Likelihood and depth-based criteria for validation of numerical simulators, from comparison with experimental data

Amandine Marrel, Héloise Velardo, Antoine Bouloré

Likelihood and depth-based criteria for comparing simulation results with experimental data, in support to validation of numerical simulators

Amandine Marrel[1,3], Héloise Velardo[1], Antoine Bouloré[2]

[1]CEA, DES, IRESNE, DER, Cadarache F-13108 Saint-Paul-Lez-Durance, France

[2]CEA, DES, IRESNE, DEC, Cadarache F-13108 Saint-Paul-Lez-Durance, France

[3]Institut de Mathématiques de Toulouse, 31062, Toulouse, France

Abstract

Within the framework of Best-Estimate-Plus-Uncertainty approaches, the assessment of model parameter uncertainties, associated with numerical simulators, is a key element in safety analysis. The results (or outputs) of the simulation must be compared and validated against experimental values, when such data is available. This validation step, as part of the broader Verification, Validation and Uncertainty Quantification process, is required to ensure a reliable use of the simulator for modeling and prediction. This work aims to define quantitative criteria to support this validation for multivariate outputs, while taking into account modeling uncertainties (uncertain input parameters) and experimental uncertainties (measurement uncertainties). For this purpose, different statistical indicators, based on likelihood or statistical depths, are investigated and extended to the multidimensional case. First, the properties of the criteria are studied, either analytically or by simulation, for some specific cases (Gaussian distribution for experimental uncertainties, identical distributions of experiments and simulations, particular discrepancies). Then, some natural extensions to multivariate outputs are proposed, with guidelines for practical use depending on the objectives of the validation (strict/hard or average validation). From this, transformed criteria are proposed to make them more comparable and less sensitive to the dimension of the output. It is shown that these transformations allow for a fairer and more relevant comparison and interpretation of the different criteria. Finally, these criteria are applied to a code dedicated to nuclear material behavior simulation. The need to reduce the uncertainty of the model parameters is thus highlighted, as well as the outputs on which to focus.

Keywords— Uncertainty quantification, Model validation, Statistical criteria, Experimental results, Likelihood, Depth statistics, Multivariate output.

1 Introduction

For several decades, numerical simulators have become fundamental tools for understanding, modeling and predicting physical phenomena. Large simulation models (or computer codes) implement complex mathematical models and have been successfully used in risk and safety assessments, in design optimization, or performance assessment of industrial systems. For nuclear engineering applications,

physical experiments are often costly, limited or even sometimes impossible, therefore simulation is of prime interest. Confidence in the simulation result, in the sense of the credibility of the prediction with respect to phenomenological reality, relies among other things on the fidelity of the physical modeling, the validity of the mathematical algorithms implemented to build the numerical model, and finally on the management of the uncertainties of the simulator's input parameters. These different components are taken into account in VV&UQ (Verification, Validation and Uncertainty Quantification) processes [National Research Council, 2012]. More precisely, the verification [Oberkampf and Roy, 2010] aims to determine whether the numerical model correctly implements the mathematical description and provides a sufficiently accurate approximation to the theoretical solution of the physical equations. Then, the validation process [Oberkampf and Roy, 2010] raises the question of whether the numerical simulator faithfully reproduces the reality that it models, with respect to the model's intended uses [ASME, 2009, 2019, Oberkampf and Trucano, 2002].

Moreover, simulation models, even the most representative and faithful to physical reality, often take a large number of uncertain (or not well known) input parameters. These parameters can characterize the studied phenomenon or be related to its physical and numerical modeling. Validation is therefore closely linked to the issue of assessing the uncertainties (also called Uncertainty Quantification) of simulator's input parameters. Besides, this UQ step is a key element in safety analysis for nuclear power plants, and has become of prime importance in the so-called Best-Estimate-Plus-Uncertainty (BEPU) methodology [Baccou et al., 2020, Wilson, 2013]. UQ aims to quantify how uncertainties in model input parameters affect simulation results, and more specifically to quantify the resulting uncertainty in the quantities of interest predicted by the simulator and related to decision-making issues. For this, the UQ process is based on two main steps: the identification of input uncertainties and their propagation within the simulator. To carry out these steps, most approaches used in engineering rely on statistical inference and call for Monte-Carlo methods. The most generic of these consists in drawing a random sample of the probability distribution of the inputs, and launching the corresponding simulations to obtain a sample for the output(s). This UQ process is in fact part of a more general framework for dealing with uncertainties in numerical simulations, the key steps of which are summarized by Figure 1, extracted from Iooss [2019]. The diagram explains the interconnections between the various steps, which include the V&V processes. This global approach is now widely adopted in engineering for dealing with uncertainties and deploying the VV&UQ process [Baccou et al., 2020, De Rocquigny et al., 2008, Ghanem et al., 2017].

Let's take a brief overview of the main steps. First, the problem is specified: this consists in defining the system under study (model, simulator or measurement process), identifying uncertain or fixed input variables, as well as the system response quantities of interest. Step B then aims at quantifying the input uncertainties. This quantification of input uncertainties can be supported by expert opinion or available data, or be determined by solving an inverse calibration or assimilation problem. Inverse methods can also be used to calibrate a model (Step B', performed simultaneously or not with the quantification step). This process consists in adjusting a set of input parameters in order to maximize the agreement of the simulated predictions with corresponding experimental data. Then, at Step C, the input uncertainties are propagated: the objective is to quantify how input uncertainties affect the output(s) predicted by the simulator, and more precisely the quantity of interest. Complementary to Step C, sensitivity analysis aims at studying the impact of each sources of input uncertainties on the quantities of interest.

Hence, to allow a reliable use, the simulators, including their uncertain inputs, have to undergo a thorough, rigorous, and extensive V&V process (which appears in Step A' and B', respectively). We focus here on the validation step which is mostly based on comparison with experimental data. Moreover, the term "simulator" will refer in all that follows to the calculation code with its uncertain model parameters (and thus their associated variation ranges), these uncertainties having been quantified in

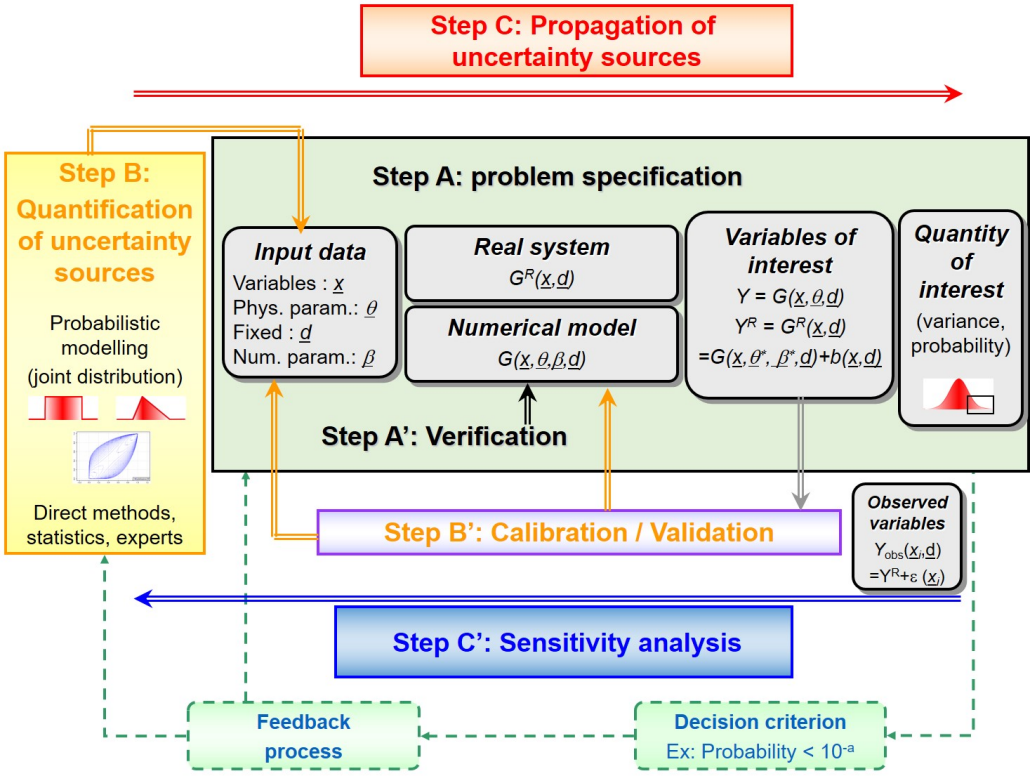


Figure 1: General scheme for the methodology of uncertainty treatment in numerical simulation, slightly modified from [Iooss, 2019], courtesy of the author Bertrand Iooss (EDF R&D).

step B, and possibly refined in a calibration step. In this framework, the assessment of consistency between simulations and reality must first take into account two types of uncertainty: that of the simulator inputs and that of the experimental measurements. Second, the comparison between simulations and experiments under uncertainties raises the question of defining quantifiable validation (or consistency) indicators. These indicators must make it possible to go beyond a sometimes subjective and complex graphical analysis, particularly in the case of multivariate or functional outputs. They must be adaptable to the different types and dimensions of experimental data available. In addition, it requires a clear definition of the meaning and purpose of the validation, which may differ depending on the application and the context considered. Should the simulator "encompass" the experiments (including their uncertainties), or, conversely, should it rather produce simulations that would have a high probability of being observed experimentally? The second objective will be addressed here, as explained in the following, but the proposed indicators can be directly adapted for the first objective (cf. Section 2.1). The objective may also be to compare two simulation models, without necessarily considering experimental data (one of the simulators would be the reference simulator). It is therefore necessary to define precisely the problem and the objective of the validation process, and to adapt the indicators proposed in this work accordingly. Besides, several versions of the indicators will be proposed in this work but only some of them will be used for the considered application.

1.1 Validation by comparison with experimental results and associated issues

First, we recall the following definitions proposed by the SAPIUM report [Baccou et al., 2020] on good practice guidance:

- "Validation (of input uncertainties): the process involving a comparison between the results of input uncertainty propagation and experimental data to determine the degree to which input uncertainties are compatible with an intended use.
- Validation (of simulation model): the process involving a comparison between the results of a simulation model and the experimental data to determine the degree to which a simulation model is compatible with an intended use."

We focus here on the first definition and associated notions. Furthermore, we consider a probabilistic framework where these uncertainties are modeled by fully or partially known probability distributions [Helton, 1997, Oberkampf et al., 2001]. In practice, information on the distributions of simulations and experiments usually differs. The former can be based on available data, expert opinions or bibliographic databases, while the latter are mostly known only through random sampling. More precisely, the uncertain inputs of the simulator are here randomly drawn according to their (assumed) probabilistic distributions and the corresponding simulator output(s) are computed (uncertainty propagation based on Monte-Carlo approach).

Among the available literature on validation indicators [Liu et al., 2011], the OECD/NEA SAPIUM project highlights several drawbacks and shortcomings of usual metrics [Baccou et al., 2020]. For example, [Oberkampf and Barone, 2006] simply proposes to use regression techniques to quantify the difference between experimental results whose measurement is assigned a random error and the simulated response, over a certain validation domain. Other authors address the validation question by checking whether the experimental data falls within the uncertainty intervals of simulations, regardless of the position of the experimental value within the interval. So-called *calibration indicators* are thus computed. These indicators therefore check that the experiments are plausible with respect to (w.r.t.) the simulations, and not the reverse. Alternatively, the simulations interval can be divided into subintervals and the uniform location of the experimental data can be assessed using hypothesis tests such as the χ^2-test. Based on the same idea of comparing the probability distribution of the simulations with that of the experimental data, statistical tests of Goodness-Of-Fit which estimates a discrepancy (or dissimilarity) measure [Cha, 2007] between both distributions can be used. In the same vein, area metric indicators [Ferson and Oberkampf, 2009] can be built. However, comparing the two probability distributions in this way is not necessarily desirable or relevant to the validation objective considered here. For example, if the uncertainty around the measurement completely encompasses the (much sharper) distribution of the simulations, a dissimilarity measure would lead to reject the adequacy of both distributions while the expected conclusion should be that there is an agreement between the experiments and simulations. Indeed, in this situation, the simulations are consistent with the available experimental data: the information provided by the latter is probably too imprecise and does not allow to detect a discordance and/or an inaccurate simulated model.

Moreover, the dissimilarity-measures-based approaches do not provide a ranking of simulations by order of agreement with the experimental results. This possibility is of particular interest for identifying a group of highly consistent simulations and, on the contrary, some incompatible simulations. Furthermore, the extension of most of the aforementioned indicators or metrics to multivariate or functional data, or data of different types, in order to provide a unique aggregated result is not straightforward.

Finally, as previously indicated, the information on the distributions of simulations and experiments differs in our application context: the first one is sampled and must be estimated, while the second one is often assumed to be a given parametric analytical model (e.g. Gaussian centered on the measured value, with a standard deviation given by the accuracy of the measuring equipment). Validation indicators must therefore be adapted to these differing types of knowledge.

1.2 Objectives and scope

This work aims to build validation indicators that overcome some of the aforementioned limitations and that are adapted to the validation issues in our application context. As in Baccou et al. [2020], a validation indicator is defined as a mathematical operator that compares the quantity of interest predicted by the simulator with its associated uncertainty (uncertainty coming from the quantified uncertainties of the simulator inputs) with the same quantity of interest but coming from experimental data. For this, [Marie et al., 2019] have recently proposed to use a likelihood-based indicator, and applied it to the validation of sodium fast reactor simulation tools. The work proposed here aims to go further by studying indicators based on the notions of likelihood but also statistical depth, and to assess how they can be adapted for validation with multivariate outputs. A key element in this work is also that it does not consider in a symmetric way the experimental and simulation uncertainties. The angle chosen here is that the experimental probability distribution constitutes the reference distribution that defines the admissible simulations, and those too extreme to model reality correctly. From this, we want to answer the following question: "Is the simulation plausible w.r.t. the information provided by the experiments?". The objective is therefore to assess the consistency of the simulated outputs, individually and conjointly, with the experimental data.

Our aim is to answer these questions within a very precise framework, which now needs to be more specified. First of all, with regard to the experimental data, we consider that the measured quantities are the same as those simulated, and that the experimental uncertainties relate only to these quantities of interest. In other words, any errors in the experimental conditions (also known as control variables) are not taken into account. It is also assumed that the experiment is repeatable, and that there is no post-processing of the measured data (and therefore no additional error generated): the quantities of interest are measured directly. Moreover, as previously mentioned, a probabilistic modeling is adopted. Inherent issues such as the choice between probabilistic or extra-probabilistic modeling, the presence or absence of data to quantify the experimental error, or metrological concerns are beyond the scope of this paper and are not addressed here.

Concerning now the uncertainty of simulated data, we do not take into account the randomness resulting from rounding errors, numerical resolution schemes, or the use of Monte Carlo-type algorithms. We only consider uncertainties directly imputable to model input parameters. For a given set of parameters, the simulation result is therefore assumed to be constant and that there is no point in repeating a simulation point. In practice, even if this is not strictly the case, this simplifying assumption is generally acceptable, the code verification step having already been achieved. The uncertainties relating to the numerical precision and resolution can then be considered of negligible influence compared to the uncertainties of the model parameters. The only exception might be the case of stochastic simulators, used in neutronic simulation for instance. However, this type of uncertainty will be implicitly conveyed in a Monte Carlo approach used to generate the inputs/output(s) sample of simulations (which is the case here).

1.3 Organization of the paper and notations

The rest of the document is organized as follows. Section 2 proposes some criteria which rely on the statistical likelihood of simulations. These criteria are first defined for the one-dimensional case before being extended to the multivariate case. Another type of criteria based on the notion of depth statistics is studied in Section 3 and adapted to our validation purpose. In particular, transformed versions are proposed for an easier comparison and analysis of likelihood- and depth-based criteria, especially regardless of the dimension of the output. Section 4 proposes an application of some of the criteria to a nuclear test case that simulates the behaviour of a nuclear material under irradiation.

Before that, we introduce a few notations for numerical simulator and experimental results. The

numerical model is represented by the relation:

$$\begin{aligned}\mathcal{M} : \mathcal{X} &\longrightarrow \mathcal{Y} \\ \mathbf{X} &\longmapsto \mathcal{M}(\mathbf{X}) = Y_{sim},\end{aligned}$$

where $\mathbf{X} = (X_1, \ldots, X_l)^\top$ and Y_{sim} are respectively the l uncertain inputs and the output. As part of the probabilistic approach, the l inputs are assumed to be continuous random variables with known joint probability distributions. Consequently, Y_{sim} is also a random variable defined in a measurable space $\mathcal{Y}_{sim}$ with probability distribution denoted $\mathbb{P}_{Y_{sim}}$. It is also assumed to be continuous with probability density function (PDF) denoted $f_{Y_{sim}}$. This distribution is unknown and in the framework of a Monte-Carlo approach, only observations (or realizations) of $\mathcal{M}$ are available. It is therefore assumed that we have a sample of size n of inputs and associated outputs $\left(\mathbf{X}^{(m)}, Y_{sim}^{(m)}\right)_{1 \leq m \leq n}$ where $Y_{sim}^{(m)} = \mathcal{M}(\mathbf{X}^{(m)})$ for $m = 1, \ldots, n$. The output can be 1-D or multidimensional (vector of d components), denoted Y_{sim} and $\mathbf{Y_{sim}}$, respectively.

In addition, it is assumed that some experimental results are available, including quantities of interest similar to the outputs computed by the simulator. These experimental data are also uncertain due to measurement error and are denoted Y_{exp} and $\mathbf{Y_{exp}}$ for the 1-D and multivariate cases, respectively. Evolving in measurable space $\mathcal{Y}_{exp}$, their probability distribution denoted $\mathbb{P}_{Y_{exp}}$ is assumed to be known, continuous and of PDF $f_{Y_{exp}}$.

2 Criteria based on likelihood of simulations

For a one dimensional output, a graphical comparison can be done by comparing each simulation y_{sim} with the PDF of the experimental data $f_{Y_{exp}}$. Ideally, it is hoped that a large proportion of y_{sim} has a high probability of being observed experimentally. From this point of view, considering a likelihood-based criterion seems relevant to reflect the capability of the simulator to correctly predict the reality, even if some of its input parameters are not well known.

2.1 Initial formulation for 1-D output

To quantify the likelihood (and compatibility) of each possible simulation y_{sim} w.r.t. the experimental distribution, Marie et al. [2019] have proposed the following criterion defined on $[0, 1]$:

$$C(y_{sim}|\mathbb{P}_{Y_{exp}}) = \text{Proba}\left[f_{Y_{exp}}(Y_{exp}) \leq f_{Y_{exp}}(y_{sim})\right] = \int_{\mathcal{Y}_{exp}} \mathbf{1}_{f_{Y_{exp}}(y) \leq f_{Y_{exp}}(y_{sim})}(y) f_{Y_{exp}}(y) dy, \qquad (1)$$

where $\mathbf{1}_A(y)$ is the indicator function defined by $\mathbf{1}_A(y) = 1$ if $y \in A$ or y satisfies A, and 0 otherwise. $C(y_{sim}|\mathbb{P}_{Y_{exp}})$ is the estimated probability that Y_{exp} takes a value less likely (*i.e.* "less probable") than y_{sim}. A very low $C(y_{sim}|\mathbb{P}_{Y_{exp}})$, e.g. lower than 5%, is a significant presumption that a simulation result is unlikely to be physically observed. Conversely, $C(y_{sim}|\mathbb{P}_{Y_{exp}})$ is one when y_{sim} corresponds to the most probable experimental value, which translates a high compatibility between the simulations and the experimental results. This criterion can be generalized to any type of variable Y_{exp} (and y_{sim}): scalar, vectorial, functional, etc. as long as a probability distribution is defined to characterize its uncertainty.

Note that C can be more generally used to compare any couple of variables with given probability distributions. It could otherwise be formulated as $C(y_{exp}|\mathbb{P}_{Y_{sim}})$ for assessing the concordance of an experimental result with the distribution predicted by the simulator. This last formulation differs from Eq. (1), $f_{Y_{exp}}$ being replaced by $f_{Y_{sim}}$. It can be considered in the context of the qualification of experimental results by simulation (validation of an experimental device by simulation via a digital twin, for instance).

2.2 1-D Criterion for a set of simulations and C^α-trimmed regions

As Y_{sim} is a random variable, so is $C(Y_{sim}|\mathbb{P}_{Y_{exp}})$. To summarize its distribution into a global quantitative indicator of validation, the median value denoted $C^{\text{med}}_{Y_{sim}|\mathbb{P}_{Y_{exp}}}$ can be considered. Other global indicators can obviously be derived from the distribution of $C(Y_{sim}|\mathbb{P}_{Y_{exp}})$, according to the purpose and the way of thinking about the validation of the simulator with experiments (see [Marie et al., 2019]).

For $\alpha \in [0,1]$, we can then define the sets of simulations y_{sim} that have a $C(y_{sim}|\mathbb{P}_{Y_{exp}})$ criterion of at least α. This forms a first nested family of C^α-trimmed regions denoted:

$$RC^{\alpha,1}_{\mathbb{P}_{Y_{exp}}} = \{y_{sim} : C(y_{sim}|\mathbb{P}_{Y_{exp}}) \geq \alpha\}. \tag{2}$$

$RC^{\alpha_{\max},1}_{\mathbb{P}_{Y_{exp}}}$ with $\alpha_{\max}$ the maximal value obtained on the set of simulations therefore defines the set of more likely simulations. More generally, $RC^{\alpha,1}_{\mathbb{P}_{Y_{exp}}}$ provides a $C(\bullet|\mathbb{P}_{Y_{exp}})$-based order statistic for the simulations and induces an outlyingness function.

Some $RC^{\alpha,1}_{\mathbb{P}_{Y_{exp}}}$-based regions might contain all the simulations while others contain none. To avoid this and further ranking the observed simulations, another trivial solution to define nested regions is to rank the set of n simulations according to the value of their criterion:

$$RC^{\alpha,2}_{\mathbb{P}_{Y_{exp}}} = \{\tilde{y}^{(\lfloor \alpha n \rfloor+1)}_{sim}, \tilde{y}^{(\lfloor \alpha n \rfloor+2)}_{sim}, \ldots, \tilde{y}^{(n)}_{sim}\}, \tag{3}$$

where $\lfloor x \rfloor$ denotes the integer part of x and $\left(\tilde{y}^{(m)}_{sim}\right)_{m\in\{1,\ldots,n\}}$ the ordered values of $(y^m_{sim})_{m\in\{1,\ldots,n\}}$ such that $C(\tilde{y}^{(1)}_{sim}|\mathbb{P}_{Y_{exp}}) \leq \ldots \leq C(\tilde{y}^{(n)}_{sim}|\mathbb{P}_{Y_{exp}})$.

2.3 Analytical computation and distribution for some specific cases

First of all, if we consider the particular case where Y_{sim} and Y_{exp} are identically distributed, it can be demonstrated that the criterion $C(Y_{sim}|\mathbb{P}_{Y_{exp}})$ follows a uniform distribution on $[0,1]$, (under some assumptions on the distribution of $f_{Y_{exp}}(Y_{sim})$). See A.1 for the demonstration and details.

Considering now the case of two Gaussian distributions, we obtain:

$$C(y_{sim}|\mathbb{P}_{Y_{exp}}) = 1 - F_{\chi^2_{(1)}}\left[\left(\frac{y_{sim}-\mu_{exp}}{\sigma_{exp}}\right)^2\right] \tag{4}$$

where μ_{exp} and σ_{exp} are respectively the mean and standard deviation of Y_{exp}, and $F_{\chi^2_{(1)}}$ being the cumulative density function (CDF) of the chi-squared distribution with one degree of freedom. From this, the CDF of the criterion, denoted F_C, can then be expressed as follows:

$$F_C(x) = 1 - F_{\chi^2_{(1),(\mu_{sim}-\mu_{exp})^2}}\left[\frac{\sigma^2_{exp}}{\sigma^2_{sim}} \times F^{-1}_{\chi^2_{(1)}}(1-x)\right] \tag{5}$$

where μ_{sim} and σ_{sim} are respectively the mean and standard deviation of Y_{sim}, and $F_{\chi^2_{(1),(\mu_{sim}-\mu_{exp})^2}}$ is the CDF of a non-central chi-squared distribution with non-centrality parameter $(\mu_{sim}-\mu_{exp})^2$ and with one degree of freedom. See A.2 for the demonstration. Eq (5) makes it possible to better understand (in the Gaussian case) the evolution of the criterion distribution according to a shift between the means (model bias) or a dilation between the variances (impact of modeling uncertainty and/or measurement errors).

2.4 Proposed extensions to multivariate output

In the multivariate case, the output consists of d variables of interest forming a d-dimensional vector $\mathbf{Y_{sim}} = (Y_{sim,1}, \ldots, Y_{sim,d})^\top$ and $\mathbf{Y_{exp}} = (Y_{exp,1}, \ldots, Y_{exp,d})^\top$, for the simulation and the experiment respectively. This vector can be composed of a similar physical variable but observed for different experimental conditions, or of a set of several physical variables of different units and order of magnitude. The criterion $C(\mathbf{y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}})$ can be naturally extended with the multivariate density of $f_{\mathbf{Y_{exp}}}$ and will be given for every observation $\mathbf{y_{sim}} \in \mathbb{R}^d$ by:

$$C_d(\mathbf{y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}}) = \text{Proba}\left[f_{\mathbf{Y_{exp}}}(\mathbf{Y_{exp}}) \leq f_{\mathbf{Y_{exp}}}(\mathbf{y_{sim}})\right]. \tag{6}$$

This first criterion will thus address this question: is my simulation consistent with the experimental results on all the variables of interest? It is easy to understand that a single output that deviates completely from the experiment would seriously degrade the value of the criterion (this simulation having a very low statistical likelihood w.r.t. the experimental PDF).

If one wishes to further assess whether the results are correctly represented on average, it is worth considering other criteria. Moreover, since in most cases the components of $\mathbf{Y_{exp}}$ can be assumed to be independent random variables[1], we propose to consider the one-dimensional criteria calculated for each quantity of interest and then to aggregate the information that they provide. This can be done for example by considering the mean value of the one-dimensional criteria:

$$C_{mean}(\mathbf{y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}}) = \frac{1}{d}\sum_{i=1}^{d} c_i \tag{7}$$

where $c_i = C(y_{sim,i}|\mathbb{P}_{Y_{exp,i}})$ denotes the one-dimensional criterion computed for the ith component. Other aggregation functions, not considered here for the sake of brevity, can obviously be considered depending on the purpose of the validation (*min*, product or weighted product function, e.g.). As in the one-dimensional case, global indicators, such as the median value, can then be considered to summarize the distribution of C_d or C_{mean}. Nested family of C_d- or C_{mean}-trimmed regions can also be defined (in a similar way to Eqs. (2, 3)).

2.5 Distribution of multivariate criteria for some specific cases and proposed transformations

Considering the reference case where $\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}$, C_d still follows a uniform distribution as in the one-dimensional case (Figure 2(a) in blue solid line). If we further assume that the random variables $(Y_{exp,i})_{I\in\{1,\ldots,D\}}$ are independent, we can show that C_{mean} follows a Bates distribution [Johnson et al., 1994] defined on interval $[0, 1]$, as a mean of d independent uniform variables. While its mean is constant, $\mathbb{E}[C_{mean}(\mathbf{Y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}})] = 0.5$, its variance depends on dimension d since $\mathbb{VAR}[C_{mean}(\mathbf{Y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}})] = \frac{1}{\sqrt{12d}}$. This may not be desirable for a fair comparison of the predictive quality (on average) of two groups of quantities of interest, that are of different dimension d for instance. Figure 2(b) illustrates this: the PDF of C_{mean} are drawn in solid lines for different d from 5 to 40.

In addition, the result obtained in the case of a divergence between $\mathbf{Y_{sim}}$ and $\mathbf{Y_{exp}}$ is also plotted on the graphs (Figures 2(a) and 2(b), in dotted lines). More precisely, $\mathbf{Y_{sim}}$ still follows a centered Gaussian distribution but differs from $\mathbf{Y_{exp}}$ by the variance over 20% of its components. A dilation of 100%, i.e. a standard deviation twice as large as that of $\mathbf{Y_{exp}}$, is considered for the probability distribution of these components. In the case of an average validation objective for a group of outputs,

[1]Most of the time, the different experimental quantities of interest (temperature, pressure, etc.) are evaluated using different measuring instruments, the resulting measurement errors are thus independent. This assumption can also be made for the same physical quantity measured at different points or times for example.

it would be desirable to have, for the same rate of disturbed components, similar distributions of criteria (and thus a similar deviation from the reference case) regardless of d. It is clear that C_d does not exhibit this behavior at all, which is logical and consistent with what this criterion controls (see Section 2.4). Concerning C_{mean}, even though the dissimilarity between the PDFs of the perturbed and reference cases increases less rapidly, it remains dependent on d for the same rate of perturbed components, which is more problematic for this criterion since it is dedicated to the evaluation of the coherence on average. One might wrongly think that there is simply a translation (independent of d) between the PDFs of the perturbed and reference cases but this is not actually the case. This can be explained by the relatively low percentage of dilated components (here 20%). The higher this percentage, the more we would see a distribution increasingly different from that of Bates. In fact, it can be shown that C_{mean} in the case with dilation no longer follows a Bates distribution, as 20% of the individual c_i (those related to the dilated $\mathbb{Y}_{sim}$ components) no longer follow a uniform distribution. More precisely, their distribution contracts towards 0 as the variance of the $\mathbb{Y}_{sim}$ components is greater than that of the corresponding $\mathbb{Y}_{exp}$ components. Note that the associated PDF can be derived analytically from the CDF given by Equation 20 in A.2. So, only the difference between the means of the two distributions of C_{mean} for the reference and dilated cases is independent of d as it only depends on the percentage of dilated components.

So, to allow a PDF of C_{mean} less sensitive to d, and independent from it at least in the reference case (two identical distributions), we propose to consider two transformations of C_{mean} that allow a PDF independent of d, at least in the reference case (two identical distributions). The first one consists in applying the CDF of the Bates distribution (denoted $F_{Bates,d}$) to C_{mean}, as follows: $\widetilde{C}_{mean}^{Bates} = F_{Bates,d}(C_{mean})$. The transformed criterion therefore follows a uniform distribution, under the hypothesis $\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}$. This can be referred to as the probability integral transform[2] (or universality of the uniform). This transformation paves the way for a simplified interpretation and an easier comparison to the reference case, whatever the dimension d. Perfectly relevant when $\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}$, $\widetilde{C}_{mean}^{Bates}$ also remains well suited when most of the individual c_i criteria are high (the simulations are on average very consistent with the experiment for all the outputs of interest). On the other hand, it can become too penalizing as soon as a significant number of individual criteria are low, in the sense that it compresses the distribution too much, making comparisons and interpretation difficult. Figure 2(c) illustrates this analysis. Moreover, as for C_d, the PDF of $\widetilde{C}_{mean}^{Bates}(\mathbf{Y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}})$ for the same rate of perturbed $\mathbf{Y_{sim}}$ components still depends on d, but to a lesser extent. The use of $\widetilde{C}_{mean}^{Bates}$ is therefore not recommended for comparing groups of different and large dimension, especially in the case of large discrepancies between simulations and experimental results.

To address larger deviations while trying to get rid of the dimension, another solution is to transform C_{mean} to only make its mean and variance (for the reference case) independent of d, and not its entire distribution. For this, we propose the following linear transformation:

$$\widetilde{C}_{mean}^{scal}(\mathbf{y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}}) = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} c_i - \frac{\sqrt{d}-1}{2}. \tag{8}$$

The distribution of $\tilde{C}_{mean}$ under $\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}$ thus obtained is therefore very close for any d, with a constant mean and variance, respectively $\mathbb{E}[\widetilde{C}_{mean}^{scal}(\mathbf{Y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}})] = \frac{1}{2}$ and $\mathbb{VAR}[\tilde{C}_{mean}^{scal}(\mathbf{Y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}})] = \frac{1}{12}$. Note that the support of the distribution is $\left[-\frac{\sqrt{d}-1}{2}, \frac{\sqrt{d}+1}{2}\right]$ and still depends on d (see Figure 2(d), in solid lines). Moreover, we can see that the perturbation of 20% of components yields a quite similar PDF for $d = 5$ or 10. The divergence with the reference case then increases for higher d ($d = 40$, e.g.) but to a much lesser extent than for the other criteria. This behavior has also been observed for other

[2]It relates to the result that i.i.d. realizations of a random variable from any given continuous distribution can be converted to i.i.d. realizations of a variable having a standard uniform distribution.

cases of divergence (e.g. on the mean), not shown here for the sake of brevity. Applicable to broader discrepancy cases, the criterion $\widetilde{C}_{mean}^{scal}$ therefore allows for a fairer comparison regardless of d.

In summary, several likelihood-based criteria have been proposed to deal with multivariate output: C_d as the natural extension of the 1-D criterion, to be reserved for the fine validation of all the quantities of interest, C_{mean} for a validation on average which makes sense in the case of independent experimental uncertainty, and two transformed versions of C_{mean} to allow for fair validation (still on average) and comparison between multidimensional outputs, regardless of their dimension. The former should be reserved for small divergences between simulations and experimental results, while the latter could address larger ones. Of course, a selection of the most relevant criteria has been proposed, but other versions could be considered.

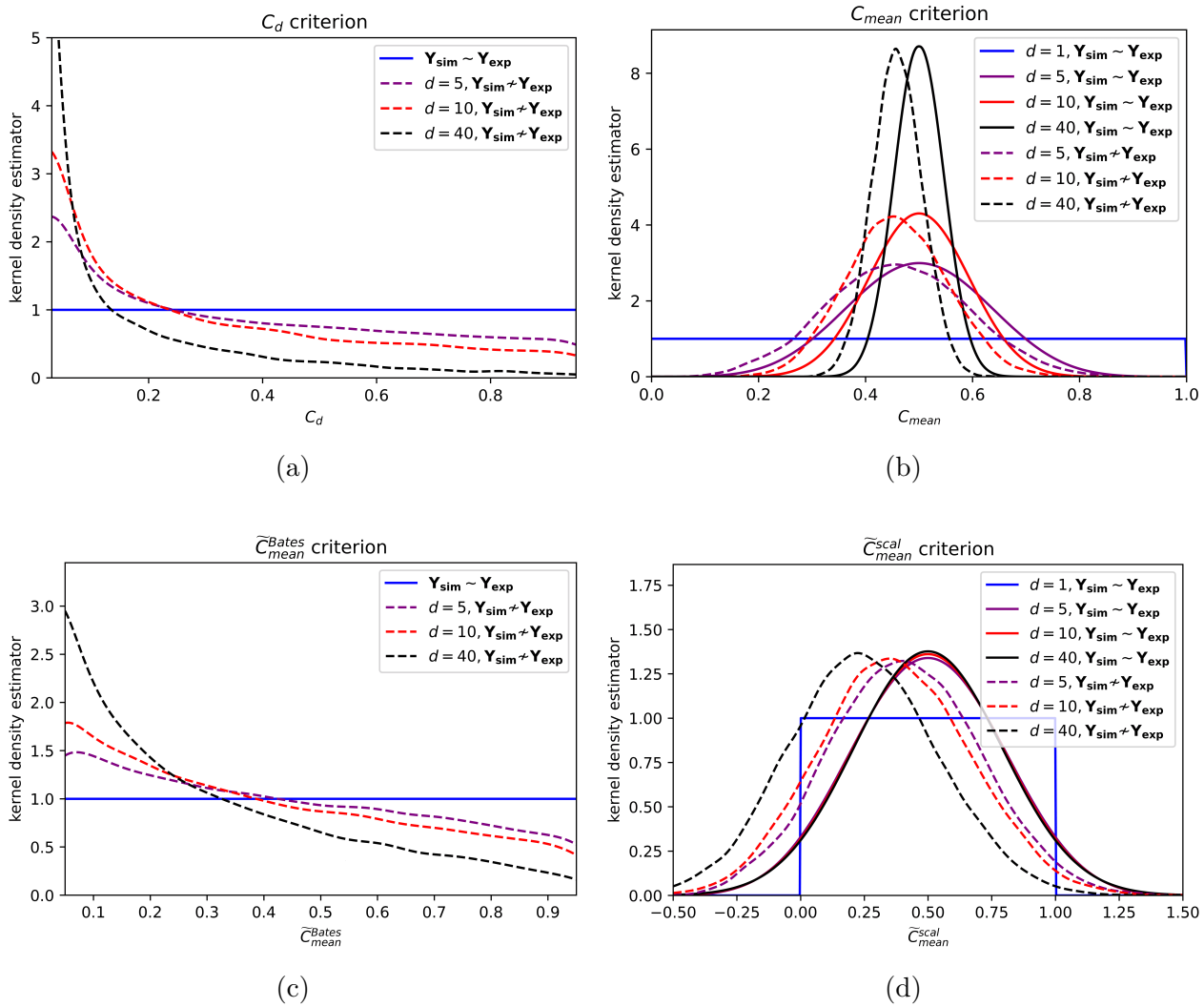


Figure 2: Estimated PDF of likelihood-based multivariate criteria for dimensions $d = 5$ to 40, and $\mathbf{Y_{exp}} \sim N_d(\mathbf{0}, I_d)$. The reference case $\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}$ is plotted in solid lines. The case where $\mathbf{Y_{sim}}$ also follows a d-normal distribution but with a different standard deviation on 20% of the components is plotted in dotted lines. The criteria represented here are C_d (Eq. 6), C_{mean} (Eq. 7), $\widetilde{C}_{mean}^{Bates}$ and $\widetilde{C}_{mean}^{scal}$ (Eq. 8).

3 Criteria based on the statistical depth of simulations

3.1 Brief review and selection of statistical depths

The notion of statistical depth was first introduced by Tukey [Tukey, 1975] as a measure of the centrality of a point among a data set in $\mathbb{R}^d$, generalizing the notion of median to the multivariate case. The concept of data depth has then been extended for the ordering of multivariate data, and different depth functions have been proposed in order to measure how "deep" a point is relative to a given data cloud. Many depth functions have been proposed for various application areas and have different characteristics regarding robustness, high dimensional computability, and ability to reflect asymmetries of the distributions (See Mosler [2013] for a complete review and a relevant classification of depth functions). Among them, we focus here on some usual distance-based and halfspaced-based depths:

- ***Mahalanobis depth*** is a distance-based depth function given by:

$$D_{Mah}(\mathbf{y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}}) = (1 + (\mathbf{y_{sim}} - \mu_{\mathbf{exp}})^\top \cdot \Sigma_{exp}^{-1} \cdot (\mathbf{y_{sim}} - \mu_{\mathbf{exp}}))^{-1}, \tag{9}$$

  where $\mu_{\mathbf{exp}} \in \mathbb{R}^d$ and $\Sigma_{exp} \in M_d(\mathbb{R})$ are the mean vector and covariance matrix of $\mathbf{Y_{exp}}$, respectively. Σ_{exp} is assumed to be nonsingular and consequently invertible.

- ***Tukey depth*** is a halfspace-based combinatorial depth which is defined as the minimum probability mass carried by any closed halfspace containing $\mathbf{y_{sim}}$:

$$D_{Tuk}(\mathbf{y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}}) = \inf\{\text{Proba}(H) : H \text{ is a closed halfspace, } \mathbf{y_{sim}} \in H\}. \tag{10}$$

- ***Spherical depth*** is defined to be the probability that the point $\mathbf{y_{sim}}$ is contained inside a random closed hyperball defined by a pair of points sampled from the reference distribution $\mathbb{P}_{\mathbf{Y_{exp}}}$. More precisely, the spherical depth is expressed as:

$$D_{Sph,init}(\mathbf{y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}}) = \text{Proba}\left[\mathbf{y_{sim}} \in S(\mathbf{Y_1}, \mathbf{Y_2})\right] \tag{11}$$

  where $\mathbf{Y_1}$ and $\mathbf{Y_2}$ are two independent random vectors in $\mathbb{R}^d$ both following $\mathbb{P}_{\mathbf{Y_{exp}}}$, and $S(\mathbf{Y_1}, \mathbf{Y_2})$ designates the unique, closed hypersphere formed by $\mathbf{Y_1}$ and $\mathbf{Y_2}$. We propose here a modified version of the spherical depth where a preliminary standardization based on the inverse square root of the covariance matrix is applied:

$$D_{Sph}(\mathbf{y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}}) = \text{Proba}\left[\Sigma_{exp}^{-1/2}\mathbf{y_{sim}} \in S(\Sigma_{exp}^{-1/2}\mathbf{Y_1}, \Sigma_{exp}^{-1/2}\mathbf{Y_2})\right]. \tag{12}$$

  In a nutshell, this transformation consists in considering the probability that $\mathbf{y_{sim}}$ belongs to the random closed ellipsoid obtained by deforming the hypersphere with the covariance matrix-based affine transformation. Consider the case where one component of $\mathbf{Y_{exp}}$, the i^{th} e.g., has a high measurement uncertainty while other components may be more precise. A simulation $\mathbf{y_{sim}}$ which is physically unlikely according to only one component $y_{sim,j}$ with $j \neq i$, will still have a high spherical depth. This problem is avoided by the standardization proposed in Eq. (12).

Note that the spherical and Tukey depths lie in $[0, 0.5]$ as opposed to $[0, 1]$ for the Mahalanobis depth. Similarly as for the likelihood-based criteria in Section 2.2, global indicators, such as the median value, can be calculated to summarize the distribution of any depth-based criterion $D(\mathbf{Y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}})$. Nested family of D^α-trimmed regions can be defined in a similar way.

It is noteworthy that a connection can be made between the likelihood-based criterion $C(y_{sim}|\mathbb{P}_{Y_{exp}})$ defined by Eq. (1) and the family of "Type D depths" defined by Zuo and Serfling [2000]. For a given

point $\mathbf{z}$ and a probability measure $\mathbb{P}_{\mathbf{X}}$, the authors define them as the minimum probability mass of $\mathbb{P}_{\mathbf{X}}$ carried by a set containing the point $\mathbf{z}$ and belonging to a given class of closed subsets in $\mathbb{R}^d$. Such kind of depth can be interpreted as the "tailedness" of $\mathbf{z}$ w.r.t. $\mathbb{P}_{\mathbf{X}}$. A direct link can therefore be made with criterion $C_d(\mathbf{y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}})$ for $\mathbf{z} = \mathbf{y_{sim}}$ and $\mathbb{P}_{\mathbf{X}} = \mathbb{P}_{\mathbf{Y_{exp}}}$ by considering for the closed subsets, the subsets of points with a PDF smaller than that of y_{sim}.

3.2 Analytical distribution for some specific cases and proposed transformations

Considering the specific case where $\mathbf{Y_{exp}}$ follows a multivariate Gaussian distribution $\mathcal{N}_d(\mu_{\mathbf{exp}}, \Sigma_{exp})$, we obtain for the Tukey depth:

$$D_{Tuk}(\mathbf{y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}}) = 1 - \Phi\left[\left\|\Sigma^{-1/2}_{\mathbf{Y_{exp}}}(\mathbf{y_{sim}} - \boldsymbol{\mu}_{\mathbf{exp}})\right\|\right],$$

where Φ denotes the CDF of the standard normal distribution (demonstration given by B). In addition, if we assume the same probability distribution for the two variables $\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}$, we obtain the following PDF for the Mahalanobis and Tukey depths:

$$f_{Mah,\mathbf{Y_{sim}}\sim\mathbf{Y_{exp}}}(x) = \frac{1}{x^2} f_{\chi^2_{(d)}}(\frac{1}{x} - 1) \qquad \forall x \in [0,1], \tag{13}$$

$$f_{Tuk,\mathbf{Y_{sim}}\sim\mathbf{Y_{exp}}}(x) = \frac{1}{f_{\mathcal{N}(0,1)}\left[\Phi^{-1}_{\mathcal{N}(0,1)}(1-x)\right]} \cdot f_{\chi_{(d)}}\left[\Phi^{-1}_{\mathcal{N}(0,1)}(1-x)\right] \qquad \forall x \in]0, 0.5], \tag{14}$$

where $f_{\mathcal{N}(0,1)}$, $f_{\chi^2_{(d)}}$ and $f_{\chi_{(d)}}$ respectively denote the PDFs of the standard normal, chi-squared and chi distribution with d degrees of freedom (for the latter two). Demonstrations are provided in B.2 and B.3 for Mahalanobis and Tukey depths, respectively. Note that for $d = 1$ (only), the Tukey depth follows a uniform distribution.

Figures 3(a) and 3(b) illustrate these PDFs for different d from 5 to 40 (in solid lines). The perturbed case (defined as in Figure 2) is also represented. Similar plots can be obtained by intensive simulation for spherical depth (not given here for sake of brevity). As previously for likelihood-based criterion C_d, the PDFs of the depth criteria have the disadvantage of having a very different shape depending on d, even when $\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}$.

To mitigate this dependence when $\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}$, the criteria can be first transformed by using their CDFs under the reference case:

$$\widetilde{D}^{CDF}_{\bullet}(\mathbf{y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}}) = F_{D_\bullet,\mathbf{Y_{sim}}\sim\mathbf{Y_{exp}}}(D_\bullet(\mathbf{y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}})), \tag{15}$$

where $F_{D_\bullet,\mathbf{Y_{sim}}\sim\mathbf{Y_{exp}}}$ is the CDF of depth criterion $D_\bullet$ when $\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}$. This CDF can be computed analytically for Mahalanobis and Tukey depths when $\mathbf{Y_{exp}}$ follows a multivariate Gaussian distribution (cf. formulas given by Eq. (24) Eq. (25) in B.2 and B.3, respectively), or otherwise by intensive simulation. Note that under the assumption of a Gaussian distribution for $\mathbf{Y_{exp}}$, the transformed Tukey depth given by Eq. (15) is actually equivalent to the criterion C_d defined by Eq. (6). Demonstration is given in C. This result is true for any given simulation vector $\mathbf{y_{sim}} \in \mathbb{R}^d$ and any dimension d. Unfortunately, the $\widetilde{D}^{CDF}_{\bullet}$ transformation does not alleviate the impact of d when $\mathbf{Y_{sim}}$ follows a perturbed distribution with the same percentage of perturbed components, as illustrated for Mahalanobis depth by Figure 3(c). As soon as d increases, the PDF of $\widetilde{D}^{CDF}_{\bullet}$ compresses rapidly around very low values of the criterion. This also raises a problem of numerical accuracy on high-dimensional data and with very large deviations. This behavior, also observed for the other CDF-based transformed depths, is understandable as depth is a measure of centrality. So the more the simulation deviates on a large number of components, the less central the simulation becomes. For this reason,

it is questionable whether the depth criteria should be modified to remove their d-dependence and cope with larger dissimilarities. At least, such modifications should be reserved for comparison (of two groups of variables for example or the predictions of two simulators) rather than for validation stricto sensu. To this end, we have considered several more or less successful modifications. One of them, relevant enough for the Mahalanobis depth, is given by:

$$\widetilde{D}^{scal}_{Mah}(\mathbf{y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}}) = \frac{d+1}{2}\, D_{Mah}(\mathbf{y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}}). \tag{16}$$

This transformation is motivated by the analysis of the PDF of D_{Mah} under the reference case and the approximation of its moments. The corrective factor $\frac{d+1}{2}$ has been chosen so that the mean value tends towards 0.5 with d, when $\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}$. As illustrated by Figure 3(d), the mean value of $\widetilde{D}^{scal}_{Mah}$ is thus less sensitive to d for the reference case, as well as for the case with dilation.

Several validation criteria from depth measures have been proposed for the validation of a simulator with multidimensional outputs, some of which have been adapted (spherical depth). They allow for a joint validation of all the outputs by assessing whether, for a given simulation, all the predicted values are central to the experimental distribution. They can naturally be used on the one-dimensional case. These depth-based criteria are geometric in nature and differ from the likelihood criteria. But it is clear that these two approaches will be very close when $\mathbf{Y_{exp}}$ follows a unimodal distribution, and in particular a Gaussian distribution. We notably show in this case the equivalence between C_d and some transformation of Tukey depth. In the more general case, we have also established relevant connections between the likelihood-based criterion C_d and family of "Type D depths". Moreover, some transformations of the depth-based criteria have been proposed for a fairer validation and comparison regardless of the dimension of the output. However, both the original and the transformed depth-based criteria assess strict validation of all predicted outputs: they will quickly become null when d increases in the case of significant deviations of the simulation from the experimental results, even on a small number of components.

4 Application to the simulation of the behavior of a nuclear material under irradiation conditions

To illustrate the practical application and value of some of the indicators previously proposed, we consider here a simulator of the behaviour of a material in irradiation conditions. This simulator models the various physical phenomena occurring in the nuclear material and provides output quantities characteristic of its evolution. For reasons of industrial confidentiality, the simulator (which will be referred to as $\mathcal{M}_{NuclMat}$) and the modeled phenomena are not detailed. The $\mathcal{M}_{NuclMat}$ model depends on several modeling parameters, some of which cannot be determined experimentally. We consider here about ten uncertain conceptual parameters without prior information on their probability distribution. Only a variation range and a uniform distribution over this range are assumed for each input parameter.

To validate the $\mathcal{M}_{NuclMat}$ simulator in steady state conditions and within a BEPU approach (i.e. including its uncertain parameters), an experimental database of about 40 experimental objects is considered. Post irradiation examinations performed at CEA give 3 types of physical quantities measured. The results of these quantities are not available for all the experimental objects. More precisely, we have a total of 94 variables of interest consisting of 40, 41, and 13 measures of type 1, 2, and 3 respectively. In addition, a measurement uncertainty is associated with each observed value: independent truncated normal distributions with known mean and standard deviation are assumed for each observed variable. The three corresponding random vectors are noted $\mathbf{Y_i}$ with $i = 1 \dots 3$ and $\mathbf{Y_{i,j}}$

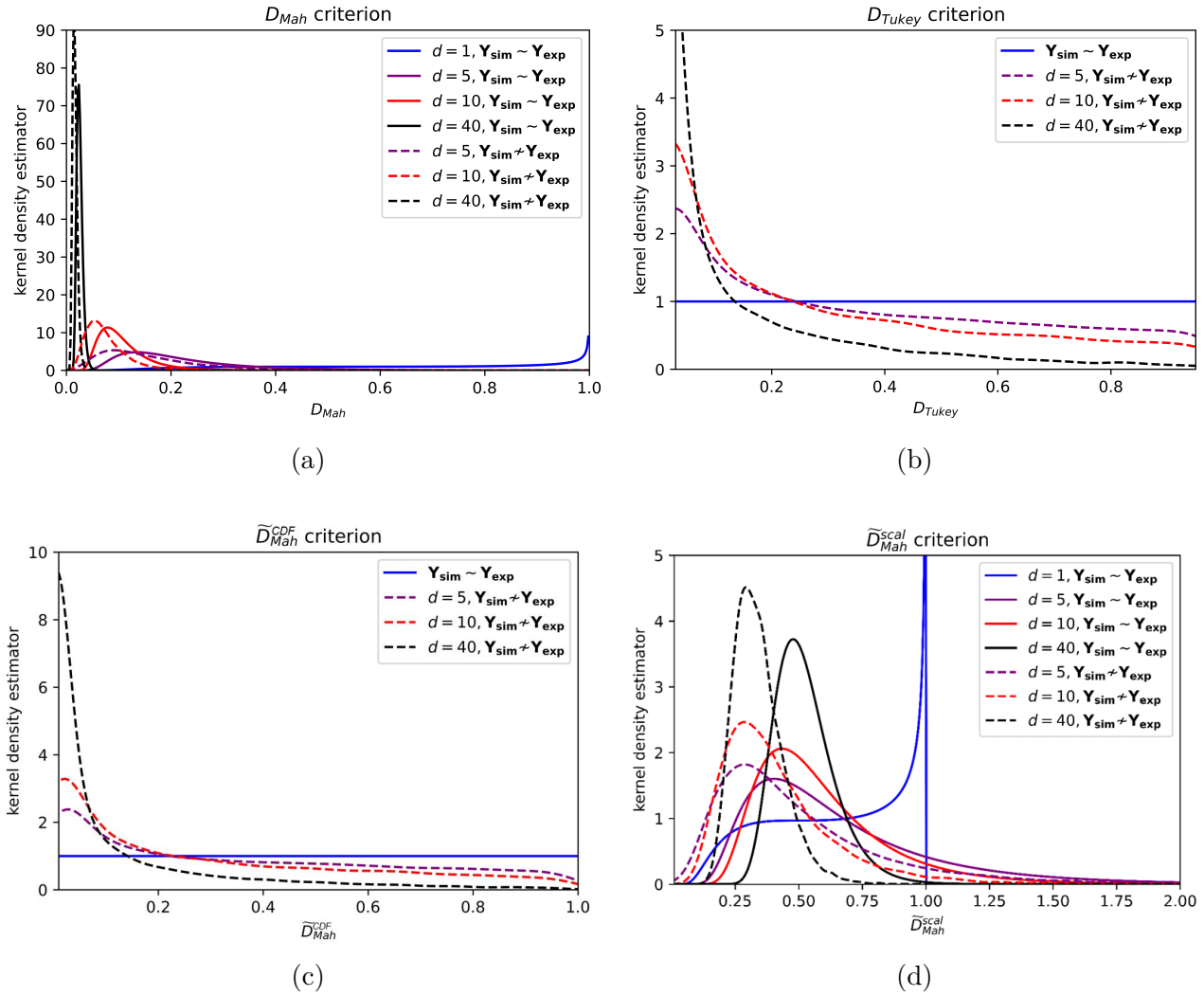


Figure 3: Estimated PDF of depth-based criteria for dimensions from $d = 5$ to 40, and $\mathbf{Y_{exp}} \sim N_d(\mathbf{0}, I_d)$. The reference case $\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}$ is plotted in solid lines. The case where $\mathbf{Y_{sim}}$ also follows a d-normal distribution but with a different standard deviation on 20% of the components is plotted in dotted lines. The criteria represented here are D_{Mah} (Eq. 9), D_{Tuk} (Eq. 10), $\widetilde{D}^{CDF}_{Mah}$ (Eq. 15) and $\widetilde{D}^{scal}_{Mah}$ (Eq. 16).

corresponds to the jth variable of group i. On the other hand, a sample of $n = 200$ simulations of the $\mathcal{M}_{NuclMat}$ simulator is available: the uncertain inputs are randomly drawn using a Latin hypercube design (Loh [1996], Park [1994]), and for every tuple of inputs, the 94 variables of interest listed above are computed by the simulator. No observed or predicted values are provided here, and all plotted values will be normalized, again for confidentiality.

The objective of the validation process of the $\mathcal{M}_{NuclMat}$ simulator ($\mathcal{M}_{NuclMat}$ code + uncertain inputs) is two-fold. Firstly, it is necessary to evaluate whether improving the knowledge and reducing the uncertainty of the inputs is required. The faithful modeling of the phenomenology by the calculation code (only) is already acquired, and it is rather a question of determining whether the uncertainty on the modeling parameters is acceptable to well represent the experimental results, or if on the contrary, a better characterization (and/or reduction) of their uncertainty is necessary. This reduction could be done via a calibration of the parameters (deterministic or Bayesian Kennedy and

O'Hagan [2001]). Secondly, we also wish to rank the 3 groups of variables of interest ($\mathbf{Y_1}$, $\mathbf{Y_2}$ and $\mathbf{Y_3}$) according to whether they are well represented or not by the simulator, and perhaps identify the group on which to focus the calibration efforts.

4.1 Graphical analysis

Before applying validation criteria, a graphical comparison between the distribution of simulations and experimental results is made. An illustration is given by Figure 4. We observe very variable results depending on the predicted output, even within the same group. Simulations are sometimes central to the experimental PDF (cases (b,d,f,h)), with a slight or small bias (cases (d) and (i) respectively), centered on the observed values but too spread out and with a very high proportion of values outside the experimental PDF (case (a)), or even completely out of line with the experimental results (very large bias, like plots (c,e,g)). This preliminary analysis highlights the complexity and the need to summarize all these very different results as best as possible.

4.2 Computation of one-dimensional criteria

The different validation criteria proposed in Sections 2 and 3 are first applied on the $\mathcal{M}_{NuclMat}$ use-case for each one-dimensional variable of interest alone. Results are given in D by Figures 6, 7 and 8 for criterion C and transformed depth-based criteria $\widetilde{D}^{CDF}_{Mah}$ and $\widetilde{D}^{CDF}_{Sph}$, respectively. All these criteria have the property of following a uniform distribution if $\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}$, which facilitates their comparison. Note that $\widetilde{D}^{CDF}_{Tuk}$ is not shown because the results are almost identical to C: the criteria are equivalent in the case of an experimental Gaussian distribution and the presence here of truncated Gaussians for some variables does not result in a significant difference.

For a given output, all 1D-criteria give similar results, which is explained by the fact that the experimental distributions here are unimodal and mostly Gaussian. The observed differences (as for $Y_{1,10}$, $Y_{1,11}$, and $Y_{1,12}$, e.g.) are explained by the truncation at 0 of the experimental Gaussian distribution, which has a greater impact on the criterion C. The obtained results also confirm the great variability between outputs. Most of the $\mathbf{Y_2}$ outputs are relatively well fitted (except for $Y_{2,5}$), while the simulated $\mathbf{Y_1}$ outputs appear inconsistent with the experimental values. The results are more heterogeneous for the $\mathbf{Y_3}$ outputs with some simulations physically very likely ($Y_{3,5}$ or $Y_{3,7}$) and others much less so ($Y_{3,9}$ and $Y_{3,12}$).

4.3 Computation of multidimensional criteria

The different multidimensional criteria are now applied to each group of outputs $\{\mathbf{Y_i}\}_{i=1...3}$, as well as to the whole set. The empirical mean of C and depth-based criteria are given by Tables 1 and 2, respectively. The variation range of criteria is recalled, as it can depend on d. In this case (i.e. for $\widetilde{C}^{scal}_{mean}$ and $\widetilde{D}^{scal}_{Mah}$), the criteria are calculated only for the three groups for an analysis of their relative value. In addition, the distributions of C-based criteria are shown as boxplots on Figure 5.

First of all, regarding the criteria on the group composed of all outputs, C_{mean} has a mean value below 0.5 (0.37 exactly): the simulator on average gives predictions that are not very consistent with experimental results. A significant proportion of outputs takes values that are physically unlikely. This naturally results in zero values for all other criteria that control the strict validation of all outputs. The $\mathcal{M}_{NuclMat}$ simulator used with no informative prior uncertainty on the modeling parameters does not yield reliable predictions. The validity of the $\mathcal{M}_{NuclMat}$ calculation code being already established, this means that a better quantification (or refinement) of the uncertainties on the modeling parameters is required.

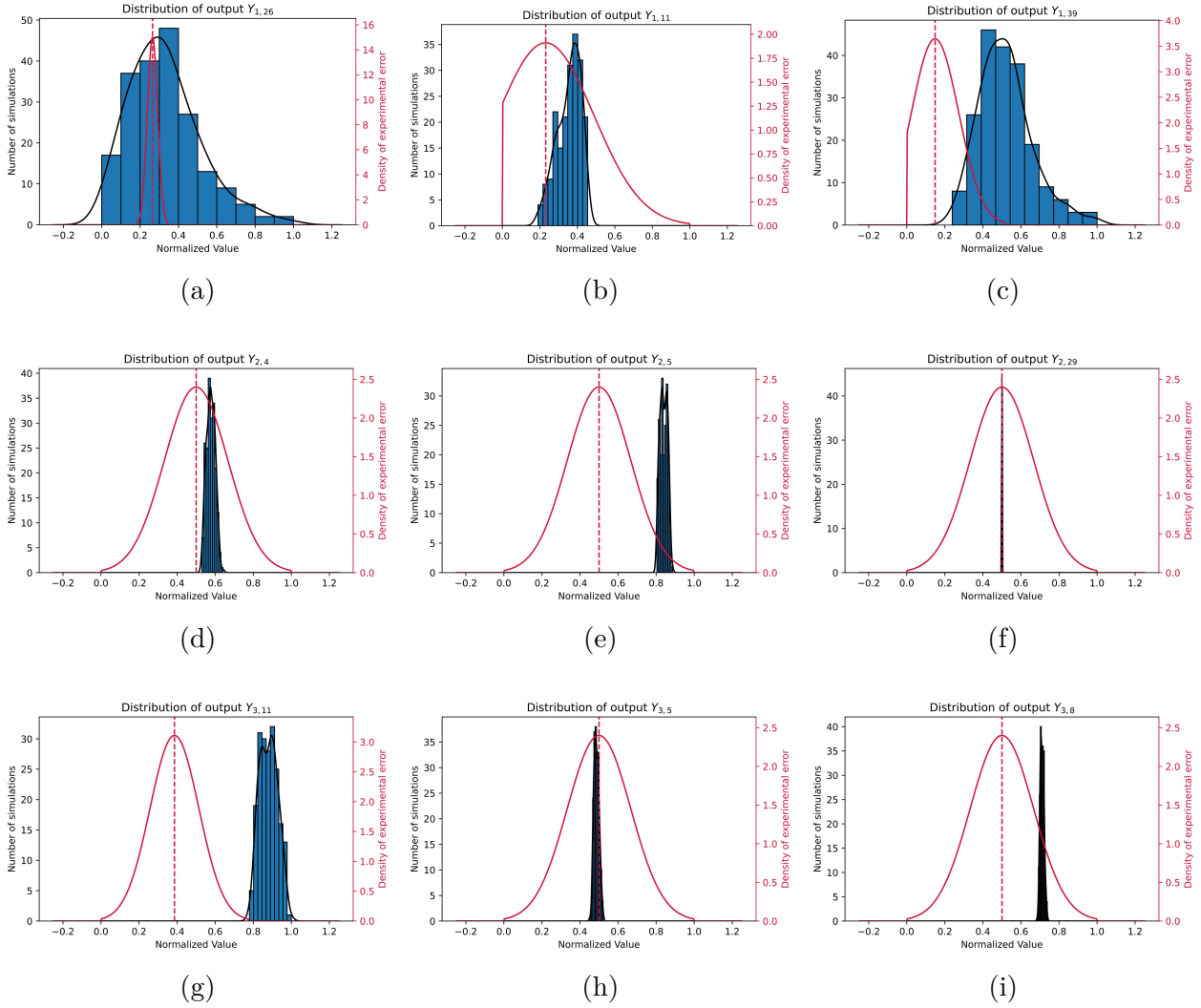


Figure 4: Comparison of the distributions of $\mathcal{M}_{NuclMat}$ simulations and experimental results, for some variables of each of the three groups ($\mathbf{Y_1}$, $\mathbf{Y_2}$ and $\mathbf{Y_3}$). A kernel density estimator is also plotted (in black solid line), for the $n = 200$-size random sample of $\mathcal{M}_{NuclMat}$ simulations. The measured experimental values (resp. the associated PDF) are indicated by a red dotted (resp. solid) line.

If we now look at the different group of outputs individually, we can observe that only the $\mathbf{Y_2}$ group is faithfully predicted with a C_{mean} equal to 0.61 (higher than 0.5 which is the mean value obtained if $\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}$), and CDF-transformed depths $\widetilde{D}^{CDF}_{Mah}$ and $\widetilde{D}^{CDF}_{Tuk}$ close to one. The latter values illustrate the interest of these transformations, the original depths being close to zero due to the very large dimension d. Concerning the two other groups, $\mathbf{Y_3}$ and $\mathbf{Y_1}$, results are far less good with a C_{mean} much lower than 0.5 and depths close to zero. Even if better results are obtained for $\mathbf{Y_3}$ with C_{mean}, this does not allow to conclude on the ranking between $\mathbf{Y_3}$ and $\mathbf{Y_1}$ because of the penalizing impact of the dimension (dimension of $\mathbf{Y_1}$ being much higher than $\mathbf{Y_3}$). To mitigate this impact, one can turn to the CDF-transformed criteria. Unfortunately, their near-zero values do not allow a robust conclusion. Criteria $\widetilde{C}^{scal}_{mean}$ and $\widetilde{D}^{scal}_{Mah}$ then offer a relevant alternative. Their analysis clearly reveals that the $\mathbf{Y_1}$ group is the most discordant with the experimental results, regardless of its larger dimension. This is confirmed by the boxplot of $\widetilde{C}^{scal}_{mean}$ given by Figure 5.

In conclusion, efforts to reduce uncertainty (e.g. through calibration) should be focused first and foremost on the $\mathbf{Y_1}$ group.

	C_d $[0,1]$	C_{mean} $[0,1]$	$\widetilde{C}_{mean}^{Bates}$ $[0,1]$	$\widetilde{C}_{mean}^{scal}$ $\left[\frac{1-\sqrt{d}}{2}, \frac{1+\sqrt{d}}{2}\right]$
$\mathbf{Y_1}$, $d = 40$	0	0.13	0	-1.85
$\mathbf{Y_2}$, $d = 41$	0.99	0.61	0.99	1.23
$\mathbf{Y_3}$, $d = 13$	0	0.36	0.04	-0.01
$\{\mathbf{Y_1}, \mathbf{Y_2}, \mathbf{Y_3}\}$, $d = 94$	0	0.37	0	

Table 1: Empirical mean of multidimensional likelihood-based criteria computed on the $n = 200$-size sample of $\mathcal{M}_{NuclMat}$ simulations. Results are given for each group of outputs and for the whole set (last line). The theoretical interval of possible variation for each criterion is also recalled (under the name of the criterion).

	D_{Mah} $[0,1]$	$\widetilde{D}_{Mah}^{CDF}$ $[0,1]$	$\widetilde{D}_{Mah}^{scal}$ $\left[0, \frac{d+1}{2}\right]$	D_{Tuk} $[0,0.5]$	$\widetilde{D}_{Tuk}^{CDF}$ $[0,1]$	D_{sph} $[0,0.5]$	$\widetilde{D}_{sph}^{CDF}$ $[0,1]$
$\mathbf{Y_1}$, $d = 40$	0	0	0.02	0	0	0	0
$\mathbf{Y_2}$, $d = 41$	0.05	0.99	1.02	0	0.99	0.44	0.68
$\mathbf{Y_3}$, $d = 13$	0.02	0	0.16	0	0	0	0
$\{\mathbf{Y_1}, \mathbf{Y_2}, \mathbf{Y_3}\}$, $d = 94$	0	0		0	0	0	0

Table 2: Empirical mean of multidimensional depth-based criteria computed on the $n = 200$-size sample of $\mathcal{M}_{NuclMat}$ simulations. Results are given for each group of outputs and for the whole set (last line). The theoretical interval of possible variation for each criterion is also recalled (under the name of the criterion).

5 Conclusions and prospects

In support to the Best-Estimate-Plus-Uncertainty (BEPU) methodology or more generally to the Verification, Validation and Uncertainty Quantification (VVUQ) approach, this paper has addressed the problem of a quantified validation of simulation tools, with uncertain input parameters, and from the comparison with available experimental results. The objective is therefore to assess the consistency of the simulated outputs, individually and conjointly, with the experimental data. To meet this objective, we have proposed and investigated different statistical indicators, based on the concepts of likelihood and depth statistics. We have extended them to the multidimensional case (i.e. when there are several scalar experimental results, and corresponding simulated outputs). The proposed indicators (or criteria) can be applied to the result (output(s)) of a single simulation or to a random sample of simulations. In the latter case, each criterion is itself a random variable, as a function of the random output of the simulator. It thus yields a global BEPU validation of the simulator, as well as a ranking of simulations according to their consistency with experiments.

For each proposed criterion, its probability distribution was first studied analytically or by simulation, for some specific cases: Gaussian experimental distribution, identical probability distribution of

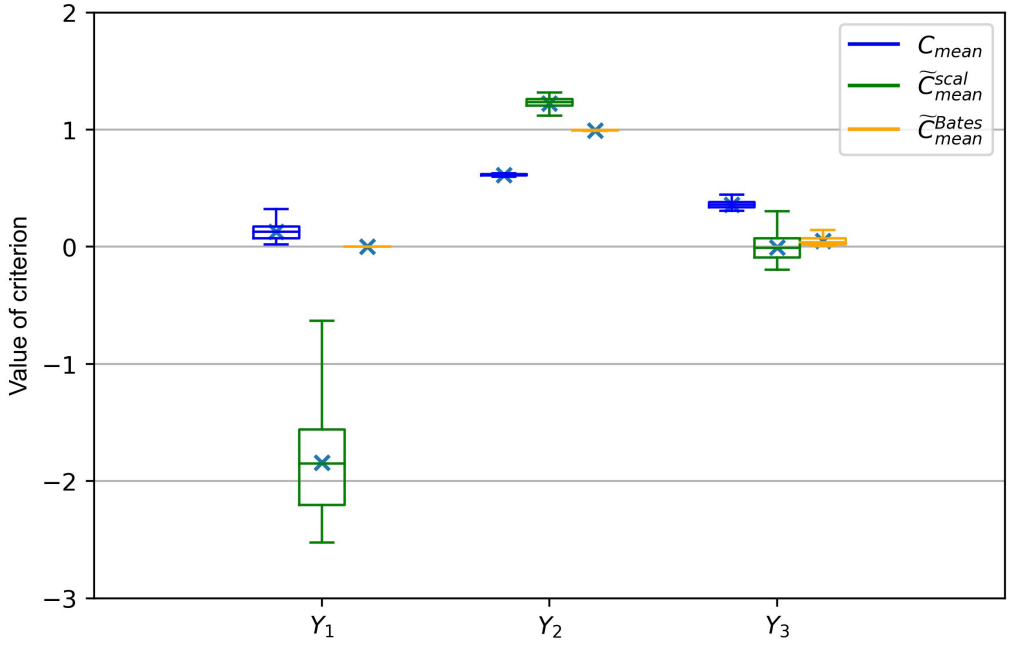


Figure 5: Boxplot of the multidimensional likelihood-based criteria C_{mean}, $\widetilde{C}^{scal}_{mean}$ and $\widetilde{C}^{Bates}_{mean}$ for the 3 different groups of outputs, computed on the sample of $\mathcal{M}_{NuclMat}$ simulations. Empirical mean and median are indicated by a cross and a horizontal line, respectively.

experiments and simulations (reference case), or for a specific divergence between both distributions. Then, some natural extensions to a multivariate output were proposed. Their behavior was analyzed, in particular as a function of the dimension of the output variable, and recommendations for their use were formulated with regard to the objectives of the validation (strict or average validation).

From there, transformed criteria were proposed either to "homogenize" the criteria by ensuring that they all follow a uniform distribution in the reference case, or to mitigate in a more general case the impact of the output dimension. These transformations (or standardization in a nutshell) allows a fairer comparison of the different criteria, independently of the dimension, and w.r.t. the reference case. Some of them should be reserved for small divergences between simulations and experimental results, while others, which have only a relative interpretation, could address and compare larger divergences.

Finally, the validation criteria were applied to a test case with a simulator modeling the behaviour of a nuclear material. This simulator is based on a validated release of the $\mathcal{M}_{NuclMat}$ code, and on a set around ten uncertain model parameters. As they cannot be directly measured, only variation ranges of these parameters are first assumed. The objective here was to evaluate the accuracy of the simulator and more precisely the relevance of the assumed non-informative uncertainty on the model parameters. For this, we relied on an experimental database composed of 94 variables of interest, grouped into 3 types. On the simulation side, a sample of $n = 200$ simulations was available: each simulation corresponds to a random draw of the input model parameters and leads to a prediction of the 94 variables of interest. For any given output variable (among the 94), the different unidimensional criteria gave similar results. By contrast, a large variability was observed between the outputs: most of the outputs of one group are relatively well fitted, while the simulated outputs of the other two groups are much more inconsistent with the experimental values. The multidimensional criteria applied to the whole set of outputs showed that the simulator used with non-informative prior uncertainty as currently defined, is not validated. As the validity of the $\mathcal{M}_{NuclMat}$ calculation code has

been established, it is the uncertainty of the model parameters that must be better characterized and more precisely reduced. The transformed criteria then offered a relevant alternative to compare the 3 output groups with each other, independently of their dimensions. It was clearly revealed that the first group of variables is the most discordant with the experimental results.

Efforts to reduce model uncertainties should therefore focus primarily on improving the modeling and prediction of the the first group of outputs. To this end, a Bayesian calibration of the model parameters is currently ongoing and should allow a more accurate representation of the outputs. In addition, to better understand the role played by the model parameters in the faithful representation of reality, it would be interesting to perform a sensitivity analysis of the validation criteria themselves. The objective would be to identify the influential (and non-influential) parameters on the criteria and especially on the occurrence of a low value (target sensitivity analysis, see [Marrel and Chabridon, 2021]). For this purpose, HSIC dependence measures ([Gretton et al., 2005]) seem to be very relevant because they allow to capture a large spectrum of dependencies. They are well adapted to the size of the sample ($n = 200$) and conditional versions have recently been proposed by [Marrel and Chabridon, 2021]. This sensitivity analysis could be carried out for each group of outputs or for the whole. In addition, the marginal distributions of the main influential parameters, a priori and conditional on low values of the criteria, could also be compared and interpreted.

As far as the validation criteria themselves are concerned, their extension to the case of functional outputs (temporal or spatial outputs, for example) is currently being studied. This extension could be based either on functional dimension reduction for likelihood-based criteria, or on functional depth statistics, such as the band-depth measure of [López-Pintado and Romo, 2009] or the h-mode depth ([Cuevas et al., 2007]).

Acknowledgments

We warmly thank Guillaume Damblin (research engineer at CEA) for very useful technical discussions on this work, and Bertrand Iooss (Senior researcher at EDF R&D) for sharing his expertise in VV&UQ and for kindly allowing us to use one of his illustrative diagrams.

A Demonstrations of formulas for the likelihood-based criteria

A.1 Analytical distribution of criterion $C(Y_{sim}|\mathbb{P}_{Y_{exp}})$ when $Y_{sim} \sim Y_{exp}$

In the general case when $Y_{sim} \sim Y_{exp}$, we can prove that $C(Y_{sim}|\mathbb{P}_{Y_{exp}})$ follows a uniform distribution $\mathcal{U}_{[0,1]}$. Let denote f_Y the probability density function (PDF) of both Y_{exp} and Y_{sim}, and F_C the CDF of $C(Y_{sim}|\mathbb{P}_{Y_{exp}})$. More generally, F_W will denote the CDF of any variable W. We have $\forall x \in [0,1]$:

$$\begin{aligned}
F_C(x) &= \text{Proba}\left[C(Y_{sim}|\mathbb{P}_{Y_{exp}}) \leq x\right] \\
&= \text{Proba}\left[\text{Proba}\left[f_Y(Y_{exp}) \leq f_Y(Y_{sim}) \mid Y_{sim}\right] \leq x\right] \\
&= \text{Proba}\left[\mathbb{E}\left[\mathbf{1}_{f_Y(Y_{exp}) \leq f_Y(Y_{sim})} \mid Y_{sim}\right] \leq x\right] \\
&= \text{Proba}\left[\mathbb{E}\left[\mathbf{1}_{W \leq W'} \mid Y_{sim}\right] \leq x\right] \text{ where } W = f_Y(Y_{exp}) \text{ and } W' = f_Y(Y_{sim}) \\
&= \text{Proba}\left[\mathbb{E}\left[\mathbf{1}_{W \leq W'} \mid W'\right] \leq x\right] \\
&= \text{Proba}\left[\text{Proba}\left[W \leq W' \mid Y_{sim}\right] \leq x\right] \\
&= \text{Proba}\left[F_W(W') \leq x\right]. \qquad (17)
\end{aligned}$$

W and W' are i.i.d. with cumulative density function (CDF) F_W and therefore $F_W(W') \sim \mathcal{U}_{[0,1]}$, provided F_W is continuous and strictly increasing. The first condition is verified as Y and therefore $f_Y(Y)$ are continuous variables. For the second condition, if we denote $f_{Y,min}$ (resp. $f_{Y,max}$) the minimal (resp. maximal) value of f_Y on its support[3], we first use the fact that F_W is strictly increasing on $[f_{Y,min}; f_{Y,max}]$ since $f_Y(Y)$ is continuous and from the intermediate value theorem. Secondly, W as well as W' are defined on $[f_{Y,min}; f_{Y,max}]$, by definition.

Hence Eq. (17) becomes $F_C(x) = \text{Proba}\left[F_W(W') \leq x\right] = x$. It follows that $C(Y_{sim}|\mathbb{P}_{Y_{exp}}) \sim \mathcal{U}_{[0,1]}$, when $Y_{sim} \sim Y_{exp} \sim Y$ and if $f_Y(Y)$ has a continuous and strictly increasing CDF.

A.2 Analytical expression of one-dimensional criterion for the Gaussian case

If Y_{exp} is normally distributed with $Y_{exp} \sim \mathcal{N}(\mu_{exp}, \sigma^2_{exp})$, we have:

$$\begin{aligned}
C(y_{sim}|\mathbb{P}_{Y_{exp}}) &= \text{Proba}\left[(f_{Y_{exp}}(Y_{exp}) \leq f_{Y_{exp}}(y_{sim})\right] \\
&= \text{Proba}\left[\frac{1}{\sqrt{2\pi}\sigma_{exp}} e^{\frac{-(Y_{exp}-\mu_{exp})^2}{2\sigma^2_{exp}}} \leq \frac{1}{\sqrt{2\pi}\sigma_{exp}} e^{\frac{-(y_{sim}-\mu_{exp})^2}{2\sigma^2_{exp}}}\right] \\
&= 1 - \text{Proba}\left[\left(\frac{Y_{exp}-\mu_{exp}}{\sigma_{exp}}\right)^2 \leq \left(\frac{y_{sim}-\mu_{exp}}{\sigma_{exp}}\right)^2\right] \\
&= 1 - F_{\chi^2_{(1)}}\left[\left(\frac{y_{sim}-\mu_{exp}}{\sigma_{exp}}\right)^2\right] \text{ since } \frac{Y_{exp}-\mu_{exp}}{\sigma_{exp}} \sim \mathcal{N}(0,1). \qquad (18)
\end{aligned}$$

[3]The set-theoretic support of a density f defined on a set X is defined as the set of points where f is non-zero: $\text{supp}(f) = \{x \in X \,:\, f(x) \neq 0\}$.

If Y_{sim} is also normally distributed with $Y_{sim} \sim \mathcal{N}(\mu_{sim}, \sigma^2_{sim})$ then the global criterion $C(Y_{sim}|\mathbb{P}_{Y_{exp}})$ is a random variable defined as follows:

$$\begin{aligned} C(Y_{sim}|\mathbb{P}_{Y_{exp}}) &= 1 - F_{\chi^2_{(1)}}\left[\left(\frac{Y_{sim}-\mu_{exp}}{\sigma_{exp}}\right)^2\right] \\ &= 1 - F_{\chi^2_{(1)}}\left[\frac{\sigma^2_{sim}}{\sigma^2_{exp}}\left(\frac{Y_{sim}-\mu_{exp}}{\sigma_{sim}}\right)^2\right] \\ &= 1 - F_{\chi^2_{(1)}}\left[\alpha^2 \cdot \widetilde{Y}\right] \text{ where } \alpha = \frac{\sigma_{sim}}{\sigma_{exp}} \text{ and } \widetilde{Y} = \left(\frac{Y_{sim}-\mu_{exp}}{\sigma_{sim}}\right)^2. \end{aligned} \tag{19}$$

We have $\widetilde{Y} \sim \chi^2_{(1)(\mu_{sim}-\mu_{exp})^2}$ where $\chi^2_{(1)(\lambda)}$ denotes the non-central chi-squared distribution of non-centrality parameter λ). We then obtain for the CDF of $C(Y_{sim}|\mathbb{P}_{Y_{exp}})$:

$$\begin{aligned} \forall x \in [0,1], \quad F_C(x) &= \text{Proba}\left[1 - F_{\chi^2_{(1)}}\left[\alpha^2 \cdot \widetilde{Y}\right] \leq x\right] \\ &= 1 - F_{\widetilde{Y}}\left[\frac{1}{\alpha^2} F^{-1}_{\chi^2_{(1)}}(1-x)\right] \\ &= 1 - F_{\chi^2_{(1)(\mu_{sim}-\mu_{exp})^2}}\left[\frac{\sigma^2_{exp}}{\sigma^2_{sim}} \cdot F^{-1}_{\chi^2_{(1)}}(1-x)\right]. \end{aligned} \tag{20}$$

Note that in the particular case where $Y_{sim} \sim Y_{exp}$ we find the result of A.1:

$$C(Y_{sim}|\mathbb{P}_{Y_{exp}}) = 1 - F_{\chi^2_{(1)}}\left[F^{-1}_{\chi^2_{(1)}}(1-x)\right] = x.$$

B Demonstrations for depth statistics with multivariate Gaussian distribution

B.1 Formulation of the Tukey depth when $\mathbf{Y_{exp}} \sim \mathcal{N}_d(\boldsymbol{\mu}_{\mathbf{exp}}, \Sigma_{exp})$

First, we consider $D_{Tukey}(\mathbf{z}|\mathbb{P}_{\mathbf{X}})$ when $\mathbf{X} \sim \mathcal{N}_d(\mathbf{0}, I_d)$. In this case, the closed halfspace that contains the smallest number of data points with boundary through $\mathbf{z}$ is delimited by the hyperplane that contains $\mathbf{z}$ and is orthogonal to the vector $\mathbf{z} = (z_1, \ldots, z_d)^\top$. The equation of this hyperplane can be written as $\sum_{i=1}^d \alpha_i x_i - \beta = 0$ with α_i being the coordinates of $\mathbf{z}$ in the natural basis, *i.e.* $\alpha_i = \frac{z_i}{\|\mathbf{z}\|}$ and $\beta = \|\mathbf{z}\|$. The Tukey depth can therefore be expressed as following:

$$\begin{aligned} D_{Tukey}(\mathbf{z}|\mathbb{P}_{\mathbf{X}}) &= \inf\{\text{Proba}(H) : H \text{ is a closed halfspace, } \mathbf{z} \in H\} \\ &= \text{Proba}\left[\sum_{i=1}^d \alpha_i X_i - \beta \geq 0\right] \\ &= 1 - \text{Proba}\left[\sum_{i=1}^d \alpha_i X_i \leq \beta\right]. \end{aligned} \tag{21}$$

Since $\sum_{i=1}^d \alpha_i^2 = 1$ and $\mathbf{X} \sim \mathcal{N}_d(\mathbf{0}, I_d)$, we have $\sum_{i=1}^d \alpha_i X_i \sim \mathcal{N}(0,1)$ and:

$$\begin{aligned} D_{Tukey}(\mathbf{z}|\mathbb{P}_{\mathbf{X}}) &= 1 - \Phi(\beta) \quad \text{with } \Phi \text{ the CDF of } \mathcal{N}(0,1) \\ &= 1 - \Phi(\|\mathbf{z}\|). \end{aligned} \tag{22}$$

If we now generalize to the case $\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}} \sim \mathcal{N}_d(\boldsymbol{\mu_{exp}}, \Sigma_{exp})$, we have $\Sigma_{exp}^{-1/2}(\mathbf{Y_{exp}} - \boldsymbol{\mu_{exp}}) \sim \mathcal{N}_d(\mathbf{0}, I_d)$, and since the Tukey depth satisfies the property of linear invariance, we obtain:

$$D_{Tukey}(\mathbf{y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}}) = D_{Tukey}\left(\Sigma_{exp}^{-1/2}(\mathbf{y_{sim}} - \boldsymbol{\mu_{exp}})|\mathbb{P}_{\Sigma_{exp}^{-1/2}(\mathbf{Y_{exp}} - \boldsymbol{\mu_{exp}})}\right) = 1 - \Phi\left[\left\|\Sigma_{exp}^{-1/2}(\mathbf{y_{sim}} - \boldsymbol{\mu_{exp}})\right\|\right]. \tag{23}$$

B.2 Distribution of the Mahalanobis depth when $\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}} \sim \mathcal{N}_d(\mu_{\mathbf{exp}}, \Sigma_{exp})$

Let $\mathbf{W} = \Sigma_{exp}^{-1/2}(\mathbf{Y_{sim}} - \mu_{\mathbf{exp}})$, we have $W \sim \mathcal{N}_d(\mathbf{0}, I_d)$. The CDF of $D^{Mah}(\mathbf{Y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}})$ is expressed $\forall x \in [0,1]$ by:

$$\begin{aligned}
F_{Mah,\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}}(x) &= \text{Proba}\left[D^{Mah}(\mathbf{Y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}}) \leq x\right] \\
&= \text{Proba}\left[\left(1 + (\mathbf{Y_{sim}} - \mu_{\mathbf{exp}})^\top \Sigma_{exp}^{-1}(\mathbf{Y_{sim}} - \mu_{\mathbf{exp}})\right)^{-1} \leq x\right] \\
&= \text{Proba}\left[(1 + \mathbf{W}^\top \mathbf{W})^{-1} \leq x\right] \\
&= \begin{cases} 1 - F_{\chi^2_{(d)}}\left(\frac{1}{x} - 1\right) & \text{for } x > 0 \text{ since } \mathbf{W}^\top \mathbf{W} \sim \chi^2_{(d)} \\ 0 \text{ otherwise} \end{cases},
\end{aligned} \tag{24}$$

where $F_{\chi^2_{(d)}}$ denotes the CDF of the chi-squared distribution with d degrees of freedom. The associated PDF is given $\forall x \in [0,1]$ by:

$$f_{Mah,\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}}(x) = \begin{cases} \frac{1}{x^2} f_{\chi^2_{(d)}}\left(\frac{1}{x} - 1\right) & \text{for } x > 0, \\ 0 \text{ otherwise,} \end{cases}$$

with $f_{\chi^2_{(d)}}$ the PDF of the chi-squared distribution with d degrees of freedom.

B.3 Distribution of the Tukey depth when $\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}} \sim \mathcal{N}_d(\mu_{\mathbf{exp}}, \Sigma_{exp})$

Keeping the previous notation $\mathbf{W} = \Sigma_{exp}^{-1/2}(\mathbf{Y_{sim}} - \mu_{\mathbf{exp}}) \sim \mathcal{N}_d(\mathbf{0}, I_d)$, we have $\|\mathbf{W}\| \sim \chi_{(d)}$. From Eq. (23), we obtain the CDF of $D_{Tukey}(\mathbf{Y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}})$, given $\forall x \in]0, 0.5]$ by:

$$\begin{aligned}
F_{Tukey,\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}}(x) &= \text{Proba}\left[D_{Tukey}(\mathbf{Y_{sim}}|\mathbb{P}_{\mathbf{Y_{exp}}}) \leq x\right] \\
&= \text{Proba}\left[1 - \Phi(\|\mathbf{W}\|) \leq x\right] \quad \text{with } \Phi \text{ being the CDF of } \mathcal{N}(0,1) \\
&= \text{Proba}\left[\Phi^{-1}(1-x) \leq \|\mathbf{W}\|\right] \quad \text{since } \Phi \text{ is a strictly increasing function} \\
&= 1 - F_{\chi_{(d)}}\left(\Phi^{-1}(1-x)\right) \quad \text{since } \|\mathbf{W}\| \sim \chi_{(d)}
\end{aligned} \tag{25}$$

where $F_{\chi_{(d)}}$ denotes the CDF of the chi distribution with d degrees of freedom. For the left bound, we have $F_{Tukey,\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}}(0) = 0$.
Hence the PDF can be deduced $\forall x \in]0, 0.5]$:

$$f_{Tukey,\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}}(x) = \frac{1}{f_{\mathcal{N}(0,1)}\left[\Phi^{-1}(1-x)\right]} \cdot f_{\chi_{(d)}}\left[\Phi^{-1}(1-x)\right]$$

with $f_{\chi(d)}$ the PDF of the chi distribution with d degrees of freedom and $f_{\mathcal{N}(0,1)}$ the PDF of standardized Gaussian distribution. For the left bound, we have $f_{Tukey,\mathbf{Y_{sim}} \sim \mathbf{Y_{exp}}}(0) = 0$ (and continuity of the PDF). Note that for $d = 1$, we can easily show that the PDF obtained is that of the uniform distribution.

C Link between the Tukey depth based on CDF transformation and the likelihood-based criterion

We will demonstrate in the following that if the experimental distribution is Gaussian, the CDF-transformed Tukey depth given Eq. (15) is strictly equivalent to the criterion C_d (Eq. (6)) for any given simulation vector $\mathbf{z} \in \mathbb{R}^d$. So, assuming that $\mathbf{Y_{exp}} \sim \mathcal{N}_d(\mu_{\mathbf{exp}}, \Sigma_{exp})$ and denoting $\mathbf{w} = \Sigma_{exp}^{-1/2}(\mathbf{z} - \mu_{\mathbf{exp}})$, we first easily obtain as in Eq. (18) that:

$$\begin{aligned} C_d(\mathbf{z}|\mathbb{P}_{Y_{exp}}) &= 1 - F_{\chi^2_{(d)}}\left[\mathbf{w}^\top \mathbf{w}\right] \text{ since } \Sigma_{exp}^{-1/2}(\mathbf{Y_{exp}} - \mu_{\mathbf{exp}}) \sim \chi^2_{(d)} \text{ under the Gaussian assumption} \\ &= 1 - F_{\chi^2_{(d)}}\left[\|\mathbf{w}\|^2\right]. \end{aligned} \tag{26}$$

Secondly, under the same assumptions, the CDF-based transformed Tukey depth given by Eq. (15) becomes

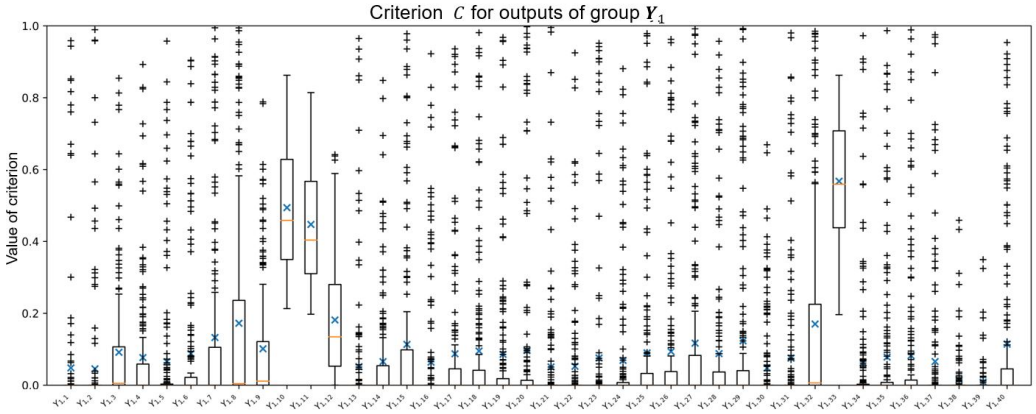
$$\begin{aligned} \widetilde{D}_{Tukey}(\mathbf{z}|\mathbb{P}_{\mathbf{Y_{exp}}}) &= 1 - F_{\chi_{(d)}}\left(\Phi^{-1}(1 - D_{Tukey}(\mathbf{z}|\mathbb{P}_{\mathbf{Y_{exp}}}))\right) \text{ from Eq. (25)} \\ &= 1 - F_{\chi_{(d)}}\left(\Phi^{-1}(\Phi(\|\mathbf{w}\|))\right) \text{ from Eq. (23)} \\ &= 1 - F_{\chi_{(d)}}\left(\|\mathbf{w}\|\right). \end{aligned} \tag{27}$$

Since the standardized chi distribution with d degrees of freedom is the distribution followed by the square root of a chi-squared random variable, we have $F_{\chi_{(d)}}(x) = F_{\chi^2_{(d)}}(x^2)\ \forall x \geq 0$ and therefore the equality of the two criteria:

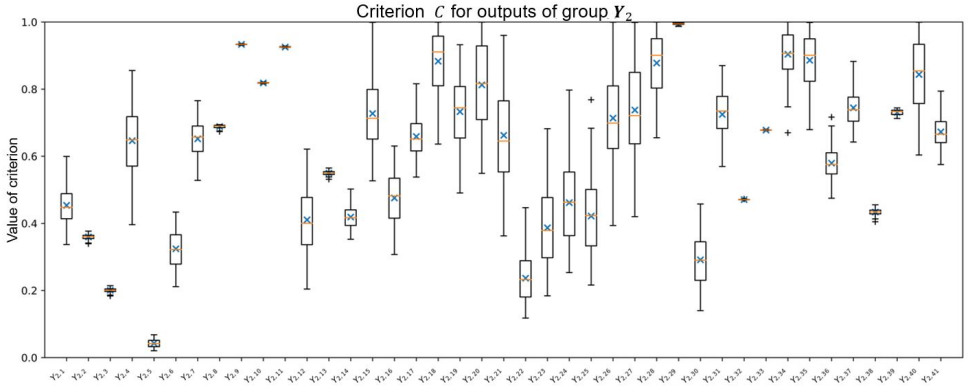
$$\text{if } \mathbf{Y_{exp}} \sim \mathcal{N}_d(\mu_{\mathbf{exp}}, \Sigma_{exp}), \text{ then } C_d(\mathbf{z}|\mathbb{P}_{Y_{exp}}) = \widetilde{D}_{Tukey}(\mathbf{z}|\mathbb{P}_{\mathbf{Y_{exp}}}). \tag{28}$$

D Additional results on the application case

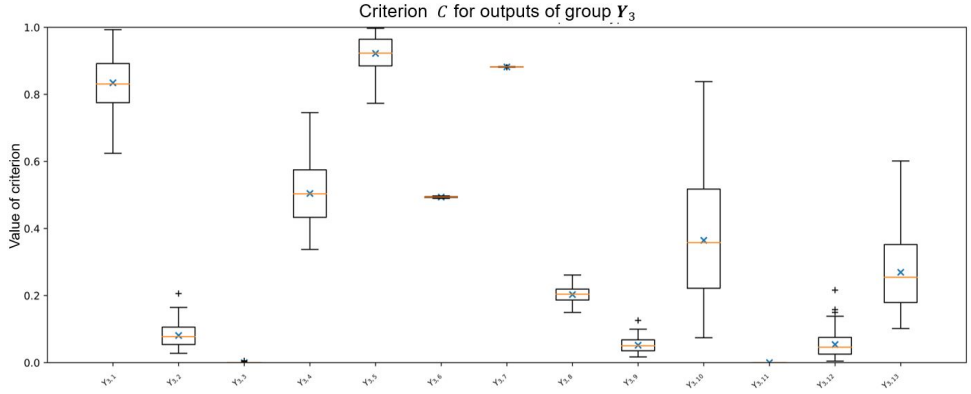
The results obtained for each of the 94 outputs for likelihood and two depth-based criteria are given by Figures 6, 7 and 8.



(a)

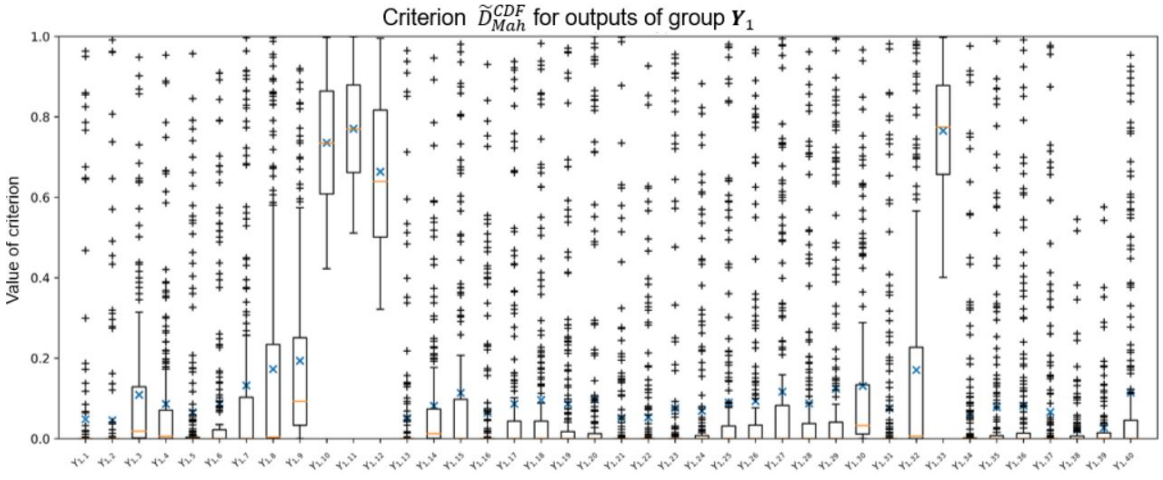


(b)

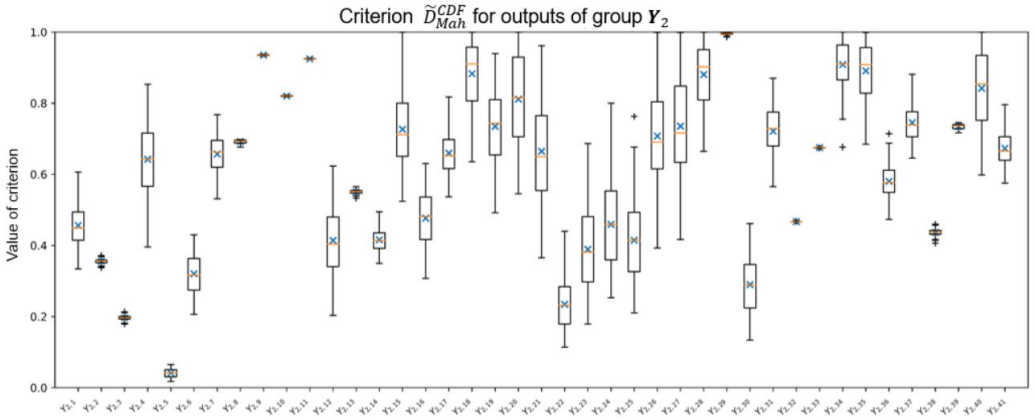


(c)

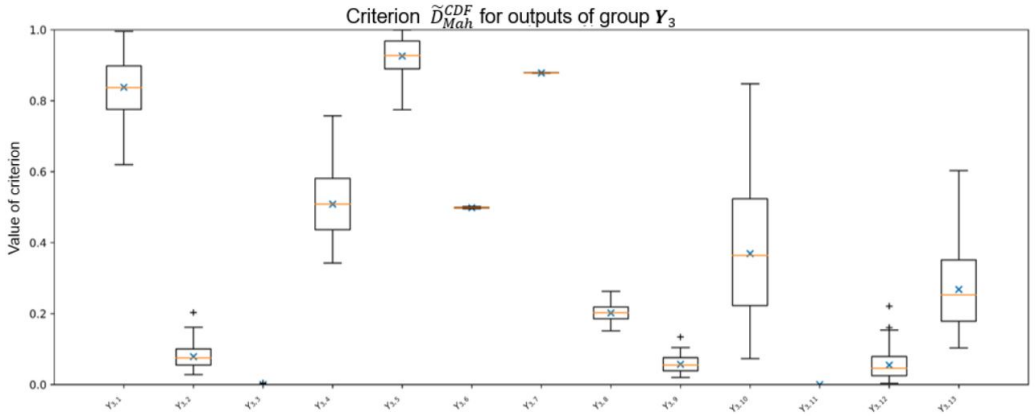
Figure 6: Boxplots of one-dimensional criterion C (Eq. 1), computed from the $n = 200$-size sample of $\mathcal{M}_{NuclMat}$ simulations. Results are plotted for each output of the 3 groups of outputs: $\mathbf{Y_1}$ (a), $\mathbf{Y_2}$ (b) and $\mathbf{Y_3}$ (c). Empirical mean and median are indicated by a blue cross and a red line, respectively.



(a)

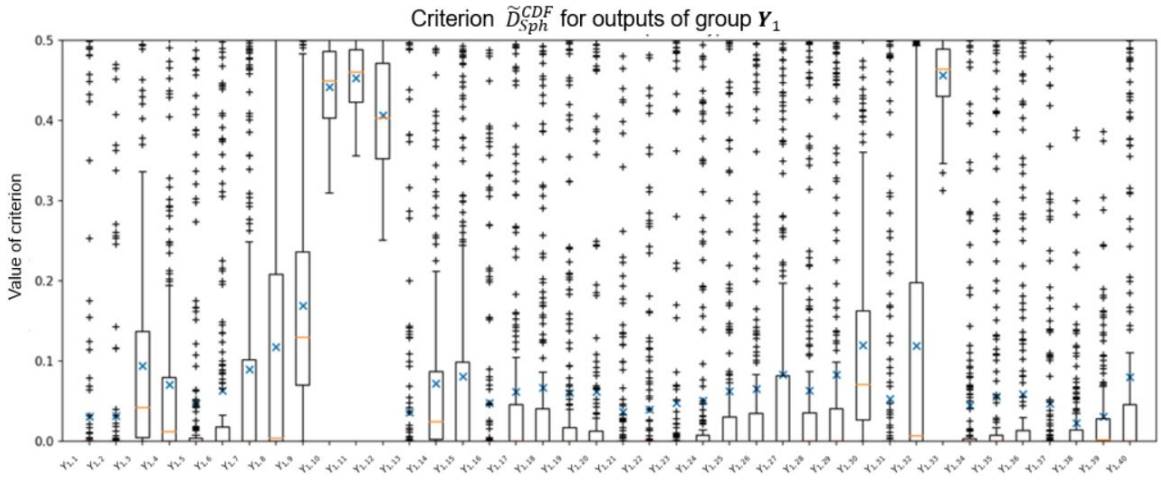


(b)

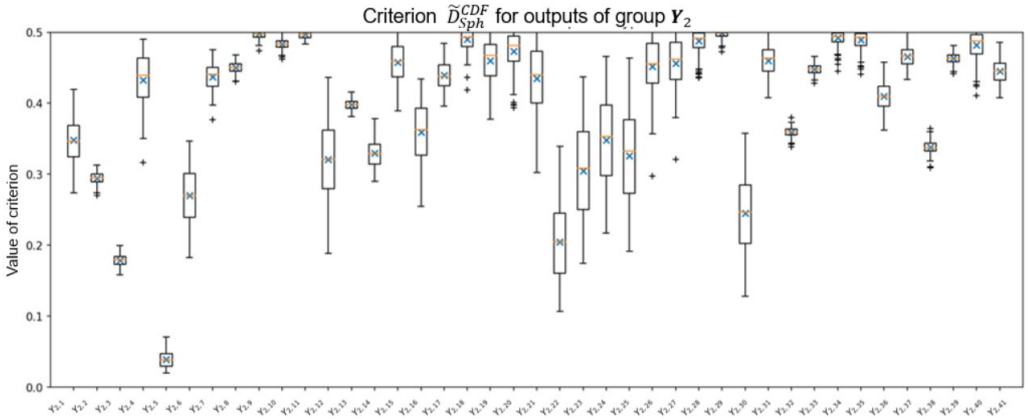


(c)

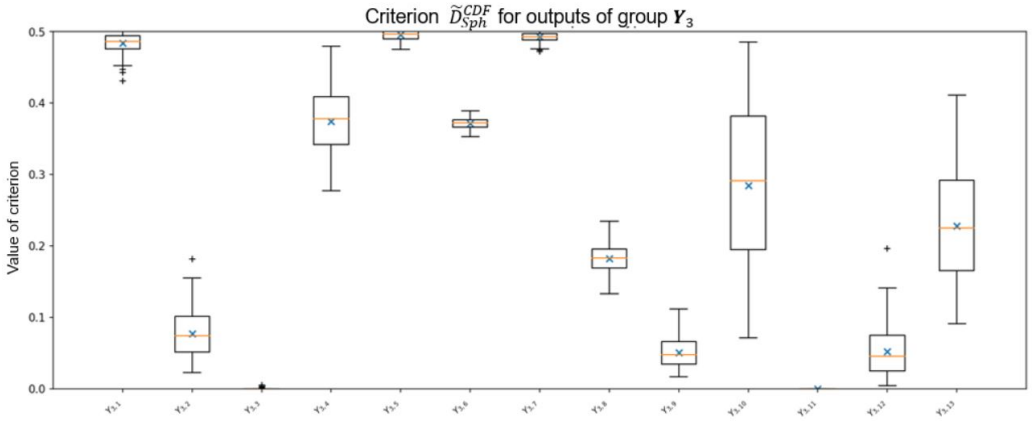
Figure 7: Boxplots of one-dimensional transformed Mahalanobis depth (Eq. 15 applied to D_{Mah}), computed from the $n = 200$-size sample of $\mathcal{M}_{NuclMat}$ simulations. Results are plotted for each output of the 3 groups of outputs: $\mathbf{Y_1}$ (a), $\mathbf{Y_2}$ (b) and $\mathbf{Y_3}$ (c). Empirical mean and median are indicated by a blue cross and a red line, respectively.



(a)



(b)



(c)

Figure 8: Boxplots of one-dimensional transformed spherical depth-based criteria (Eq. 15 applied to D_{Sph}), computed from the $n = 200$-size sample of $\mathcal{M}_{NuclMat}$ simulations. Results are plotted for each output of the 3 groups of outputs: $\mathbf{Y_1}$ (a), $\mathbf{Y_2}$ (b) and $\mathbf{Y_3}$ (c). Empirical mean and median are indicated by a blue cross and a red line, respectively.

References

ASME (2009). *Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfers*. ASME V&V 20-2009. American Society of Mechanical Engineers.

ASME (2019). *Standard for Verification and Validation in Computational Solid Mechanics*. ASME V&V 10-2019. American Society of Mechanical Engineers.

Baccou, J., Zhang, J., Fillion, P., Damblin, G., Petruzzi, A., Mendizábal, R., Reventos, F., Skorek, T., Couplet, M., Iooss, B., Oh, D.-Y., Takeda, T., and Sandberg, N. (2020). Sapium: A generic framework for a practical and transparent quantification of thermal-hydraulic code model input uncertainty. *Nuclear Science and Engineering*, 194(8-9):721–736.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 4:300–307.

Cuevas, A., Febrero, M., and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496.

De Rocquigny, E., Devictor, N., and Tarantola, S., editors (2008). *Uncertainty in industrial practice*. Wiley.

Ferson, S. and Oberkampf, W. (2009). Validation of imprecise probability models. *International Journal of Reliability and Safety*, 3:3–22.

Ghanem, R., Higdon, D., and Owhadi, H., editors (2017). *Springer Handbook on Uncertainty Quantification*. Springer.

Gretton, G., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *Proceedings Algorithmic Learning Theory*, pages 63–77. Springer-Verlag.

Helton, J. (1997). Uncertainty and sensitivity analysis in the presence of stochastic and subjective uncertainty. *Journal of Statistical Computation and Simulation*, 57(1-4):3–76.

Iooss, B. (2019). Methods and issues for the analysis of complex systems by numerical simulation at EDF. Presentation at Industry Day of ICIAM2019, Valencia, Spain,July, 15-19.

Johnson, N., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions, Volume 2*. Wiley & Sons.

Kennedy, M. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society*, 63(3):425–464.

Liu, Y., Chen, W., Arendt, P., and Huang, H. (2011). Toward a better understanding of model validation metrics. *Journal of Mechanical Design*, 133(7).

Loh, W.-L. (1996). On Latin hypercube sampling. *Annals of Statistics*, 24:2058–2080.

López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734.

Marie, N., Marrel, A., and Herbreteau, K. (2019). Statistical methodology for a quantified validation of sodium fast reactor simulation tools. *Journal of Verification, Validation and Uncertainty Quantification*, 4(3).

Marrel, A. and Chabridon, V. (2021). Statistical developments for target and conditional sensitivity analysis: application on safety studies for nuclear reactor. *Reliability Engineering and System Safety*, 214:107711.

Mosler, K. (2013). Depth statistics. In Becker, C., Fried, R., and Kuhnt, S., editors, *Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather*, pages 17–34. Springer Berlin Heidelberg.

National Research Council (2012). *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*. The National Academies Press, Washington, DC.

Oberkampf, W., Helton, J., and Sentz, K. (2001). Mathematical representation of uncertainty. In *19th AIAA Applied Aerodynamics Conference*, page 1645.

Oberkampf, W. L. and Barone, M. F. (2006). Measures of agreement between computation and experiment: Validation metrics. *Journal of Computational Physics*, 217(1):5–36.

Oberkampf, W. L. and Roy, C. J. (2010). *Verification and Validation in Scientific Computing*. Cambridge University Press.

Oberkampf, W. L. and Trucano, T. G. (2002). Verification and validation in computational fluid dynamics. *Progress in Aerospace Sciences*, 38(3):209–272.

Park, J.-S. (1994). Optimal Latin-hypercube designs for computer experiments. *Journal of Statistical Planning and Inference*, 39:95–111.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver*, volume 2, pages 523–531.

Wilson, G. E. (2013). Historical insights in the development of best estimate plus uncertainty safety analysis. *Annals of Nuclear Energy*, 52:2–9.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, 28(2):461–482.