SCIENTIFIC TEST & ANALYSIS TECHNIQUES
CENTER OF EXCELLENCE

Uncertainty Quantification: An Overview

December 2022

Joseph Lazarus, Ctr
Corinne Weeks, Ctr
Nicholas Jones, Ctr
Kyle Provost, Ctr
Gina Sigler, Ctr

To develop and apply independent, tailored Scientific Test & Analysis Techniques solutions to Test and Evaluation that deliver insight to inform better decisions.

About this Publication:
This work was conducted by the Scientific Test & Analysis Techniques Center of Excellence under contract FA8075-18-D-0002, Task FA8075-21-F-0074.

For more information:
Visit, www.AFIT.edu/STAT
Email, AFIT.ENS.STATCOE@us.af.mil
Call, 937-255-3636 x4736

Technical Reviewers:
Corinne Weeks, Ctr
Gina Sigler, Ctr

Scientific Test & Analysis Techniques Center of Excellence
2950 Hobson Way
Wright-Patterson Air Force Base, Ohio

Version: 1, FY23

Table of Contents

Introduction

Uncertainty Quantification (UQ) is the science of the characterization and reduction of uncertainties (Saouma & Hariri-Ardebili, 2021). UQ is not a standalone field of study, but it is incorporated within related fields such as, but not limited to, mathematics, statistics, and computer science and engineering. Thus, definitions and terms used in UQ tend to vary depending on the discipline employing its use. The Scientific Test and Analysis Techniques Center of Excellence (STAT COE) is uniquely positioned to offer UQ as a cross-disciplinary approach to help test teams understand different principles of uncertainty to better inform decisions. As a result, this paper presents the STAT COE's stance on UQ in relation to two pertinent technical disciplines: 1. Modeling & Simulation (M&S) and 2. Machine Learning (ML). Two techniques are also offered to assist in the implementation of UQ: 1. Design of Experiments (DOE) and 2. Monte Carlo (MC) methods. This paper concludes by summarizing concepts and limitations of UQ.

Background

UQ in the context of predictive science involves the quantification of uncertainty and errors in models, simulations, and experiments. UQ is a crucial field that helps to identify and address the sources of uncertainty that affect predictions and improve their accuracy (Smith, 2014). The sources of uncertainty within UQ are categorized into two types: aleatoric and epistemic. Understanding the differences between these two types is critical to improving the quality of UQ.

Aleatoric uncertainty is uncertainty arising from inherently random effects, which is prevalent in modeling a stochastic process in M&S. In experiments or physical data, aleatoric uncertainty can arise from the process or sensor used to capture the data (Amini et al., 2019). It is inherent random noise associated with that process or sensor that cannot be reduced. However, different processes may have different aleatoric uncertainty. For example, data collected by humans in the cockpit using stopwatches and pens may have greater noise compared to using onboard computer sensors, which would eliminate the potential for human error and, thus, the aleatoric uncertainty.

Epistemic uncertainty, on the other hand, arises from a lack of knowledge and can be reduced by obtaining additional information. This type of uncertainty can be prevalent in experiments or physical data where there is a lack of knowledge of the system or uncertainty in a measurement, such as a limited number of reportable significant figures. In M&S and ML models, epistemic uncertainty can be due to a lack of data supporting the model, numerical approximations, and uncertainty in the model itself. All models introduce uncertainty since they approximate real-world properties or behaviors, which can lead to high-epistemic uncertainty in ML models, indicating that the model is not confident in its prediction.

See a comparison between aleatoric and epistemic uncertainty in Table 1.

Table 1

Summary of Aleatoric and Epistemic Uncertainty for Different Fields of Study

Field	Aleatoric Uncertainty	Epistemic Uncertainty
Overall	Irreducible uncertainty due to inherently random effects	Reducible uncertainty due to lack of knowledge
Physical Data	Random noise associated with a process	Unknown behavior or measurement error
M&S	Inherent variation in a quantity, characterized by a probability density function	Unknown input values, numerical approximation, and model form
ML	Noise in the input or training data	Lack of training data; confidence in model prediction

In M&S, UQ frameworks commonly quantify input uncertainty and propagate that uncertainty through the model to allow output uncertainty to be quantified and understood. UQ supports model validation, which establishes trust in the model to represent the real world. In ML, Bayesian UQ methods are used to quantify uncertainty and confidence in model predictions. Many tools support UQ and vary depending on the discipline to which they are applied. The following sections will cover UQ in M&S, ML, and two statistical techniques for calculating UQ.

UQ in Modeling & Simulation

Uncertainty in M&S is typically classified as aleatoric, epistemic, a mixture of both aleatoric and epistemic, and can also be labeled by the source of the uncertainty. The source of the uncertainty can be in model inputs, numerical approximations, or the model form. Model input uncertainties can be aleatoric, epistemic, or both, and commonly appear in system conditions, environmental conditions, model parameters, or other inputs. Numerical approximation uncertainty is introduced when discretization or iterative convergence is required to obtain a solution. Finally, a model will always introduce uncertainty since a model assumes a form which does not perfectly replicate reality—also known as epistemic uncertainty. This epistemic uncertainty can be reduced through a more representative model form; however, bias can only be quantified by comparing to an authoritative referent in validation. Uncertainty should be quantified at the source and propagated through the model (Figure 1) to allow uncertainty in the output to be quantified (Roy & Oberkampf, 2011).

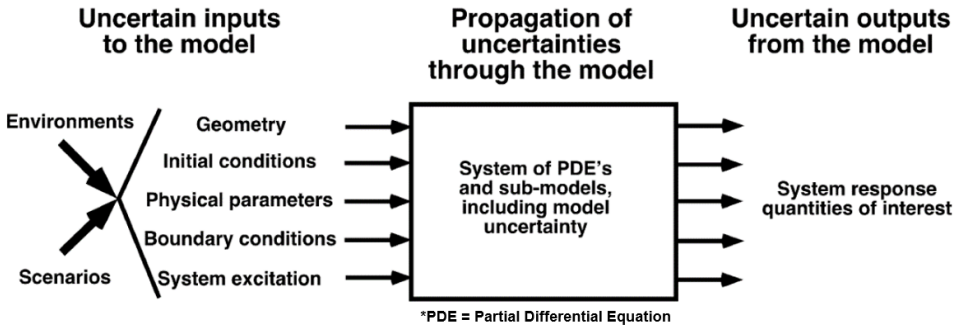

Figure 1

Propagation of Uncertainties Through a Model (Roy & Oberkampf, 2011)

UQ is considered a key part of the model Verification and Validation (V&V) process. Model uncertainties should be quantified as discussed in sections above. Similarly, uncertainties in referent or physical data should be quantified and classified as being aleatoric, epistemic, or as a mixture of both. The model and referent are independently assessed through UQ, and then followed by the process of validation which assesses the degree to which M&S is an accurate representation of the real world from the perspective of the intended use (DODI 5000.61). In other words, validation assesses bias, as opposed to uncertainty in the model outputs or data itself. Additionally, rigorous V&V should consider all uncertainties that can contribute to the risk in using model results (e.g., uncertainty that referent data is representative of the operational environment); however, this exceeds what is traditionally within the scope of UQ.

The STAT COE is currently developing Model Validation Levels (MVLs) which yield an objective metric quantifying how much trust can be placed in the results of a model to represent the real world. MVLs incorporate UQ as part of the input into the validation process by using measures of aleatoric and epistemic uncertainties for both the model and the referent. Model-referent fidelity is graded on the closeness of model and referent responses relative to the amount of uncertainty present. Fidelity is additionally graded by how well the model uncertainty represents the uncertainty present in the referent. MVLs consider fidelity in addition to other factors which affect trust in the results of a model (Provost et al., 2022).

UQ in Machine Learning

UQ in ML pertains to identifying when the predictions of a ML model can be trusted. One of the reasons for UQ in ML is the gap between how models are trained and their operational environment. Laboratory conditions tend to differ from the real-world environment which can propagate biases from training. It is difficult to present the model with every condition it will encounter. The model might see too many edge cases in the real world. It is a near impossible task to mimic the operational space in such a way to eliminate all potential biases.

Models are also susceptible to failure when presented with out-of-distribution data. For instance, a classifier trained to identify bacteria based on genomic sequences tends to fail on real-world data. This is because bacteria are ever evolving. Thus, when a classifier encounters genomes from unseen classes, it will inevitably fail (Ren & Lakshminarayanan, 2019).

Critical to understanding UQ in ML is that obtained probabilities should never be mistaken as confidence. Neural Networks (NNs) use activation functions to transfer the weights and biases of networks (Rumelhart et al., 1986). In the final layer of an image classification, NN activation functions convert the outputs of the NN to probability scores used for prediction. To get a probabilty score a special type of activation function is used and places constraints on the output. Each output must be greater than zero and the sum of the probability must sum to 1. To demonstrate these concepts consider Google's™ Bird or Bicycle image classification challenge (Brown & Olsson, 2018).

The core of this challenge is to train a model that correctly labels images of either a bird or bicycle. Figure 2 is a representation NN classifier that might be used in Google's™ Bird or Bicycle challenge. This figure depicts a picture which needs to be classified as either a bird or a bicycle. The output is a probability that belongs to each category and those outputs, probability bird and probability bicycle, must sum to 1.

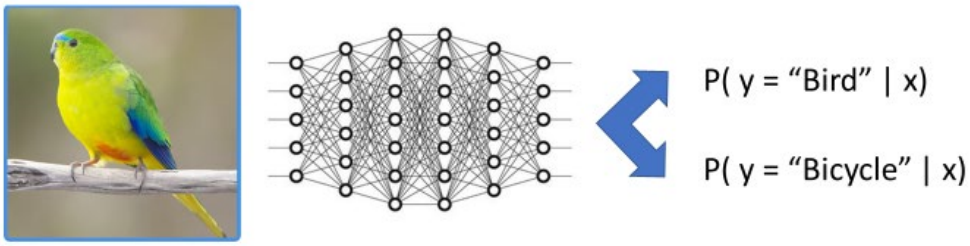

Figure 2
Image Classification Using a Neural Network (Brown & Olsson, 2018)

Training the NN in this manner captures probabilties. However, probabilities are not the same as confidence. Even if a model recieves an input unlike anything it has seen before, say an airplane, the model must give an output. Infact the model must output two things: the probability that this image is a bird and the probability that this image is a bicycle. Since these probabilities must sum to 1, the output likelihood will be unreliable. When the input is unlike anything seen during training, it is called being out of distribution.

Training the model to give an indication of how certain it is about a prediction is the goal of UQ in ML. When a model recieves a valid input it must output a prediction, even if the input is out-of-distribtuion. The key to UQ is to accompany an out-of-distribution prediction with lower confidence or uncertainty.

Bayesian neural networks (BNNs) and evidential deep learning are two UQ techniques that attempt to capture the uncertainty in ML model predictions. ML frameworks without UQ priors are placed over the data and assume that the underlying properties of the distribution that the model predicts from can be learned from the data. Bayesian techniques place priors over the weights in the neural network. In this approach, data is not assumed to come from a single distribution. Evidential neural networks place priors over the likelihood function and attempt to learn the underlying properties of the distribution.

Bayesian Neural Networks

One approach to estimating epistemic uncertainty is using BNNs. In a typical neural network, the weights associated with the nodes are a fixed number. In this standard approach the neural network is deterministic. Given the same input passed into the model several times will yield the same result. This deterministic approach does not allow for a chance to understand uncertainty. Instead of a deterministic neural network with fixed weights, BNNs instantiates a model where the weights are represented as different probability distributions. Instead of modeling a single number for every weight, BNNs attempt to capture a full distribution over every weight and use this to measure the epistemic uncertainty in the model (Amini, 2019).

Evidential Deep Learning

Instead of sampling to determine the variance and mean of the epistemic uncertainty, evidential learning tries to learn the underlying parameters of the higher-order evidential distribution. Evidential deep learning techniques seek to probabilistically estimate those parameters (Figure 3) (Amini, 2019).

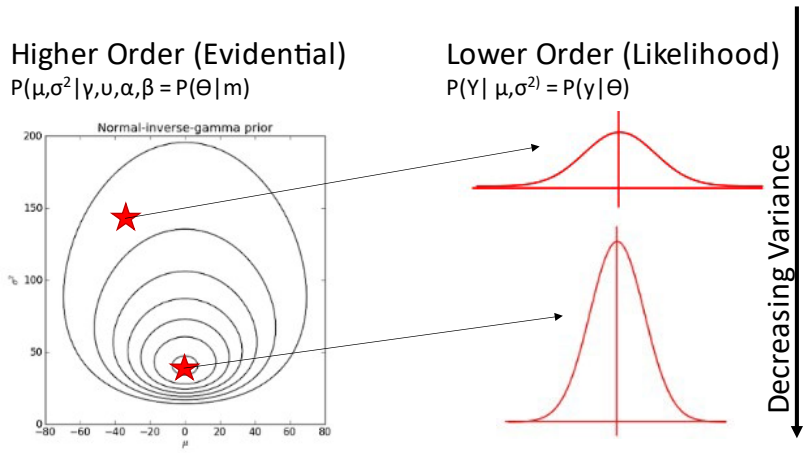

Figure 3
Inverse-Gamma Distribution

Sampling from any point in the higher-order evidential distribution (the normal-inverse-gamma distribution) corresponds to its own Gaussian distribution that is defined by their own mu (μ) and sigma (σ). The network is going to try and predict what this distribution is for any given input. These are called evidential distributions since they have greater density in areas where there are more evidence in support of a given likelihood. The distributions can change, and categorical likelihood functions will change as well. Thus, sampling from a point in the normal-inverse-gamma distribution with greater density corresponds to sampling from a Gaussian distribution that has less uncertainty (Amini, 2019).

UQ Implementation Techniques

Overall, statistical techniques play a critical role in UQ by providing a foundation for making quantitative predictions and characterizing uncertainty in complex systems. These techniques, along with computational simulations, experiments and domain knowledge, and statistics provide an approach to understand and mitigate uncertainty in a wide range of applications. In the following sections, this paper will explore two statistical UQ implementations techniques: Design of Experiments (DOE) and Monte Carlo (MC) methods.

Design of Experiments

DOE is used to explain how to sample over an operational region. It is generally desired to achieve the most information with the lowest computational and experimental costs. Characterizing a response across an operational area increases knowledge about the correct model form and decreases epistemic uncertainty in models derived from an experiment. In instances where unexpected results are observed, sequential testing can capture the necessary data to characterize the response. The DOE methodology provides a means of conducting UQ by creating interval estimates in the operational space. According to the test objectives, DOE concepts allow you to reduce and quantify the sources of uncertainty.

Due to the deterministic nature of computer experiments, more traditional DOE methods such as factorial or fractional factorial designs are not as useful. Factorial designs may not provide sufficient resolution to allow complex relationships between inputs and outputs to be discovered. Additionally, the hidden replication of these designs is wasted on a deterministic response where there will never be variability in a response for the same input conditions. Instead, more advanced designs such as Latin Hypercubes (LHCs) are used. LHCs divide each input axis into

several sections and ensure that there is at least one design point in each section to maximize information across the space.

Additional research has been conducted that shows how DOE can be used to characterize measurement uncertainty (a form of epistemic uncertainty) by constructing a test to quantify the effect of various measurement techniques on the resulting measured value. This is typically executed using a technique called Analysis of Variance (ANOVA) to calculate measurable variability but could be extended to other analysis techniques depending on the types of designs used (Adatrao, 2022).

Monte Carlo Methods

MC methods are frequently used in UQ as tools to estimate uncertainty through the artificial generation of random data samples. A common application in UQ is to apply MC methods when the uncertainty of the response is unknown, but uncertainties around related factors are known or more easily investigated. By expressing the factor of interest as a function of dependent factors, these methods aim to propagate the uncertainties of the dependent factors forward into the independent factor. Forward propagation of uncertainties is done through random sampling from the input factors. Each input factor is given a value randomly selected according to the uncertainty associated with it. The factor of interest is then calculated using those input values and stored as one sample observation. The input values are then randomized, and the factor calculation repeated many times over, until a sufficiently large sample is generated. The distribution of this simulated data is then taken to be reflective of the true distribution of our factor of interest.

MC methods are a powerful tool, but they are only one step in larger UQ processes. To apply MC methods, all relevant sources of uncertainty need to first be identified as input conditions and then quantified. Furthermore, MC methods need follow-on verification to ensure that all sources of uncertainty are accounted for and to ensure that the methods used to propagate uncertainties forward correctly captured the relationship between the inputs and the factor of interest.

Conclusion

UQ characterizes and quantifies the uncertainties present in a model. It is used to separate the sources of uncertainty and, where possible, reduce uncertainty. In instances where uncertainty cannot be reduced, understanding the source uncertainty provides is valuable. Once the uncertainty in a model is characterized, deciding what to do with that information requires a wide range of supporting topics. These topics include aspects of statistics, analysis, and human-machine interfacing. As the reliance on models continues to grow, the desire for UQ will grow with it. Effective implementation of UQ will continue to require cross-discipline teams to deliver the full benefit of UQ.

It is important to note that UQ does not tell if a model is “right” or “wrong.” It simply quantifies the variability in a model due to different sources. The “correctness” of a model can only be assessed relative to some source of truth: an activity known as validation. Furthermore, any referent used for validation will itself have some uncertainty due to its inherent variability and the ability to measure it. UQ provides the means to quantify the uncertainty present in a model or a referent, but it doesn’t directly compare them, or decide if they are true or valid. UQ supports model validation by providing an understanding of the reasonable level of disagreement that a model and referent might have, based on their uncertainties, while still supporting a conclusion that the model is valid or “correct.”

References

Alexander Amini, Wilko Schwarting, Ava P. Soleimany, & Daniela Rus. (2019). Deep Evidential Regression. *Neural Information Processing Systems*, *33*, 14927–14937.

Adatrao, S., van der Velden, S., van der Meulen, M-J., Cruellas Bordes, M., & Sciacchitano, A. (2022). Design of experiments: A statistical tool for PIV uncertainty quantification. *Measurement Science and Technology*, *34*(1), [015201]. https://doi.org/10.1088/1361-6501/ac9541

Brown, T. D., & Olsson, C. (2018, September 13). *Introducing the unrestricted adversarial examples challenge*. Google AI Blog. Retrieved February 10, 2023, from https://ai.googleblog.com/2018/09/introducing-unrestricted-adversarial.html.

Provost, K., Weeks, C., Jones, N., Sieck, V. (2022, August). Elements of a Mathematical Framework for Model Validation Levels. http://Afit.edu/STAT

Roy, C. J., & Oberkampf, W. L. (2011). A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering*, *200*(25-28), 2131–2144. https://doi.org/10.1016/j.cma.2011.03.016

Smith, R. C. (2014). *Uncertainty quantification: Theory, implementation, and applications*. Society for Industrial and Applied Mathematics.

U.S. Department of Defense. (2018). Department of Defense Instruction 5000.61

Saouma, V. E., & Hariri-Ardebili, M. A. (2021). *Aging, shaking, and cracking of infrastructures: From mechanics to concrete dams and nuclear structures*. Springer.

Ren, J., & Lakshminarayanan, B. (2019, December 17). [web log]. Retrieved February 10, 2023, from https://ai.googleblog.com/2019/12/improving-out-of-distribution-detection.html

Rumelhart, D., Hinton, G., & Williams R. (1986). Learning Representations by Back Propagation Errors. Nature, 323, 533-536.