Bayesian calibration and probability bounds analysis solution to the Nasa 2020 UQ challenge on optimization under uncertainty

A. Gray, A. Wimbush, M. DeAngelis, P. O. Hristov, E. Miralles-Dolz, D. Calleja

Institute for Risk and Uncertainty, University of Liverpool, UK.
E-mail: {ander.gray, a.wimbush, marco.de-angelis, p.hristov2, enmidol, d.calleja}@liverpool.ac.uk

R. Rocchetta

Department of Mathematics and Computer Science, Technical University of Eindhoven, The Netherlands.
E-mail: roberto.rocchetta@tue.nl

Uncertainty quantification is a vital part of all engineering and scientific pursuits. Some of the current most challenging tasks in UQ involve accurately calibrating, propagating and performing optimisation under aleatory and epistemic uncertainty in high dimensional models with very few data; like the challenge proposed by Nasa Langley this year. In this paper we propose a solution which clearly separates aleatory from epistemic uncertainty. A multidimensional 2nd-order distribution was calibrated with Bayesian updating and used as an inner approximation to a p-box. A sliced normal distribution was fit to the posterior, and used to produce cheap samples while keeping the posterior dependence structure. The remaining tasks, such as sensitivity and reliability optimisation, are completed with probability bounds analysis. These tasks were repeated a number of times as designs were improved and more data gathered.

Keywords: Bayesian calibration, 2nd-order distribution, probability bounds analysis, uncertainty propagation, uncertainty reduction, epistemic uncertainty, optimization under uncertainty

1. Introduction

In this paper, it will be assumed that the reader is familiar with the Nasa challenge manifesto (Crespo and Kenny, 2020). In this challenge a control system is emulated using a black-box model. Efficient propagation of uncertainty through such a model would provide a flexible solution to the study of uncertainty in many fields of science and engineering. A distinction between aleatory and epistemic uncertainty is made, thus the two kinds will be treated differently. Such a distinction carries strong practical implications, which is expressed in the adopted solution strategy. The aleatory uncertainty is modelled with probability theory, thus for the problem under consideration, by means of continuous probability distributions. The epistemic uncertainty is modelled by means of bounded sets, whose shape is defined by the dependence structure of the epistemic variables. Because the aleatory model is not given, and must therefore be constructed solely from the data provided by the hosts, additional epistemic uncertainty is injected in the model to account for the provided limited data, and to mitigate the choice of a specific probabilistic family for such model. A double-loop uncertainty propagation scheme is at the core of the strategy to ensure separation between the two kinds of uncertainty.

2. Problem A: Model Calibration & Uncertainty Quantification

Part A of the challenge is to calibrate the model's input against the limited experimental data that is given. A calibrated uncertainty model will constitute a five-dimensional probability distribution for A and a reduced four-dimensional hyperbox for E. Most of the existing methods for uncertainty characterisation, such as Bayesian, interval, and frequentist model updating are usually suited for one and not both of these tasks (Nagel and Sudret, 2016). The dimension of the model's output also presents a challenge. A single nine dimensional vector of model inputs produces a 5001 discretised time series. The calibration data provided represents 100 discretised time series; which can be viewed as a stochastic process, with samples from a distribution at each time point. After inspection it can be seen that the provided data is both non-Gaussian and non-stationary. In this section, the calibration of the uncertainty model UM is performed against the first set of data $D_1 = \left\{y^{(i)}(t)\right\}$. As a result of this calibration, the initiating nine dimensional hyperbox is updated into the uncertainty model UM-y. After the second calibration on the set of data $D_2 = \left\{z_{1:2}^{(i)}(t)\right\}$, the uncertainty model will be denoted as UM-z.

Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference
Edited by Piero Baraldi, Francesco Di Maio and Enrico Zio
Published by Research Publishing, Singapore.
ISBN: 978-981-14-8593-0; doi:10.3850/978-981-14-8593-0

2.1. *Creation and Calibration of an Uncertainty Model*

2.1.1. *Probabilistic model for A*

The problem of simultaneously calibrating A and E may be overcome by selecting a parametric model for A. The parameters of this probabilistic model may then be included with the four epistemic inputs and standard calibration technique may then be used. The Beta distribution is a two parameter distribution which produces a wide variety of shapes, and is bounded to the $[0, 1]$ interval, which is favourable since A is bounded to $[0, 2]$. The Beta distribution may be scaled to this range (nis, 2020). Five beta distributions were selected for the marginal distributions of A, one per dimension; adding ten (two per dimension) extra parameters to be included in the calibration. Once parametric families for the marginal distributions of A have been selected, a multivariate distribution may be created by selecting a parametric copula family (Joe, 2006). A Gaussian copula family was used to construct the dependency structure for A, which takes a 5x5 correlation matrix as an input, the off diagonal elements of which may be varied to create different dependence structures. A valid correlation matrix must be semi-positive definite. It has however been shown in Joe (2006) that a semi-positive definite correlation matrix may be parametrised in terms of partial correlations, which all independently take values in $[-1, 1]$. It is shown that this parametrisation reaches all possible correlation matrices. Thus ten partial correlations, which may all take values between $[-1, 1]$ may be included in the calibration procedure to define the dependence structure for A. Under this parametrisation of f_A 20 extra dimensions are added to the original four Epistemic inputs to the model. The complete set of parameters are summarised in Table 1. The 24-dimensional input vector $e_{1:24} \in E \times E_B$, where $E_B \subseteq \mathbb{R}^{20}$ defines the epistemic space for the Beta model, uniquely defines a five-dimensional probability distribution and a point in model's epistemic input. Figure 1 shows three random 5 dimensional distributions sampled using this scheme.

2.1.2. *Propagation*

Under this parametrization aleatory and epistemic uncertainty are treated separately, and may be propagated separately by means of double loop Monte Carlo. The propagation procedure is outlined in the following steps:
*Generate a random sample of $e_{1:24}$
*Create a $F_A(a|e_{5:24})$ using $e_{5:24}$
*Generate samples of $a \sim F_A(a|e_{5:24})$
*Propagate samples a with point $e_{1:4}$

This sampling process, first releasing a random probabilistic model and then evaluating its samples in the model, produces a probability distribution at each point in time in the output domain a.k.a. stochastic process. Repeating this sampling process leads to a 2nd-order distribution in the output domain, or a distribution of distributions, at each point in time. 2nd-order distributions are very difficult to work with in practice, and so they are often approximated as probability boxes (p-box), by taking the envelope over the 2nd-order distributions CDFs (Ferson and Troy Tucker, 2006). Although the internal structure of the 2nd-order distribution is lost, the p-box contains bounds on all of properties of the 2nd-order distribution (i.e. bounds on the moments, bounds on the tails etc.). This is what the authors of the challenge question refer to as the inner approximation to the p-box. In both the cases, epistemic and aleatory uncertainty are clearly separated by the p-box and the 2nd-order distribution. The aleatory uncertainty is encoded by each of the singular distributions of the 2nd-order distribution, or by the distribution which the p-box is bounding; whilst the epistemic uncertainty is encoded by the difference of each individual distributions, or by the width of the p-box. As epistemic uncertainty is reduced, the distributions become more similar, or the bounds of the p-box tighten. With zero epistemic uncertainty, the bounds of the p-box meet creating a singular distribution. With a reduction in aleatory uncertainty, the bounds of the p-box steepen; eventually becoming an interval.

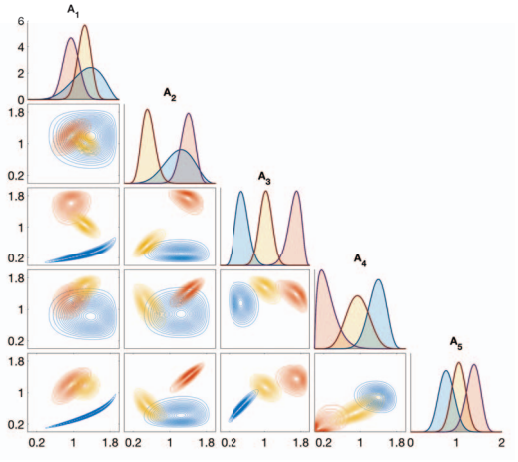

Fig. 1. Three random realisations of the probabilistic model

2.1.3. *Calibration*

Under this parametrisation, a standard calibration technique may be performed over $e_{1:24}$. In Bayesian model updating, one would like to construct a probability distribution over $e_{1:24}$ that is reflective of the provided experimental data. For

Table 1. Summary of parameters.

	Epistemic inputs to the model	Parameters for marginals of F_A	Partial correlations of F_A
Parameter	$e_{1:4}$	$e_{5:14}$	$e_{15:24}$
Prior bounds	$[0, 2]$	$[0, 100]$	$[-1, 1]$

this, a uniform distribution was used as a prior with bounds defined in Table 1, which reflects the analysts knowledge about the parameter space $e_{1:24}$ before observing any data, and altering or updating it such that the updated distribution (a posterior) matches the experimental data in the output domain.

2.1.4. *Stochastic area metric*

Accurate estimation of the predictive density is very costly, and would require a large amount of samples and possibly a density estimator such as Kernel Density Estimation. We therefore follow a similar approach outlined in Bi et al. (2019), where the likelihood is replaced by an approximate likelihood based on a stochastic distance metric. A stochastic distance metric defines distances between probability distributions (ie. defines how dissimilar two are). This framework fits well in our problem, since both our predictive distribution (for a fixed $e_{1:24}$) and the experimental data are distributional. We may therefore define a higher likelihood for points in $e_{1:24}$ which produce similar distributions to D, and penalise points which produce distributions which are "far" to D w.r.t our metric. We define an approximate likelihood based on the Gaussian distribution for each time point, as:

$$\text{likelihood}(D|e_{1:24}) \propto exp\{-\frac{d^2}{\epsilon^2}\} \qquad (1)$$

Where, d is the area metric and ϵ is called the "width factor" which defines the spikiness of the posterior. The d is the stochastic area metric is defined as:

$$d(F_P, F_D) = \int_{-\infty}^{+\infty} |F_P(x) - F_D(x)|\, dx \qquad (2)$$

where F_P and F_D are the empirical cumulative distribution functions of the prediction and data respectively. The area metric has the favourable property that it works well for distribution functions which are far away as well as near, as opposed to other metrics such as K-S distance or the Bhattacharyya distance which require a large overlap between F_D and F_P. We can therefore perform the Bayesian updating in a single step, instead of the 2-step updating pursued by Bi et al. (2019) using the Bhattacharyya distance.

The likelihood Eq.(1) is computed at every time point in the output, then the likelihood for the sample $e_{1:24}$ may be computed by taking the product of every time point. This would require 5001 evaluations of the integral Eq.(2), and would add significant cost to the Bayesian updating. To reduce the number of comparisons that must be done, this procedure may be done in the Fourier domain. It can be seen that with about 30 harmonics the output signal is well represented. Therefore, if a Fourier transformation is performed of the experimental data and of the predictive distribution only 60 (30 real and 30 imaginary) evaluations of Eq.(1) and Eq.(2).

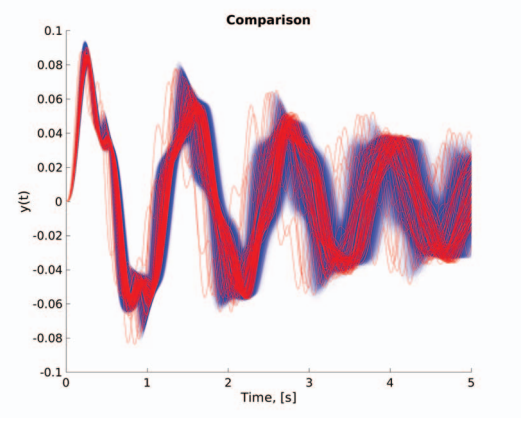

Fig. 2. Shows the calibrated model in blue, with the experimental data in red. All of the samples of the 2nd order distribution have been plotted on the same axis.

With a prior and a likelihood now defined, samples of the posterior may now be simulated by an Markov Chain Monte Carlo (MCMC) algorithm. For this work Transitional Markov Chain Monte Carlo Ching and Chen (2007) was used, an MCMC algorithm which works well for high dimensional problems such as this. 520 samples of the posterior were drawn. Figure 2 shows the calibrated model output against the provided data. The output of the calibrated model is now a p-box at each point in time. Figure 3 shows this predictive p-box at the time slice t = 2s compared to the experimental data. If you inspect the p-box at other points in time (not shown), it can be seen that the tails of the data are captured by most, but not all, of the predictive p-boxes. This is one of the main drawbacks from using 2nd - order Monte Carlo as a propagation method: that a very large number of samples are needed for high dimensions. We have however proposed a cheap method for generating a large number of samples of this

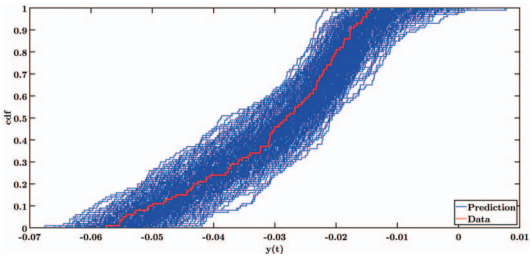

Fig. 3. Shows a comparison of the predictive p-box (blue) to the ecdf of the experimental data (red) at t = 2s.

24-dimensional posterior, which we will cover in the next subsection. The second reason may be that our choice of stochastic metric in Eq.(2) is not very sensitive to the tails of distributions. If a metric which gives higher weight to the tails of distributions we believe the tails would be better captured even with a low number of samples.

2.2. *Sliced Normals for Propagation*

Due to the difficulty of sampling in the beta-model epistemic space through Bayesian Updating, a means of efficiently generating additional samples of the posterior is required. Sliced normal distributions are a flexible means of modelling multivariate distributions with non-Gaussian dependence (Crespo et al., 2019). The approach used here to calculate the optimal sliced normal parameters matches that of Colbert et al. (2020). The pseudo code for generating 10^4 additional samples from the beta-model epistemic space using a sliced normal distribution is as follows:

(1) Samples from beta-model epistemic space $e_B^{(1:N_e)} \subseteq E_B$ Bayesian posterior are taken as the physical-space input.
(2) 2-Degree of freedom monomial vector is generated for each $e_B^{(i)}$, producing an expanded 324-dimensional feature space dataset $Z_B^{(1:N_e)}$.
(3) Mean μ_Z and unscaled covariance Σ_Z^* calculated for feature space dataset.
(4) Hyper-Ellipse fit over physical-space input data to generate support samples $s_B^{(1:N_b)} \subseteq E_B$ for estimation of normalisation constant.
(5) Support samples generated within hyper-ellipse, expanded into 2-degree of freedom monomial vector to produce support set in feature space.
(6) Feature space support and input data set likelihoods maximised through optimisation of scaling parameter γ to produce scaled covariance Σ_Z.
(7) 10^4 Markov chains initialised from random seeds within $s_B^{(1:N_b)}$ to produce first sample set.
(8) Final $e_B^{(1:10^4)} \subseteq E_B$ samples generated from unnormalised sliced normal with parameters μ_Z and Σ_Z using TMCMC algorithm. Sample likelihoods are evaluated using the unnormalised sliced normal distribution.

Generation of 10^4 sliced-normal samples takes roughly 90 minutes when parallelised using a 4-core Intel i7-7700HQ processor.

2.3. *Justification of uncertainty model*

Whilst a single Beta-marginal multivariate distribution is likely not capable of representing the 'true' aleatory distribution, resorting to an imprecise Beta model with imprecise correlation among the marginals may give us the flexibility that is needed to rigorously describe the observations, and produce a structure that can be generalised to a p-box containing the distribution in the aleatory space. Generating samples from the sliced normal approximation of the epistemic distribution allows for a more accurate estimation of the aleatory p-box bounds without the heavy computational cost of producing further samples through Bayesian updating. The geometry for E reflects the degree to which we were able to update the UM given the limited evidence and the dependence structure between these parameters. The implementation of sliced-normals, as discussed in the previous section, will enable us to preserve the posterior shape for E, while improving the efficiency of sampling E.

3. Problem B: Uncertainty Reduction

The task of ranking the *epistemic parameters according to their ability to improve the predictive ability of the computational model* and consequently determine their reductions is a challenging one, because it cannot be done using naïve implementations of variance-based sensitivity analysis. In response to this challenge, we need a metric that can quantify the amount of both variability and incertitude characterising the UM, and can simultaneously score the output against the observed data, as well as deal with high-dimensional output.

3.1. *Ranking of epistemic parameters*

The ranking of the epistemic parameters is done via scoring each of the 8 possible refinements $\{e_i^-, e_i^+\}$, $i = 1:4$, where e_i^- denote a request for the lower bound of the e_i to be increased, whilst e_i^+ denote a request for the upper bound of e_i to be decreased. For UM-0: $e_i^- = [0, 1]$ and $e_i^+ = [1, 2]$. The ranking is done by comparing the amount of epistemic uncertainty carried by the output $y(t)$ before and after the refinement,

and simultaneously checking that the data D_1 still fall within the bounds produced by the epistemic uncertainty. The comparison is enabled by the area metric of Eq.(2), which quantifies the amount of *area* contained between two distributions.

For the purpose of the sensitivity analysis, the calibrated uncertainty model of Problem A, consisting of correlated 2nd-order probability distribution, is converted in its equivalent bounded counterpart. So the 24-dimensional epistemic space is *boxed* in an hyper-rectangle containing all the samples resulting from the Bayesian updating. This has allowed us to treat as probability boxes the calibrated inputs for the sake of the uncertainty propagation. When the metric is applied to the two bounding distributions of a probability box (p-box), the resulting area metric quantifies the amount of epistemic uncertainty carried by the p-box itself. After the refinement, the area of the p-box shrinks (at best does not change) to reflect the fact that less epistemic uncertainty is projected to the output following refinement. Denoting with $[F_Y]_0$ and $[F_Y]_{e_i^*}$ the p-box of the output $y(t)$ before and after the refinement $e_i^* = \{e_i^-, e_i^+\}$, and with W_0 and $W_{e_i^*}$ their respective area metrics, we compute the sensitivity indices as:

$$S_{e_i} = 1 - \frac{W_{e_i^-} + W_{e_i^+}}{2\,W_0}. \tag{3}$$

In order for the computation of Eq.(2) to succeed the p-boxes $[F_Y]_{e_i}$ need to be fully nested in $[F_Y]_0$, i.e. $[F_Y]_{e_i} \subseteq [F_Y]_0$. When this does not hold, negative sensitivity indexes can show up. When the uncertainty propagation is not rigorous it is not obvious to avoid this problem. A meta-model strategy has been developed within this work to ensure that the nesting always occurs. An additional step is endured to check that the empirical CDF–carrying confidence bounds–of the data D_1 falls within the reduced p-box.

The ranking of the epistemic parameters is done by sorting S_{e_i}, computed as in Eq.(3) in descending order, with the most sensitive parameter carrying the highest sensitivity index. This analysis is commonly referred to in literature as value-of-information (VoI) sensitivity (Ferson and Troy Tucker, 2006). For comparison, variance-based sensitivity analysis was also conducted on both prior and posterior space. Main- and total-effect sensitivity indexes were computed as in Sobol (2001). The main effect index weighs the importance of each individual input, while the total effect provides information about the significance of input interactions. Sobol' indices were obtained using the updated posterior, drawing $N_e = 5000$ samples from the sliced-normals of F_e. Since the model of the subsystem has a time-varying output, the Sobol sensitivity indices were obtained at each of the 5001 time steps, whilst the VoI S_{e_i} were computed in the reduced

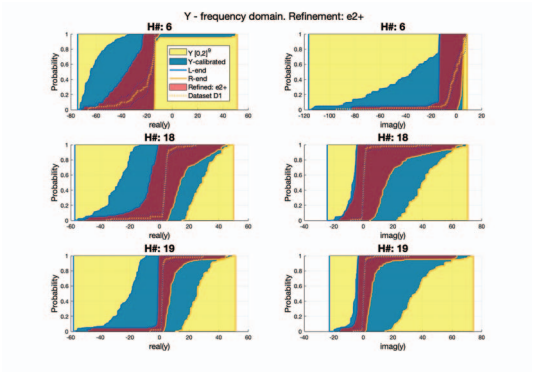

Fig. 4. Refinement e_{2_+} for harmonics 6, 18, 19.

Fourier space looking at the first 85 harmonics. A summary of the ranking is provided in Table 2. In Table 2, UM-0 and UM-y are respectively the UMs before and after the calibration.

Table 2. Summary of sensitivity ranking. Value-of-info' v. main- and total-effect Sobol indices.

	UM-y			UM-0	
	VoI	Total	Main	VoI	Total
e_2	0.2758	0.1706	0.0264	0.2296	0.3535
e_3	0.1675	0.0131	0.0063	0.0756	0.0455
e_1	0.1447	0.0179	0.0092	0.0785	0.1401
e_4	0.0524	0.0002	0.0045	0.0565	0.0011

Both VoI and variance-based method largely agree on the top-ranked parameters e_2 carrying the highest sensitivity indices. A slight difference between the two methods holds for the ranking of parameters e_3 and e_1, which are ranked differently however above e_4.

3.2. *Parameter refinement*

The parameter refinements are scored as in Eq.(4) by computing the relative amount of *reduced* uncertainty that each carries compared to the total.

$$S_{e_i^*} = 1 - \frac{W_{e_i^*}}{W_0}, \quad * = \{-, +\} \tag{4}$$

Eq.(4), where $S_{e_i^-} + S_{e_i^+} = S_{e_i}$, weighs the importance of each partial index $S_{e_i^*}$ towards the whole index S_{e_i}. The results are summarised in Table 3.

Table 3. VoI weights for each refinement.

$S_{e_i^*}$	e_1	e_2	e_3	e_4
e_i^-	0.0828	0.1348	0.0294	0.0130
e_i^+	0.0618	0.1410	0.1381	0.0394

Out of 8 possible refinements and a maximum of allowed $k \leq 4$, we have requested one reduction for this problem. The updated UM of δ incorporating the *reduction* provided by NASA will be used in the following sections.

4. Problem C: Reliability Analysis

The challenge of problem C is to balance the computational cost with the precision of the failure probability estimators. The non-linearity of the system complicates this task. Moreover, the severity estimator is very sensitive to the UM in the tails. It is also really difficult to obtain precise lower failure probability bounds due to the epistemic uncertainty at the tails of our calibrated model. We question the interpretation of small failure probabilities ($< 10^{-4}$) yielded by advanced Monte Carlo methods such as importance sampling, given the paucity of data $n_1 = 100$, and if explicitly accounting for the epistemic uncertainty can allow us to bounds such small failure probability.

4.1. *Range of failure probability for individual and all requirements*

The range of failure probability R, for each individual requirement $g_{1:3}$ for the baseline design, θ_{base}, are estimated via double loop sampling. The uncertainty propagation is done as follows:

- Outer loop: N_e epistemic realizations are obtained and each sample $e^{(i)}$ uniquely define a probabilistic model F_A.
- Inner loop: the probability operator $\mathbb{P}[\cdot]$ is estimated sampling N_a aleatory realizations from F_A.

For instance, the failure probability for the integrated system is estimated in the inner loop as $\frac{1}{N_a}\sum_{i=1}^{N_a} \mathbf{1}_{w^{(i)}>0}$, where $w^{(i)} = w(\theta, \delta^{(i)})$ and $\mathbf{1}_{w^{(i)}>0}$ is the indicator function for the condition $w^{(i)} > 0$ (any failure). In the outer loop the minimum (maximum) operator is approximated via the samples as $\min_{e^{(1:N_e)}}\{\cdot\} \approx \min_{e\in E}\{\cdot\}$.

The number of samples limits the precision that can be obtained for R. Moreover, strong non-linearities of the $g_{1:3}$ combined with poor coverage of the epistemic domain can lead to inner approximations of the bounds of these operators. The inner approximation of these bounds is reduced by: (*) sampling the sliced normal of the posterior distributions in $E \times E_B$ for better coverage; (*) loosening the Beta assumption by adopting a focal element propagation scheme for counterbalancing inner approximations.

In Table 4, we show the results of the above analysis with respect to R. The reliability performance of θ_{base} is estimated using $N_e = 10^4$ epistemic samples and $N_a = 200$ aleatory samples from UM-y. The most uncertain reliability scores for θ_{base} are with respect to, in order, settling time [0, 0.99], stability [0, 0.67], and energy consumption [0, 0.27].

4.2. *Rank of epistemic uncertainty on R*

In this section, the epistemic parameters are ranked according to the contraction of $R(\theta)$ that may result from reducing their uncertainty. To do so we use a value of information metric, which measures the reduction in area of the p-box of $g_{1:3}$ for a fixed value of $e_{1:4}$, relative to the area obtained with all parameters varying in their posterior range. Additionally, any reduction in the range of R was also taken into account. The performance requirements $g_{1:3}$ showed high sensitivity towards the epistemic uncertainty in the UM. Only modest contraction of $R_{1:3}$ were computed, with parameters e_1 and e_2 ranking highest.

4.3. *Identify realizations of δ with large likelihood near the failure domain*

We isolated a focal element for each failing requirement $g_{1:3} = 0$, so three in total, and identified the realizations in the input domain falling inside these three focal elements. Equivalently, we looked at the p-box lower bound of each requirement, and isolated the samples near the failure domain. We then identified three examples of time responses for each type of failure. We then conducted a parallel-plot study to highlight the 24-dimensional vector of epistemic realization of δ that were closer to the failure domain, as well as plotted 2-d projections of the failure region for each pair of $\delta_{1:9}$.

4.4. *Severity of requirement violation*

The severity of each individual requirement and joint requirements is estimated via double loop Monte Carlo. The accuracy of this estimator is more sensitive to the number of aleatory samples if compared to R, i.e. an outlier can alter substantially the value of the severity. The quality of a design θ is evaluated on the basis of the reliability indicators presented in Crespo et al. (2019). Additionally, the following severity indicator for the integrated system is considered $s_w(\theta) =$

$$\max_{e\in E}\{\mathbb{E}[w(\theta,\delta)|w(\theta,\delta) > 0] \times \mathbb{P}[w(\theta,\delta) > 0]\}$$

$s_w(\theta)$ is also referred to as the severity score for the integrated system and the design. If the value of w when $w > 0$ is a measure of the severity of failures, the analyst might want to control not only the failure probability but also the shape and

thickness of the upper tail of w. The conditional expectation $\mathbb{E}[w|w>0]$ is estimated by sampling N_a times the aleatory space as follows:

$$\mathbb{E}[w(\theta,\delta)|w(\theta,\delta)>0] = \frac{\sum_{i=1}^{N_a} w^{(i)}\mathbf{1}_{w^{(i)}>0}}{\sum_{i=1}^{N_a}\mathbf{1}_{w^{(i)}>0}}$$

The requirements ranking for the most severe failures are reversed in order compared to the reliability ones, see Table 4, with the energy consumption potentially leading to the most severe failures $s_3(\theta_{base}) = 0.7154$. Figure 5 shows the composite failure regions and safe regions on 2-d projections of the space of uncertain factors.

5. Problem D: Reliability-Based Design

5.1. *Optimality criteria*

Failure probability and severity of the integrated system are selected as suitable metrics to define a new candidate design. To address the RBDO problem we focused our attention on two optimization programs defined as follows:

$$\langle\theta_R^\star\rangle = \arg\min_{\theta\in\Theta}\{\max_{e\in E}\mathbb{P}[w(\theta,\delta)>0]\} \quad (5)$$

$$\langle\theta_s^\star\rangle = \arg\min_{\theta\in\Theta}\{s_w(\theta)\} \quad (6)$$

where Θ is the design space, which is not provided in the challenge. Two suitable choices can be made for Θ: (a) consider the hypercube $[-10,10]^9$ centred on the baseline design, or (b) consider Θ to be unbounded and new tentative designs are randomly generated around the baseline at a distance that is a fraction (typically a half) of the baseline absolute value. Program (5) seeks a design $\theta_R^\star$ which minimizes the upper bound on the failure probability approximated via sampling. Differently, program (6) seeks a design $\theta_s^\star$ that minimizes severity function for the integrated system, i.e., a combination of failure probability and mass of w in the failure region. Both objective functions are non-convex in the space of θ and local optimization methods ineffective. Furthermore, $\mathbb{P}[w(\theta,\delta)>0]$ is step-wise discontinuous function in A. For this reasons, we selected a global gradient-free optimization strategy to search for an optimal candidate design θ_{new}.

5.2. *Computational approach*

Genetic Algorithm is the computational viable optimization method selected. This is because individual requirements are competitive and failures not mutually exclusive. Hence, a contraction of one failure region generally leads to the expansion of another. Out-of-sample errors for the severity are higher. Failure probability and severity estimators are stepwise discontinuous functions.

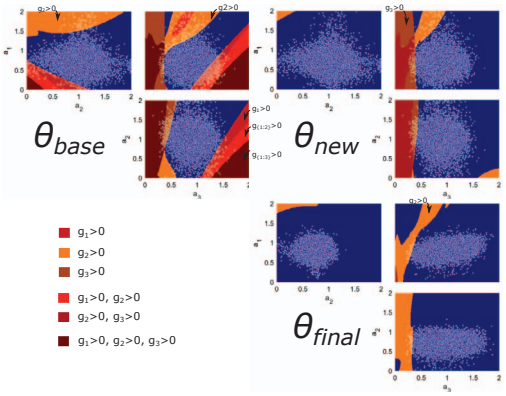

Fig. 5. Projections of the individual failure regions and safe regions (blue) induced by θ_{base} (3 panels on the top-left), θ_{new} (3 panels on the top-right), and θ_{final}. White markers show realization of a_1, a_2 and a_3 for the proposed models UM-y (top panels) and UM-z (bottom panel).

5.3. *Analysis for θ_{new}*

Table 4 presents the results of the optimization. The reliability metrics are obtained as described for the baseline design. It can be observed a substantial improvement of the severity and reduction of upper bound on the system reliability. Figure 5 shows projections of the failure (red) and safe (blue) regions on the uncertainty space. The transparent markers are aleatory and epistemic samples obtained from the model of the uncertainty. Although this is only a 2-d projection of a 9 dimensional space, it can be observed that the failure regions has been pushed away from the samples, i.e., away from the high-probability mass of the proposed uncertainty model.

6. E: Model Update & Design Tuning

Upon finding a new design point θ_{new}, the last problem of the challenge is to perform a final improvement to the UM and design. After providing θ_{new} to the challenge hosts, 100 realisations of the subsystems Z_1 and Z_2 were provided for calibration. The calibration approach pursued here is identical to that described in section 2, where the new data was used alongside the original data to improve the UM. A sensitivity analysis identical to that in section 3 was done to request the 3 more parameter refinements to E. These refinements were selected with VoI sensitivity analysis performed with respect the subsystems Z_1, Z_2 and Y. Table 5 shows the overall VoI, whose magnitude led to choose e_1^-, e_1^+ and e_3^+ as refinements. Once the challenge hosts returned with 3 reductions to E, the UM was calibrated a final time over this reduced space. The updated model was then used within the RBDO described in Section 5 to obtain θ_{final}. Table 4 presents the reliability scores for

Table 4. Summary of RBDO results for UM-y and UM-z estimated using $N_e = 10^4$ and $N_a = 200$.

Design	$R_w(\theta)$	$R_1(\theta)$	$R_2(\theta)$	$R_3(\theta)$	$s_w(\theta)$	$s_1(\theta)$	$s_2(\theta)$	$s_3(\theta)$
θ_{base} (UM-y)	[0,0.99]	[0,0.67]	[0,0.99]	[0, 0.269]	0.7393	0.1197	0.0171	0.7154
θ_{base} (UM-z)	[0,0.7]	[0,0.7]	[0,0.51]	[0,0.19]	0.4511	0.1522	0.0106	0.3649
$\theta^{\star}_{new}$ (UM-y)	[0,0.980]	[0,0.359]	[0,0.976]	[0,0.0348]	0.0572	0.0527	0.0070	0.0170
$\theta^{\star}_{new}$ (UM-z)	[0,0.12]	[0,0]	[0,0.12]	[0,0.045]	0.0126	0	3.69×10^{-4}	0.0125
$\theta^{\star}_{final}$ (UM-y)	[0,0.985]	[0,0]	[0,0.985]	[0,0]	0.0039	0	0.0039	0
$\theta^{\star}_{final}$ (UM-z)	[0,0.075]	[0,0]	[0,0.075]	[0,0]	1.27×10^{-4}	0	1.27×10^{-4}	0

Table 5. Summary of VoI sensitivity ranking for UM-z.

		$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_3$	$\mathbf{e}_4$
Y,Z_1,Z_2	e^-	0.406	0.306	0.208	0.137
	e^+	0.679	0.169	0.429	0.230

θ_{final} and Figure 5 shows a 2-d projection of the individual failure regions. Similarly to θ_{new}, the most uncertain reliability score for θ_{final} is with respect to the settling time requirement. However, the epistemic uncertainty affecting this score decreased substantially with the new model of the uncertainty, i.e., from [0,0.985] to only [0,0.075].

Conclusions

In this paper we have presented a summary of results addressing the Nasa UQ challenge. The Nasa UQ problems are designed to be on the edge of what is possible with current UQ methods, to focus efforts in the UQ community and push forward the state of the art. Our proposed strategy is therefore a collection of methods which have appeared recently in literature, and some of which have been adapted to the stringent challenge requests: a Bayesian method with stochastic distances for calibrating aleatory and epistemic uncertainty, sliced normal distributions for high dimensional density estimation and sampling, an information based sensitivity analysis with probability boxes, and an efficient reliability analysis and design with probability bounds analysis; all performed on a computational model with a high dimensional output. Given the broad formulation of the challenge the proposed set of methods can find a broad range of applications. **Disclaimer**: Given eight-page limit, figures that would have significantly improved clarity of the manuscript could not be presented in the paper. Because of the open challenge, we were not allowed to disclose the reduced epistemic parameters and the design points leading to the results of Table 4.

Acknowledgements

This research is also funded by the Engineering & Physical Sciences Research Council with grant no. EP/R006768/1, "*Digital twins for improved dynamic design*". The authors would also like to acknowledge the gracious support of this work through the local authorities under grant agreement "ITEA-2018-17030-Daytime". This work has been carried out within the EUROfusion Consortium and Euratom research and training programme under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

References

(2020). NIST/SEMATECH e-Handbook of Statistical Methods. https://www.itl.nist.gov/div898/handbook/, accessed 11-02-2020.

Bi, S., M. Broggi, and M. Beer (2019). The role of the bhattacharyya distance in stochastic model updating. *Mechanical Systems and Signal Processing 117*, 437 – 452.

Ching, J. and Y.-C. Chen (2007). Transitional markov chain monte carlo method for bayesian model updating, model class selection, and model averaging. *Journal of engineering mechanics 133*(7), 816–832.

Colbert, B. K., L. G. Crespo, and M. M. Peet (2020). Improving the uncertainty quantification of sliced normal distributions by scaling the covariance matrix. In *Proceedings of the American Control Conference*.

Crespo, L. G., B. K. Colbert, S. P. Kenny, and D. P. Giesy (2019). On the quantification of aleatory and epistemic uncertainty using sliced-normal distributions. *Systems & Control Letters 134*, 104560.

Crespo, L. G. and S. P. Kenny (2020). The nasa langley uq challenge on optimization under uncertainty. *ESREL 2020*.

Ferson, S. and W. Troy Tucker (2006). Sensitivity analysis using probability bounding. *Relaibility Engineering and System Safety 91*(10-11), 1435–1442.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis 97*(10), 2177–2189.

Nagel, J. B. and B. Sudret (2016). A unified framework for multilevel uncertainty quantification in bayesian inverse problems. *Probabilistic Engineering Mechanics 43*, 68–84.

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation 55*(1-3), 271–280.