2017;67:85–95.

[47] Bourinet J-M. Anisotropic-kernel-based support vector regression for reliability assessment. In *Proc. 12th international conference on structural safety and reliability (ICOSSAR), August 6-10, 2017*, Vienna, Austria, 2017.

[48] Chojazyk AA, Teixeira AP, Neves LC, Cardoso JB, Guedes Soares C. Review and application of artificial neural networks models in reliability analysis of steel structures. Struct Saf 2015;52:78–89.

[49] Kroetz HM, Tessari RK, Beck AT. Performance of global metamodeling techniques in solution of structural reliability problems. Adv Eng Softw 2017;114:394–404.

[50] Sundar V, Shields MD. Surrogate-enhanced stochastic search algorithms to identify implicitly defined functions for reliability analysis. Struct Saf 2016;62:1–11.

[51] Gomes WJS. Structural reliability analysis using artificial neural networks. ASCE-ASME J Risk Uncertain Eng Syst B: Mech Eng 2019;5:1–8.

[52] Schöbi R, Sudret B, Marelli S. Rare event estimation using Polynomial-Chaos-Kriging. ASCE-ASME J Risk Uncertain Eng Syst A: Civ Eng 2016;3(2). http://dx.doi.org/10.1061/AJRUA6.0000870, D4016002.

[53] Sadoughi MK, Li M, Hi C, Mackenzie CA. High-dimensional reliability analysis of engineered systems involving computationally expensive black-box simulations. In *Proc. asme 2017 international design engineering technical conferences and computers and information in engineering conference, August 6-9, 2017*, Cleveland, Ohia, USA, 2017.

[54] Li M, Wang Z. Deep learning for high-dimensional reliability analysis. Mech Syst Signal Process 2020;139:1–18.

[55] Wagner P-R, Marelli S, Papaioannou I, Straub D, Sudret B. Rare event estimation using stochastic spectral embedding. Struct Saf 2021.

[56] Gaspar B, Teixeira AP, Guedes Soares C. Adaptive surrogate model with active refinement combining kriging and a trust region method. Reliab Eng Syst Saf 2017;165:277–91.

[57] Zhao H, yue Z, Liu H, Gao Z, Zhang Y. An efficient reliability method combining adaptive importance sampling and kriging metamodel. Appl Math Model 2015;39:1853–66.

[58] Dubourg V, Sudret B, Bourinet J-M. Meta-model-based importance sampling for reliability sensitivity analysis. In *Proc. 11th asce specialty conference on probabilistic mechanics and structural reliability*, Notre Dame, USA, 2012.

[59] Cadini F, Santos F, Zio E. An improved adaptive kriging-based importance technique for sampling multiple failure regions of low probability. Reliab Eng Syst Saf 2014;139:109–17.

[60] Balesdent M, Morio J, Marzat J. Kriging-based adaptive importance sampling algorithms for rare event estimation. Struct Saf 2013;44:1–10.

[61] Razaaly N, Congedo PM. Novel algorithm using active metamodel learning and importance sampling: Application to multiple failure regions of low probability. J Comput Phys 2018;368:92–114.

[62] Yang X, Liu Y, Mi C, Wang X. Active learning kriging model combining with kernel-density estimation-based importance sampling method for the estimation of low failure probability. J Mech Des 2018;140:1–9.

[63] Zhang J, Taflanidis AA. Adaptive kriging stochastic sampling and density approximation and its application to rare-event estimation. ASCE-ASME J Risk Uncertain Eng Syst A: Civ Eng 2018;4:1–17.

[64] Liu F, Wei P, Zhou C, Yue Z. Reliability and reliability sensitivity analysis of structure by combining adaptive linked importance sampling and kriging reliability method. Chin J Aeronaut 2019;33:1218–27.

[65] Zhang X, Wang L, Sorensen JD. AKOIS: An adaptive kriging oriented importance sampling method for structural system reliability analysis. Struct Saf 2020;82:1–13.

[66] Au SK, Beck JL. Subset simulation and its application to seismic risk based on dynamic analysis. J Eng Mech 2003;129(8):901–17.

[67] Zhang J, Xiao M, Gao L. An active learning reliability method combining kriging constructed with exploration and exploitation of failure region and subset simulation. Reliab Eng Syst Saf 2019;188:90–102.

[68] Ling C, Lu Z, Feng K, Zhang X. A coupled subset simulation and active learning kriging reliability analysis method for rare failure events. Struct Multidiscip Optim 2019;60:2325–41.

[69] Li L, Bect J, Vazquez E. Bayesian subset simulation: a kriging-based subset simulation algorithm for the estimation of small failure probabilities. In: 11th international probabilistic assessment and management conference (psam11) and the annual european safety and reliability conference (esrel 2012), Helsinki : Finland (2012). Curran; 2012.

[70] Bect J, Li L, Vazquez E. BayesIan subset simulation. SIAM/ASA J Uncertain Quant 2017;5:762–86.

[71] Tong C, Sun Z, Zhao Q, Wang Q, Wang S. A hybrid algorithm for reliability analysis combining kriging and subset simulation importance sampling. J Mech Sci Tech 2015;29:3183–93.

[72] Guo Q, Liu Y, Chen B, Zhao Y. An active learning kriging model combined with directional importance sampling method for efficient reliability analysis. Probl Eng Mech 2020;60:1–9.

[73] Bo X, HuiFeng T. A robust and efficient structural reliability method combining radial-based importance sampling and kriging. Sci China Technol Sci 2018;61:724–34.

[74] Lv Z, Lu Z, Wang P. A new learning function for kriging and its applications to solve reliability problems in engineering. Comput Math Appl 2015;70:1182–97.

[75] Peijuan Z, Ming W, Zhouhong Z, Liqi W. A new active learning method based on the learning function u of the AK-MCS reliability analysis method. Eng Struct 2017;148:185–94.

[76] Wen Z, Pei H, Liu H, Yue Z. A sequential kriging reliability analysis method with characteristics of adaptive sampling regions and parallelizability. Reliab Eng Syst Saf 2016;153:170–9.

[77] Sun Z, Wang J, Li R, Tong C. LIF: A New kriging based learning function and its application to structural reliability analysis. Reliab Eng Syst Saf 2017;157:152–65.

[78] Tong C, Wang J, J. L. A kriging-based active learning algorithm for mechanical reliability analysis with time-consuming and nonlinear response. Math Probl Eng 2019;2019:1–14.

[79] Hu Z, Mahadevan S. Global sensitivity analysis-enhanced surrogate (GSAS) modeling for reliability analysis. Struct Multidiscip Optim 2016;53:501–21.

[80] Xiao N-C, Zuo MJ, Guo W. Efficient reliability analysis based on adaptive sequential sampling design and cross-validation. Appl Math Model 2018;58:404–20.

[81] Jiang C, Qiu H, Yang Z, Chen L, Gao L, Li P. A general failure-pursuing sampling framework for surrogate-based reliability analysis. Reliab Eng Struct Saf 2019;183:47–59.

[82] Zhang X, Wang L, Sørensen JD. REIF: A Novel active-learning function toward adaptive kriging surrogate models for structural reliability analysis. Reliab Eng Syst Saf 2019;185:440–54.

[83] Jian W, Zhili S, Qiang Y, Rui L. Two accuracy measures of the Kriging model for structural reliability analysis. Reliab Eng Syst Saf 2017;167:494–505.

[84] Moustapha M, Sudret B, Bourinet J-M, Guillaume B. Quantile-based optimization under uncertainties using adaptive kriging surrogate models. Struct Multidiscip Optim 2016;54(6):1403–21.

[85] Fauriat W, Gayton N. AK-SYS: AN application of the AK-MCS method for system reliability. Reliab Eng Struct Saf 2017;123:137–44.

[86] Dubourg V, Sudret B, Deheeger F. Metamodel-based importance sampling for structural reliability analysis. Probl Eng Mech 2013;33:47–57.

[87] Marelli S, Sudret B. UQLab: A framework for uncertainty quantification in Matlab. In *Vulnerability, uncertainty, and risk (proc. 2nd int. conf. on vulnerability, risk analysis and management (icvram2014)*, Liverpool, United Kingdom, 2014, p. 2554–63.

[88] Rozsas A, Slobbe A. Repository and black-box reliability challenge 2019. 2019, https://gitlab.com/rozsasarpi/rprepo/ (Accessed: 2021-05-04).