Singapore Nuclear Research and Safety Initiative

NATIONAL RESEARCH FOUNDATION

Approximate Bayesian Computation

An Introduction to ABC towards Stochastic Model Updating

By: Adolphus Lye

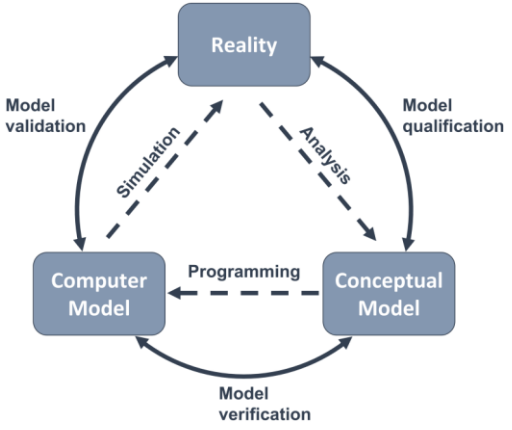
Date: 5-April-2024

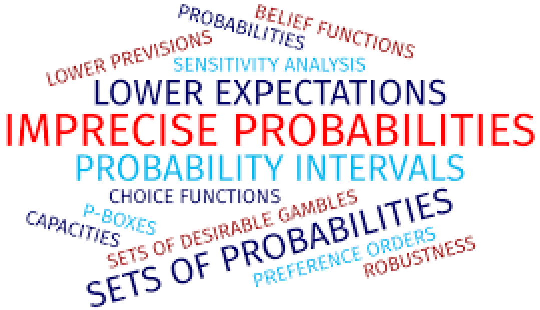
Risk, Reliability and Uncertainty Quantification

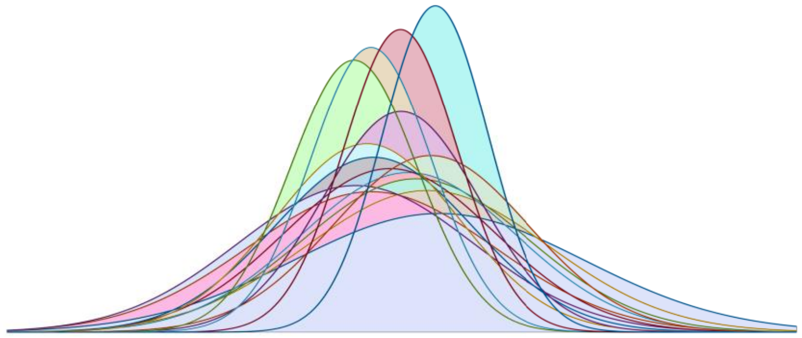
R2UQ

Overview

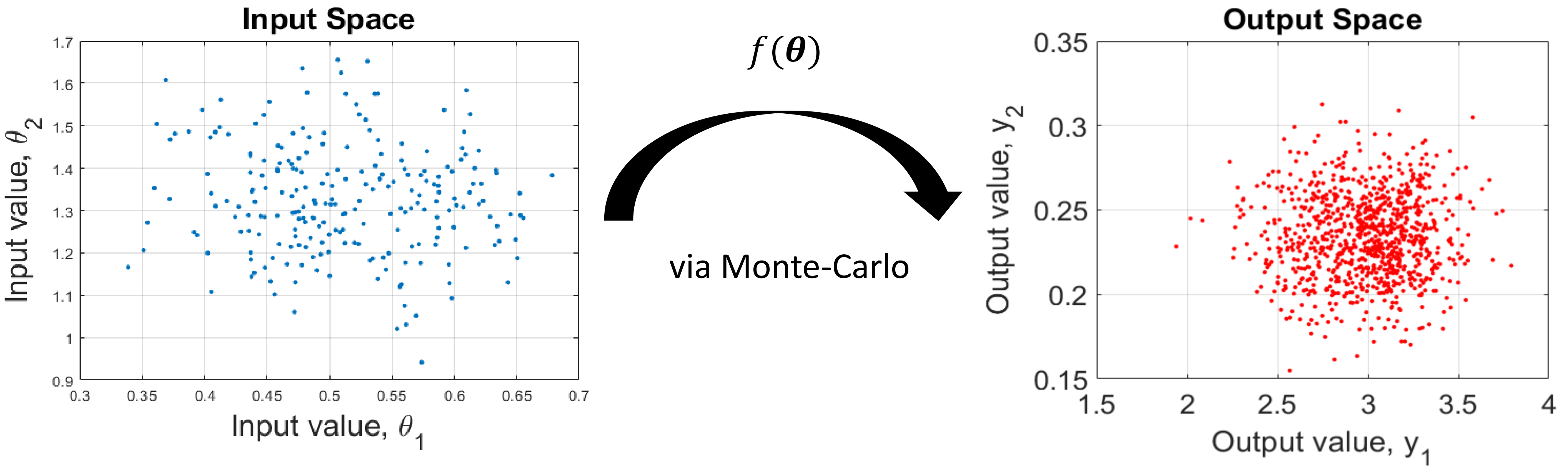
- Introduction to Model Updating
- Motivation behind Model Updating
- Introduction to Bayesian Model Updating
- Introduction to Approximate Bayesian Computation
- Application: SANDIA Thermal Challenge Problem
- Concluding Remarks







The Forward Problem:

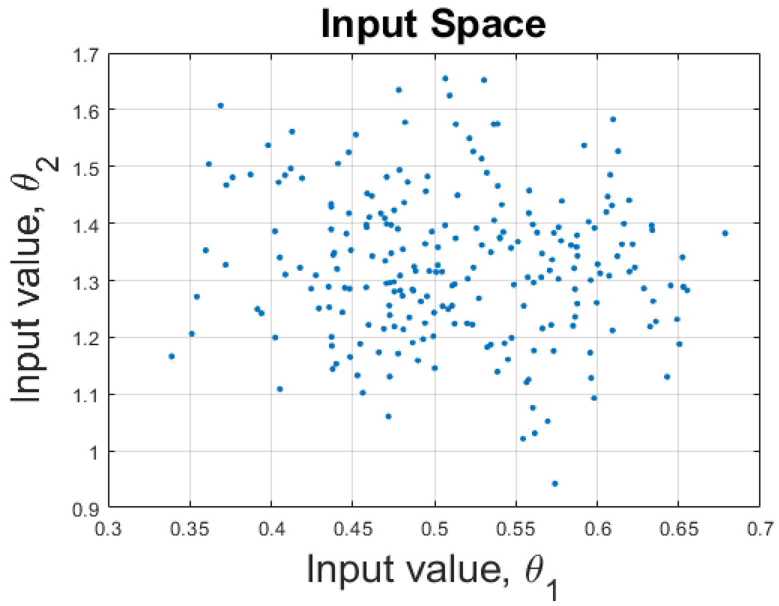


Introduction to Model Updating

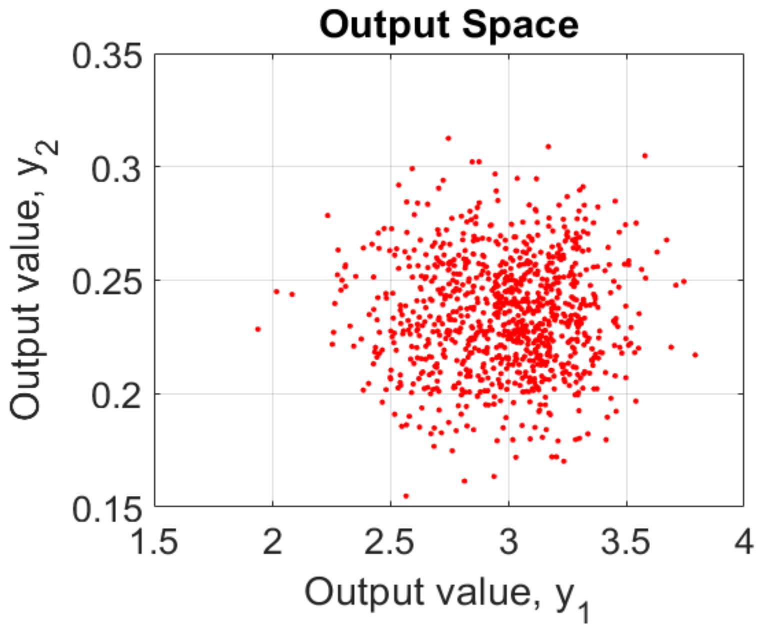
The Inverse Problem:

Not necessarily available

$$\boxed{f^{-1}(\boldsymbol{y})}$$



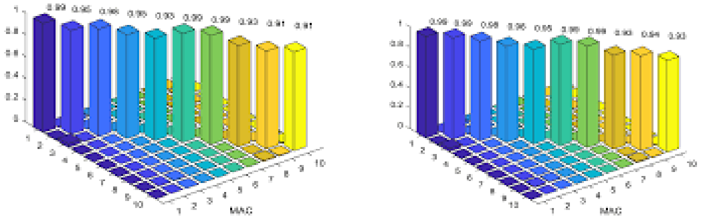
via ???

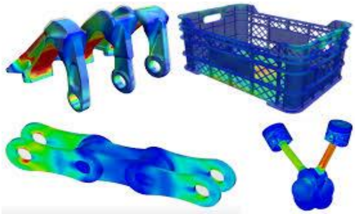


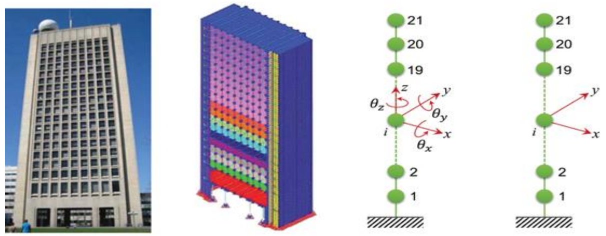
The Answer: Model Updating

Motivation behind Model Updating

- Mathematical models are used to describe virtual behaviours of engineering structures under operational and extreme conditions;
- For the model to produce output representative of the structure’s response, there is a need to update the model’s input parameter(s);
- This seeks to minimise the difference between the model output and the measured response of the system;
- This, however, is often done under uncertainty due to limited training data which brings the need to characterise such uncertainty;
- At times, the model itself can be stochastic due to input parameters being random variables;
- This brings forth the need for Stochastic Model Updating.


(a) (b)





(a) Green Building (b) Full scale model (c) 126 DOFs (d) 42 DOFs

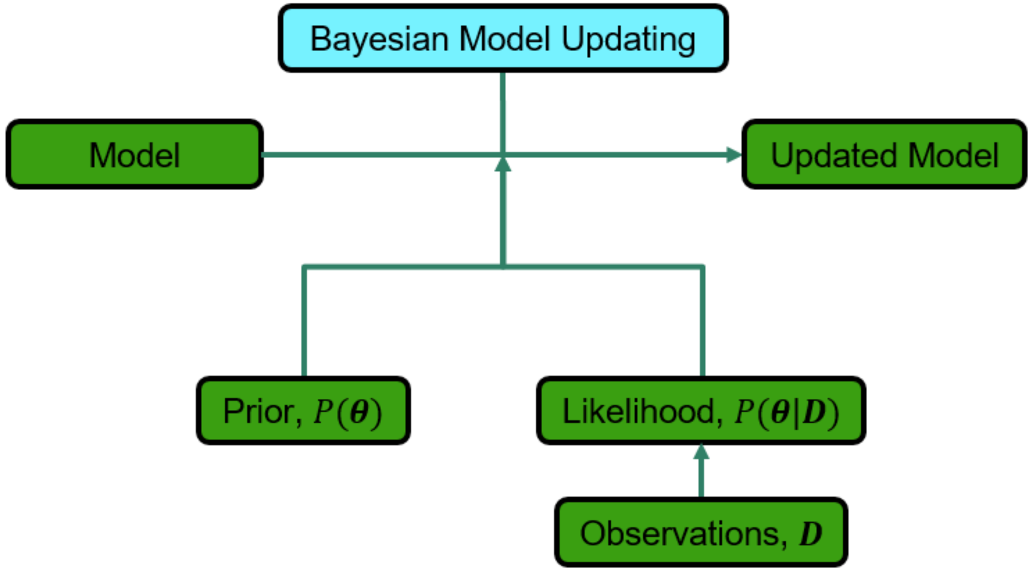
Introduction to Bayesian Model Updating

Risk, Reliability and Uncertainty Quantification
R2UQ

- Stochastic Model Updating technique based on Bayes' Inference:

$$P(\boldsymbol{\theta}|\boldsymbol{D}, M) = \frac{P(\boldsymbol{\theta}|M) \cdot P(\boldsymbol{D}|\boldsymbol{\theta}, M)}{P(\boldsymbol{D}|M)}$$

Symbol	Description
$P(\boldsymbol{\theta}\|M)$	Prior
$P(\boldsymbol{D}\|\boldsymbol{\theta}, M)$	Likelihood
$P(\boldsymbol{\theta}\|\boldsymbol{D}, M)$	Posterior
$P(\boldsymbol{D}\|M)$	Evidence
$\boldsymbol{\theta}$	Epistemic parameters
$\boldsymbol{D}$	Observations
M	Model



- The Evidence term is a normalisation constant and is a numerical constant. Hence:

$$P(\boldsymbol{\theta}|\boldsymbol{D}, M) \propto P(\boldsymbol{\theta}|M) \cdot P(\boldsymbol{D}|\boldsymbol{\theta}, M)$$

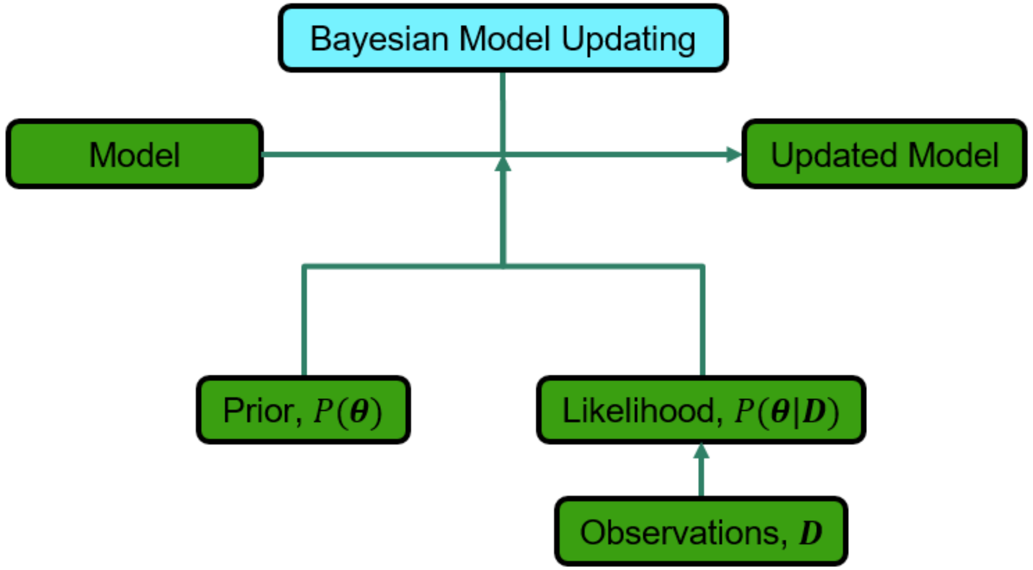
Introduction to Bayesian Model Updating

- An important aspect in Bayesian model updating is the definition of the Likelihood function $P(\boldsymbol{D}|\boldsymbol{\theta}, M)$ [1];

$$P(\boldsymbol{\theta}|\boldsymbol{D}, M) \propto P(\boldsymbol{\theta}|M) \cdot P(\boldsymbol{D}|\boldsymbol{\theta}, M)$$

- Often computed while assuming independence between the individual data points:

$$P(\boldsymbol{D}|\boldsymbol{\theta}, M) = \prod_{k=1}^{N_{\text{obs}}} P(\boldsymbol{D}_k|\boldsymbol{\theta}, M) = \prod_{k=1}^{N_{\text{obs}}} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\boldsymbol{D}_k - M(\boldsymbol{\theta}))^2}{2\,\sigma^2}\right)$$



- BUT doing so would lead to three key problems:
  1) Loss of information over the dependencies between the data points and the features within the data;
  2) The computation assumes an analytical form of likelihood function is available. Even if it is not the case, a Kernel density estimate is used which requires there to be a lot of data and the density estimate adds to the computational cost;
  3) At times, the full likelihood function itself can become computationally expensive if the model M is itself computationally expensive.

Introduction to Bayesian Model Updating

Risk, Reliability and Uncertainty Quantification
R2UQ

- The problem becomes more pronounced when dealing with stochastic models – models that contain random variables as inputs;

- In such case, we now deal with models whose output follows a certain distribution when implemented repeatedly;

- The problem involves identifying what distribution model those random variable(s) input follow and inferring the shape parameter of such distribution model based on the (often limited) data provided;

- Both Aleatory and Epistemic uncertainties present – polymorphic/hybrid uncertainty;

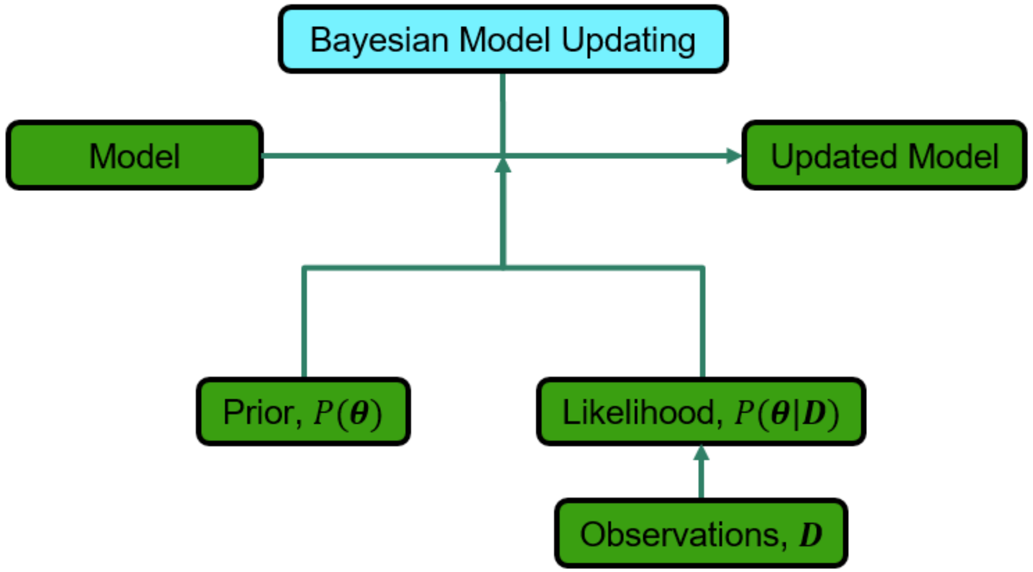
- This begets two questions:
  1) How do we account for such stochasticity in the model output in the Likelihood function?
  2) What metric should we incorporate within the Likelihood function to optimise the model input parameter(s) of interest?

$$P(\boldsymbol{\theta}|\boldsymbol{D}, M) \propto P(\boldsymbol{\theta}|M) \cdot P(\boldsymbol{D}|\boldsymbol{\theta}, M)$$

Introduction to Approximate Bayesian Computation

- To address these problems, one could use an approximate Gaussian likelihood function and define a statistical distance function to quantify the statistical difference between the data $\boldsymbol{D}$ and the model output $\boldsymbol{D}_{\text{sim}} = M(\boldsymbol{\theta})$ [1];

$$P(\boldsymbol{\theta}|\boldsymbol{D}, M) \propto P(\boldsymbol{\theta}|M) \cdot P(\boldsymbol{D}|\boldsymbol{\theta}, M)$$



$$P(\boldsymbol{D}|\boldsymbol{\theta}, M) = \exp\left(-\frac{d(\boldsymbol{D}, \boldsymbol{D}_{\text{sim}})^2}{\varepsilon^2}\right)$$

where ε is a width factor determined by the analyst;

- A general distance function fulfils the following properties:
  1) Non-negativity: $d(x_1, x_2) \geq 0$
  2) Symmetry: $d(x_1, x_2) = d(x_2, x_1)$
  3) Triangle Inequality: $d(x_1, x_2) \leq d(x_1, y) + d(y, x_2)$
  4) Identity of Indiscernible: $d(x_1, x_2) = 0,\ \text{if } x_1 = x_2$

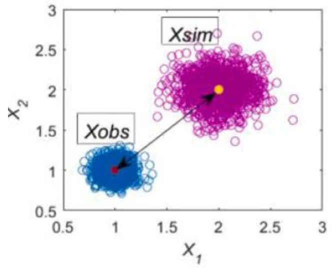
- Examples:
  1) Euclidean distance;
  2) Bhattacharyya distance – violates property #3;
  3) Bray-Curtis distance – violates property #3;
  4) Jenson-Shannon divergence;
  5) 1-Wasserstein distance;

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

Introduction to Approximate Bayesian Computation

Euclidean distance:

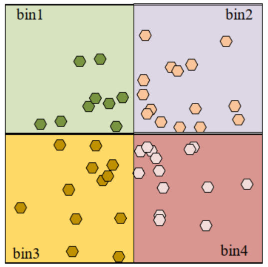
$$d_E = \sqrt{(\overline{\boldsymbol{D}} - \overline{\boldsymbol{D}}_{\text{sim}}) \cdot (\overline{\boldsymbol{D}} - \overline{\boldsymbol{D}}_{\text{sim}})^{\mathrm{T}}}$$



Bray-Curtis distance:

$$d_{\text{BC}} = \frac{\sum_{i_m=1}^{n_{\text{bin}}} \dots \sum_{i_1=1}^{n_{\text{bin}}} \left| p_D(b_{i_1,\dots,i_m}) - p_{\text{sim}}(b_{i_1,\dots,i_m}) \right|}{\sum_{i_m=1}^{n_{\text{bin}}} \dots \sum_{i_1=1}^{n_{\text{bin}}} p_D(b_{i_1,\dots,i_m}) + p_{\text{sim}}(b_{i_1,\dots,i_m})}$$

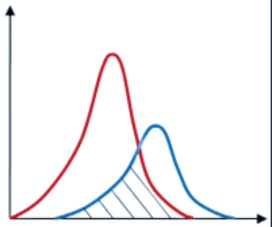
where p_D and p_{sim} are the PMFs of $\boldsymbol{D}$ and $\boldsymbol{D}_{\text{sim}}$ respectively.



Bhattacharyya distance:

$$d_B = -\log\left\{\sum_{i_m=1}^{n_{\text{bin}}} \dots \sum_{i_1=1}^{n_{\text{bin}}} p_D\left(b_{i_1,\dots,i_m}\right) \cdot p_{\text{sim}}\left(b_{i_1,\dots,i_m}\right)\right\}$$

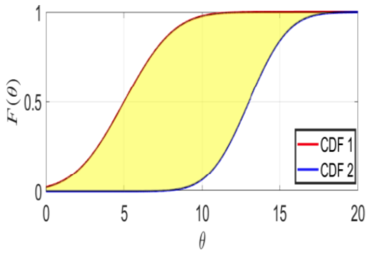
where p_D and p_{sim} are the PMFs of $\boldsymbol{D}$ and $\boldsymbol{D}_{\text{sim}}$ respectively.



1-Wasserstein distance:

$$d_{1W} = \int_{\infty}^{\infty} \left| F_D(\boldsymbol{x}) - F_{\text{sim}}(\boldsymbol{x}) \right| \cdot d\boldsymbol{x}$$

where F_D and F_{sim} are the CDFs of $\boldsymbol{D}$ and $\boldsymbol{D}_{\text{sim}}$ respectively.



Jenson-Shannon divergence:

$$d_{\text{JS}} = \frac{1}{2} \sum_{i_m=1}^{n_{\text{bin}}} \dots \sum_{i_1=1}^{n_{\text{bin}}} \left(p_D\left(b_{i_1,\dots,i_m}\right) \cdot \log\left(\frac{p_D\left(b_{i_1,\dots,i_m}\right)}{T\left(b_{i_1,\dots,i_m}\right)}\right) + p_{\text{sim}}\left(b_{i_1,\dots,i_m}\right) \cdot \log\left(\frac{p_{\text{sim}}\left(b_{i_1,\dots,i_m}\right)}{T\left(b_{i_1,\dots,i_m}\right)}\right)\right)$$

where $T\left(b_{i_1,\dots,i_m}\right) = \frac{1}{2}\left(p_D\left(b_{i_1,\dots,i_m}\right) + p_{\text{sim}}\left(b_{i_1,\dots,i_m}\right)\right)$

Note: This is just a brief mathematical introduction to the different distance functions implemented for ABC. For details, feel free to refer to:

A. Lye, S. Ferson, and S. Xiao (2024). Comparison between distance functions for Approximate Bayesian Computation towards Stochastic model updating and Model validation under limited data. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems Part A: Civil Engineering*, ***10***.

SANDIA Thermal Challenge Problem

SANDIA Thermal Challenge Problem

Risk, Reliability and Uncertainty Quantification
R2UQ

- Involves a slab material that could be used in the construction of the nuclear reactor vessel;

- Its dimensions are illustrated in the figure on the right and whose physics-based temperature model is described as:

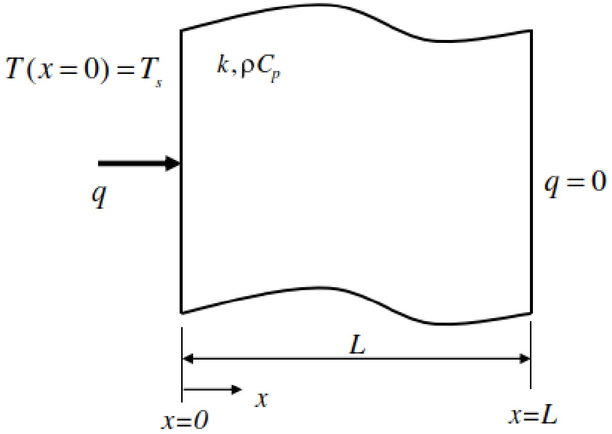
$$M_T(x,t) = \begin{cases} T_i & \text{, for } t = 0s \\ T_i + \frac{qL}{k}\left[\frac{(k/\rho C_p)t}{L^2} + \frac{1}{3} - \frac{x}{L} + \frac{1}{2}\left(\frac{x}{L}\right)^2 - \frac{2}{\pi^2}\sum_{n=1}^{6}\frac{1}{n^2}\exp\left(-n^2\pi^2\frac{(k/\rho C_p)t}{L^2}\right)\cos\left(n\pi\frac{x}{L}\right)\right] & \text{, for } t > 0s \end{cases}$$

- Research objectives:

  a) To characterise the random variable model inputs of the slab material's temperature model;

  b) To validate the calibrated temperature model against a set of experimental data;

- Full details to the challenge are found in [2].



Parameter	Description	Value
L	Full thickness of slab $[cm]$	2.54
x	Thickness coordinate of slab $[cm]$	$[0, L]$
q	Heat flux $[W/m^2]$	1000
T_i	Initial ambient temperature $[^oC]$	25.0
k	Thermal conductivity $\left[\frac{W}{m\,^oC}\right]$	Random Variable
ρC_p	Volumetric heat capacity $\left[\frac{J}{m^3\,^oC}\right]$	Random Variable

[2] K. J. Dowding, M. Pilch, and R. G. Hills (2008). Formulation of the thermal problem. *Computer Methods in Applied Mechanics and Engineering*, 197(29-32), 2385–2389.

SANDIA Thermal Challenge Problem

Risk, Reliability and Uncertainty Quantification
R2UQ

- To characterise the variability of k and ρC_p, 20 experimental data points are provided to calibrate their distributions;
- An ensemble of 4 experimental data set, each with 11 data points, are used to perform an ensemble validation of the temperature model.

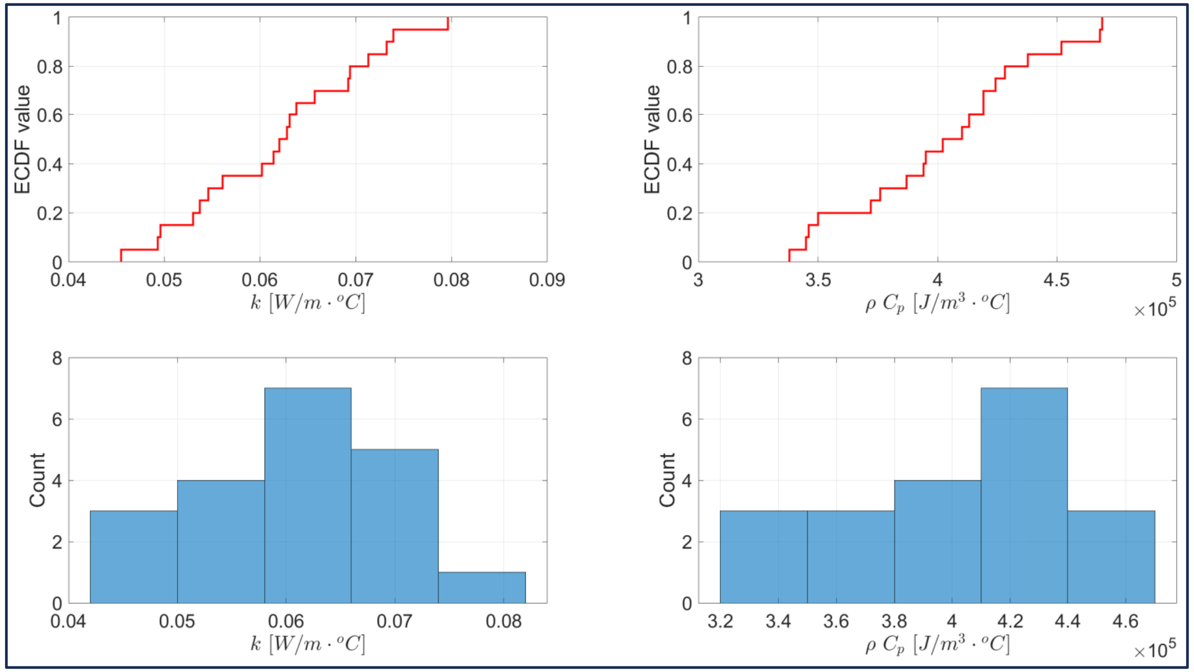


Fig. 1(a): The Empirical CDF and histograms to the 20 data points to be used to characterise the distribution over k and ρC_p. Data obtained from [2].

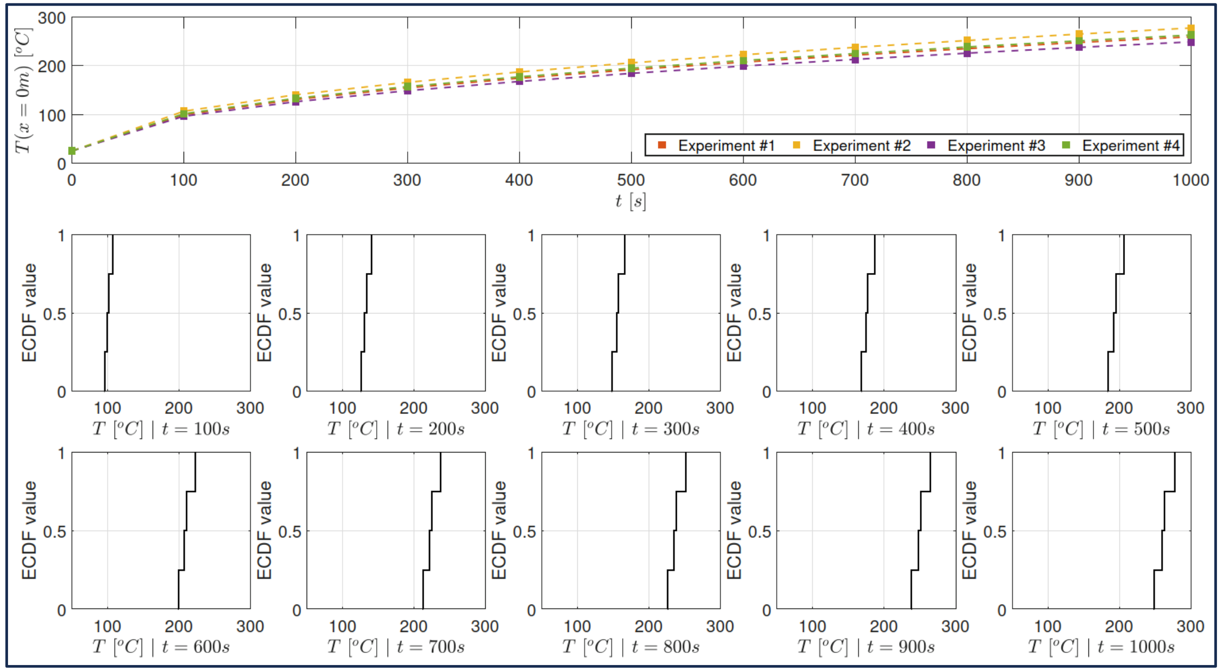


Fig. 1(b): The ensemble experimental data on the surface temperature of the slab material over time used to validate the calibrated temperature model. Data obtained from [2].

[2] K. J. Dowding, M. Pilch, and R. G. Hills (2008). Formulation of the thermal problem. *Computer Methods in Applied Mechanics and Engineering*, 197(29-32), 2385–2389.

SANDIA Thermal Challenge Problem

Risk, Reliability and Uncertainty Quantification

R2UQ

Methodology to Calibrate the distributions for k and ρC_p:

- The challenge problem makes the following three assumptions:
  1) the diagnostic variability or uncertainty is negligible compared to specimen-to-specimen variability;
  2) the estimates of thermal conductivity and that of the volumetric heat capacity are independent between one another;
  3) the estimates of thermal conductivity and that of the volumetric heat capacity are temperature independent.
- Based on the previous work in [3], a Normal distribution was fitted on both quantities based on the distribution profile of their ECDFs;
- The goal of the work is to compare the ABC-based Stochastic model updating performance using different distance functions in this case.

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

SANDIA Thermal Challenge Problem

Risk, Reliability and Uncertainty Quantification
R2UQ

Methodology to Calibrate the distributions for k and ρC_p:

- The Bayesian model updating set-up follows:

a) Uniform prior is used on the shape parameters of the Normal distributions to k and ρC_p:

Shape parameter	Description	Prior bounds	Units
μ_k	Mean of k	$[0.02, 0.08]$	$[W/m^o C]$
σ_k	Standard deviation of k	$[0.001, 0.100]$	$[W/m^o C]$
$\mu_{\rho C_p}$	Mean of ρC_p	$[2.00 \times 10^5, 8.00 \times 10^5]$	$[J/m^3\ ^o C]$
$\sigma_{\rho C_p}$	Standard deviation of ρC_p	$[1.00 \times 10^4, 8.00 \times 10^4]$	$[J/m^3\ ^o C]$

Table 1(a): The bounds to the Uniform prior distribution associated with the shape parameters of the Normal distribution to k and ρC_p. See [2].

b) The respective values to ε of the approximate Gaussian likelihood function:

Aleatory parameter	d_E	d_B	d_{BC}	d_{W_1}
k	1.0×10^{-3}	0.06	0.06	1.5×10^{-3}
ρC_p	2.5×10^3	7.0×10^{-2}	3.6×10^{-2}	4.0×10^3

Table 1(b): The values to the width parameter ε to the approximate Gaussian likelihood function for the different distance functions. See [2].

[4] A. Lye, S. Ferson, and S. Xiao (2024). Comparison between distance functions for Approximate Bayesian Computation towards Stochastic model updating and Model validation under limited data. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems Part A: Civil Engineering*.

SANDIA Thermal Challenge Problem

Risk, Reliability and Uncertainty Quantification
R2UQ

ABC-based Stochastic model updating Results:

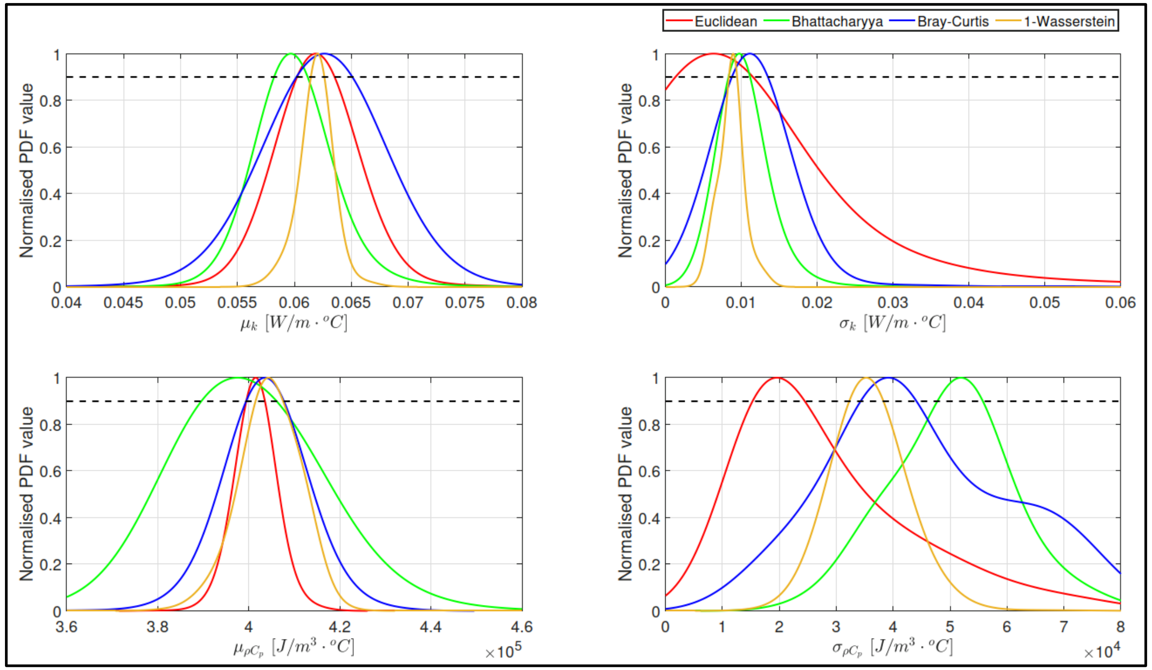


Fig. 2(a): The posterior-based fuzzy set over the estimates of the shape parameters of the Normal distribution over k and ρC_p. Results presented in [4].

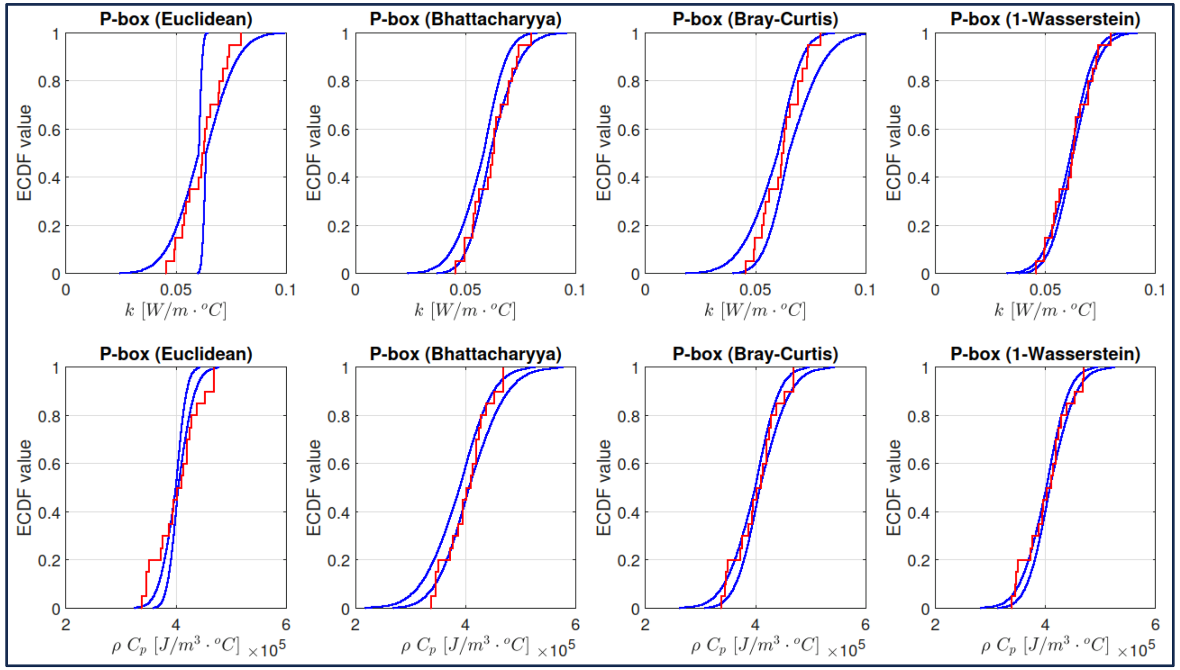


Fig. 2(b): The probability boxes to characterise the imprecise distribution over k and ρC_p given the different distance functions used in the ABC procedure. Results presented in [4].

[4] A. Lye, S. Ferson, and S. Xiao (2024). Comparison between distance functions for Approximate Bayesian Computation towards Stochastic model updating and Model validation under limited data. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems Part A: Civil Engineering*.

SANDIA Thermal Challenge Problem

Risk, Reliability and Uncertainty Quantification
R2UQ

Methodology to Validate the Temperature model given the imprecise distributions for $\boldsymbol{k}$ and $\boldsymbol{\rho C_p}$:

- An ensemble of 4 experimental data set, each with 11 data points, are used to perform an ensemble validation of the temperature model;
- A total of 500 Normal distributions on k and ρC_p is obtained from 500 samples obtained from the resulting updated epistemic interval on the respective shape parameters from the Bayesian model updating;
- For each Normal distribution realization on k and ρC_p, they serve as input to the Temperature model to yield an output Temperature distribution for a given time t;
- Repeat this procedure 500 times to obtain a collection of Temperature distribution at each given time t;
- This collection of distribution constitute a probability pox (i.e., P-box).

SANDIA Thermal Challenge Problem

Risk, Reliability and Uncertainty Quantification R2UQ

Stochastic model validation Results:

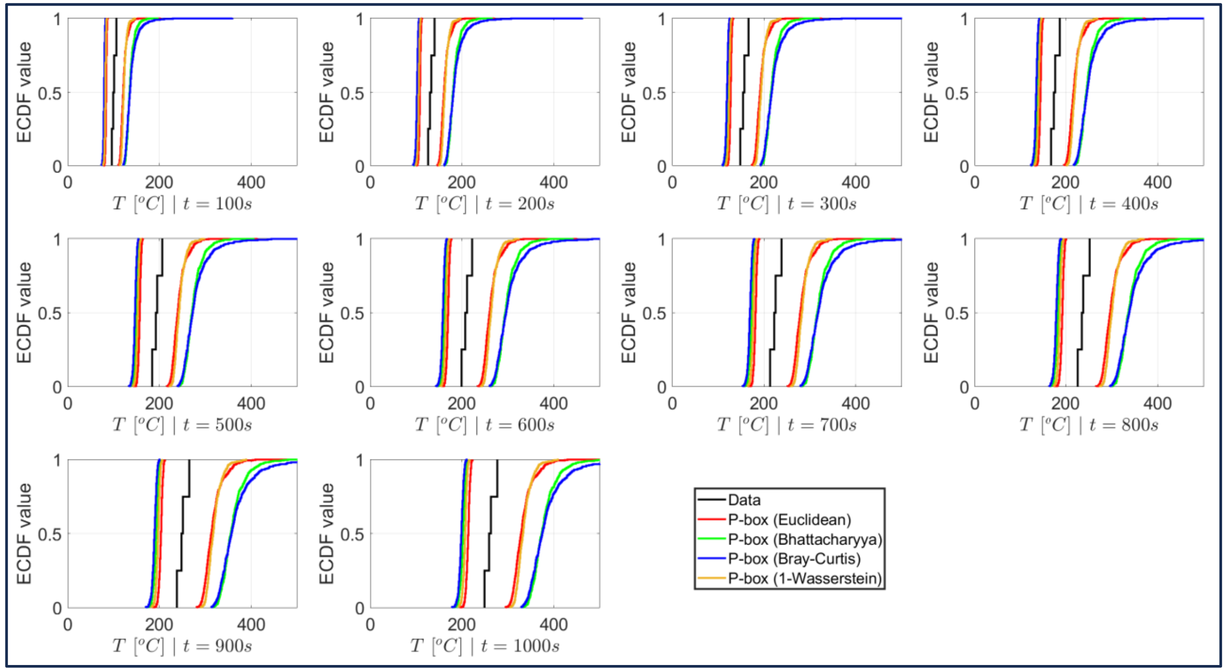


Fig. 3(a): The resulting probability boxes of the calibrated model against the ensemble experimental data given the different distance functions used in the ABC procedure. Results presented in [4].

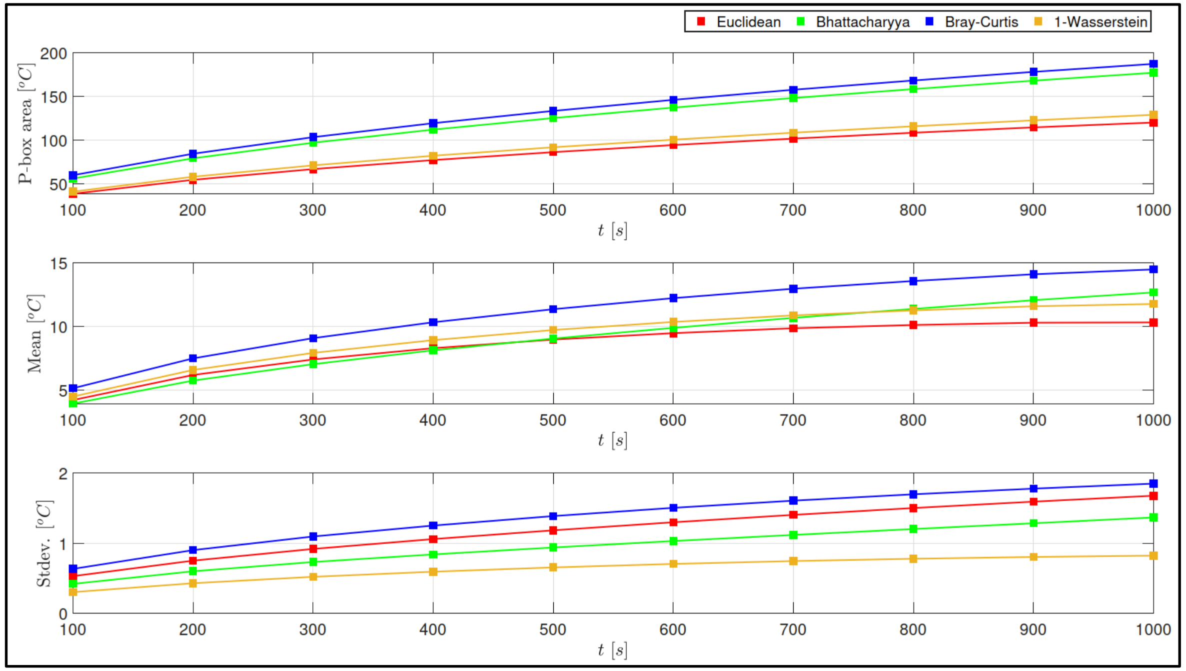


Fig. 3(b): Statistics to the validation performance of the calibrated model given the respective distance functions used for the ABC procedure. Results presented in [4].

[4] A. Lye, S. Ferson, and S. Xiao (2024). Comparison between distance functions for Approximate Bayesian Computation towards Stochastic model updating and Model validation under limited data. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems Part A: Civil Engineering*.

Concluding Remarks

Risk, Reliability and Uncertainty Quantification
R2UQ

Research Conclusions:

- Evaluated the strengths and limitations of the different distance functions used in approximate Bayesian computation for Stochastic model updating;

- It was found that the 1-Wasserstein distance function is most useful when one if performing ABC to optimise a given distribution’s shape parameters against an empirical distribution;

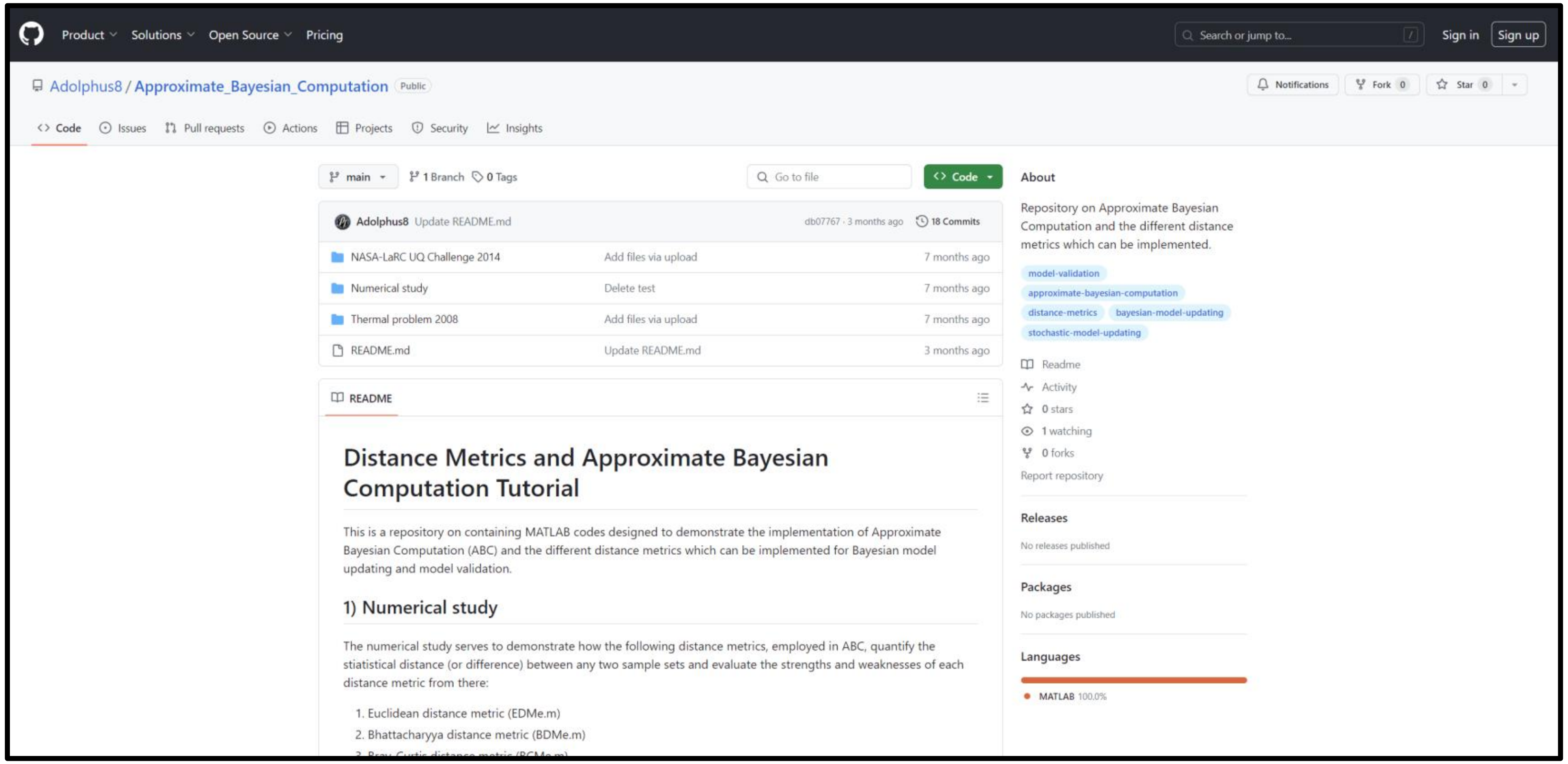
- The calibrated model via ABC using the Euclidean distance, interestingly, provides for the best model validation performance against the ensemble validation data;

- This is despite the Euclidean distance encoding the least of statistical information in that it only considers the difference in mean between the data and the model output and not the other moments such as the variance, skewness, and kurtosis.

Concluding Remarks

Risk, Reliability and Uncertainty Quantification
R2UQ

Tutorials Available:
https://github.com/Adolphus8/Approximate_Bayesian_Computation.git



State-of-the-Art Review

ASCE

Comparison between Distance Functions for Approximate Bayesian Computation to Perform Stochastic Model Updating and Model Validation under Limited Data

Adolphus Lye[1]; Scott Ferson[2]; and Sicong Xiao[3]

Abstract: The paper reviews the distance functions that have been or could be used to perform stochastic model updating via approximate Bayesian computation, namely: (1) Euclidean distance; (2) Bhattacharyya distance; (3) Bray–Curtis distance; and (4) 1-Wasserstein distance functions. For each of these distance functions, details on their mathematical formalism and implementation are provided along with an evaluation of their respective advantages and disadvantages illustrated though a numerical study. Subsequently, to provide a basis for comparison in the stochastic model updating performance by the respective distance functions, two case studies in the form of model validation problems are presented. The first problem is based on the 2014 NASA-LaRC Multidisciplinary Uncertainty Quantification Challenge. The second problem is based on the 2008 thermal problem presented by Sandia National Laboratories. Both problems provide a relatively complex and realistic setting to assess the strengths and robustness of each distance function in performing approximate Bayesian computation under polymorphic uncertainty and limited data for a one-dimensional output space. The comparison will be done based on the following: (1) the precision of the inference results on the inferred parameters; and (2) the precision of the corresponding stochastic model output against a set of validation data. **DOI: 10.1061/AJRUA6.RUENG-1223.** *© 2024 American Society of Civil Engineers.*

Introduction

Within the field of modern engineering, the need for numerical models and simulations have grown significantly in recent years in such a manner that they are seemingly irreplaceable in the study of engineered systems and components (Bi et al. 2023). This characteristic is especially so in cases whereby it becomes challenging to obtain data through experimental campaigns resulting in the limited availability of experimental data (Lye et al. 2022c, 2023b). The numerical models should ideally represent the physical system in a manner such that there is no discrepancy between the resulting simulated model output and the experimental measurement.

However, such discrepancy exists in reality because the simulated output by the numerical models do not take into account the effects of the measurement "noise" (Helton 1997; Helton et al. 2004) as well as the other sources of uncertainties such as the model form uncertainty and the model parameter uncertainty. Such uncertainties are classified into two broad categories (Oberkampf et al. 2004; Roy and Oberkampf 2011): (1) aleatory uncertainty and (2) epistemic uncertainty. Aleatory uncertainty refers to the fluctuations and the statistical uncertainty of a given variable due to its inherent randomness and variability. Such uncertainty is irreducible and is usually described with a probability distribution. Epistemic uncertainty, conversely, refers to the lack of knowledge about the parameter(s) or variable(s) of interest. Unlike in the case of aleatory uncertainty, epistemic uncertainties are reducible with increased study or observations, and such uncertainties are often conceptualized as a fixed value within a bounded set whose breadth reflects the level of knowledge on the parameter.

Model updating techniques have been developed to calibrate the model parameters, and hence the model itself, such that the discrepancy between the simulated model output and the experimental measurement is minimized (i.e., as close to zero as possible) (Mottershead and Friswell 1993; Mottershead et al. 2021). The process of model updating can also be viewed as model calibration and has been widely implemented in model verification and validation (Worden 2001; Simoen et al. 2015). Model verification refers to the process of ensuring that the numerical model is consistent and its solution agrees with the underlying physics-based or mathematical equations. Model validation refers to the process of evaluating how accurate the numerical model represents the physical system being studied based on physical experiments conducted from the perspective of the intended user(s) (Bi et al. 2023). The latter concept serves as the basis of study of the paper. Broadly speaking, model updating can be categorized into two distinct types (Lye et al. 2021): (1) deterministic model updating; and (2) stochastic model updating.

In a deterministic model updating framework, the calibration of the model is done without accounting for any form of uncertainty associated with the inferred model parameter(s). Through such an approach, the model calibration procedure is performed based on a single set of test data (Patelli et al. 2017). This yields a single set of crisp values on the inferred model parameter(s) resulting in a single model prediction with maximum fidelity given the single set of test data. Examples of such approaches toward model updating include the following: (1) the linear least-squares minimization; and (2) sensitivity-based model updating (Patelli et al. 2017). Although the deterministic model updating approach can be

[1]Research Fellow, Singapore Nuclear Research and Safety Initiative, National Univ. of Singapore, 1 CREATE Way, Singapore S138602 (corresponding author). ORCID: https://orcid.org/0000-0002-1803-8344. Email: snrltsa@nus.edu.sg

[2]Professor and Director of the Institute for Risk and Uncertainty, Institute for Risk and Uncertainty, Univ. of Liverpool, Liverpool L697ZF, UK. Email: ferson@liverpool.ac.uk

[3]Senior Research Fellow and Principal Investigator, Singapore Nuclear Research and Safety Initiative, National Univ. of Singapore, 1 CREATE Way, Singapore S138602. Email: snrxs@nus.edu.sg

Note. This manuscript was published online on March 18, 2024. Discussion period open until August 18, 2024; separate discussions must be submitted for individual papers. This paper is part of the ***ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering***, © ASCE, ISSN 2376-7642.

NUS National University of Singapore
Singapore Nuclear Research and Safety Initiative

Thank You Very Much! Connect. Contact. Collaborate.

Risk, Reliability and Uncertainty Quantification
R2UQ

Risk, Reliability and Uncertainty Quantification
R2UQ

Scan here for more information on the R2UQ research framework.

NUS National University of Singapore | Singapore Nuclear Research and Safety Initiative

ADOLPHUS LYE

Research Fellow

E: snrltsa@nus.edu.sg
T: @adolphus_lye
W: sites.google.com/view/adolphus-lye

Address: 1 CREATE Way, #04-01 CREATE Tower, National University of Singapore | Singapore S(138602)