13th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP13
Seoul, South Korea, May 26-30, 2019

Towards a Bayesian framework for model validation

Ander Gray
PhD Student, Inst. for Risk and Uncertainty, Univ. of Liverpool, Liverpool, UK

Edoardo Patelli
Professor, Inst. for Risk and Uncertainty, Univ. of Liverpool, Liverpool, UK

ABSTRACT: In this paper we discuss some concepts and a methodology of a Bayesian framework for model validation under uncertainty, which produces a probabilistic value for a models validity and may be used in the design of "validation experiments. By using a stochastic metric as a measure of the distance between experiment and prediction, we update a validation distribution. We show this in practice using a simple numerical experiment and discuss the current shortcomings of the method. We finally discuss the role of information entropy in designing validation experiments.

1. INTRODUCTION

Following the greater availability of resources, computational models are being increasingly used by the engineer to address the performance and reliability of his/her designs and by the scientist to perform prediction and inferences. Recently also, physics and engineering codes describing different phenomena are being integrated to produce multi-physical and multi-scale systems simulations[1]. It is evident that if these unions are to be trusted uncertainty quantification must play a central role within the constituents and at the interface; and that a concrete framework for addressing their reliability and predictive capability is needed. This task falls to Verification and Validation (V&V). The ASME (2006)[2] provides a widely accepted formal definition and guide for V&V. They can be summarised as follows:

- Verification: How well the computational model is mathematically representative of its conceptual design.
- Validation: How well the computational model is physically representative for its intended use.

Verification can be subdivided further into *code verification* as the removal of logical errors in the implementation, and *calculation verification* as removal of errors associated with any numerical approximation, such as discretization of continuous quantities or the truncation of infinite sequences. It is evident that this should be performed prior to any comparison or improvement to the model by physical data.

It is validation which is the definitive test for whether a model is fit for its intended application. It is here when the model is directly compared to physical data. At this time the model should not be changed or improved. If model calibration is performed it should come prior to the validation test, and any data used in calibration should not be reused in validation. It is argued that the experiment used in validation should be specially designed to either prove or disprove that the model is acceptable (Oberkampf et al. (2004)). Designing validation experiments can be difficult, particularly when obtaining experimental samples is costly. In this paper we adopt concepts and propose a Bayesian framework for model validation which, along with providing a probabilistic value for the validity of a model, may be used by the analyst in the design of

[1] The coupling of neutronics and thermal–hydraulic codes is a good example that will soon become common practice in the civil nuclear industry. Avramova and Ivanov (2010)

[2] American Society of Mechanical Engineers

validation experiments. This paper follows from recent published work on a frequentest probabilistic metric for validation (Dvurecenska et al. (2018)): this work may act as its Bayesian compliment.

2. PROPOSED APPROACH

The proposed methodology is based on the following assumptions:

- Both the validation data and model are uncertain in the subjective probabilistic sense: for example model uncertainty may come from a previous Bayesian calibration which provides an uncertain model; or by other methods for quantifying uncertainty. The assumption of probabilistic uncertainty may be relaxed however.
- A stochastic metric is chosen to determine the distributional distance between the data and model.
- A tolerance has been provided by the engineer or scientist for what is an acceptable distance between model and data.

It should be noted for the last point that we provide no advice on how to define an acceptable tolerance for a models intended use, this will most likely need to be discussed between the computational modeller and the experimentalist.

In the next section we will discuss a stochastic metric useful for validation. Following this we will discuss a Bayesian scheme for updating a validation distribution. Finally we will discuss the role of information entropy in designing validation experiments.

2.1. A STOCHASTIC METRIC FOR VALIDATION

When performing a comparison between an uncertain model and uncertain data, some notion of distance is required. This is what a stochastic metric will provide. There are some characteristics for a distance metric that are favourable for validation:

1. *Retains the physical units of the distributions*

Most stochastic metrics result in some statistical measure that is both difficult to interpret and can be unfamiliar to the modeller. A metric which preserves physical units would provide a statistical distance that is familiar to the engineer. It also allows for the definition of the minimum tolerable distance between prediction and experiment to be defined in physical units.

2. *That the metric may be used to compare the various output dimensions simultaneously*

In the case where one would like to address the complete predictive capacity of a computational model, it is necessary to perform the distance comparison on multiple dimensions of the models output space at once. If the physical unit is preserved in distance evaluation, then output dimensions with different physical units will need to be projected onto a universal unitless scale.

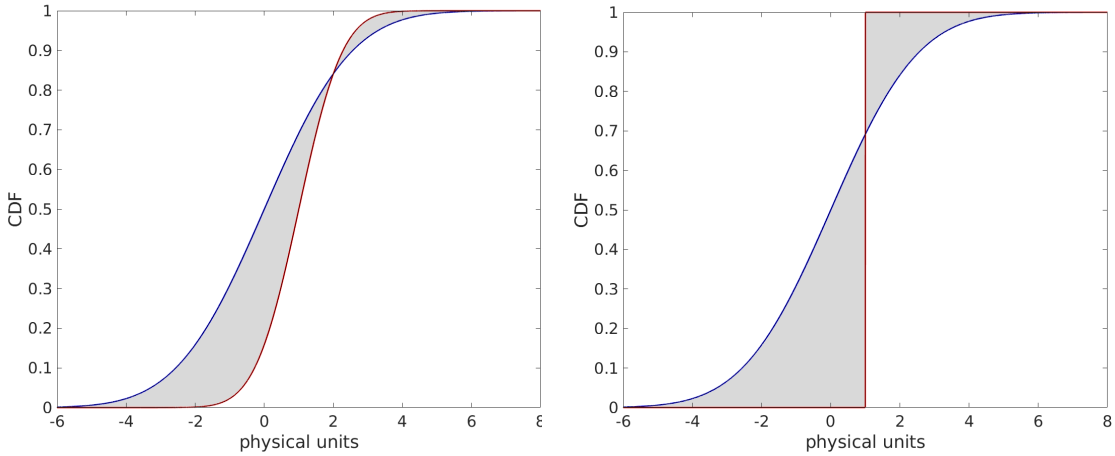
3. *Holds the mathematical definition of a metric*

Most stochastic metrics do not hold the mathematical definition of a metric. A formal mathematical metric d holds the following properties (Fréchet (1906)):

1. Non-negativity
   - $d(x,y) \geq 0$,
2. Symmetry
   - $d(x,y) = d(y,x)$,
3. Triangle inequality
   - $d(x,z) \leq d(x,y) + d(y,z)$, and
4. Identity of indiscernibles
   - $d(x,y) = 0 \iff x = y$.

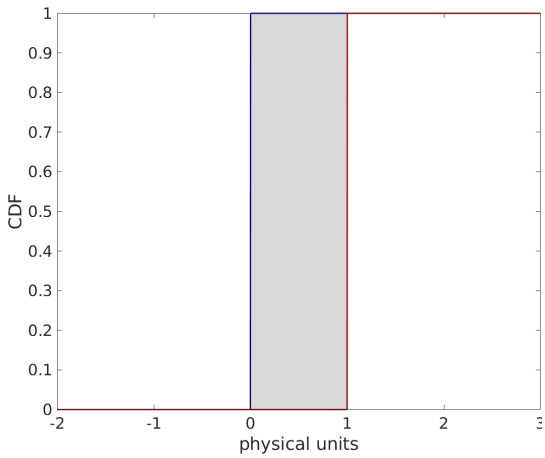
It should also hold that as the two distributions under comparison tend further and further apart that the metric should tend to infinity.

4. *That the metric may be used for comparing data that is not only distributional*

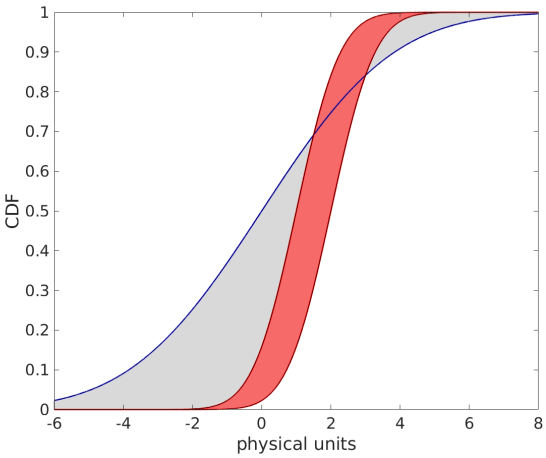


(a) A comparison of two distributions

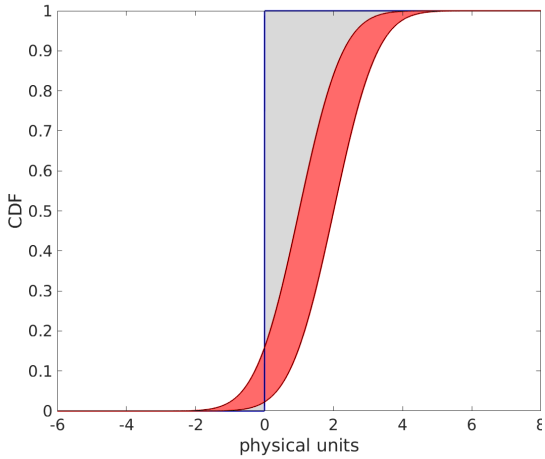
(b) Data point and a distribution



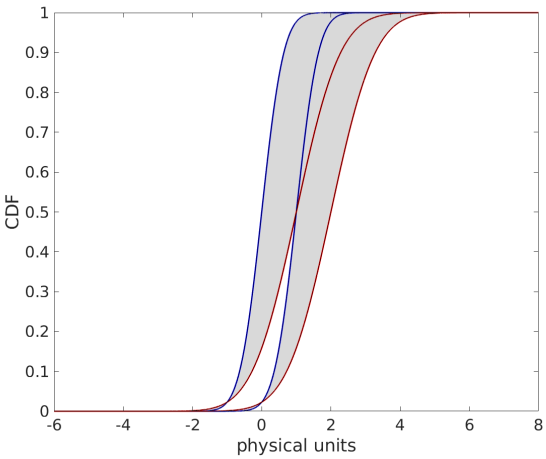
(c) Two data points



(d) Distribution and an imprecise distribution



(e) A point and an imprecise distribution



(f) Two imprecise distributions

Figure 1: A comparison of various types of uncertain data using the area metric. The resulting distances are 1.1666u, 1.7911u, 1u, 1.5622u, 1.0918u and 2.017u respectively and where u is the physical unit of the abscissa axis.

We may sometimes be presented with a validation case where our data may be a mix of certain and probabilistic, or that the model or data may be uncertain but in a non-probabilistic[3] sense. For example we may have a model which produces an interval prediction as apposed to a probability distribution, and we would like to compare it with distributional experimental data. Or it may also be the case where we have an uncertain distribution, like a probability box (Ferson et al. (2015)), that we would like to compare. A stochastic metric which is robust to any type of presented data is favourable.

Ferson et al. (2008) proposed a stochastic metric which holds all of the above characteristics. They suggested to use the area between the predictive and data distributions as a measure of their comparison. This, what they call the area metric, is the horizontal integral of the two distributions commutative distribution functions (CDF), or:

$$d(P,E) = \int_{-\infty}^{+\infty} |P(x) - E(x)| dx, \tag{1}$$

where P and E are the predictive and experimental distributions, respectively. Figure 1 shows the area metric being used for various types of data points. Notably, when the metric is performed with two scalars, as it is in (c), the metric returns their difference. When comparing an imprecise distribution, as in (d), the distance is taken as the area between the distribution and the envelopes of the imprecise distribution, where any part of the distribution that falls within the envelopes (the red area in d) will score a zero in the area metric. The distance between the distribution and an imprecise distribution is found to be:

$$\frac{A_1 + A_2 - A_3}{2}, \tag{2}$$

where A_1 and A_2 is the area between the distribution and the two envelopes and A_3 is the area between the envelopes. If the entire distribution falls within the envelopes, then the area metric returns a distance of zero.

[3]Non-probabilistic here meaning other than probabilistic rather than not-probabilistic. There exist types of uncertainty that are best modelled by theories that are extensions of probability theory (Klir (2005))

The area metric may also be used to compare multiple output dimensions at once. This is done by transforming samples from the experimental distribution onto the CDF axis using the CDF of the predictive distribution. The distribution of transformed samples from all outputs may now be compared to a uniform distribution. By definition, if the experimental samples are distributed according to the predictive distribution then the transformed samples should be uniformly distributed. This technique, called u-pooling, is very well described by Ferson et al. (2008) in their original area metric paper.

In the next section we will discuss our proposed Bayesian framework where the area metric may be used to provide a probabilistic value for model validity.

2.2. *UPDATING A VALIDATION DISTRIBUTION*

Without losing generality, uncertain experimental data and model may be presented as (Kennedy and O'Hagan (2001)):

$$y(x) = \tilde{y}(x) + e(x) + \delta_1(x) \tag{3}$$

and

$$f(\theta; x), \tag{4}$$

where $\tilde{y}(x)$ is the true physical response both the experiment and model are aiming for. y(x) is the experimentally measure response, with the measurement noise e(x) being a random variable of mean y(x) and δ_1(x) corresponding to experimental bias, which is some unknown function. x here is known as a system variable, a variable which the system response depends on. It could for example be energy in a spectrum or time in a time series. $f(\theta; x)$ is our computational model, which attempts to predict the true system response $\tilde{y}(x)$. θ are uncertain inputs to our model, which if have been previously characterised by calibration will provide a probabilistic prediction for our model. In general, a model bias is also included, however Kennedy and O'Hagan (2001) provide a well established Bayesian framework for calibrating both input parameters and model bias.

The aim of this work is that for a given f and y to provide a probabilistic value for the degree to which they agree. For this we propose updating a validation distribution, with a range between 0 and 1 corresponding to the probability that the model is valid for the given data. Bayes' law applied to the validation problem is:

$$P(V|y,f) = \frac{P(y|V,f)P(V,f)}{P(y,f)}. \tag{5}$$

Here $P(V,f)$ is a prior probability distribution for our model validity, which lies between 0 and 1, and where the value V corresponds the agreement of the model and data. If no prior knowledge is known about the models validity then we propose that a uniform distribution (or a non-informative prior if available) should be used between 0 and 1; or if one is available, the posterior of a previous validation analysis may also be used.

$P(V|y,f)$ is the posterior: a probability distribution with the range of the prior but has been altered by data y. Not only does this provide the most likely value for validity, the mode of the posterior, but also provides the uncertainty associated with this estimate: the distributions dispersion. The evidence $P(y,f)$ is a constant which ensures that the posterior is normalised.

$P(y|V,f)$ is referred to as the likelihood in Bayesian updating. For our validation case it is here that the stochastic metric, data and model are cast probabilistically to be used in the updating procedure. As is often the case in Bayesian updating, constructing a likelihood that is rigorous and true to our updating problem is the main technical difficulty in this scheme.

We propose that the beta distribution be used in the updating, a common distribution used in Bayesian updating. For this initial work, we use the fact that the binomial and beta distributions are prior conjugates, allowing for an analytic solution to the updating. We create very simple binomially distributed data by comparing the models predictive distribution with the experimental distribution using the area metric at the locations of the experimental points in the input space. If the distance between experiment and prediction falls outside the provided tolerance, then the model is considered to fail at this location, and returns a 0. Otherwise the model passes and a 1 is generated. This series of passes and failures may then be used to update the prior beta distribution analytically to solve for the posterior.

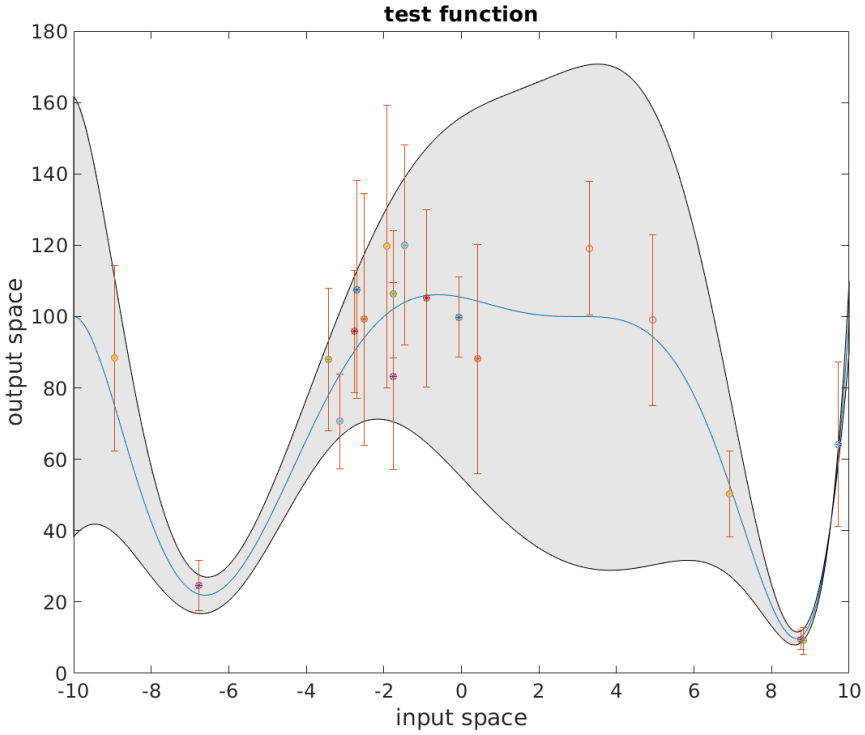
2.3. NUMERICAL EXAMPLE

Figure 2 shows an example of this framework being used for a simple test function, with Gaussian predictions, and synthetic experimental points also normally distributed. Both the grey envelope of the model and the error bars of the experiment show a single standard deviation. The data points that have passed the tolerance test have been marked with a cross. The resulting posterior distribution is also shown. 20 data points were used with a tolerance which was provided as a fraction of the models mean prediction.

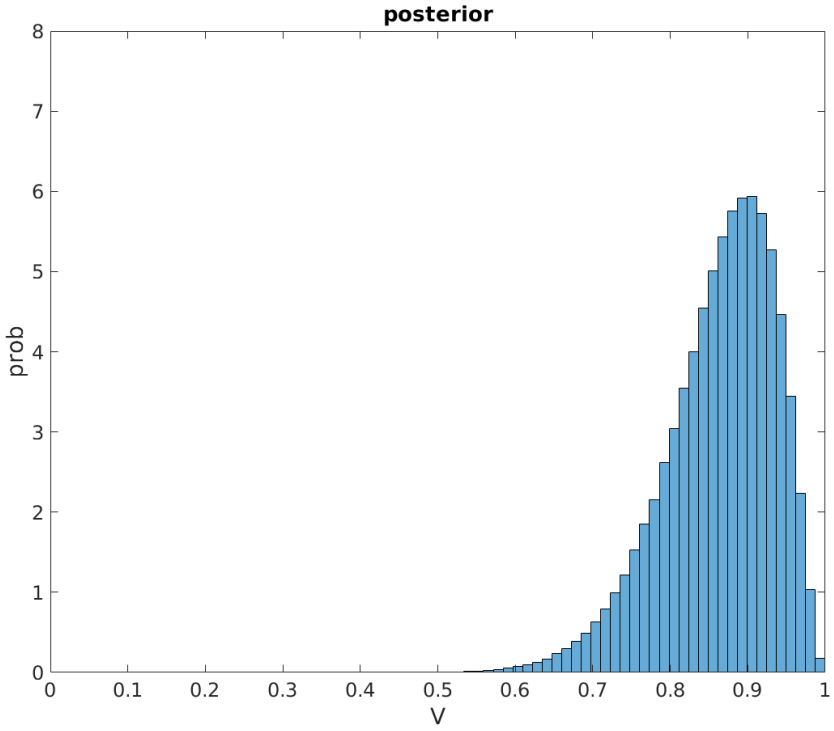
The test function[4] is a simple polynomial, with the uncertainty band around it produced using a trigonometric function. The 20 experimental points are randomly selected: their location is uniformly sampled on the $[-10,10]$ input space. Their mean is then randomly perturbed about the mean of the test function, with a random variance also selected. Since the analytical solution is available for the updating, the computational cost of producing the posterior is negligible. Although not realistic, it serves as as an initial test case for this framework providing distributional data and model. In future, we would like to further test our framework on a real case validation problem, specifically the 2014 Sandia labs V&V challenge problem (Hu and Orient (2016)).

There are some clear shortcomings with the presented framework thusfar. For example, each experimental data point is given equal weighting in the updating. We believe that ideal scenario is that the weighting of each point should be a function of the models and data uncertainty, and their corresponding distance. Another consequence of having equal weighting for each validation test is that

[4]Test function, area metric implementation and Bayesian updating code available for matlab. Results reproducible: `https://github.com/AnderGray/bayesianModelValidation`



(a) An uncertain test function with synthetic experimental points.



(b) Posterior of the validation distribution. Its mean and deviation are 0.8637 and 0.0715 respectively

Figure 2: Example of the validation framework for a test function

the posterior is a direct function of the number of validation points used. That is that the more you experimental points used the posterior will become more narrow, a general feature of Bayesian updating. The ideal scenario is that one should be able to determine the number of experimental points need to validate the model. As it stands, only by providing more experimental points will one achieve a more accurate validation estimate. Notice also the two failed data points around 3 and 5 in the input space of figure 2 (a). These two have been marked as failures although they fall within the predictive distribution. It is debatable whether they are failures or not. On one hand, one could argue that they should be passes since they fall within the models uncertainty envelope. Another argument however could be that the predictive distribution is too uncertain to provide a prediction for experiments at their accuracy, and so the model should be considered to fail at these points. The solution to this ambiguity most likely depends on the models intended use.

We believe that if these shortcomings can be solved, then this framework will provide a rigorous method for validating uncertain models. If these issues are overcome, we also believe that this framework may be used to design validation experiments using the principle of information entropy. We will discuss this concept in the following section.

2.4. INFORMATION ENTROPY: THE VALUE OF AN EXPERIMENTAL POINT

We believe that information entropy, often used in Bayesian updating to produce prior distributions of maximum uncertainty under some known constraints, can be used in this framework to design validation experiments. Information entropy may be regarded as the uncertainty that is stored within a distribution (Klir (1991)). The larger a distributions dispersion, the greater its information entropy is and hence the predictability of a sample drawn from this distribution falls: its associated uncertainty is greater. In his book on generalised information theory Klir (2004) discusses the role of uncertainty in information theory and outlines the intimate duality information and uncertainty have.

When designing a validation experiment with a limited budget for obtaining experimental samples, we believe that this principle is useful. We propose that the experimental sample that reduces the information entropy the greatest between the prior and posterior distribution contains the most information value for validation. If one could optimise in the input space to locate the points of greatest information value, then these points could be selected to design experiments within the experimental cost for each point. Ideally, one would also provide the minimum experimental uncertainty necessary for this task. Using this technique, one would not only be able to determine the minimum number of experiments necessary to prove or disprove a model but also their locations and could incorporate this with experimental cost. Clearly for this type of optimisation to be available, one would need to overcome the shortcomings outlined in the previous section, particularly that all experimental samples have equal weighting.

3. CONCLUSIONS

In this paper we present the concepts and foundations for a Bayesian framework for validation, where when presented with an uncertain model and experimental data one can provide a probabilistic value for the models validity. For this we discuss the favourable characteristics of a stochastic metric for comparing model and experimental outputs, along with the metric we have selected for this work. We present the implementation of this metric in a Bayesian framework for updating a validation distribution, which has range $[0, 1]$ and whose value gives the degree to which the model and data agree. By measuring the distance between experiment and prediction at experimental points and checking if they lie within a predefined tolerance, we update a Beta distribution analytically to solve for a posterior validation distribution. We show this being used with a test function, and discuss the issues that need to be resolved for this to work in practice. Namely that each experimental point has equal weighting in the updating and there is sometimes ambiguity on whether a point/prediction combination returns a pass or a failure in the test. Upon overcoming these issues, we believe that information entropy may be used to design validation experiments. By selecting the points in the input

space which reduces the information entropy between the prior and posterior the greatest contain the most information value for the validation problem.

4. REFERENCES

ASME (2006). "Guide for verification and validation in computational solid mechanics." ASME.

Avramova, M. N. and Ivanov, K. N. (2010). "Verification, validation and uncertainty quantification in multi-physics modeling for nuclear reactor design and safety analysis." *Progress in Nuclear Energy*, 52(7), 601–614.

Dvurecenska, K., Graham, S., Patelli, E., and Patterson, E. A. (2018). "A probabilistic metric for the validation of computational models." *Royal Society Open Science*, 5(11), 180687.

Ferson, S., Kreinovich, V., Grinzburg, L., Myers, D., and Sentz, K. (2015). "Constructing probability boxes and dempster-shafer structures." *Report no.*, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

Ferson, S., Oberkampf, W. L., and Ginzburg, L. (2008). "Model validation and predictive capability for the thermal challenge problem." *Computer Methods in Applied Mechanics and Engineering*, 197(29-32), 2408–2430.

Fréchet, M. M. (1906). "Sur quelques points du calcul fonctionnel." *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 22(1), 1–72.

Hu, K. T. and Orient, G. E. (2016). "The 2014 sandia verification and validation challenge: Problem statement." *Journal of Verification, Validation and Uncertainty Quantification*, 1(1), 011001.

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

Klir, G. J. (1991). "Generalized information theory." *Fuzzy sets and systems*, 40(1), 127–142.

Klir, G. J. (2004). "Generalized information theory: aims, results, and open problems." *Reliability Engineering & System Safety*, 85(1-3), 21–38.

Klir, G. J. (2005). *Uncertainty and information: foundations of generalized information theory*. John Wiley & Sons.

Oberkampf, W. L., Trucano, T. G., and Hirsch, C. (2004). "Verification, validation, and predictive capability in computational engineering and physics." *Applied Mechanics Reviews*, 57(5), 345–384.

Patelli, E. (2016). "Cossan: a multidisciplinary software suite for uncertainty quantification and risk management." *Handbook of uncertainty quantification*, 1–69.

Patelli, E., Broggi, M., Tolo, S., and Sadeghi, J. (2017a). "Cossan software: A multidisciplinary and collaborative software for uncertainty quantification." *Proceedings of the 2nd ECCOMAS thematic conference on uncertainty quantification in computational sciences and engineering, UNCECOMP*.

Patelli, E., Govers, Y., Broggi, M., Gomes, H. M., Link, M., and Mottershead, J. E. (2017b). "Sensitivity or bayesian model updating: a comparison of techniques using the dlr airmod test data." *Archive of Applied Mechanics*, 87(5), 905–925.

Rocchetta, R., Broggi, M., Huchet, Q., and Patelli, E. (2018). "On-line bayesian model updating for structural health monitoring." *Mechanical Systems and Signal Processing*, 103, 174–195.