Computational methods for the robust optimization of the design of a dynamic aerospace system in the presence of aleatory and epistemic uncertainties

Computational Methods for the Robust Optimization of the Design of a Dynamic Aerospace System in the Presence of Aleatory and Epistemic Uncertainties

Nicola Pedroni

[1] Energy Department, Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino, 10129, Italy

Email: nicola.pedroni@polito.it

Abstract

In this paper, we consider the computational model of a dynamic aerospace system and address the issues posed by the NASA Langley Uncertainty Quantification Challenge on Optimization Under Uncertainty, which comprises six tasks. Subproblem A deals with the model calibration and (aleatory and epistemic) uncertainty quantification of a subsystem by means of a limited number of observations. A simple, two-step approach based on Maximum Likelihood Estimation (MLE) is proposed to address this task. Subproblem B requires the identification and ranking of those (epistemic) parameters that are more effective in improving the predictive ability of the computational model of the subsystem. Two approaches are compared: the first is based on a sensitivity analysis within a factor prioritization setting, whereas the second employs the Energy Score (ES) as a multivariate generalization of the Continuous Rank Predictive Score (CRPS). Since the output of the subsystem is a function of time, both subproblems are addressed in the space defined by the orthonormal bases resulting from a Singular Value Decomposition (SVD) of the subsystem observations. Subproblem C requires identifying the (epistemic) reliability (resp., failure probability) bounds of a given system design. The issue is addressed by an efficient combination of: (i) Monte Carlo Simulation (MCS) to propagate the aleatory uncertainty described by probability distributions; (ii) Genetic Algorithms (GAs) to solve the optimization problems related to the propagation of epistemic uncertainty by interval analysis; and (iii) fast-running Artificial Neural Networks (ANNs) to reduce the computational time related to the repeated model evaluations. In Subproblem D, system reliability is improved by identifying a new design point within an iterative robust optimization framework. In Subproblem E both the uncertainty model and the design obtained are tuned using additional data. Finally, a risk-based design is carried out in Subproblem F by neglecting "outliers" (i.e., less likely values of some epistemic parameters) in the design optimization.

Keywords: Singular Value Decomposition, Model Calibration, Uncertainty Propagation, Sensitivity Analysis, Energy Score, Design Optimization.

1 Introduction

The quantitative analyses of the phenomena occurring in safety-critical (e.g., civil, nuclear, aerospace and chemical) engineering systems are based on mathematical models [1]. In practice, not all the characteristics of the system under analysis can be captured in the model: thus, uncertainty is present in the values of the input parameters and in the model hypotheses and structure. This is due to: (i) the intrinsically random nature of several of the phenomena occurring during system operation (aleatory uncertainty); (ii) the incomplete knowledge about some phenomena and operating conditions, often due

to the scarcity of quantitative data available, which may be either very sparse or prohibitively expensive to collect (epistemic uncertainty). Such uncertainty propagates within the model and causes uncertainty in its outputs [2]. The characterization and quantification of this output uncertainty is of paramount importance for: (i) making robust decisions in safety-critical systems applications; (ii) optimally designing and operating such systems; and (iii) driving resource allocation for uncertainty reduction by the identification of the model parameters and hypotheses that contribute the most to the output uncertainty [3].

Within this framework, we tackle the issues raised by the NASA Langley Uncertainty Quantification (UQ) Challenge on Optimization Under Uncertainty [4]. In Task (A), the mathematical (black-box) model of a (sub)system is considered, which includes nine inputs and one time-dependent output. The inputs are uncertain and divided into five purely aleatory variables (described by a possibly joint probability distribution) and four purely epistemic parameters (described by intervals). The Challengers provide a limited number (i.e., 100) of observations of the physical (sub)system (notice such observations are given in the form of discrete time histories). On this basis, an Uncertainty Model (UM) for the (nine) input variables/parameters to the subsystem model should be created. A straightforward, two-step (parametric) approach based on *Maximum Likelihood Estimation* (MLE) is here employed to address this task. In the first step, a multivariate Gaussian Mixture (GM) [5] is chosen as the functional form of the joint probability distribution of the aleatory variables and the corresponding parameters are calibrated by MLE. In the second step, the UM for the pure epistemic parameters is defined as the smallest hyper-rectangular set enveloping their joint four-dimensional $\alpha \cdot 100\%$ Confidence Interval (CI) (in this paper, $\alpha = 0.99$). Since the output of the subsystem is a function of time, the approach is applied in the space defined by the orthonormal bases resulting from a *Singular Value Decomposition* (SVD) of the subsystem observations: in other words, a multivariate dynamic problem in the real domain is translated into a multivariate static problem in the SVD space. In addition, the likelihood of the data is evaluated by Kernel Density Estimation (KDE) techniques in the SVD space [6-8].

The Task of uncertainty reduction (B) is tackled in two ways. In the first (namely, *sensitivity analysis* within a 'factor prioritization' setting), we rank the epistemic input parameters according to degree of *reduction* in the output epistemic uncertainty, which one could hope to obtain by refining their (epistemic) uncertainty models, i.e., by reducing the epistemic uncertainty range. A sensitivity index is adopted in analogy with variance-based Sobol indices [9, 10]: in this view, the most important epistemic parameters in the ranking are those that give rise to the *highest expected reduction* in the amount of epistemic uncertainty contained in the model output, when the corresponding parameter values are considered *constant* (i.e., when the amount of their epistemic uncertainty is reduced to zero). Notice that the 'amount' of epistemic uncertainty is here defined as the *volume* of the *convex hull* enveloping the realizations of the model output in the orthonormal SVD space [11]. In the second approach, the use of the *Energy Score* (ES) (computed in the orthonormal space) is proposed as a multivariate generalization of the Continuous Rank Predictive Score (CRPS) to assess the probabilistic predictive capability of the subsystem model [12]. The idea is to rank the epistemic parameters according to their capability to *improve* the *predictive ability* of the model, i.e., to *decrease* the ES, when their epistemic uncertainty is *reduced*.

The Task (C) of reliability analysis is tackled by solving the (optimization) problem of identifying the epistemic reliability (resp., failure probability) bounds for a given system design point. The solution of the corresponding nonlinear, constrained optimization problems is efficiently tackled by resorting to *heuristic* approaches (i.e., *Evolutionary Algorithms-EAs*): such methods deeply explore the search space by evaluating a large number (i.e., a population) of candidate solutions in order to find a near-optimal solution [13]. Notice that the population-based nature of such evolutionary algorithms allows an efficient exploration and characterization of abrupt and disconnected search spaces, which is the case of the present challenge. During the optimization search, the aleatory uncertainty described by probability distributions is propagated by *Monte Carlo Simulation* (MCS). Also, the original (black-box) mathematical model of the

system is replaced by a fast-running, surrogate regression model based on Artificial Neural Networks (ANNs), in order to reduce the computational cost associated to the analysis [14].

In Task (D) the system's reliability has to be improved by identifying a new design point. The problem is here addressed within a *robust design* framework, where the objective is to minimize the (epistemic) upper bound of the system failure probability. An *iterative* optimization algorithm (combining MCS, EAs and ANNs) is implemented to efficiently deal with the *unbounded* nature of the design variables (which can range over the entire real axis) [15].

In Subproblem (E), the UM calibrated in Task (A) and the design point found in Task (D) have to be updated and tuned by means of (100) additional data/observations coming from the integrated system under analysis. The model of the integrated system includes nine inputs and two time-dependent outputs. The two-step *parametric* MLE-based approach of Task A and the iterative algorithm of Task D are employed.

Finally, in Subproblem F the design of the system has to be improved by accepting a small *risk* (namely, risk-based design). The task is here addressed by neglecting some "outliers" in the design optimization process [16]: in particular, portions of the epistemic parameter space with comparatively *small likelihood* are ignored so as to *maximize* a properly defined *gain* in the system performance. In this paper, we seek to maximize the *relative* decrease in the (epistemic) upper bound of the system failure probability.

The remainder of the paper is organized as follows. In Section 2, the main characteristics of the mathematical system models under analysis are outlined; in Section 3, the NASA Challenge is addressed: the approaches adopted to tackle the problems are described in detail and the results obtained are reported; finally, conclusions are drawn in the last Section.

2 The System

The system of interest is modelled as a set of interconnected subsystems. The uncertain parameter δ is concentrated onto a single subsystem. This subsystem is modelled by the function $y(\boldsymbol{a}, \boldsymbol{e}, t)$, where $\boldsymbol{a}$ is a n_a-dimensional vector of aleatory variables ($n_a = 4$), $\boldsymbol{e}$ is a n_e-dimensional vector of epistemic parameters ($n_e = 4$) and t is time. The Uncertainty Model (UM) for $\boldsymbol{a}$ is denoted as $f_{\boldsymbol{a}}$, where $f_{\boldsymbol{a}}$ is a joint density supported in the set $A_0 = [0, 2]^{na}$. In contrast, the UM for $\boldsymbol{e}$ is denoted as E, where E is a hyper-rectangular set included in $E_0 = [0, 2]^{ne}$. Hence, the UM of is fully prescribed by the pair $\langle f_{\boldsymbol{a}}, E \rangle$. The integrated system is instead modeled by $\boldsymbol{z}(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}, t) = \{z_1(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}, t), z_2(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}, t)\}$, where $\boldsymbol{\theta}$ is an n_ϑ-dimensional vector of design variables ($n_\vartheta = 9$). Hence, the output of the subsystem is a function of time, whereas the output of the integrated system are two functions of time. Each of these functions will be given as a discrete time history, e.g., $y(t) = [y(0), y(dt), \ldots, y(N_T \cdot dt)]$, with $N_T = 5000$ and $N_T \cdot dt = T = 5$ s.

The (possibly competing) reliability requirements for the system are defined by $n_g = 3$ performance functions $\boldsymbol{g}(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) = \{g_1(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}), g_2(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}), g_3(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta})\}$. In particular, $g_1(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) < 0$ is needed for system stability, $g_2(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) < 0$ with

$$g_2(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) = \max_{t \in [T/2, T]} |z_1(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}, t)| - 0.02 \tag{1}$$

for the settling time of $z_1(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}, t)$ to be sufficiently fast, and $g_3(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) < 0$ with

$$g_3(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) = \max_{t \in [0, T]} |z_2(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}, t)| - 4 \tag{2}$$

for the energy consumption to be acceptable.

For fixed values of $\boldsymbol{\theta}$ and $\boldsymbol{e}$, the set of $\boldsymbol{a}$ points where $\boldsymbol{g}(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) < 0$ is called the safe domain, whereas its complement set is called the failure domain. The worst-case performance function, defined as

$$w(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) = \max_{i = 1, \ldots, n_g = 3} \{g_i(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta})\}, \tag{3}$$

enables defining the safe and failure domains in terms of a single inequality, i.e., the safe domain is given by the $\boldsymbol{a}$ points where $w(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) < 0$. Further details can be found in Ref. [4].

3 Approaches to the NASA Langley Uncertainty Quantification (UQ) Challenge on Optimization Under Uncertainty

The approaches used to tackle the NASA Langley UQ Challenge on Optimization Under Uncertainty Problems are presented together with the corresponding results obtained: in particular, Sections 3.1-3.6 deal with Subproblems A (namely, Model Calibration & UQ of the Subsystem), B (namely, Uncertainty Reduction), C (namely, Reliability Analysis of Baseline Design), D (namely, Reliability-Based Design), E (namely, Model Update and Design Tuning) and F (namely, Risk-Based Design), respectively.

3.1 Subproblems A: Model Calibration and Uncertainty Quantification (UQ) of the Subsystem

Given the time-dependent (i.e., multivariate) nature of the function $y(\boldsymbol{a}, \boldsymbol{e}, t)$, the calibration and uncertainty quantification of parameter δ (fully prescribed by the pair $<f_{\boldsymbol{a}}, E>$) are carried out in the space of the orthogonal basis vectors resulting from Singular Value Decomposition (SVD) of the data $\boldsymbol{D}_1 = \{y^{(i)}(t)\}$, $i = 1, 2, \ldots, n_1 = 100$, $t = 0, 1, \ldots, N_T$ [6, 7]. In particular, the following steps are performed to pre-process the data before the model calibration and uncertainty quantification tasks:

1. Evaluate the (time-dependent) sample mean $m(t)$ of the dataset $\boldsymbol{D}_1$ as $m(t) = 1/n1 \cdot \sum_{i=1}^{n1} y^{(i)}(t)$.
2. Subtract the mean $m(t)$ from the available time-series $\{y^{(i)}(t)\}$ to obtain the *centered* data $\boldsymbol{D}_1^* = \{y^{*(i)}(t)\} = \{y^{(i)}(t)\} - m(t)$, $i = 1, 2, \ldots, n_1 = 100$, $t = 0, dt, \ldots, N_T dt$.
3. Perform an SVD of the centered data $\{y^{*(i)}(t)\}$. If $\boldsymbol{D}_1^*$ is the (n_1 x N_T) centered matrix of the system realization, then it can be expressed as $\boldsymbol{U}^*\boldsymbol{S}^*\boldsymbol{V}'$, where: $\boldsymbol{U}$ is the (n_1 x n_1) matrix of left singular vectors; $\boldsymbol{S}$ is the (n_1 x N_T) diagonal matrix of the nonnegative singular values s_i, $i = 1, 2, \ldots, n_1$, in decreasing order; and $\boldsymbol{V}$ is the (N_T x N_T) matrix of right singular vectors: in other words, the columns of $\boldsymbol{V}$ contains N_T orthonormal N_T-dimensional (eigen)vectors $\boldsymbol{v}_k$, $k = 1, 2, \ldots, N_T$, constituting an orthonormal basis ß for $\boldsymbol{D}_1^*$.
4. In order to reduce the dimensionality of the problem (while accounting for the overall variability of the model output of interest), select a proper number $n_B(y) < N_T$ of basis vectors to be retained in the analysis. In this work, $n_B(y)$ is selected so that at least $\varepsilon = 99\%$ of the variance associated to the $n_1 = 100$ observed time histories is explained. In details:

   $$\sum_{i=1}^{n_B(y)} \frac{s_i^2}{\sum_{j=1}^{n_1} s_j^2} \geq \varepsilon = 0.99, \quad (4)$$

   noting that $\sum_{j=1}^{n_1} s_j^2$ equals the overall variance of the entire dataset. In this case, the value of $n_B(y)$ turns out to be 10.
5. Project the centered dataset $\boldsymbol{D}_1^*$ onto the orthonormal basis defined by the $n_B(y)$ eigenvectors $\boldsymbol{v}_k$ corresponding to the $n_B(y)$ largest singular values s_k, $k = 1, 2, \ldots, n_B(y)$, of matrix $\boldsymbol{S}$. In particular, the (n_1 x $n_B(y)$) matrix $\boldsymbol{C_1} = \{c_{ik}\}$ containing the projections of the centered data $\boldsymbol{D}_1^*$ onto the orthonormal basis ß = $\{\boldsymbol{v}_k, k = 1, 2, \ldots, n_B(y)\}$, is obtained as:

   $$\boldsymbol{C_1} = \boldsymbol{D_1}^* \cdot \boldsymbol{V}[1{:}n_B(y)] = \boldsymbol{D_1}^* \cdot \left[\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{n_B(y)}\right]. \quad (5)$$

   The coefficients/projections $\{c_{1,ik}$: $i = 1, 2, \ldots, n_1$, $k = 1, 2, \ldots, n_B(y)\}$ are equivalently expressed as:

   $$c_{1,ik} = \sum_{l=1}^{N_T} y^{*(i)}(l \cdot dt) \cdot v_k(l) \ . \quad (6)$$

The idea is to perform the calibration and uncertainty quantification tasks in the (static multivariate) *projected* space (i.e., in the space defined by the orthonormal basis ß) rather than in the (dynamic multivariate) time domain.

The functional form of $f_{\boldsymbol{a}}$ (representing the Uncertainty Model-UM of the aleatory variables $\boldsymbol{a}$) is chosen as a multivariate Gaussian Mixture (GM). A GM model is a probabilistic model that assumes the data are drawn from a mixture of a n_G (multivariate) Gaussian distributions with unknown hyper-parameters. Such hyper-parameters and their corresponding weights are prescribed by minimizing a measure of the offset between the data and the prediction. The Probability Density Function (PDF) of a GM model is given by:

$$f_a^{GM}(\boldsymbol{a}|\boldsymbol{w},\boldsymbol{\mu},\boldsymbol{\Sigma}) = f_a^{GM}(\boldsymbol{a}|\boldsymbol{\varphi_a}) = \sum_{l=1}^{n_G} w_l \cdot \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l), \quad (7)$$

where the *l*-th component of the mixture is characterized by a multivariate Normal distribution with weight $w_l \in [0, 1]$, means $\boldsymbol{\mu}_l$ and covariance matrix $\boldsymbol{\Sigma}_l$ (notice that in *each* multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ of the mixture, the covariance matrix $\boldsymbol{\Sigma}_l$ is able to capture and describe *only linear dependences* between the aleatory variables $\boldsymbol{a}$). For the sake of compact notation, the ensemble of calibration hyper-parameters of the joint aleatory PDF $f_{\boldsymbol{a}}$ is indicated as $\boldsymbol{\varphi_a} = \{w_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l : l = 1, 2, \ldots, n_G\}$. This functional form has been chosen because of its *relatively high flexibility*: by adjusting the corresponding hyper-parameters, a GM can be forced to assume different (*unimodal* and *multi-modal*) *shapes* and to describe a large variety of dependence structures among the aleatory variables [5].

The calibration approach adopted is based on the evaluation of the joint multivariate likelihood of the data in the projected space defined by the orthonormal basis ß, $L^{GM}(\boldsymbol{C_1}|\ \boldsymbol{\Phi}) = L^{GM}(\boldsymbol{C_1}|\boldsymbol{\varphi_a}, \boldsymbol{e}) = L^{GM}(\boldsymbol{C}_1|\, w_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l : l = 1, 2, \ldots, n_G; e_1, e_2, e_3, e_4)$, where $\boldsymbol{\Phi}$ is a vector containing the ensemble of *all* the calibration parameters (i.e., the hyper-parameters of the GM model and the epistemic parameters $\boldsymbol{e}$). The likelihood is here approximated by Monte Carlo Simulation (MCS) of the model $y(\boldsymbol{a}, \boldsymbol{e}, t)$, SVD of the corresponding response and finally Kernel Density Estimation (KDE) in the projected space ß. The detailed algorithm is as follows:

Inputs: $\boldsymbol{\Phi} = [\boldsymbol{\varphi_a}, \boldsymbol{e}]$, $\boldsymbol{C_1}$

Output: $L^{GM}(\boldsymbol{C_1}|\ \boldsymbol{\Phi})$

1. Generate N_{like} realizations of $\boldsymbol{a}$ ($\boldsymbol{a}_q$, $q = 1, 2, \ldots, N_{like}$) by random sampling from the corresponding PDF $f_a^{GM}(\boldsymbol{a}|\boldsymbol{\varphi_a})$ (7). In this work, $N_{like} = 100000$.
2. For each $\boldsymbol{a}_q$ evaluate the model response $y(\boldsymbol{a}_q, \boldsymbol{e}, t)$, $q = 1, 2, \ldots, N_{like}$. Let $\boldsymbol{Y^{\Phi}}$ be the (N_{like} x N_T) matrix containing such responses.
3. Subtract the mean $m(t)$ of the *dataset* $\boldsymbol{D}_1$ $(1/n_1 \cdot \sum_{i=1}^{n1} y^{(i)}(t))$ from the simulated time-series $y(\boldsymbol{a}_q, \boldsymbol{e}, t)$, $q = 1, 2, \ldots, N_{like}$, to obtain the model responses $\boldsymbol{Y^{*\Phi}}$, "*centered*" with respect to the *mean value* of the *real data* (i.e., with respect to the mean value of the "true" system response). Then, project $\boldsymbol{Y^{*\Phi}}$ onto the basis ß = $\{\boldsymbol{v}_k, k = 1, 2, \ldots, n_B(y)\}$ found above, in order to obtain the (N_{like} x n_B) matrix $\boldsymbol{H^Y} = \{h^Y_{qk}\}$ of the corresponding coefficients (projections) $\boldsymbol{H^Y} = \boldsymbol{Y^{*\Phi}} \cdot \boldsymbol{V}[1: n_B(y)] = \boldsymbol{Y^{*\Phi}} \cdot [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{nB(y)}]$.
4. Based on the (projected) model responses $\boldsymbol{H^Y} = \{h^Y_{qk}\}$ and relying on KDE techniques, estimate the likelihood $L^{GM}_{\boldsymbol{H^Y}}(\boldsymbol{C_1}|\boldsymbol{\Phi})$ in the space defined by the orthonormal basis ß. In this paper, a *multivariate* product Gaussian kernel is employed, so that $L^{GM}_{\boldsymbol{H^Y}}(\boldsymbol{C_1}|\boldsymbol{\Phi})$ becomes:

$$L^{GM}_{\boldsymbol{H^Y}}(\boldsymbol{C_1}|\boldsymbol{\Phi}) = \prod_{i=1}^{n_1} \frac{1}{N_{like} b_1 b_2 \ldots b_{nB(y)}} \sum_{q=1}^{Nlike} \prod_{k=1}^{nB(y)} K\left(\frac{c_{1,ik} - h^Y_{qk}}{b_k}\right). \quad (8)$$

In formula (8) b_k, $k = 1, 2, \ldots, n_B(y)$, are the one-dimensional bandwidths of the kernel-smoothing windows, calculated applying the Silverman's rule to the N_{like} simulated output projections $\boldsymbol{H^Y} = \{h^Y_{qk}\}$ (on the *k*-th basis): in particular, $b_k = \hat{\sigma}_k \cdot \left\{\frac{4}{(n_B(y)+2)N_{like}}\right\}^{1/(n_B(y)+4)}$, where $\hat{\sigma}_k$ is the sample

standard deviation of the simulated output projections $\{h^Y_{qk}\}$, q = 1, 2, …, N_{like}; and $K(\cdot)$ is a *one-dimensional* Gaussian kernel function. It is worth noting that even if $L^{GM}_{\boldsymbol{H}^Y}(\boldsymbol{C_1}|\boldsymbol{\Phi})$ (8) uses a product of one-dimensional kernels, this does *not* imply that the $n_B(y)$ variables are *independent* (this is obviously due to the fact that we employ a *sum* of N_{like} products of $n_B(y)$ one-dimensional kernels). As a final remark, notice that the choice of the Silverman's rule above minimizes the mean integrated square error in the non-parametric inference, but it should be used with care. Indeed, it may yield widely inaccurate (for instance, strongly over-smoothed) estimates if the data come from a strongly skewed or multimodal distribution [17].

Two considerations are in order. Centering the simulated time-series $y(\boldsymbol{a}_q, \boldsymbol{e}, t)$ with respect to the mean $m(t)$ of the real data (step 3 above) should guarantee that the likelihood $L^{GM}_{\boldsymbol{H}^Y}(\boldsymbol{C_1}|\boldsymbol{\Phi})$ thereby generated drives the calibration process to match also the true mean of the system response (besides the 99% of the variance, as specified in (4)). Also, notice that the likelihood $L^{GM}_{\boldsymbol{H}^Y}(\boldsymbol{C_1}|\boldsymbol{\Phi})$ (8) is here computed by randomly sampling with replacement the dataset $\boldsymbol{C_1}$ (N_{boot} = 25) and averaging the results: this *bootstrapping* procedure should limit the problem of *overfitting* in the presence of scarce data.

The likelihood $L^{GM}_{\boldsymbol{H}^Y}(\boldsymbol{C_1}|\boldsymbol{\Phi})$ (8) is employed to carry out model calibration and uncertainty quantification by a two-step Maximum Likelihood Estimation (MLE)-based approach, which is aimed at finding: (i) the MLE point estimates $\boldsymbol{\Phi}^{\boldsymbol{MLE}}$ of the ensemble of the calibration parameters $\boldsymbol{\Phi}$, and (ii) the hyper-rectangular set E representing the UM for the pure epistemic parameters $\boldsymbol{e}$. The steps of the algorithm are the following:

1. Perform a Maximum Likelihood Estimation of the ensemble of *all* the calibration parameters $\boldsymbol{\Phi} = [\boldsymbol{\varphi_a}, \boldsymbol{e}] = [w_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma_l}: l = 1, 2, \ldots, n_G; e_1, e_2, e_3, e_4]$ as:

$$\boldsymbol{\Phi}^{MLE} = arg\max_{\Phi}\left\{log\left(L^{GM}_{\boldsymbol{H}^Y}(\boldsymbol{C_1}|\boldsymbol{\Phi})\right)\right\}. \quad (9)$$

2. Fixing the MLE estimates of $\boldsymbol{\varphi_a}$ at $\boldsymbol{\varphi}_a^{MLE}$ quantify the epistemic uncertainty in $\boldsymbol{e}$:
   a) Build a likelihood of the data *only* as a function of the pure epistemic parameters $\boldsymbol{e}$, i.e., $L^{GM}_{\boldsymbol{H}^Y}(\boldsymbol{C_1}|\boldsymbol{e}, \boldsymbol{\varphi_a} = \boldsymbol{\varphi}_a^{MLE})$ (notice that this function is *four-dimensional*). In practice, select (deterministically or stochastically) N_e possible values $\boldsymbol{e}^k$, k = 1, 2, …, N_e, of the epistemic parameters $\boldsymbol{e}$ (in this paper, N_e = 1000 samples are uniformly drawn within the respective ranges). Evaluate the likelihood of the data in correspondence of such realizations of the epistemic parameters to obtain $L^{GM}_{\boldsymbol{H}^Y}(\boldsymbol{C_1}|\boldsymbol{e}^k, \boldsymbol{\varphi_a} = \boldsymbol{\varphi}_a^{MLE})$, k = 1, 2, …, N_e. Construct an *approximation* $\tilde{L}^{GM}_{\boldsymbol{H}^Y}(\boldsymbol{C_1}|\boldsymbol{e}, \boldsymbol{\varphi_a} = \boldsymbol{\varphi}_a^{MLE})$ to $L^{GM}_{\boldsymbol{H}^Y}(\boldsymbol{C_1}|\boldsymbol{e}, \boldsymbol{\varphi_a} = \boldsymbol{\varphi}_a^{MLE})$, e.g., by fitting a (quick-running) *response surface* to the discrete points $L^{GM}_{\boldsymbol{H}^Y}(\boldsymbol{C_1}|\boldsymbol{e}^k, \boldsymbol{\varphi_a} = \boldsymbol{\varphi}_a^{MLE})$, k = 1, 2, …, N_e (in this paper, an Artificial Neural Network regression is employed to this purpose).
   b) Based on the approximation $\tilde{L}^{GM}_{\boldsymbol{H}^Y}(\boldsymbol{C_1}|\boldsymbol{e}, \boldsymbol{\varphi_a} = \boldsymbol{\varphi}_a^{MLE})$ to the real likelihood $L^{GM}_{\boldsymbol{H}^Y}(\boldsymbol{C_1}|\boldsymbol{e}, \boldsymbol{\varphi_a} = \boldsymbol{\varphi}_a^{MLE})$ (which is a function only of the n_e = 4 pure epistemic parameters), define the UM E as the smallest hyper-rectangle enveloping the *joint* four-dimensional α·100% Confidence Interval (CI) of $\boldsymbol{e}$ (in this paper, α = 0.99). In practice, *normalize* $\tilde{L}^{GM}_{\boldsymbol{H}^Y}(\boldsymbol{C_1}|\boldsymbol{e}, \boldsymbol{\varphi_a} = \boldsymbol{\varphi}_a^{MLE})$ over the original domain of variation E_0 of the epistemic parameters, in order to provide it with the properties of a Probability Density Function (PDF): $Q \cdot \int_{E_0} \tilde{L}^{GM}_{\boldsymbol{H}^Y}(\boldsymbol{C_1}|\boldsymbol{e}, \boldsymbol{\varphi_a} = \boldsymbol{\varphi}_a^{MLE})d\boldsymbol{e} = 1$, where Q is the normalization constant. In this paper, likelihood normalization is carried out by *discretization* of the epistemic space and simple *numerical integration*, thanks to the comparatively low

dimensionality of the domain (in particular, equally spaced bins of 0.01 width are adopted for all the epistemic parameters e_i, i = 1, 2, 3, n_e = 4). Once the approximation $Q \cdot \tilde{L}_{\boldsymbol{H}^Y}^{GM}(\boldsymbol{C_1}|\boldsymbol{e}, \boldsymbol{\varphi_a} = \boldsymbol{\varphi}_a^{MLE})$ to the normalized likelihood is available, the *joint* four-dimensional α·100% CI of $\boldsymbol{e}$ can be computed numerically by applying the definition $P[\boldsymbol{e} \in (\alpha \cdot 100\%\ CI)] \geq \alpha$ to the previously discretized epistemic domain. Notice that also *sampling-based* procedures could be used to build an *empirical* CI for $\boldsymbol{e}$. In fact, the likelihood $L_{\boldsymbol{H}^Y}^{GM}(\boldsymbol{C_1}|\boldsymbol{e}, \boldsymbol{\varphi_a} = \boldsymbol{\varphi}_a^{MLE})$ can be readily sampled using, e.g., Markov Chain Monte Carlo (MCMC) methods, for which knowing Q is not required [18]. On the one hand, this approach is useful in the presence of high-dimensional spaces, because MCMC does not suffer the curse of dimensionality. On the other hand, it often requires significant computational efforts, since numerous (hundreds of thousands) model evaluations are needed to obtain reliable PDF estimations and robust CI evaluations.

The two-step approach presented has the advantages of being simple and based on a well-established and sound technique for model calibration (i.e., MLE). However, notice that the separate calibration of the aleatory distribution parameters $\boldsymbol{\varphi_a}$ (step 1) and of the intervals E for the epistemic parameters (step 2) may lead to an issue. As indicated by the likelihood $L_{\boldsymbol{H}^Y}^{GM}(\boldsymbol{C_1}|\boldsymbol{e}, \boldsymbol{\varphi_a} = \boldsymbol{\varphi}_a^{MLE})$, the final shape and size of E is *intrinsically conditional* on the (sub-)optimal (in this case, *parametric*) aleatory model identified at step 1: thus, if the parametric probabilistic model fails to capture the true relationships between the aleatory variables (in particular, their complex dependences), the resulting box E may focus on wrong portions of the epistemic space and/or possibly underestimate the corresponding uncertainty. In this paper, we try to limit these drawbacks by adding *conservatism* in two ways: (i) during calibration, *all* the epistemic uncertainty is "loaded" on $\boldsymbol{e}$ (different from classical Bayesian approaches, no epistemic uncertainty is associated to the other calibration parameters $\boldsymbol{\varphi_a}$ in step 1); (ii) a relatively large confidence level α is chosen (i.e., 0.99), which may give us the possibility to rigorously *envelop* the observations, and produce a structure that can be generalized to a probability box *containing* the *true distribution* in the aleatory space.

A GM model with n_{GM} = 2 multivariate Gaussian distributions is chosen. Notice that also the option with n_{GM} = 3 has been tested: however, in spite of the *significant* increase in the number of calibration hyper-parameters (i.e., from 45 to 66, respectively), a *negligible* improvement in the (log-)likelihood $log\left(L_{\boldsymbol{H}^Y}^{GM}(\boldsymbol{C_1}|\boldsymbol{\Phi})\right)$ of the observations is registered (i.e., from 1137.2 to 1142.0, respectively). In addition, the reader should notice that further increasing the hyper-parameter (search) space would make the calibration process (i.e., the maximization task in (9)) hardly tractable even for global (meta-heuristic) optimization tools. In the light of these results and considerations, additional tests with larger values of n_{GM} (i.e., n_{GM} > 3) have not been performed. Parameters $\boldsymbol{\Phi}$ range in the following intervals (hyper-rectangles): the weights $\boldsymbol{w}$ in [0, 1], the Gaussian means $\boldsymbol{\mu}$ in $A_0 = [0, 2]^{na}$, the Gaussian variances $\boldsymbol{\sigma}^2$ in $[0.0025, 25]^{na}$ and the Pearson correlation coefficients ρ_{ij}, i = 1, 2, …, n_a-1, $j = i$+1, …, n_a (used to build the covariance matrices $\boldsymbol{\Sigma}$), in [-1, 1]; the epistemic parameters $\boldsymbol{e}$ are defined in $E_0 = [0, 2]^{ne}$. The total number of decision variables is thus equal to 45. Due to the large-sized and multi-modal nature of the parameter space, the MLE optimization problem (9) is tackled by resorting to a *population-based*, *heuristic* optimization technique, i.e., a Genetic Algorithm (GA) [13]. Such method deeply explores the search space by evaluating many candidate solutions in order to find a near-optimal solution. Although GA is a *global* optimizer, in some problems (characterized by massive multimodality of the objective function to be optimized), it may converge to *local* optima. The performance of GA depends on its ability to thoroughly explore the search space (i.e., to maintain a sufficient "genetic diversity" in the population of candidate solutions), while attempting to drive the search efficiently and intelligently towards the "interesting region" of the search space, i.e., towards the global optimum. A thorough exploration of the search space (i.e., a sufficient "genetic diversity") is guaranteed by the following strategies: (i) GA is *repeated* few times (say, five times)

with different random seeds (i.e., *different* random *initial populations*) and only the best result over all the simulation is retained; (ii) some of the GA parameters are properly set, mainly based on the experience of the author in the use of GAs: for example, a relatively high population size (i.e., N_{pop} = 200), high mutation rates (i.e., p_{mut} = 0.025) and a large number of generations (N_{gen} = 1000) are employed [19].

The pictorial result of Step 1 of the calibration process (i.e., the PDF $f_a^{GM}(\boldsymbol{a}|\boldsymbol{\varphi}_{\boldsymbol{a}}^{MLE})$) is not reported here due to space limitations: only the *final* aleatory UM will be presented in Section 3.5. Instead, an *overall quantitative* evaluation of the degree of dependency between the parameters $\boldsymbol{a}$ is provided by means of a *rank* correlation matrix $\boldsymbol{R}_{\boldsymbol{a},Spear}$ (in particular, based on the Spearman measure). The choice of a rank correlation coefficient is justified by its *nonparametric* nature and its *invariance* with respect to the marginals of the dependent random variables analyzed. Notice that the Spearman correlation between two variables is equal to the Pearson correlation between the *rank values* of those two variables; while Pearson's correlation assesses *linear relationships*, Spearman's correlation assesses *monotonic relationships* (whether *linear or not*). Such dependency evaluation has been made by means of a sample of 100000 realizations of $\boldsymbol{a}$, generated from the PDF $f_a^{GM}(\boldsymbol{a}|\boldsymbol{\varphi}_{\boldsymbol{a}}^{MLE})$. The Spearman *rank* correlation matrix $\boldsymbol{R}_{\boldsymbol{a},Spear}$ is as follows:

$$\boldsymbol{R}_{a,Spear}=\begin{bmatrix} 1 & 0.394 & 0.110 & -0.094 & 0.068 \\ 0.394 & 1 & -0.162 & 0.229 & -0.150 \\ 0.110 & -0.162 & 1 & 0.035 & 0.851 \\ -0.094 & 0.229 & 0.035 & 1 & -0.081 \\ 0.068 & -0.150 & 0.851 & -0.081 & 1 \end{bmatrix} \quad (10)$$

The sign of the Spearman correlation indicates the direction of association between variables a_i and a_j. If a_i tends to increase when a_j increases (resp., decreases), the Spearman correlation coefficient is positive (resp., negative). Also, the Spearman correlation increases in magnitude as a_i and a_j become closer to being perfectly monotonically related (see, e.g., a_3 and a_5, whose rank correlation coefficient is 0.851). It must be acknowledged that while $\boldsymbol{R}_{\boldsymbol{a},Spear}$ (10) represents a synthetic and easily interpretable measure of the *strength* of *correlations*, it may obviously fail to fully and accurately describe the possibly *complex* and *nonlinear* patterns of *dependence* between aleatory variables.

The uncertainty model (hyper-rectangle) E = $[\underline{e}_1, \overline{e}_1]$ x $[\underline{e}_2, \overline{e}_2]$ x $[\underline{e}_3, \overline{e}_3]$ x $[\underline{e}_4, \overline{e}_4]$ for the epistemic parameters $\boldsymbol{e}$ chosen according to Step 2 of the calibration algorithm is as follows: $[\underline{e}_1, \overline{e}_1]$ = [0, 0.3719], $[\underline{e}_2, \overline{e}_2]$ = [0.1910, 0.7273], $[\underline{e}_3, \overline{e}_3]$ = [0, 0.8543] and $[\underline{e}_4, \overline{e}_4]$ = [0, 2]. The corresponding MLEs are 0.0704, 0.5176, 0.0411 and 1.9059, respectively. Figure 1 shows the calibrated model output against the data provided. Let us denote this base-case, initial Uncertainty Model (UM) as "UM-0(y)". In Figure 1 (left) we report the time series observations (red solid lines) along with the extreme upper and lower *bounds* (blue dashed lines) resulting from the propagation through the model function $y(\boldsymbol{a}, \boldsymbol{e}, t)$ of 500000 configurations $(\boldsymbol{a}_i, \boldsymbol{e}_i)$, i = 1, 2, …, 500000 (i.e., N_a = 1000 aleatory samples for N_e = 500 epistemic vectors, including the *vertices* of the box E); also, the overall *mean* of the calibrated output (averaged over the aleatory PDF $f_a^{GM}(\boldsymbol{a}|\boldsymbol{\varphi}_{\boldsymbol{a}}^{MLE})$ and over epistemic space E) is shown as black crosses. The data is *captured* and *enveloped* quite tightly, at most of the time steps. Finally, in Figure 1 (right) the same calibration results are represented in the orthonormal SVD space projected on the basis pair (h^Y_1, h^Y_2): again, the calibrated model (blue points) envelops quite tightly the data provided (red crosses). Based on these results, we can argue that the produced UM structure is likely to represent a probability box *possibly containing* (with the prescribed confidence α) the *true distribution* in the aleatory space. Also, the level of conservatism injected in the calibration process (see above) presumably allows the UM to properly withstand (i.e., "envelop") *most* aleatory and epistemic uncertainties, including *unexpected* and *extreme* events possibly occurring in the life of the system.

A final remark is in order. The solution to (9) entails the *repeated* evaluation of the mathematical model $y(\boldsymbol{a}, \boldsymbol{e}, t)$ (i.e., the likelihood estimation step based on N_a samples) for *every* possible solution $\boldsymbol{\Phi}$ proposed by the heuristic optimization tool during the search. As a consequence, the total number of model

evaluations can easily reach tens/hundreds millions, which makes the proposed approach impractical also in the presence of mathematical models that take even only few seconds/minutes to run. In such cases, two main options can be considered within the parametric calibration framework proposed: i) employ *approximate* versions of the likelihood function, based on stochastic distance metrics computed only on few relevant statistics, e.g., means and/or quantiles [20, 21]; ii) replace the original (possibly long-running) system model $y(\boldsymbol{a}, \boldsymbol{e}, t)$ by a *surrogate* (cheaper-to-evaluate) regression model, able to reproduce *functional data* (e.g., time series, like in the present case) [22, 23].

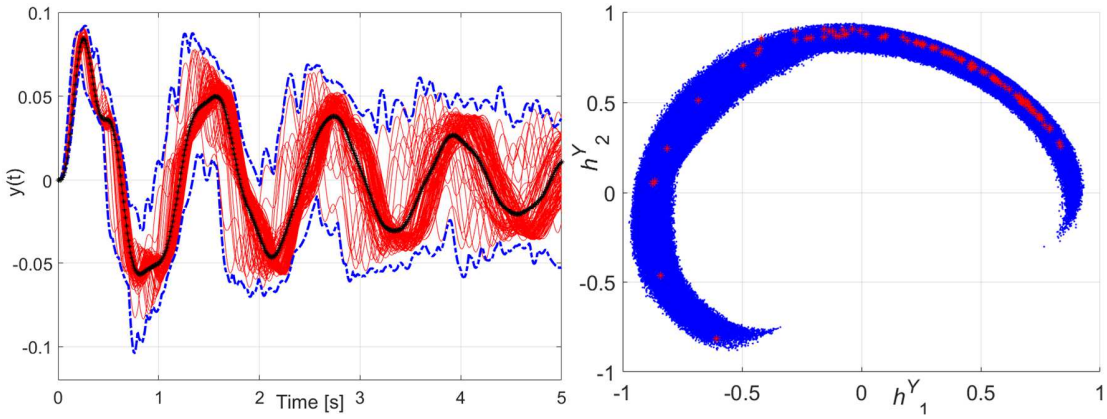

Figure 1. Left: Bounds on the calibrated model output y(a, e, t) (blue dashed lines) against experimental data (red solid lines). Right: calibration result in the SVD space projected on the basis pair (h^Y_1, h^Y_2)

3.2 Subproblem B: Uncertainty Reduction

The epistemic parameters should be ranked according to their ability to improve the *predictive capability* of the computational model of the subsystem. In other words, as specified by the challengers, the epistemic parameters leading to the largest *reduction* in the output's *spread* should be identified. Two approaches have been developed to address this task: the first is based on *sensitivity analysis* within a 'factor prioritization' setting in analogy with variance-based Sobol' indices (Section 3.2.1); the second relies on the evaluation of the energy score, i.e., a multivariate generalization of the *Continuous Rank Predictive Score* (CRPS), to measure the predictive capability of a stochastic computational model (Section 3.2.2).

3.2.1 Sensitivity Analysis in a 'Factor Prioritization' Setting

This approach is aimed at assessing how much less epistemic uncertainty the model output of interest (resp., higher predictive capability) would have if extra knowledge about an input were available. This can be done by comparing the epistemic uncertainty before and after '*pinching*' an input, i.e., replacing it with a value without (or with less) epistemic uncertainty. Quantifying this effect assesses the contribution by the input epistemic uncertainty to the overall epistemic uncertainty in the output of interest [24, 25].

Let $U_{\boldsymbol{e}}$ be an indicator of the 'amount' of epistemic uncertainty contained in the output of the computational model $y(\boldsymbol{a}, \boldsymbol{e}, t)$. The subscript '$\boldsymbol{e}$' suggests that indicator $U_{\boldsymbol{e}}$ is computed over all the (epistemic) input parameters $\boldsymbol{e}$ (and over the corresponding space of variation E). The indicator $U_{\boldsymbol{e}}$ could be obviously measured in different ways (e.g., straightforwardly by the overall *variance* of y). In this paper, coherently with the approach presented in Section 3.1, the uncertainty in the subsystem model output y is measured by the overall *spread* of y in the projected space defined by the orthonormal bases ß = $\{\boldsymbol{v}_k, k = 1, 2, \ldots, n_B(y)\}$. Such spread is here quantified by the *volume* $\mathcal{G}$ of the *convex hull* able to envelop the projections of $y(\boldsymbol{a}, \boldsymbol{e}, t)$ onto ß (i.e., $\mathcal{G}_{\boldsymbol{e}} = U_{\boldsymbol{e}}$). Obviously, the larger the volume of the convex hull, the larger the spread

of the output y, i.e., the larger the (epistemic) uncertainty in y [11]. Notice that the computational burden related to the calculation of the convex hull volume in n_B = 10 dimensions may be prohibitive. Thus, the strategy proposed in Ref. [11] is adopted: an approximation $\mathcal{G}'$ to the volume $\mathcal{G}$ is obtained by *projecting* the full n_B-dimensional convex hull onto subspaces of *lower dimensionality* n_B' ($n_B' << n_B$): the smaller n_B', the smaller the computational effort. In particular, $\binom{n_B}{n_B'} = \binom{10}{2}$ = 45 2-dimensional projections of the full 10-dimensional convex hull are here employed, which reduces the computational time by more than two orders of magnitude [11]. For example, the 10-dimensional hull is projected on the basis pairs (h^Y_1, h^Y_2), (h^Y_2, h^Y_5) and so on. The corresponding 2-dimensional volumes $\mathcal{G}'_i$, i = 1, 2, …, 45, are computed and the overall volume $\mathcal{G}$ is roughly approximated by $\sum_{i=1}^{45} \mathcal{G}'_i$. Further details can be found in Ref. [11].

We want to assess - by means of a sensitivity index $S_i(U_{\boldsymbol{e}})$ - the effect that a refinement of the uncertainty model of the generic epistemic input e_i (i.e., a reduction in its epistemic uncertainty) has on the amount of epistemic uncertainty $U_{\boldsymbol{e}}$ of the model output. In order to address this issue, a sensitivity index is used in analogy with variance-based Sobol' indices [10]. Imagine that we fix e_i at a particular value e_i^* in $[\underline{e_i}, \overline{e_i}]$. Let $U_{\boldsymbol{e}}(e_i = e_i^*)$ be the resulting amount of epistemic uncertainty in $y(\boldsymbol{a}, \boldsymbol{e}_{-i}, t)$, taken over all parameters $\boldsymbol{a}$ and $\boldsymbol{e}_{-i}$ and keeping parameter e_i fixed at e_i^* (instead, all the other epistemic parameters $\boldsymbol{e}_{-i}$ are allowed to range in their corresponding space of variation E_{-i}). We would imagine that having frozen one potential source of epistemic uncertainty (e_i), the resulting indicator $U_{\boldsymbol{e}}(e_i = e_i^*)$ will be lower than the corresponding total (or unconditional) one $U_{\boldsymbol{e}}$. One could therefore conceive of using $U_{\boldsymbol{e}}(e_i = e_i^*)$ as a measure of the relative importance of e_i, reasoning that the smaller $U_{\boldsymbol{e}}(e_i = e_i^*)$, the greater the influence of e_i. However, notice that this approach makes the sensitivity measure dependent on the position of the point e_i^* for each input factor. Thus, we take the average of the measure $U_{\boldsymbol{e}}(e_i = e_i^*)$ over all the possible points e_i^* in $[\underline{e_i}, \overline{e_i}]$, which removes the dependence on e_i^*. The resulting indicator is then written synthetically as $E_{ei}[U_{\boldsymbol{e}}(e_i)]$ and represents the *expected* amount of epistemic uncertainty contained in the output y when e_i is fixed to a constant value (i.e., when the amount of its epistemic uncertainty is reduced to zero). Obviously, the lower $E_{ei}[U_{\boldsymbol{e}}(e_i)]$, the more important e_i: in other words, the most important parameter is the one which *on average*, once *fixed*, causes the greatest reduction in the epistemic uncertainty of y. Finally, the sensitivity $S_i(U_{\boldsymbol{e}})$ of the output y to the epistemic uncertainty of e_i can be synthesized with an expression like

$$S_i(U_e) = 1 - \frac{E_{ei}[U_e(e_i)]}{U_e}. \tag{11}$$

Index (11) is an estimate of the *value* of *additional* empirical information about the input e_i in terms of the *fractional reduction* in epistemic uncertainty that might be achieved in y when the input parameter is replaced by a better estimate obtained from future empirical study. This 'pinching' procedure can be applied to each input quantity in turn and the results used to rank the inputs in terms of their sensitivities (i.e., in terms of their capability of reducing the output spread).

In this paper, the sensitivity index $S_i(U_{\boldsymbol{e}})$ (11) related to the generic parameter e_i is straightforwardly estimated as follows [10]:

1. letting $\boldsymbol{e}$ range within the entire space of variation E, propagate the mixed aleatory and epistemic uncertainty from the inputs $\boldsymbol{a} \sim f_{\boldsymbol{a}}$ and $\boldsymbol{e}$, respectively, to the output of interest $y(\boldsymbol{a}, \boldsymbol{e}, t)$ and evaluate the resulting (total, unconditional) amount of epistemic uncertainty $U_{\boldsymbol{e}}$.
2. select (deterministically or stochastically) N_e values e_i^k, k = 1, 2, …, N_e, of the epistemically uncertain parameter e_i under analysis within its interval of variation $[\underline{e_i}, \overline{e_i}]$. These N_e realizations of epistemic uncertainty e_i^k, k = 1, 2, …, N_e, should be chosen in such a way to evenly cover the corresponding interval $[\underline{e_i}, \overline{e_i}]$: in this paper, a sequence of 20 equally spaced points is adopted to this aim.
3. fixing the value of e_i to e_i^k, k = 1, 2, …, N_e, and letting all the other epistemically uncertain parameters $\boldsymbol{e}_{-i}$ vary within E_{-i}, propagate the mixed aleatory and epistemic uncertainty from the inputs $\boldsymbol{a} \sim f_{\boldsymbol{a}}$ and $\boldsymbol{e}_{-i}$ to the output of interest $y(\boldsymbol{a}, \boldsymbol{e}, t)$ and evaluate the resulting (conditional) amount of

epistemic uncertainty $U_{\boldsymbol{e}}(e_i = e_i^k) = \mathcal{G}_{\boldsymbol{e}}\left(e_i = e_i^k\right)$ in y. The propagation of the mixed aleatory and epistemic uncertainty is carried out with N_a = 10000 samples.

4. estimate the index (11) as $S_i(U_e) = 1 - 1/N_e \cdot \sum_{k=1}^{N_e} \frac{U_e(e_i = e_i^k)}{U_e}$.

The total number of model evaluations required by the method is thus $N_a \cdot N_e \cdot n_e$ (= 800000 in this case). The approach provides a satisfactory *global* indication of the *overall capability* of the epistemic parameters to *reduce the output spread*. However, it has two drawbacks: (i) being *moment-dependent* (i.e., variance-based), it does *not* guarantee that after pinching one parameter, the predictive capability of the model remains acceptable (e.g., that the mean is still matched or that the corresponding p-box still envelops the observations); (ii) it provides *no direct indication* on which *side* of the interval to refine. The values obtained for $S_i(U_{\boldsymbol{e}})$ are shown in Table 1, together with the corresponding parameter ranking. The author's choice is to refine parameters e_2 and e_3: the importance of e_2 is more than ten times larger than the other parameters; also, the importance of e_3 is twice larger than that of e_1. The side of the interval to refine is determined as the one leading to the *largest contraction* of the interval, while still *including* the MLEs: in this case, the lower bound of e_2 should be increased, whereas the upper bound of e_3 should be reduced.

	Variance-based Sensitivity Analysis – Factor Prioritization	
	Base Uncertainty Model UM-0(y) (after Subproblem A)	
Parameter (MLEs)	$S_i(U_e)$	Ranking
e_1 (0.0704)	0.0125	3
e_2 (0.5176)	0.1048	1
e_3 (0.0411)	0.0235	2
e_4 (1.9059)	$7.65 \cdot 10^{-6}$	4

Table 1. Ranking of the capability of the e_i's to reduce the output spread obtained using UM-0(y)

3.2.2 The Energy Score (ES) as a multivariate generalization of the Continuous Rank Predictive Score (CRPS)

The Continuous Rank Predictive Score (CRPS) is arguably the most versatile scoring rule for probabilistic forecasts of a univariate scalar variable. It measures the *distance* between the CDF of the provided data (i.e., realizations/measurements of the real system of interest) and the CDF of the forecast data, i.e., data generated based on the predictive model. To assess probabilistic forecasts of a *multivariate* quantity (like the one of interest in the present Challenge), the use of the *Energy Score* (ES) (computed in the orthonormal space ß) is proposed [12]. The idea is to rank the epistemic parameters according to their capability to *improve* the *predictive ability* of the model (i.e., to *decrease* the ES), when their uncertainty is *reduced*.

Let $\boldsymbol{D}_1$ be the available dataset in the original measurement space and $\boldsymbol{C_1} = \{\boldsymbol{c_{1,i}}: i = 1, 2, \ldots, n_1 = 100\} = \{c_{1,ik}: i = 1, 2, \ldots, n_1 = 100, k = 1, 2, \ldots, n_B\}$ the matrix containing the corresponding projections onto the orthonormal space ß. Let $y^{(q)}(\boldsymbol{a}, \boldsymbol{e}, t)$, $\boldsymbol{a} \sim f_{\boldsymbol{a}}$, $\boldsymbol{e} \in E$, $q = 1, 2, \ldots, N_s$, be a collection of N_s randomly generated realizations of the subsystem model, whose uncertainty is prescribed by the results of Subproblem A, and $\boldsymbol{H^Y} = \{\boldsymbol{h^Y}_q: q = 1, 2, \ldots, N_s\} = \{h^Y_{qt}: q = 1, 2, \ldots, N_s, k = 1, 2, \ldots, n_B(y)\}$ the matrix containing corresponding projections onto the orthonormal space ß. The evaluation of the predictive capability of model $y(\boldsymbol{a}, \boldsymbol{e}, t)$ (characterized by the UM $\boldsymbol{a} \sim f_{\boldsymbol{a}}$, $\boldsymbol{e} \in E$) with respect to the generic projected datum $\boldsymbol{c}_{1,i}$, $i = 1, 2, \ldots, n_1 = 100$, is:

$$ES\left(\langle f_{\boldsymbol{a}}, E\rangle, \boldsymbol{c}_{1,i}\right) = \frac{1}{N_s}\sum_{q=1}^{N_s}\left\|\boldsymbol{h}_{\boldsymbol{q}}^{\boldsymbol{Y}} - \boldsymbol{c}_{1,i}\right\| - \frac{1}{2N_s^2}\sum_{q=1}^{N_s}\sum_{j=1}^{N_s}\left\|\boldsymbol{h}_{\boldsymbol{q}}^{\boldsymbol{Y}} - \boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{Y}}\right\|. \quad (12)$$

where $||\cdot||$ is the Euclidean norm. It can be demonstrated that if the number n_B of dimensions equals 1, then Eq. (12) reduces to the well-known CRPS [26]. Also, it is straightforward to notice that the smaller (12), the smaller the average distance between the (projected) model predictions $\boldsymbol{h^Y}_q$ and the (projected) datum $\boldsymbol{c}_{1,i}$, i.e., the better the predictive capability of the model.

To have a global measure of the *overall* predictive capability of the model, an average of (12) over the entire dataset $\boldsymbol{C}_1$ is carried out:

$$ES(\langle f_{\boldsymbol{a}}, E\rangle) = \frac{1}{n_1}\sum_{i=1}^{n_1} ES\left(\langle f_{\boldsymbol{a}}, E\rangle, \boldsymbol{c}_{1,i}\right). \tag{13}$$

To assess the capability of the epistemic parameters $\boldsymbol{e}$ to *improve* the *predictive ability* of the model (i.e., to *decrease* the ES, when their epistemic uncertainty is *reduced*), a procedure similar to that outlined in the previous Section 3.2.1 is adopted. Each e_i is fixed at different values e_i^k, k = 1, 2, …, N_e = 20, within its range of variation $[\underline{e}_i, \overline{e}_i]$. In correspondence of each e_i^k a (*conditional*) value of the ES is computed as $ES\left(\langle f_{\boldsymbol{a}}, E_{-i}|e_i = e_i^k \rangle\right)$ according to (12) (for the sake of compact notation, let the *conditional* Energy Score $ES\left(\langle f_{\boldsymbol{a}}, E_{-i}|e_i = e_i^k \rangle\right)$ be indicated as $ES\left(e_i = e_i^k\right) = ES(e_i)$. In this paper, the evaluation of (13) is obtained by resorting to N_s = 10000 randomly generated realizations of the subsystem model (for each epistemic value e_i^k). Then, the parameter characterized by the *highest* ability to improve the predictive capability of the computational model is the one with the *minimum* value of $ES(e_i) = ES\left(e_i = e_i^k\right)$ computed over its range $[\underline{e}_i, \overline{e}_i]$. In this view, the corresponding sensitivity indicator $S_i(ES)$ is computed as the ratio between the minimum ES obtained for the "reduced" epistemic model, i.e., $\min_{ei}\{ES(e_i)\}$, and the ES associated to the full epistemic model $ES(\langle f_{\boldsymbol{a}}, E\rangle)$:

$$S_i(ES) = \frac{\min_{ei}\{ES(e_i)\}}{ES(\langle f_{\boldsymbol{a}}, E\rangle)} \tag{14}$$

Obviously, the smaller (14), the higher the capability of e_i of improving the predictive capability of the model. The total number of model evaluations required by the method is $N_s \cdot N_e \cdot n_e$ (= 800000 in this paper).

The evolution of quantity $ES(e_i) = ES\left(e_i = e_i^k\right)$, i = 1, 2, 3, 4, obtained using UM-0(y) (i.e., after Subproblem A) is shown in Figure 2 for different (fixed) values of the epistemic parameters (normalized between 0 and 1 for the sake of illustration). As for the previous approach, the author's choice is to refine parameters e_2 and e_3. The side of the epistemic interval to refine is determined as the one leading to the *largest contraction* of the interval, while *including* the point value leading to the minimum ES: coherently with the sensitivity-based approach, the lower bound of e_2 should be increased, whereas the upper bound of e_3 should be reduced. The refined epistemic intervals provided by the Challengers are e_2 = [0.4064, 0.7664] and e_3 = [0.0330, 0.3330] and the corresponding resulting uncertainty model is denoted as E_1. The updated epistemic box $E \subseteq E_1$ is then selected as $E = [\underline{e}_1, \overline{e}_1]$ x $[\underline{e}_2, \overline{e}_2]$ x $[\underline{e}_3, \overline{e}_3]$ x $[\underline{e}_4, \overline{e}_4]$ = [0, 0.3719] x [0.4064, 0.7664] x [0.0330, 0.3330] x [0, 2]. Let us denote this first refined uncertainty model as UM-1(y).

In passing, notice that the value ES(UM-1(y)) of the energy score for the full uncertainty model UM-1(y) is 21.28, whereas the ES of UM-0(y) (i.e., the initial not refined model) turns out to be 22.49, meaning that the refinement has led to an improvement in the predictive capability of the 5.38%.

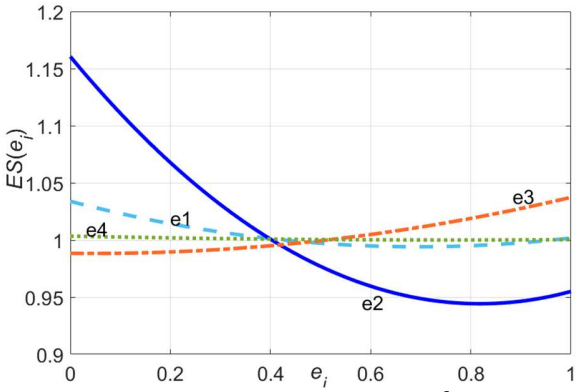

Figure 2. Conditional Energy Score $\boldsymbol{ES(e_i) = ES\left(e_i = e_i^k\right)}$ *obtained using UM-0(y)*

3.3 Subproblem C: Reliability Analysis of the Baseline Design

The objective is the evaluation of the (epistemic) range $R_i(\boldsymbol{\theta})$ of the failure probability $P[g_i(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) \geq 0]$ (15) for each individual stability requirement g_i, $i = 1, \ldots, n_g = 3$, given the baseline system design $\boldsymbol{\theta} = \boldsymbol{\theta}_{base}$. Also, it is requested to calculate the range $R(\boldsymbol{\theta})$ of the failure probability $P[w(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) \geq 0]$ (16) for all requirements:

$$R_i(\boldsymbol{\theta}) = \left[\min_{e \in E} P[g_i(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) \geq 0], \max_{e \in E} P[g_i(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) \geq 0]\right], \; i = 1, \ldots, n_g = 3 \quad (15)$$

$$R(\boldsymbol{\theta}) = \left[\min_{e \in E} P[w(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) \geq 0], \max_{e \in E} P[w(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) \geq 0]\right], \quad (16)$$

where $\boldsymbol{\theta} = \boldsymbol{\theta}_{base}$. These problems are addressed by an efficient combination of: (i) GAs to deeply search the epistemic space E and solve the optimization problems related to the propagation of epistemic uncertainty by interval analysis; (ii) MCS to propagate aleatory uncertainty and estimate the failure probabilities; and (iii) fast-running regression models to reduce the computational time related to the repeated model evaluations required by uncertainty propagation [10].

GAs are run for $N_{gen} = 200$ generations with a population of $N_{pop} = 100$ individuals; also, $N_a = 100000$ random samples are used to estimate the failure probabilities by MCS. The regression model is constructed on the basis of a *finite* set D_{tr} of N_{tr} data representing *examples* of the input/output nonlinear relationships underlying the original system model. The generation of this data set D_{tr} entails running the original system mathematical model $\boldsymbol{g_i(a, e, \theta)}$ a predetermined number of times N_{tr} for specified values $\{(\boldsymbol{a}_t, \boldsymbol{e}_t)\ t = 1, 2, \ldots, N_{tr}\}$ of the input variables $(\boldsymbol{a}, \boldsymbol{e})$ and collecting the corresponding values $\{\boldsymbol{g}_t: t = 1, 2, \ldots, N_{tr}\}$ of the outputs $\boldsymbol{g} = \{g_i: o = 1, \ldots, n_g = 3\}$ of interest. Then, statistical techniques (for example, regression error minimization procedures) are employed for calibrating/adapting the internal *parameters/coefficients* of the regression model to fit the input/output data $D_{tr} = \{(\boldsymbol{a}_t, \boldsymbol{e}_t, \boldsymbol{g}_t): t = 1, 2, \ldots, N_{tr}\}$ generated in the previous step and to capture the underlying (possibly nonlinear and non-monotonic) relationship. Once built, the meta-model can be used for performing, in an acceptable computational time, the numerous repeated evaluations needed for an accurate estimation of the probability ranges above.

In this work, a four-layered feed-forward Artificial Neural Network (ANN) regression model is considered [27]. From a mathematical viewpoint, ANNs consist of a set of nonlinear (e.g., sigmoidal) basis functions with adaptable parameters that are adjusted by a process of *training* (on many different input/output data examples), i.e., an iterative process of regression error minimization. ANNs have been demonstrated to be universal approximants of *continuous* nonlinear functions (under mild mathematical conditions) [28], i.e., in principle, an ANN model with a properly selected architecture can be a consistent estimator of any continuous nonlinear function. Further details about ANN regression models are not reported here for brevity; the interested reader may refer to the cited references and the copious literature in the field. Notice that the recommendation of using ANN regression models is mainly based on (i) theoretical considerations about the (mathematically) demonstrated capability of ANN regression models of being *universal* approximants of continuous nonlinear functions [28] and (ii) the experience of the authors' in the use of ANN regression models for propagating the uncertainties through model codes of safety systems [14].

We train four ANN regression models (one for each stability requirement and one for the worst-case performance metric) using a set of input/output data examples of size $N_{train} = 150000$. A Latin Hypercube Sample (LHS) of the inputs is drawn to give the vectors $(\boldsymbol{a}, \boldsymbol{e})_t$, $t = 1, 2, \ldots., N_{tr}$. Then, the original model $\boldsymbol{g_i(a, e, \theta)}$ is evaluated on the inputs $(\boldsymbol{a}, \boldsymbol{e})_t$, $t = 1, 2, \ldots., N_{tr}$, to obtain the corresponding outputs $\boldsymbol{g}_t = \boldsymbol{g(a, e, \theta)}$, $t = 1, 2, \ldots, N_{tr}$, and build the data sets $D_{tr,g} = \{(\boldsymbol{a}_t, \boldsymbol{e}_t, \boldsymbol{g}_t): t = 1, 2, \ldots, N_{tr}\}$ and $D_{tr,w} = \{(\boldsymbol{a}_t, \boldsymbol{e}_t, w_t): t = 1, 2, \ldots, N_{tr}\}$. Finally, the adjustable internal parameters of the ANN regression model are calibrated to fit the generated data: in particular, the common error back-propagation algorithm is implemented to *train* the

ANN. In the present case study, the number of inputs to each ANN regression model is equal to $n_{inp} = n_a + n_e = 5 + 4 = 9$ (i.e., the number of aleatory and epistemic variables), whereas the number of outputs is equal to $n_{out} = 1$ (i.e., one ANN is built for each requirement). The number of nodes n_h in the two hidden layers has been set equal to 50 by trial and error. *Validation* data sets $D_{val,g} = \{(\boldsymbol{a}_t, \boldsymbol{e}_t, \boldsymbol{g}_t): t = 1, 2, \ldots, N_{val} = 40000\}$ and $D_{val,w} = \{(\boldsymbol{a}_t, \boldsymbol{e}_t, w_t): t = 1, 2, \ldots, N_{val} = 40000\}$ (different from the training sets D_{train}) are used to monitor the accuracy of the ANN model during the training procedure and to stop it in case of *overfitting*. The time needed to train the ANN is approximately 5 hours on an Intel(R) Xeon(R) E5-2637 v3 CPU@3.50GHz. For a realistic measure of the ANN model accuracy, the widely adopted coefficient of determination R^2 and the RMSE are computed for each output $\{g_i: i = 1, 2, \ldots, n_g = 3\}$ on new data sets $D_{test,g} = \{(\boldsymbol{a}_t, \boldsymbol{e}_t, \boldsymbol{g}_t): t = 1, 2, \ldots, N_{test}\}$ and $D_{test,w} = \{(\boldsymbol{a}_t, \boldsymbol{e}_t, w_t): t = 1, 2, \ldots, N_{test}\}$ of size $N_{test} = 20000$, not used during training. The values of R^2 associated to the *final* estimates of each output $\{g_l: l = 1, 2, \ldots, n_g = 3\}$ and of the worst-case requirement metric w of interest, computed on the test set D_{test} are: 0.9976, 0.9968, 0.9972 and 0.9901; the RMSEs are 0.0013, 0.0051, 0.0013 and 0.0122, respectively. The large values of R^2 (i.e., larger than 0.99), and the small values of the RMSEs lead us to assert that the accuracy of the ANN model can be considered satisfactory for the needs of capturing the global behavior of the highly nonlinear and non-monotonic functions $\boldsymbol{g_i(a, e, \theta)}$ and $\boldsymbol{w(a, e, \theta)}$ and, thus, of estimating the corresponding failure probabilities. Notice that the use of a dedicated ANN for $\boldsymbol{w(a, e, \theta)}$ is *not* strictly *necessary*: the maximum of the outputs of the ANNs built for the three g_i's could be actually used. However, our choice was motivated by the advantage of directly training an ANN using *real* model outputs, instead of *possibly cumulating* regression errors coming from three different (*approximate*) surrogate models.

In order to validate *a posteriori* the results obtained using the ANN meta-model in the optimization search relying on the first refined uncertainty model UM-1(y), the optimal (epistemic) vectors $\boldsymbol{e}$ thereby found are sent in input to the *real* system models and the corresponding intervals $R_i(\boldsymbol{\theta})$ and $R(\boldsymbol{\theta})$ are calculated using $N_a = 500000$ aleatory samples drawn from the calibrated $f_{\boldsymbol{a}} = f_{\boldsymbol{a}}^{GM}(\boldsymbol{a}|\boldsymbol{\varphi}_{\boldsymbol{a}}^{MLE})$. Finally, in order to take into account the *statistical variability* in the failure probability estimates (obtained by plain random sampling), the upper and lower bounds of the corresponding intervals are 'extended' above and below, respectively, of an amount equal to two standard deviations: the final 'conservative' estimates are reported in Table 2. For the baseline design ($\boldsymbol{\theta} = \boldsymbol{\theta_{base}}$) the requirement violated with the highest probability is g_2, i.e., the one related to the settling time (its upper bound is 0.2980). The same way of proceeding and the same approaches are adopted for the estimation of the *severity* $s_i(\boldsymbol{\theta})$ of each individual requirement violation (defined as the expected value of each requirement g_i *conditional* to failure):

$$s_i(\boldsymbol{\theta}) = \max_{\boldsymbol{e} \in E} \mathbb{E}[g_i(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) | g_i(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) \geq 0] \cdot P[g_i(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) \geq 0], \; i = 1, \ldots, n_g = 3 \tag{17}$$

The corresponding values are also reported in Table 2: for the baseline design ($\boldsymbol{\theta} = \boldsymbol{\theta_{base}}$) the requirement presenting the most severe violation is g_3, i.e., the one related to energy consumption ($s_3(\boldsymbol{\theta_{base}}) = 0.0823$).

The epistemic parameters $\boldsymbol{e}$ are then ranked according to the *reduction* in the length $L[R(\boldsymbol{\theta})]$ of $R(\boldsymbol{\theta})$ that might result from their refinement. As before, each e_i is fixed at different values e_i^k, $k = 1, 2, \ldots, N_e = 20$, within its range of variation $[\underline{e_i}, \overline{e_i}]$. In correspondence of each e_i^k the length $L[R(\boldsymbol{\theta}|e_i = e_i^k)]$ of the worst-case failure probability range is computed (for the sake of compact notation, let $L[R(\boldsymbol{\theta}|e_i = e_i^k)]$ be indicated as $L[R(\boldsymbol{\vartheta}|e_i)]$). The corresponding sensitivity indicator $S_i(L)$ for parameter e_i is then computed as:

$$S_i(L) = 1 - \frac{E_{ei}\left[L[R(\boldsymbol{\theta}|e_i)]\right]}{L[R(\boldsymbol{\theta})]}, \tag{18}$$

quantifying the *expected fractional contraction* in $R(\boldsymbol{\theta})$ that would result from a refinement in e_i. The values obtained for $S_i(L)$ are shown in Table 3, together with the corresponding epistemic parameter ranking.

	First Refined Uncertainty Model (UM-1(y), after Subproblem B)						
Design	$R_1(\boldsymbol{\theta})$	$R_2(\boldsymbol{\theta})$	$R_3(\boldsymbol{\theta})$	$R(\boldsymbol{\theta})$	$s_1(\boldsymbol{\theta})$	$s_2(\boldsymbol{\theta})$	$s_3(\boldsymbol{\theta})$
$\boldsymbol{\theta}_{base}$	[0.0577, 0.1459]	[0.0680, 0.2980]	[0.0548, 0.0827]	[0.1377, 0.3318]	0.02103	0.00247	0.0823
$\boldsymbol{\theta}_{new}$	[0, $3.0404\cdot10^{-5}$]	[0.0134, 0.0617]	[0.0018, 0.0267]	[0.0155, 0.0629]	$3.6636\cdot10^{-7}$	$1.9851\cdot10^{-4}$	0.0058
	Final Refined Uncertainty Model (UM-*final*(yz), after Subproblem E)						
Design	$R_1(\boldsymbol{\theta})$	$R_2(\boldsymbol{\theta})$	$R_3(\boldsymbol{\theta})$	$R(\boldsymbol{\theta})$	$s_1(\boldsymbol{\theta})$	$s_2(\boldsymbol{\theta})$	$s_3(\boldsymbol{\theta})$
$\boldsymbol{\theta}_{base}$	[0.1429, 0.1954]	[0.1441, 0.1771]	[0.1224, 0.1393]	[0.2578, 0.2950]	0.0335	0.0026	0.1458
$\boldsymbol{\theta}_{new}$	[$2.522\cdot10^{-6}$, $1.559\cdot10^{-4}$]	[0.0479, 0.0820]	[0.0029, 0.0220]	[0.0487, 0.0821]	$4.4809\cdot10^{-6}$	$3.6637\cdot10^{-4}$	0.0023
$\boldsymbol{\theta}_{final}$	[$1.6010\cdot10^{-4}$, 0.0039]	[0.0246, 0.0527]	[0.0091, 0.0168]	[0.0280, 0.0540]	$7.3870\cdot10^{-5}$	$1.2187\cdot10^{-4}$	0.0028

Table 2. Reliability metrics $R_i(\boldsymbol{\theta})$, $R(\boldsymbol{\theta})$ and $s_i(\boldsymbol{\theta})$, $i = 1, 2$, $n_g = 3$, for different designs $\boldsymbol{\theta}$ and UMs

	Epistemic parameters importance: $S_i(L)$ (Ranking)		
	Uncertainty Models-UMs		
Parameter	First Refined UM (UM-1(y)), $\boldsymbol{\theta}_{base}$	First Refined UM (UM-1(y)), $\boldsymbol{\theta}_{new}$	Final Refined UM (UM-*final*(yz)), $\boldsymbol{\theta}_{final}$
e_1	0.4726 (2)	0.3465 (2)	0.1506 (3)
e_2	0.6084 (1)	0.3558 (1)	0.0502 (4)
e_3	0.2725 (3)	0.3209 (3)	0.4477 (1)
e_4	0.0809 (4)	0.1326 (4)	0.3724 (2)

Table 3. Ranking of the epistemic parameters according to their ability in reducing the length of $R(\boldsymbol{\theta})$, for different design configurations $\boldsymbol{\theta}$ and uncertainty models

3.4 Subproblem D: Reliability-Based Design

The objective is to identify a new design point $\boldsymbol{\theta}_{new}$ to improve the system's reliability. The optimality criterion here chosen is that of a *robust design*, i.e., we seek to *minimize* the (epistemic) *upper bound* of the failure probability for the *worst-case* performance function $w(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta})$:

$$\boldsymbol{\theta}_{new} = arg\min_{\boldsymbol{\theta}} \left\{ \max_{e \in E} P[w(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) \geq 0] \right\}. \tag{19}$$

This choice is motivated by the fact that the resulting design should be able to properly withstand *all* the aleatory and epistemic uncertainties, including *unexpected* events possibly occurring in the life of the system. The main drawback is that such design may be penalized by *extreme* (typically *very unlikely*) events (worst-case scenarios) and by *outliers*, which may lead to an *over-conservatism* (and, in practice, to excessive costs).

Since the design variables can range over the entire real axis, the following *iterative* optimization algorithm is implemented to find $\boldsymbol{\theta}_{new}$ [15]:

1. Set the current value of the system design as $\boldsymbol{\theta}_{new}^{curr} = \boldsymbol{\theta}_{base}$.
2. Define a (*local*) optimization search space $\left[\underline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}\right]$ as a hypercube centered around $\boldsymbol{\theta}_{new}^{curr}$ whose sides has a length which is a fraction of the absolute value of the current design: in details, $\left[\underline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}\right] = \boldsymbol{\theta}_{new}^{curr} \pm k \cdot |\boldsymbol{\theta}_{new}^{curr}|$, with k typically equal to 0.2-0.5. The value of k is progressively reduced as the iterations proceed and the convergence to the optimum is attained.
3. To reduce the computational cost due to the numerous model evaluations required by the (two-level) optimization problem (19), a surrogate regression model is used. In this case, an ANN is trained on the (*local*) optimization search space $\left[\underline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}\right]$ to reproduce the relationship between the (eighteen-dimensional) input space, made of the aleatory ($\boldsymbol{a} \in \Re^5$), epistemic ($\boldsymbol{e} \in \Re^4$) and design ($\boldsymbol{\theta} \in \Re^9$) variables, and the one-dimensional worst-case performance function $w(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta})$. The behavior of $w(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta})$ in the space of the design variables $\boldsymbol{\theta}$ is stepwise, abrupt, highly nonlinear, and strongly non-monotonic, which makes ANN training difficult. Very large-

sized training, validation and test sets D_{tr} = {($\boldsymbol{a}_t$, $\boldsymbol{e}_t$, $\boldsymbol{\theta}_t$, w_t): t = 1, 2, …, N_{tr} = 600000}, D_{val} = {($\boldsymbol{a}_t$, $\boldsymbol{e}_t$, $\boldsymbol{\theta}_t$, w_t): t = 1, 2, …, N_{val} = 200000} and D_{test} = {($\boldsymbol{a}_t$, $\boldsymbol{e}_t$, $\boldsymbol{\theta}_t$, w_t): t = 1, 2, …, N_{test} = 100000}, respectively, are employed to build a four-layered ANN with 18 inputs, one output and 54 neurons for each of the two hidden layers. In this configuration the training of an ANN may last up to 10 hours on an Intel(R) Xeon(R) E5-2637 v3 CPU@3.50GHz. An additional consideration is in order. The estimation of the worst-case failure probability requires the ANN to be able to accurately discriminate between safe and failed system configurations. Such failure probability becomes smaller and smaller as the iterations of the algorithm proceed: thus, the *relative* impact of ANN regression errors may become more and more important, making the design optimization search less effective. This issue is addressed at each iteration by generating training, validation and test sets that lie preferably *across* the system *failure threshold* (i.e., where $\boldsymbol{w(a, e, \theta)}$ = 0): by so doing, the ANN becomes more and more specialized in mimicking these system configurations, thus providing very accurate predictions around the failure limit, at the expense of less satisfactory estimations in less interesting (i.e., safe) regions of the design space. The principle is inspired by Ref. [29], but it is practically implemented according to a different (less rigorous) *empirical* procedure: i) around 33% of the total (N_{tr} + N_{val} + N_{test}) = 900000 training, validation and test patterns are obtained by uniformly sampling A_0, E and $\left[\underline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}\right]$; ii) the generated patterns are sorted according to the corresponding values of the worst-case performance function; iii) the failure points (and, in general, those system configurations characterized by comparatively high values of $\boldsymbol{w(a, e, \theta)}$ close to the failure threshold) are selected as "*seeds*" to start a sort of Markov *chain* and produce the remaining (64%) patterns; iv) a *local search* is performed to produce chains of system configurations $\boldsymbol{(a, e, \theta)}$ by sampling uniform (proposal) distributions iteratively centered on the elements of the chains, and with interval length equal to twice the sample standard deviation of the seeds identified above.

4. Use GAs and the trained ANN to find an (approximate) updated optimal design point, i.e., $\boldsymbol{\theta}_{new}^{updated} = \arg \min_{\boldsymbol{\theta} \in \left[\underline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}\right]} \left\{ \max_{\boldsymbol{e} \in E} P[w(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) \geq 0] \right\}$. As before, the adoption of global, gradient-free heuristic optimization tools is motivated by the abrupt and high-dimensional nature of the (design and epistemic) search spaces. The number N_a of aleatory samples drawn for probability estimation is 100000; the size N_{pop} of the GA populations used to explore the design and epistemic spaces is 100 and 20, respectively; finally, the number of GA generations N_{gen} is set to 150 and 25 for the design and epistemic searches, respectively.
5. The solution $\boldsymbol{\theta}_{new}^{updated}$ thereby identified is checked: if at least one of the design variables lies at the boundary of the (local) search space $\left[\underline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}\right]$ (i.e., if $\theta_{i,new}^{updated} = \underline{\theta_i}$ or $\theta_{i,new}^{updated} = \overline{\theta}_i$, for at least one i = 1, 2, …, 9), then the iterative algorithm continues: set $\boldsymbol{\theta}_{new}^{curr} = \boldsymbol{\theta}_{new}^{updated}$ and go to step 2. Otherwise, the algorithm is stopped: go to step 6.
6. Set the final new design $\boldsymbol{\theta}_{new} = \boldsymbol{\theta}_{new}^{updated}$.

The new design vector resulting from the optimization after 8 iterations is $\boldsymbol{\theta}_{new}$. The corresponding reliability metrics $R_i(\boldsymbol{\theta})$, $R(\boldsymbol{\theta})$ and $s_i(\boldsymbol{\theta})$, $i = 1, 2, n_g = 3$ obtained with $\boldsymbol{\theta}_{new}$ and UM-1(y) are reported in Table 2. It can be observed a considerable reduction in the upper bound of the system failure probabilities and a substantial improvement in the requirement violation severity. R_1 is reduced by four orders of magnitude, while R_2, R_3 and R are reduced by factors 3.1-5.3. Notice that the most important contributor to system failure is still g_2. The severity of the violations has been reduced by 4 orders of magnitude for g_1 and by about 1 order of magnitude for g_2 and g_3. Again, the highest violation severity is for requirement g_3. The

epistemic parameter ranking (Table 3) is the same as that of $\boldsymbol{\theta}_{base}$. However, the indicators $S_i(L)$ for e_1, e_2 and e_3 have now a *comparable* magnitude.

Finally, to highlight the relevance of the use of ANNs, the design optimization process (19) is repeated using only the original system model. Notice however that the use of the real model with the same parameter settings as those specified at step 4. above would obviously lead to an impracticable effort. To carry out a relatively fair comparison, at *each iteration* of the algorithm the *total number* of system *model evaluations* and the overall *computational time* are kept larger than or equal to those resulting from the ANN-based procedure (i.e., N_{tr} + N_{val} + N_{test} = 900000 and 10 hours on average for ANN training, respectively). To this aim, the algorithm parameters are adjusted as follows: i) the total number of iterations is fixed to 8 (i.e., the value resulting from the iterative procedure described); ii) N_a = 250 aleatory samples $\boldsymbol{a}$ are drawn for each vector $\boldsymbol{e}$ of the epistemic parameters; iii) the epistemic space E is explored by N_e = 100 vectors $\boldsymbol{e}$ (including the *vertices* of the box E) for each design solution $\boldsymbol{\theta}$ proposed by the GA; finally, iv) the GA searches the design space evolving N_{pop} = 20 individuals for a maximum number of generations equal to N_{gen} = 20. Notice that the maximum number of model evaluations with such settings is 10^7 (larger than 900000); also, the average computational time *per iteration* results to be 13.1 hours (i.e., larger than the 10 hours approximately needed for ANN training). The optimized upper bound of $R(\boldsymbol{\theta}_{new})$ turns out to be 0.1145: in spite of the *larger* number of *model evaluations* (by a factor 11) and of the *higher computational cost* (by a factor 1.3), the *design performance* is *worse* almost by a factor 2 (i.e., 0.1145 versus 0.0629). This confirms the advantage of using ANNs for reliability-based design optimization in the presence of both aleatory and epistemic uncertainties (and of computationally intensive models).

3.5 Subproblem E: Model Update and Design Tuning

Upon finding a new design point $\boldsymbol{\theta}_{new}$, Subproblem E requires performing a final improvement to the UM and design. After providing $\boldsymbol{\theta}_{new}$ to the Challenge hosts, n_2 = 100 realizations of the responses $z_1(t)$ and $z_2(t)$ from the integrated system has been provided for calibration (dataset $\boldsymbol{D}_2 = \{z_1^{(i)}(t), z_2^{(i)}(t)\}$, i = 1, 2, …, n_2 = 100). Using the same procedure based on SVD of Section 3.1, the dataset $\boldsymbol{D}_2$ is projected onto an orthonormal basis to obtain the coefficients/projections $\boldsymbol{C}_2$. The ε = 99% of the total variance of the dataset is retained in the SVD process, which results in $n_B(z_1)$ = 14 bases for $z_1(t)$ and $n_B(z_2)$ = 4 bases for $z_2(t)$. Kernel Density Estimation (KDE) is then employed to fit the likelihood $L_{\boldsymbol{H}^Z}^{GM}(\boldsymbol{C_2}|\boldsymbol{\Phi})$ of the data in an SVD space of $n_B(z_1)$ + $n_B(z_2)$ = 17 dimensions. Since datasets $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ (resp., $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ in the SVD space) are generated independently, the overall likelihood $L_{\boldsymbol{H}^{YZ}}^{GM}(\boldsymbol{C_1}, \boldsymbol{C_2}|\boldsymbol{\Phi})$ of the entire set of n_1 + n_2 = 200 realizations is obtained as $L_{\boldsymbol{H}^Y}^{GM}(\boldsymbol{C_1}|\boldsymbol{\Phi}) \cdot L_{\boldsymbol{H}^Z}^{GM}(\boldsymbol{C_2}|\boldsymbol{\Phi})$ and employed in the two-step MLE-base algorithm of Section 3.1 to update the UM (step (E.2) of the Challenge). Notice that in this update the epistemic space E remains fixed to E = $[\underline{e}_1, \overline{e}_1]$ x $[\underline{e}_2, \overline{e}_2]$ x $[\underline{e}_3, \overline{e}_3]$ x $[\underline{e}_4, \overline{e}_4]$ = [0, 0.3719] x [0.4064, 0.7664] x [0.0330, 0.3330] x [0, 2]. The resulting change in the (aleatory) UM $f_{\boldsymbol{a}}^{GM}$ is also minor and is not pictorially shown due to space limitations. Let us denote this uncertainty model as UM-1(yz). It is worth noting that the predictive capability (Energy Score-ES) of the resulting UM-1(yz) is 21.48 with respect to dataset $\boldsymbol{D}_1$ (i.e., almost the same as UM-1(y)); also, the ES of UM-1(yz) with respect to dataset $\boldsymbol{D}_2$ is 30.47 (notice that the ES of the previous UM-1(y) tested on dataset $\boldsymbol{D}_2$ is 31.89).

The sensitivity analysis methods proposed in Section 3.2 are applied to request two additional parameter refinements for E. Based on the results reported in Table 4 (variance-base sensitivity analysis) and Figure 3 (conditional Energy Score), a reduction in the upper bound of e_1 and e_2 is requested. The refined epistemic intervals provided by the Challengers are e_1 = [0.5240, 0.9240] and e_2 = [0.4873, 0.5449]: the corresponding epistemic space is denoted as E_2.

	Variance-based Sensitivity Analysis – Factor Prioritization

	Second Refined Uncertainty Model (UM-1(*yz*), after Subproblem Task E.2)	
Parameter (MLEs)	$S_i(U_e)$	Ranking
e_1 (0.1233)	$2.38 \cdot 10^{-3}$	2
e_2 (0.5035)	$4.58 \cdot 10^{-3}$	1
e_3 (0.3204)	$8.60 \cdot 10^{-4}$	3
e_4 (1.6287)	$1.12 \cdot 10^{-5}$	4

Table 4. Ranking of the capability of the e_i's to reduce the output spread, obtained using UM-1(yz)

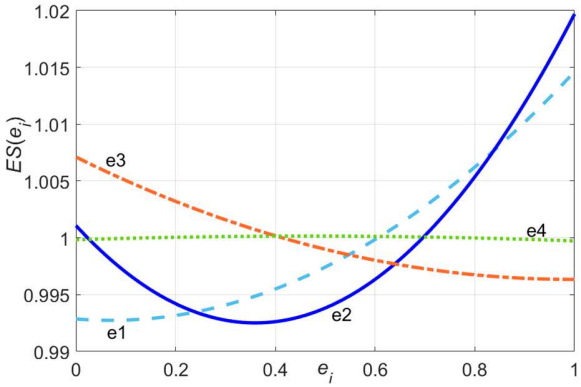

Figure 3. Conditional Energy Score $\boldsymbol{ES(e_i) = ES(e_i = e_i^k)}$, obtained using UM-1(yz)

It is evident that the range of e_1 does *not* even overlap with the one resulting from the calibration of Subproblem A and from the corresponding update of step (E.2). The two-step calibration process of Section 3.1 is repeated in the light of this new information to obtain the final UM, denoted as UM-*final*(*yz*). Figure 4 shows the result of Step 1 of the calibration process: the marginal PDFs (histograms) of a_i, i = 1, 2, …, 5, are plotted together with the two-dimensional projections of the joint (five-dimensional) PDF $f_{a,final}^{GM}(\boldsymbol{a}|\boldsymbol{\varphi}_{\boldsymbol{a,final}}^{MLE})$ for pairs (a_i, a_j). The *final* uncertainty model (hyper-rectangle) E_{final} = $[\underline{e}_1, \overline{e}_1]$ x $[\underline{e}_2, \overline{e}_2]$ x $[\underline{e}_3, \overline{e}_3]$ x $[\underline{e}_4, \overline{e}_4]$ for the epistemic parameters $\boldsymbol{e}$ chosen according to Step 2 of the calibration algorithm is modified as follows: $[\underline{e}_1, \overline{e}_1]$ = [0.5240, 0.7202], $[\underline{e}_2, \overline{e}_2]$ = [0.4873, 0.5449], $[\underline{e}_3, \overline{e}_3]$ = [0.0330, 0.3330] and $[\underline{e}_4, \overline{e}_4]$ = [0, 2] (i.e., the width of the interval of e_1 has been further reduced).

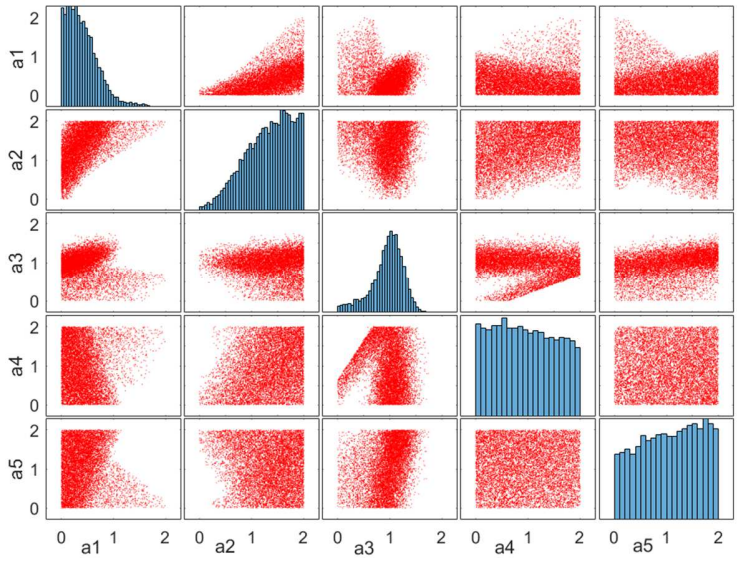

Figure 4. Final Calibrated aleatory model $f_{a,final}^{GM}(\boldsymbol{a}|\boldsymbol{\varphi}_{\boldsymbol{a,final}}^{MLE})$

Figure 5 shows the calibrated model output $\boldsymbol{z}(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}_{new}, t)$ against the data provided $\boldsymbol{D}_2 = \{z_1^{(i)}(t), z_2^{(i)}(t)\}$, $i = 1, 2, \ldots, n_2 = 100$). We report the time series observations (red solid lines) along with the upper and lower *bounds* (blue dashed lines) resulting from the propagation through the model functions $z_1(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}_{new}, t)$ (left) and $z_2(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}_{new}, t)$ (right) of 500000 configurations $(\boldsymbol{a}_i, \boldsymbol{e}_l)$, $i = 1, 2, \ldots, 500000$ (i.e., $N_a = 2000$ aleatory samples for $N_e = 250$ epistemic vectors, including the *vertices* of the box E). Also, the overall *mean* of the calibrated output (averaged over the final aleatory PDF $f_{a,final}^{GM}(\boldsymbol{a}|\boldsymbol{\varphi}_{\boldsymbol{a,final}}^{MLE})$ and over the final epistemic space E) is shown as black crosses. It can be seen that the calibrated model envelops the data provided at most time instants; however, it is also evident the over-conservatism in the assessment of the epistemic uncertainty.

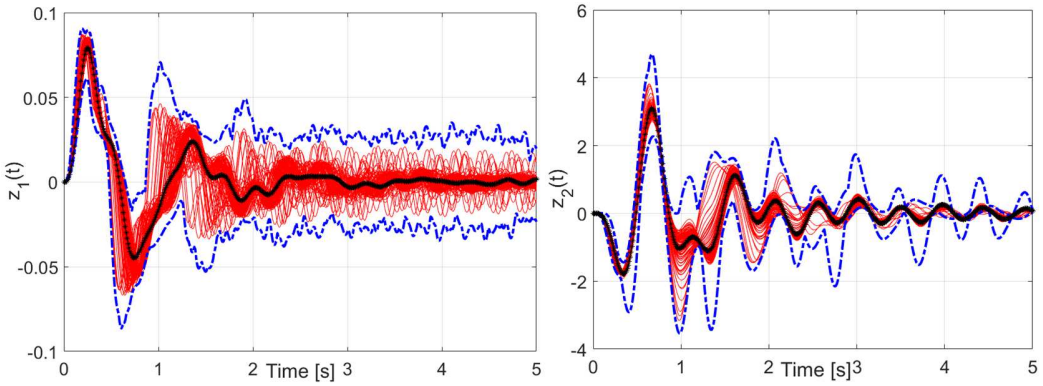

Figure 5. Bounds on the calibrated model output $z_1(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}_{new}, t)$ (left) and $z_2(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}_{new}, t)$ (right) (blue dashed lines) against experimental data (red solid lines)

It very important to notice that the resulting final model UM-*final*(*yz*) shows a comparatively satisfactory predictive capability with respect to dataset $\boldsymbol{D}_2$: the ES is equal to 28.01, which means an improvement of 8.07% with respect to UM-1(*yz*). However, it presents a poor performance with respect to dataset $\boldsymbol{D}_1$: the corresponding ES is 24.27, which is far larger than that of UM-1(*yz*) (in particular, the corresponding predictive capability decreases of about 13%). This may be due to the *combination* of two factors: (i) the proposed *parametric* aleatory model based on Gaussian Mixtures largely fails to capture the relationships and complex dependences between variables $\boldsymbol{a}$; (ii) non-negligible *discrepancies* (i.e., differences between model outputs and observation data) may affect the computational models of the subsystem and of the integrated system provided by the Challengers. The issue of discrepancy is not addressed in this work.

For the sake of *comparison*, the energy scores quantifying the predictive capability of all the different generated UMs on the two datasets $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are summarized in Table 5 (notice that obviously these figures are *not* suitable to quantify the *absolute* performance of the models). It can be seen a progressive *improvement* of the UMs with respect to the data coming from the *integrated system*, while a worsening is registered with respect to the observations from the subsystem.

	Energy Scores (ES) – Predictive capability	
Uncertainty models	**Dataset D_1**	**Dataset D_2**
UM-0(*y*) (initial)	22.49	/
UM-1(*y*) (after Subproblem B)	21.28	31.89
UM-1(*yz*) (after Subtask E.2)	21.48	30.47
UM-final(*yz*) (final, after subtask E.4)	24.27	28.01

Table 5. Energy Scores (predictive capabilities) of the different UMs

In the light of these results, the following strategy is implemented. Since the performance of UM-*final*(*yz*) with respect to $\boldsymbol{D}_2$ (*integrated system*) is comparatively *acceptable*, the remaining tasks of the Challenge are addressed with the identified final UM. Actually, since all the performance metrics (i.e., failure probabilities and violation severities) are based on the model of the integrated system $\boldsymbol{z}(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}, t)$, the results obtained in the following tasks may be meaningful (even in the presence of a *globally* unsatisfactory UM).

The UM-*final*(*yz*) $\langle f_{a,final}^{GM}(\boldsymbol{a}|\boldsymbol{\varphi}_{\boldsymbol{a,final}}^{MLE}), E_{final}\rangle$ is used within the robust design optimization approach of Section 3.4 to obtain the final design $\boldsymbol{\theta}_{final}$. The corresponding reliability metrics are reported in Table 2. There is a strong reduction in the upper bound of $R_1(\boldsymbol{\theta})$ (by 2-3 orders of magnitude) with respect to the baseline design $\boldsymbol{\theta}_{base}$. Also, the upper bounds of $R_2(\boldsymbol{\theta})$, $R_3(\boldsymbol{\theta})$ and $R(\boldsymbol{\theta})$ are reduced by factors 3.4-8.3, while the violation severities $s_i(\boldsymbol{\theta})$ are reduced by factors 450 (g_1), 21 (g_2) and 52 (g_3) with respect to $\boldsymbol{\theta}_{base}$. Finally, from the ranking of Table 3 there is radical change in the importance of the parameters, mainly due to the final refinement in the UM: in particular, e_3 and e_4 are now the most effective in reducing the epistemic uncertainty of the worst-case failure probability $R(\boldsymbol{\theta})$.

3.6 Subproblem F: Risk-Based Design

In this subproblem we seek a design point that accounts for most of the final epistemic space E_{final}. A portion corresponding to the *r*% of the volume of E_{final} has to be neglected, where *r* ranges in [0, 100) and is called *risk*. In practice, by reducing the epistemic space (i.e., by neglecting a portion of it) we are accepting that some (future) system's configurations (i.e., scenarios) may fall outside such a set, i.e., we are increasing the risk that the system is not able to withstand such "outliers". However, if this risk is relatively small and the epistemic volume to be neglected is optimally chosen, this practice might be advantageous. For example, a design based upon an epistemic box enclosing 95% of the scenarios may show better overall performance (e.g., smaller failure probability) than one enclosing 100% [16].

Let $V(E_{final})$ be the volume of the final epistemic box; $E_{final}(r)$ the final epistemic space where the *r*% of its volume has been neglected; $V(E_{final}(r))$ the volume of the reduced epistemic space. Coherently with the approach of Section 3.4, the metric proposed to quantify the gain $l(r, \boldsymbol{\theta})$ resulting from taking the risk *r* (and pertaining to the improved performance of the retained (100 – *r*)% of the epistemic space) is based on the upper bound of the worst-case failure probability $R(\boldsymbol{\theta})$. In particular, $l(r, \boldsymbol{\theta})$ evaluates the *relative decrease* that we obtain in the epistemic upper bound of the overall system failure probability thanks to the reduction in the volume of the epistemic space:

$$l(r,\boldsymbol{\theta}) = \frac{\max\limits_{\boldsymbol{e}\in E_{final}} P[w(\boldsymbol{a},\boldsymbol{e},\boldsymbol{\theta})\geq 0] - \max\limits_{\boldsymbol{e}\in E_{final}(r)} P[w(\boldsymbol{a},\boldsymbol{e},\boldsymbol{\theta})\geq 0]}{\max\limits_{\boldsymbol{e}\in E_{final}} P[w(\boldsymbol{a},\boldsymbol{e},\boldsymbol{\theta})\geq 0]} \tag{20}$$

The portion of the epistemic space to be ignored is chosen according to the following criteria:

a) The neglected portion should be characterized by comparatively *small likelihood*, i.e., the "scenarios" left out should have a low probability of occurring in the future;
b) The neglected portion should be obtained by "manipulating" those epistemic parameters that have the *strongest impact* on the (epistemic uncertainty of) the performance of the system (i.e., in its reliability/failure probability): in that case, even small variations in the bounds of one epistemic parameter are expected to have an important effect on (i.e., improvement in) the overall performance of the system (i.e., on the performance gain (20));
c) To obtain the best possible reduction of the epistemic space, the variations in the epistemic parameter bounds should be themselves optimized during the gain optimization process.

The practical impact (and the corresponding implementation) of the criteria above is given in what follows. Figure 6 shows four one-dimensional projections of the overall likelihood $L_{H^{YZ}}^{GM}\left(\boldsymbol{C_1}, \boldsymbol{C_2} | \boldsymbol{e}, \boldsymbol{\varphi} = \boldsymbol{\varphi}_{a,final}^{MLE}\right) = L_{H^{Y}}^{GM}\left(\boldsymbol{C_1} | \boldsymbol{e}, \boldsymbol{\varphi} = \boldsymbol{\varphi}_{a,final}^{MLE}\right) \cdot L_{H^{Z}}^{GM}\left(\boldsymbol{C_2} | \boldsymbol{e}, \boldsymbol{\varphi} = \boldsymbol{\varphi}_{a,final}^{MLE}\right)$ (one for each epistemic parameter). It is evident that according to criterion a) above, we would like to reduce the volume of epistemic space by decreasing the upper bound of e_1 and/or increasing the lower bounds of e_2, e_3 and e_4 (i.e., by removing the portions characterized by the smaller "likelihood mass"). In addition, it is evident from the last column of Table 3 (reporting the epistemic ranking obtained in correspondence of $\boldsymbol{\theta}_{final}$ with the final uncertainty model UM-*final*(*yz*)) that the parameters having the *largest impact* on the epistemic uncertainty of the *system failure probability* are e_3 and e_4. Finally, the sensitivity analysis reported in Figure 7 (showing the variation of the upper bound of $R(\boldsymbol{\theta}_{final})$ obtained by fixing the epistemic parameters at some values in their ranges) demonstrates that an *increase* in the lower bounds of e_3 and e_4 leads to a *considerable reduction* in the upper bound of the system failure probability, whereas the effect of e_1 and e_2 is less evident. Based on these considerations, we decide to reduce the volume of the epistemic space by *optimally manipulating* (in particular, *increasing*) *only* the position of the lower bounds of e_3 and e_4. In practice, the lower bounds $\underline{e}_3$ and $\underline{e}_4$ are *optimized* together with the design variables $\boldsymbol{\theta}$ in the gain $l(r, \boldsymbol{\theta})$ maximization process:

$$\max_{\underline{e}_3, \underline{e}_4, \boldsymbol{\theta}} \{l(r, \boldsymbol{\theta})\} = \max_{\underline{e}_3, \underline{e}_4, \boldsymbol{\theta}} \left\{ \frac{\max_{\boldsymbol{e} \in E_{final}} P[w(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) \geq 0] - \max_{\boldsymbol{e} \in E_{final}(r)} P[w(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) \geq 0]}{\max_{\boldsymbol{e} \in E_{final}} P[w(\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\theta}) \geq 0]} \right\} \tag{21}$$

The optimization problem (21) is solved by means of GAs under the following (*hard*) *inequality constraint* that defines the risk-based design: $V(E_{final}(r)) \geq V(E_{final}) \cdot [(100 - r)\%]$ (i.e., the solutions which do *not* respect the risk-based requirement are *discarded* during the optimization search).

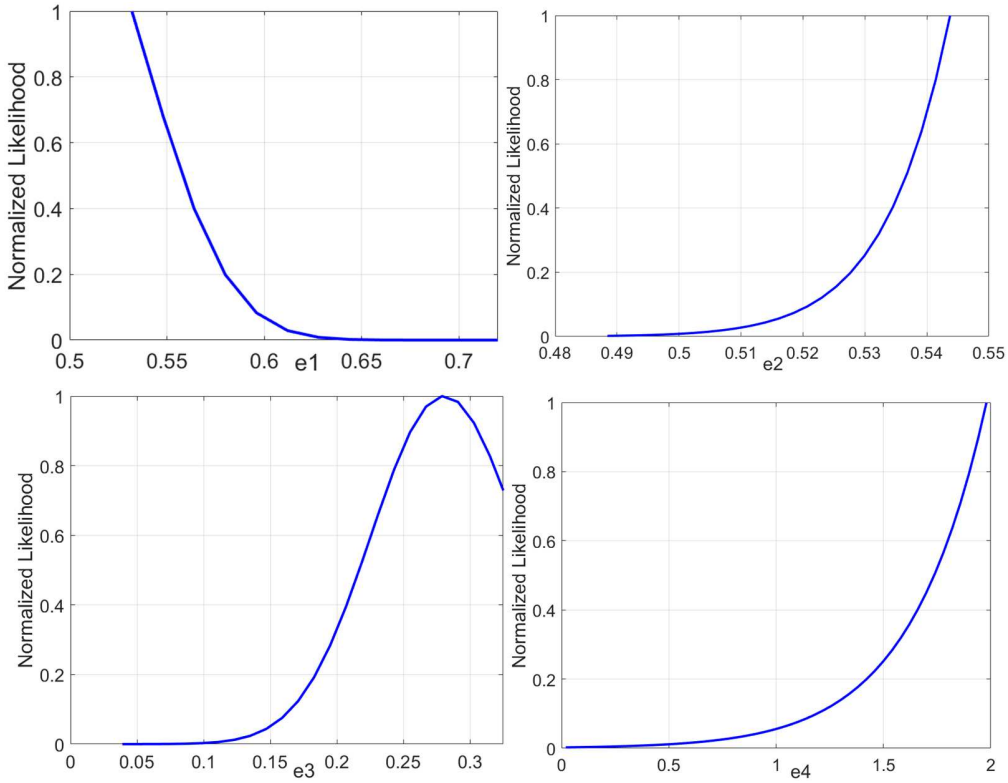

Figure 6. One-dimensional projections of the overall likelihood $L_{H^{YZ}}^{GM}\left(\boldsymbol{C_1}, \boldsymbol{C_2} | \boldsymbol{e}, \boldsymbol{\varphi} = \boldsymbol{\varphi}_{a,final}^{MLE}\right)$ as a function of each epistemic parameter e_1 (top, left), e_2 (top, right), e_3 (bottom, left) and e_4 (bottom, right)

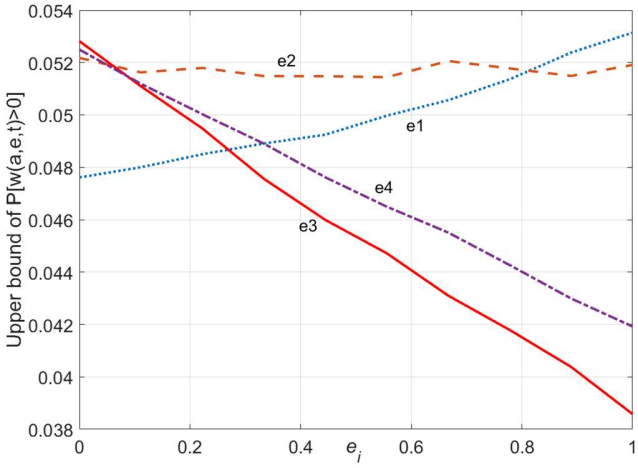

Figure 7. Variation of the upper bound of $R(\boldsymbol{\theta}_{final})$ obtained by fixing the epistemic parameters e_1, e_2, e_3 and e_4 at some values in their range (normalized between 0 and 1 for the sake of illustration)

For $\hat{r}$ = 5%, the design point $\boldsymbol{\theta}_{\hat{r}\%Risk}$ resulting from the optimization (21) is obtained. The corresponding performance metrics are compared to those of $\boldsymbol{\theta}_{final}$ in Table 6, for both the full and the reduced epistemic spaces E_{final} and $E_{final}(\hat{r})$. For the full epistemic space, the upper bound of $R(\boldsymbol{\theta}_{final})$ is 0.0540, whereas that of $R(\boldsymbol{\theta}_{\hat{r}\%Risk})$ is 0.0520: thus, the gain in performance is 3.7%. For the optimally reduced epistemic space, the upper bound of $R(\boldsymbol{\theta}_{final})$ is 0.0533, whereas that of $R(\boldsymbol{\theta}_{\hat{r}\%Risk})$ is 0.0505: in this case, the gain in performance is 5.25% (i.e., slightly larger than the fractional reduction in the volume). Overall, moving from the initial configuration (characterized by $\boldsymbol{\theta}_{final}$ and E_{final}) to the risk-based configuration (characterized by $\boldsymbol{\theta}_{\hat{r}\%Risk}$ and $E_{final}(\hat{r})$), we reduce the upper bound of $R(\boldsymbol{\theta})$ from 0.0540 to 0.0505, with a gain in performance of 6.48%.

Finally, in Figure 8 we compare the gains $l(r, \boldsymbol{\theta}_{final})$ (blue dashed line) and $l(r, \boldsymbol{\theta}_{\hat{r}\%Risk})$ (dotted red line) for different values of r = 0.05, 0.5, 1, …, 10. The black solid line indicates those points where the risk taken is equal to the gain. Notice that in determining the gain for each r, the reduced epistemic space $E_{final}(r)$ (i.e., the lower bounds of e_3 and e_4) is again optimized. The gain increases almost linearly with the risk (no optimal value is identified). However, we can consider advantageous to accept a risk-based design, when the percentage gain is *larger* than the corresponding risk. In this case, for $\boldsymbol{\theta}_{final}$ acceptable risk values range between 0.05% and 0.5%, whereas for $\boldsymbol{\theta}_{\hat{r}\%Risk}$ they lie between 0.05% and 2-3%.

	Risk-based Design						
	Final Refined Uncertainty Model (UM-*final*(*yz*), after Subproblem E) – Full Epistemic space E_{final}						
Design	$R_1(\boldsymbol{\theta})$	$R_2(\boldsymbol{\theta})$	$R_3(\boldsymbol{\theta})$	$R(\boldsymbol{\theta})$	$s_1(\boldsymbol{\theta})$	$s_2(\boldsymbol{\theta})$	$s_3(\boldsymbol{\theta})$
$\boldsymbol{\theta}_{final}$	[1.6010·10^{-4}, 0.0039]	[0.0246, 0.0527]	[0.0091, 0.0168]	[0.0280, 0.0540]	7.3870·10^{-5}	1.2187·10^{-4}	0.0028
$\boldsymbol{\theta}_{\hat{r}\%Risk}$	[1.9025·10^{-4}, 0.0068]	[0.0248, 0.0501]	[0.0079, 0.0149]	[0.0287, 0.0520]	8.5161·10^{-5}	1.1091·10^{-4}	0.0023
	Final Refined Uncertainty Model (after Subproblem E) – Reduced Epistemic space $E_{final}(\hat{r})$						
Design	$R_1(\boldsymbol{\theta})$	$R_2(\boldsymbol{\theta})$	$R_3(\boldsymbol{\theta})$	$R(\boldsymbol{\theta})$	$s_1(\boldsymbol{\theta})$	$s_2(\boldsymbol{\theta})$	$s_3(\boldsymbol{\theta})$
$\boldsymbol{\theta}_{final}$	[1.6010·10^{-4}, 0.0039]	[0.0246, 0.0517]	[0.0091, 0.0167]	[0.0280, 0.0533]	7.1389·10^{-5}	1.1821·10^{-4}	0.0028
$\boldsymbol{\theta}_{\hat{r}\%Risk}$	[1.9210·10^{-4}, 0.0075]	[0.0248, 0.0485]	[0.0079, 0.0149]	[0.0280, 0.0505]	8.3659·10^{-5}	1.0613·10^{-4}	0.0025

Table 6. Performances of $\boldsymbol{\theta}_{\hat{r}\%Risk}$ and $\boldsymbol{\theta}_{final}$ for the full and reduced epistemic spaces E_{final} and $E_{final}(\hat{r})$

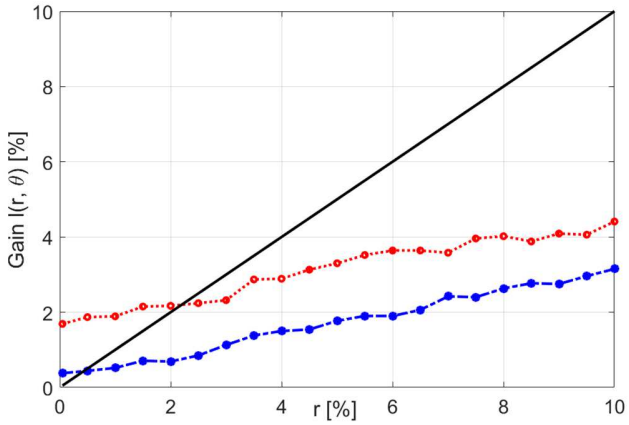

Figure 8. Gain $l(r, \boldsymbol{\theta}_{final})$ (blue dashed line) and $l(r, \boldsymbol{\theta}_{\hat{r}\%Risk})$ (dotted red line) for different values of r

4 Conclusions

In this paper, we have addressed the NASA Langley Uncertainty Quantification Challenge on Optimization under Uncertainty. We have proposed an efficient combination of methods to tackle the diverse issues posed by the Challenge:

- SVD to perform dimensionality reduction in the presence of high-dimensional (time series) data;
- a two-step MLE-based method relying on GMs for the calibration of multivariate probability distributions *and* interval sets;
- KDE for the construction of (very high-dimensional) non-parametric likelihood functions;
- Variance-based sensitivity analysis and the ES for uncertainty reduction and model refinement;
- Flexible sampling-based strategies (MCS) for the propagation of aleatory uncertainty;
- GAs as heuristic tools for solving complex, nonlinear optimization problems in the presence of abrupt, disconnected, stepwise search spaces;
- ANN metamodels for reducing the computational cost associated to uncertainty propagation and (iterative) robust design optimization.

The following issues and findings have emerged with respect to the proposed methods:

- Although flexible, sampling-based strategies are extremely computationally intensive, which can make optimization impracticable in the presence of both aleatory and epistemic uncertainties;
- Metamodel errors should be carefully controlled, in particular when they are employed for estimating small failure probabilities and/or for mapping high-dimensional spaces;
- Parametric probabilistic models likely fail to capture complex nonlinear dependences between aleatory variables;
- Robust designs may perform satisfactorily even in the presence of poorly calibrated models. On the other hand, they may be overly conservative (i.e., driven only by “outliers” and worst-case scenarios characterized by severe consequences but negligible likelihood).

Several issues and approaches are worth investigation in the future:

- Non-sampling strategies (e.g., bounding methods) [30, 31] could be adopted to estimate small probabilities at a manageable computational cost in optimization problems;
- More advanced methods should be investigated to model complex, nonlinear dependences: e.g., copulas [32, 33], fully non-parametric approaches based on KDE and Markov Chain Monte Carlo within a Bayesian framework [18] or Sliced Normal Distributions [16];

- Discrepancies between model outputs and observations from the real system should be included in the calibration process [34, 35]: failing to do so could lead to overly optimistic results;
- Rigorous methods (e.g., Scenario Theory) [36] should be embraced to design robust systems, while optimally controlling, selecting and possibly discarding outliers.

5 References

1. NASA, Risk-Informed Decision Making Handbook, NASA/SP-2010-576 – Version 1.0, April 2010.
2. J. C. Helton, J. D. Johnson, C. J. Sallaberry, C. B. Storlie, Survey of sampling-based methods for uncertainty and sensitivity analysis, Reliability Engineering & System Safety 91 (2006) 1175-1209.
3. E. Plischke, E. Borgonovo, C. L. Smith, Global Sensitivity Measures from Given Data, European Journal of Operational Research 226 (2013) 536–550.
4. L.G. Crespo, S.P. Kenny, 2021. The NASA Langley challenge on optimization under uncertainty. Mechanical Systems and Signal Processing 152, 107405.
5. A. Gaymann, M. Pietropaoli, L.G. Crespo, S.P. Kenny, F. Montomoli, Random Variable Estimation and Model Calibration in the Presence of Epistemic and Aleatory Uncertainties, SAE Int. J. Mater. Manuf. 11(4) (2018) 453-466.
6. X. Wu, T. Kozlowski, H. Meidani, Kriging-based inverse uncertainty quantification of nuclear fuel performance code BISON fission gas release model using time series measurement data, Reliab Eng Syst Saf 169 (2018) 422–36.
7. X. Wu, T. Kozlowski, H. Meidani, K. Shirvan, Inverse uncertainty quantification using the modular Bayesian approach based on Gaussian process, Part 1: Theory, Nucl Eng Des 335 (2018) 339–55.
8. R.D. Wilkinson, Bayesian Calibration of Expensive Multivariate Computer Experiments, in: L. Biegler et al. (Eds.), Large-Scale Inverse Problems and Quantification of Uncertainty, John Wiley & Sons, Chichester, UK, 2010, pp. 195–215.
9. E. Borgonovo, W. Castaings, S. Tarantola, Moment Independent Importance Measures: New Results and Analytical Test Cases, Risk Analysis 31(3) (2011) 404-428.
10. N. Pedroni, E. Zio, Hybrid Uncertainty and Sensitivity Analysis of the Model of a Twin-Jet Aircraft, Journal of Aerospace Information Systems 12 (2015) 73-96.
11. M. Faes, M. Broggi, E. Patelli, Y. Govers, J. Mottershead, M. Beer, D. Moens, A multivariate interval approach for inverse uncertainty quantification with limited experimental data, Mechanical Systems and Signal Processing 118 (2019) 534–548.
12. T. Gneiting, A:E. Raftery, Strictly Proper Scoring Rules, Prediction, and Estimation, Journal of the American Statistical Association 102(477) (2007) 359-378. Review Article.
13. X. S. Yang, Engineering optimization: an introduction with metaheuristic applications, Wiley, New York, NY, USA, 2010.
14. N. Pedroni, E. Zio, An Adaptive Metamodel-Based Subset Importance Sampling approach for the assessment of the functional failure probability of a thermal-hydraulic passive system, Applied Mathematical Modelling 48 (2017) 269-288.
15. E. Patelli, DA Alvarez, M. Broggi, M. de Angelis, Uncertainty management in multidisciplinary design of critical safety systems, Journal of Aerospace Information Systems, 12(1) (2015) 140 - 169.
16. L. G. Crespo, B. K. Colbert, S. P. Kenny, D. P. Giesy, 2019. On the quantification of aleatory and epistemic uncertainty using Sliced-Normal distributions, Systems & Control Letters 134, 104560.
17. B. W. Silverman, Density estimation for statistics and data analysis, Volume 26 of Monographs on Statistics and Applied Probability, Chapman & Hall/CRC, London, UK, 1996.
18. J. Goodman, J. Weare, Ensemble samplers with affine invariance, Communications in Applied Mathematics and Computational Science 5(1) (2010) 65–80.

19. Y.F. Li, N. Pedroni, E. Zio, A Memetic Evolutionary Multi-Objective Optimization Method for Environmental Power Unit Commitment, IEEE Transactions on Power Systems 28(3) (2013) 2660-2669.
20. C. Safta, K. Sargsyan, H. N. Najm, K. Chowdhary, B. Debusschere, L. P. Swiler, M. S. Eldred, Probabilistic Methods for Sensitivity Analysis and Calibration in the NASA Challenge Problem, Journal of Aerospace Information Systems 12(1) (2015) 219-234.
21. S. Bi, M. Broggi, M. Beer, The role of the Bhattacharyya distance in stochastic model updating, Mechanical Systems and Signal Processing 117 (2019) 437 – 452.
22. A. Marrel, N. Pérot, C. Mottet, Development of a surrogate model and sensitivity analysis for spatio-temporal numerical simulators, Stochastic Environmental Research and Risk Assessment 29(3) (2014) 959-974.
23. S. Nanty, C. Helbert, A. Marrel, N. Pérot, C. Prieur, Uncertainty quantification for functional dependent random variables, Computational Statistics 32(2) (2017) 559-583.
24. S. Ferson, W.T. Tucker, Sensitivity analysis using probability bounding, Reliability Engineering and System Safety 91 (2006) 1435–1442.
25. S. Bi, M. Broggi, P. We, M. Beer, The Bhattacharyya distance: enriching the P-box in stochastic sensitivity analysis, Mechanical Systems and Signal Processing 129 (2019) 265-281.
26. G.J. Székely, M.L. Rizzo, A New Test for Multivariate Normality, Journal of Multivariate Analysis 93 (2005) 58–80.
27. S. Nissen, Implementation of a Fast Artificial Neural Network Library (fann), Tech. rep., Department of Computer, Science University of Copenhagen (DIKU), 2003, http://fann.sf.net.
28. G. Cybenko, Approximation by superpositions of a sigmoidal function, Mathematics of Control Signals Systems 2 (1989) 303-314.
29. B. Echard, N. Gayton, M. Lemaire, AK-MCS: An active learning reliability method combining Kriging and Monte Carlo Simulation, Structural Safety 33(2) (2011) 145–154. DOI: 10.1016/j.strusafe.2011.01.002.
30. L. G. Crespo, D. P. Giesy, S. P. Kenny, Reliability-based analysis and design via failure domain bounding, Structural Safety 31 (2009) 306-315.
31. L. G. Crespo, D. P. Giesy, S. P. Kenny, Bounding of the failure probability range of polynomial systems subject to p-box uncertainties, Mechanical Systems and Signal Processing 37(1-2) (2013) 121-136.
32. J. Segers, Copulas: An Introduction, Columbia University, New York, USA, 2013.
33. I.H. Haff, Parameter estimation for pair-copula constructions, Bernoulli 19(2) (2013) 462–491.
34. M. C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63(3) (2001) 425–464.
35. PD Arendt, DW Apley, W Chen, Quantification of model uncertainty: Calibration, model discrepancy, and identifiability, J Mech Des Trans ASME 134 (2012) 1–12. https://doi.org/10.1115/1.4007390.
36. M. Campi, S. Garatti, F. Ramponi, A general scenario theory for non-convex optimization and decision making, Trans. Autom. Control 63(12) (2018) 4067 - 4078.