
Article:
Gardner, P., Lord, C. orcid.org/0000-0002-2470-098X and Barthorpe, R. orcid.org/0000-0002-6645-8482 (2019) A unifying framework for probabilistic validation metrics. Journal of Verification, Validation and Uncertainty Quantification, 4 (3). 031005. ISSN 2377-2158

https://doi.org/10.1115/1.4045296

A unifying framework for probabilistic validation metrics

Paul Gardner
Research Assistant
Dynamics Research Group
Department of Mechanical Engineering
University of Sheffield
Sheffield, S1 3JD
Email: p.gardner@sheffield.ac.uk

Charles Lord
Lecturer
Dynamics Research Group
Department of Mechanical Engineering
University of Sheffield
Sheffield, S1 3JD
Email: c.lord@sheffield.ac.uk

Robert J. Barthorpe
Lecturer
Dynamics Research Group
Department of Mechanical Engineering
University of Sheffield
Sheffield, S1 3JD
Email: r.j.barthorpe.ac.uk

Probabilistic modelling methods are increasingly being employed in engineering applications. These approaches make inferences about the distribution for output quantities of interest. A challenge in applying probabilistic computer models (simulators) is validating output distributions against samples from observational data. An ideal validation metric is one that intuitively provides information on key differences between the simulator output and observational distributions, such as statistical distances/divergences. Within the literature only a small set of statistical distances/divergences have been utilised for this task; often selected based on user experience and without reference to the wider variety available. As a result, this paper offers a unifying framework of statistical distances/divergences, categorising those implemented within the literature, providing a greater understanding of their benefits, and offering new potential measures as validation metrics. In this paper two families of measures for quantifying differences between distributions, that encompass the existing statistical distances/divergences within the literature, are analysed: f-divergence and Integral Probability Metrics (IPMs). Specific measures from these families are highlighted, providing an assessment of current and new validation metrics, with a discussion of their merits in determining simulator adequacy, offering validation metrics with greater sensitivity in quantifying differences across the range of probability mass.

1 Introduction

Validation is a crucial part of any model generation, especially for complex computer models (herein defined as simulators), without which, trust in outputs for specific input domains cannot be obtained. Traditionally, validation metrics for quantifying the simulators' level of adequacy have been deterministic, as most modelling techniques produce deterministic outputs. In this setting, distance metrics are commonly used, such as mean squared errors and L_2-norms as they provide a clear and interpretable method of validating and understanding the simulators performance. However, in recent years, best practice in validation [1, 2] has seen a move towards understanding and quantifying uncertainties within the modelling procedure; providing better information to make more robust decisions from simulators. By incorporating uncertainties, simulator outputs provide more information than just a mean (or deterministic) prediction. This presents new challenges in selecting validation metrics such that both the mean predictive performance and uncertainties are appropriately assessed.

This paper focuses on the problem of quantifying differences between probabilistic simulator outputs and observational samples, specifically the distance between two distributions from these sources. As a result, the simulator output and observational variables considered in this paper are those that can be defined as random variables, typically applying to ordered magnitude variables, e.g. stress, acceleration etc., as well as ratio variables, such a temperature in Kelvin. The Area Metric and Kolmogorov distance have been extensively applied in this scenario [2–6]. This paper provides a context

for these distances by defining their relationships within a wider range of statistical distances, specifically those related to the f-divergence and Integral Probability Metric (IPM) families of distances. Considering these broader families of distances provides not only new understanding of these established distance metrics, but also reveals measures with novel potential for application as validation metrics.

The list of validation metrics within this paper is not intended to be exhaustive, but encompass those commonly implemented within the literature. For example, the reliability metric, which has been developed for similar purposes, is not categorised by these two families [7, 8]. This is because the reliability metric assesses the probability that the Mahalanobis distance between the simulators' mean and observational data, given the simulator covariance, is less than a given tolerance (meaning it only considers low order statistical moments) and is better categorised as a type of hypothesis; with the authors linking it to Bayesian hypothesis testing [7, 8]. It is noted that although the emphasis of this paper is in validation metrics that quantify differences between distributions, each of the measures presented has its own hypothesis test which could be used to make informative decisions.

The outline of the paper is as follows. Section 2 provides a criteria for an ideal validation metric, clarifying the difference between a validation metric and the mathematical definition of a metric. Subsequently, the two families of measures, f-divergences and IPMs respectively, are introduced in Sections 3 and 4; with specific measures within these families defined and reviewed. These distance/divergence measures are demonstrated on numerical examples (Section 5) in order to demonstrate and evaluate their applicability as validation metrics. Following these discussions the measures are applied to model predictions from Bayesian History Matching (BHM) on a five storey building structure (Section 6). These provide a practical examination of the information each provides, leading to a discussion on how to use these measures in practice. Finally Section 7 offers conclusions and highlights areas for further research.

2 Validation Metrics and Metrics

This paper is concerned solely with validation metrics in a probabilistic setting, and in comparing their performance in providing a quantification of differences between distributions. The definition of a *validation metric* is a computable measure that quantifies the agreement between predictions from a simulator and observational data [2, 4, 9]. It has been stated in the literature that a validation metric should be separate from the criteria used in deciding whether to accept the simulator for a particular predictive context, and therefore a given validation metric is only required to quantify the difference [4, 9].

In order to assess the merits of particular distances/divergences as validation metrics it is appropriate to define criteria for an ideal validation metric. Combining previous criteria from the literature [3, 4, 9], and the authors' opinions, these criteria in the context of probabilistic engineering simulators are:

1. It should quantify the difference between the simulator predictions and observational data [3, 4, 9]
2. It should be interpretable and aid identifying simulator improvements
3. It should provide objective information and be consistent when applied to different probabilistic models or applications [3, 4]
4. It should account for the complete form of the distribution (and not just statistical moments) - if the underlying distribution of the observational data is unknown it should ideally have a non-parametric estimator with convergence guarantees

For clarity of terminology within this paper, the term *validation metric* is used to refer specifically to those mathematical operators that quantify the dissimilarities between predictions and observational data. The term *metric*, where used on its own, refers to the strict mathematical distance definition i.e. a distance $D(\cdot,\cdot)$ is a metric if it abides by four requirements [2]:

1. Non-negative: $D(x,y) \geq 0$
2. Identity of indiscernibles: $D(x,y) = 0$ if and only if $x = y$
3. Symmetric: $D(x,y) = D(y,x)$
4. Triangle inequality: $D(x,z) \leq D(x,y) + D(y,z)$

where x, y and z are three quantities (which for the simplest case would be points). It may be necessary for a validation metric to be a mathematical metric, the merits of this will be discussed further within this paper.

Finally, it is noted that each of the measures investigated as potential validation metrics within this paper can be formed into a frequentist hypothesis test, where the null hypothesis is that the simulator output and observational distributions are equal. By posing the problem of whether simulator outputs are adequate as a hypothesis test, a simulator can be determined inadequate, for a given significance level, if it causes the null hypothesis to be rejected (it is noted that statistically a hypothesis can never be proved, only rejected).

At a fundamental level, a hypothesis test provides a statistically rigorous framework for calculating a threshold, based on a given statistical distance, with which to make a decision about whether the simulator is invalid. The process for obtaining this threshold will be different for each measure, and will lead to different properties of the hypothesis test. In addition, the effectiveness of a given hypothesis test will depend on the distance/divergence measure it is constructed from. For these reasons, the paper focuses on the abilities of each measure investigated to quantify differences between distributions that occur anywhere within the probability mass, and does not perform hypothesis testing. If a measure is unsuccessful in quantifying dissimilarities anywhere in the probability mass, then it will not perform well as a general hypothesis test.

3 f-Divergences

The first family of distances/divergences considered are f-divergences (also known as *Csiszár's* ϕ*-divergences*). This category includes measures such as the Kullback-Leibler (KL)-divergence, and defines distances/divergences that depend on a ratio between probability measures [10]. These measures are of the form,

$$D_\phi(\mathbb{P},\mathbb{Q}) = \int_M \phi\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) \mathrm{d}\mathbb{P} \tag{1}$$

where M is a measurable space and ϕ is a convex function. $\mathbb{P}$ and $\mathbb{Q}$ are stated as probability measures, but generally will be utilised in the form of a Probability Density Function (PDF) or Cumulative Density Function (CDF). Equation (1) holds when $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$ and $-\infty$ otherwise. Different forms of the f-divergence depend on the choice of function ϕ with notable cases being the KL divergence, $\phi(t) = t\log(t)$, Hellinger distance, $\phi(t) = \left(\sqrt{t}-1\right)^2$, and Total Variation distance, $\phi(t) = |t-1|$ [10]. This family of divergence measures is widely used throughout information theory and machine learning [11].

3.1 Kullback-Leibler Divergence

The KL-divergence is the most widely used f-divergence and has many applications. A notable example is in performing variational inference as it represents a natural formulation of the ratio between two likelihood functions [12]. The KL-divergence of probability measures $\mathbb{P}$ and $\mathbb{Q}$ is,

$$D_{KL}(\mathbb{P},\mathbb{Q}) = KL(\mathbb{P}||\mathbb{Q}) = \int p(x)\log\left(\frac{p(x)}{q(x)}\right)\mathrm{d}x \tag{2}$$

where $p(x)$ and $q(x)$ are probability distributions of the random variable x, and is a measure of relative entropy [11]. It takes either the units nats or bits depending on the base of the logarithm, respectively exponential or base two. The divergence informs of the average number of extra nats (or bits) required to encode the data given that the distribution $\mathbb{Q}$ is used to model the 'true' distribution $\mathbb{P}$. More simply, it measures information lost when $\mathbb{Q}$ is used approximate $\mathbb{P}$. It is noted that a frequentist hypothesis test exists for the KL-divergence [13]. Resultantly, the hypothesis test could be used to objectively determine whether there are statistically significant differences between the simulator and observational distributions.

The KL-divergence can be difficult to estimate when the distribution form is unknown, and often proves challenging when the dimension size of samples increases (i.e. in the instants where d increases when $M = \mathbb{R}^d$). On the other hand, the divergence can be practical to compute between low-dimensional probability density functions and therefore is useful when the observational density function is known or can be accurately approximated.

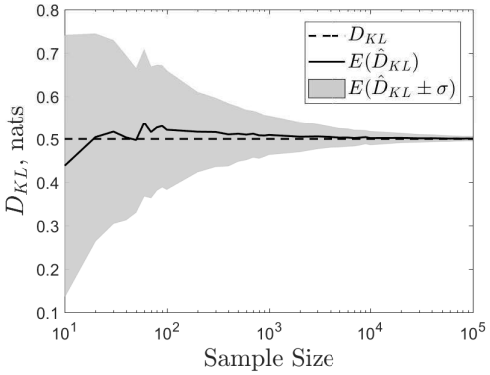


Fig. 1. Estimation of KL-divergence using data-dependent partitions where $\mathbb{P} \sim \mathcal{N}(0,1)$ and $\mathbb{Q} \sim \mathcal{N}(1,1)$. $D_{KL}(\mathbb{P},\mathbb{Q}) = 0.5$.

Empirical estimation of the KL-divergence in a non-parametric manner for continuous distributions can be performed using several approaches [14, 15]. However, often these non-parametric estimators require large sample sizes in order to converge as illustrated in Fig. 1. This example studies the convergence rate of one method for obtaining empirical estimate of the KL-divergence; calculated via data-dependent partition method proposed by Wang et al. [14]. In this example the empirical estimator is obtained when samples are drawn from two Gaussian distributions, $\mathbb{P} \sim \mathcal{N}(0,1)$ and $\mathbb{Q} \sim \mathcal{N}(1,1)$. 500 repetitions were performed at each sample size in order to demonstrate the variance of the estimator. It is clear from Fig. 1 that although the estimator will converge, this can be slow and requires a large sample size. In most engineering applications it is often not possible to obtain even hundreds of samples at each input indicating a drawback with the estimator.

3.1.1 Jenson-Shannon Distance

The KL-divergence is not a true mathematical metric as it does not meet two of the four requirements: it is neither symmetric nor does it obey the triangle inequality. A smoothed and symmetrised form of the KL-divergence is the Jenson-Shannon divergence [16], which by taking the square root becomes a metric, known as the Jenson-Shannon distance,

$$D_{JSD}(\mathbb{P},\mathbb{Q}) = \sqrt{\frac{1}{2}D_{KL}(\mathbb{P},\mathbb{M}) + \frac{1}{2}D_{KL}(\mathbb{Q},\mathbb{M})} \tag{3}$$

where $\mathbb{M} = \frac{1}{2}(\mathbb{P}+\mathbb{Q})$ and is the midpoint of the probability measures $\mathbb{P}$ and $\mathbb{Q}$. The Jenson-Shannon distance will always produce a finite result, unlike the KL-divergence as $\mathbb{P}$ and $\mathbb{Q}$ are always absolutely continuous with respect to $\mathbb{M}$ [16]. The computational overheads of the Jenson-Shannon distance are high due to the evaluation of the mixture distribution $\mathbb{M}$,

which becomes prohibitive in high dimensional data [17]. By construction it is less sensitive to scenarios when distribution $\mathbb{Q}$ contains sample values that are impossible in $\mathbb{P}$, unlike the KL-divergence, as it is bounded [16].

3.2 Hellinger Distance

The Hellinger distance is another statistical distance within the f-divergence family. It can be considered analogous to the Euclidean distance as it is the L_2-norm between two probability measures,

$$D_H(\mathbb{P},\mathbb{Q}) = \sqrt{\frac{1}{2}\int\left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2 \mathrm{d}x} \tag{4}$$

and is formed such that $D_H(\mathbb{P},\mathbb{Q}) \leq 1$. In addition, the Hellinger distance is a metric meeting all four requirements. This provides an intuitive interpretation of the distance where values of zero mean the two probability density functions are exactly equal and a distance close to one indicates very dissimilar probability density functions; however, the distance will nonlinearly change within these bounds. Frequentist hypothesis tests utilising the Hellinger distance also exist, which may aid decision making about simulator adequacy [18, 19].

3.3 Total Variation Distance

Total Variation distance is the L_1-norm equivalent to the Hellinger distance [20],

$$D_{TV}(\mathbb{P},\mathbb{Q}) = \frac{1}{2}\int |p(x)-q(x)|\mathrm{d}x \tag{5}$$

and is the only distance measure that can be classified as both an f-divergence and IPM (discussed in Section 4) [10]. In IPM form, Total Variation is written as,

$$D_{TV}(\mathbb{P},\mathbb{Q}) = \sup_{||f||_\infty \leq 1} |p(x)-q(x)| \tag{6}$$

and, like the Hellinger distance, Total Variation takes values in [0 1] aiding objectivity across applications. The metric can also be used within a frequentist hypothesis test [21].

4 Integral Probability Metrics

IPMs differ from f-divergences as they depend on the *difference* rather than *ratio* of probability measures. The general form of IPMs is defined as,

$$D_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}}\left|\int_M f\mathrm{d}\mathbb{P} - \int_M f\mathrm{d}\mathbb{Q}\right| \tag{7}$$

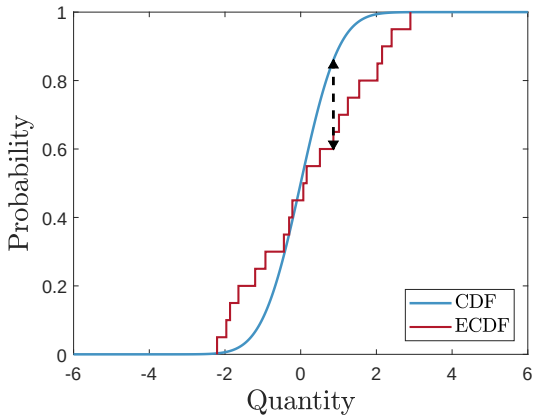


Fig. 2. An example of the Kolmogorov distance between $\mathbb{P} = \mathcal{N}(0, 0.8^2)$ and 20 samples from $\mathbb{Q} = \mathcal{T}(5)$ where $D_K(\mathbb{P},\mathbb{Q}) = 0.26$.

where $\mathcal{F}$ is a class of functions on M and sup is the supremum: the least upper bound of pointwise differences. The choice of $\mathcal{F}$ leads to various IPMs, such as the Total Variation distance where $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$; the Kolmogorov distance where $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]} : t \in \mathbb{R}_d\}$; Maximum Mean Discrepancy (MMD) where $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$ (i.e. all f that are Reproducing Kernel Hilbert Space (RKHS), $\mathcal{H}$); and the Wasserstein distance where $\mathcal{F} = \{f : ||f||_L \leq 1\}$ where L here refers to Lipschitz functions. These distances and their properties are considered in more depth below.

4.1 Kolmogorov Distance

The Kolmogorov distance is the maximum L_1-norm between two CDFs bounded [0, 1] and mathematically defined as,

$$D_K(\mathbb{P},\mathbb{Q}) = \sup_{x\in\mathbb{R}} |F_P(x) - F_Q(x)| \tag{8}$$

where $F_P(x)$ is a CDF for the probability measure $\mathbb{P}$ over the random variable x. The Kolmogorov distance is simply the largest vertical difference between the two CDFs and is most commonly used in hypothesis testing [22].

Figure 2 illustrates an example of the distance for a set of samples (forming an Empirical Cumulative Density Function (ECDF)[1]) $\hat{F}_Q(x)$ and a known distribution $F_P(x)$. Note, however, the distance holds if either $\mathbb{P}$ or $\mathbb{Q}$ are known or empirical. This is an advantage of the Kolmogorov distance, meaning it has the ability to handle a mixture of empirical and/or known CDFs, making it a flexible non-parametric tool for validation purposes.

The Kolmogorov distance is closely related to the Total Variation distance, described in Section 3.3. If the probability function is non-decreasing then Total Variation will provide the same solution as the Kolmogorov distance [23]. Furthermore, Total Variation is an upper bound on the Kolmogorov distance i.e. $D_K(\mathbb{P},\mathbb{Q}) \leq D_{TV}(\mathbb{P},\mathbb{Q})$ [20].

[1] An ECDF is mathematically defined as $\hat{F}_N(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_i \leq x)$.

4.2 Maximum Mean Discrepancy Distance

MMD is a measure of the maximum distance between the mean embeddings of two sample sets in a RKHS; projected using the function class $\mathcal{F}$, where the function f is called a reproducing kernel $k(\cdot,\cdot)$ [24]. The distance is defined as,

$$D_{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}} |\mathbb{E}_x(f(x)) - \mathbb{E}_y(f(y))| \tag{9}$$

where x and y are samples from $\mathbb{P}$ and $\mathbb{Q}$ respectively. There are several kernel types that can be chosen within the MMD metric, with a popular choice being the radial basis kernel [24, 25], defined as,

$$k(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right) \tag{10}$$

where σ is an associated hyperparameter that controls the width of the kernel. It is noted that most kernels will have some set of hyperparameters that need to be determined. A common approach for determining these hyperparameters is to use the median pairwise distance among the joint data [26]. The choice of kernel should reflect the prior belief about the smoothness of the underlying distribution and is often selected in a heuristic manner. However, Gretton et al. proposed an optimisation methodology for large sample sets in [27] whereby for a given α level (the significance level of a hypothesis test [24, 25, 28]) the technique selects linear combinations of kernels that minimise the probability of Type II errors and thus maximise the test power when used as the metric for a two sample hypothesis test [24]. In this paper by Gretton et al. the method is shown to perform well in the context of large data sets, where estimating the hyperparameter via a median heuristic approach and kernel selection via selecting the kernel with the largest MMD (i.e. choosing the conservative kernel) fails. In contrast, most validation tasks present the converse problem of involving small sample sizes where limited data could pose challenges to implementing this procedure.

MMD is a frequentist statistic and thus can be empirically estimated in both unbiased and biased forms, depending on whether the sample means are calculated using the U-statistics (unbiased),

$$D^2_{MMDu}(\mathbb{P},\mathbb{Q}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k(x_i,x_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(y_i,y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i,y_j) \tag{11}$$

or V-statistics (biased),

$$D^2_{MMDb}(\mathbb{P},\mathbb{Q}) = \frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i,x_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i,y_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i,y_j) \tag{12}$$

where m and n are the number of points in the samples X and Y respectively. These two forms of the statistic will both be zero when $\mathbb{P}=\mathbb{Q}$ and large when the distributions are far apart. MMD is a non-parametric technique, meaning that the form of the distribution does not need to be known before estimation.

4.2.1 Maximum Mean Discrepancy Witness Function

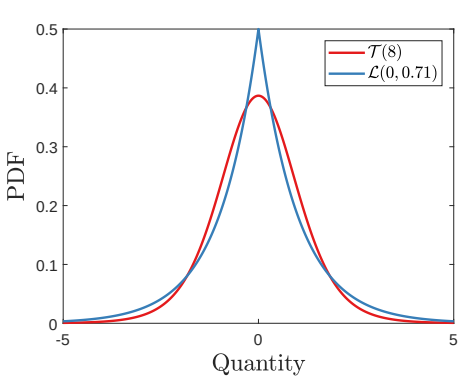
MMD, defined in Eq. (9), provides a key additional benefit in that the kernel embedding can be applied over a variable t in order to visualise the behaviour of the RKHS embeddings. This produces the witness function, f^*. An empirical estimation of the witness function can be defined as,

$$f^*(t) \propto \frac{1}{m}\sum_{i=1}^{m} k(x_i,t) - \frac{1}{n}\sum_{i=1}^{n} k(y_i,t) \tag{13}$$

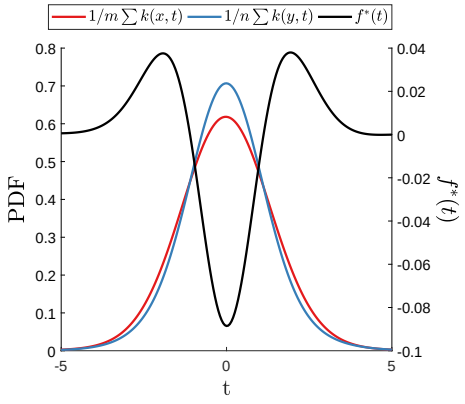
and used to provide a method for visually determining the dissimilarities between two distributions. The witness function is zero intuitively where the two distributions are the same, positive when $\mathbb{P}$ is larger than $\mathbb{Q}$, and negative when $\mathbb{Q}$ is greater than $\mathbb{P}$, as far as the smoothness constraint allows.

To demonstrate the effectiveness of the witness function, a one-dimensional example is presented in Fig. 3. The scenario considers the difference between a Student's t distribution with eight degrees of freedom and a Laplace distribution, $\mathcal{L}(0, 0.71)$. 10000 samples were drawn from each distribution and the MMD distances (both biased and unbiased) calculated using a radial basis kernel with $\sigma = 0.85$; $D_{MMDu} = D_{MMDb} = 0.11$. Visually, the witness function in Fig. 3 highlights where key differences in the probability mass occur.

The witness function can be implemented as a tool for locating the differences between distributions and helping diagnose model inadequacies. For example, if in Fig. 3 X are simulator predictions and Y observations, it can be easily identified that more probability mass is located around zero from the sample set Y than is modelled by X; this is indicated by negative values in the witness function. In addition, X has more probability mass in both tails, indicated by the positive values in the witness function. A near symmetric witness function indicates that the mean predictions are very similar. The witness function in this example would diagnose a conservative simulator output, where a distribution with a steeper probability mass decay from the mode would



(a)



(b)

Fig. 3. An example of a witness function between 10000 samples from $X \sim \mathcal{T}(8)$ and $Y \sim \mathcal{L}(0,0.71)$, $D_{MMDu} = D_{MMDb} = 0.11$, where $\mathcal{L}(\cdot,\cdot)$ and $\mathcal{T}(\cdot)$ are Laplace and Student's t distributions. A radial basis kernel where σ is inferred from the median heuristic is implemented. Panel (a) are the PDFs of the distributions from which the finite samples are drawn and panel (b) are the mean kernel embeddings of the two samples and the witness function over a space t.

improve the prediction. In this one dimensional case this information may appear obvious, however this will not always be the case in more complex and bespoke distributions. Furthermore, in higher dimensional spaces it becomes challenging to compare two PDFs. The witness function potentially provides a very useful, low dimensional, interpretable diagnostic for such scenarios.

4.3 Area Metric

The Area Metric, proposed by Ferson et al. [3], is a popular validation metric in engineering for assessing the difference between two distributions [2, 4–6]. The Area Metric is the area of the L_1-norm between two CDFs,

$$D_{Area}(\mathbb{P},\mathbb{Q}) = \int |F_P(x) - F_Q(x)| \mathrm{d}x \tag{14}$$

and is illustrated in Fig. 4.

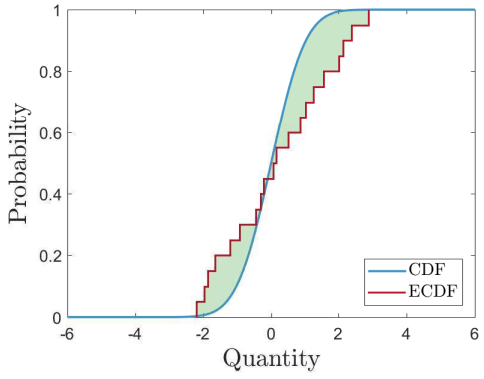


Fig. 4. An example of the Area Metric (the shaded region) between $\mathbb{P} = \mathcal{N}(0,0.8^2)$ and 20 samples from $\mathbb{Q} = \mathcal{T}(5)$. In this case $D_{Area}(\mathbb{P},\mathbb{Q}) = 0.64$.

The metric also represents the distance between quantile functions (inverse CDFs) i.e. $\int |F_P^{-1}(p) - F_Q^{-1}(p)| dp$ where p is a probability [2]. This is the definition of a *Kantorovich metric*, i.e. $D_W(\mathbb{P},\mathbb{Q}) = \int |F_P(x) - F_Q(x)| \mathrm{d}x = \int |F_P^{-1}(p) - F_Q^{-1}(p)| dp$ where F^{-1} is the inverse function of the general distribution function F [29, 30]. This means that the Area Metric is part of the Wasserstein (or Kantorovich) distances, and is, in fact, the univariate case. As a result the Wasserstein distance hypothesis tests [31] could be applied to the Area Metric such that decisions could be made about the statistically significant differences between simulator predictions and observational data. More generally, the Area Metric is part of a family of metrics, known as the L_p metrics, where the L_p-norm is taken rather than L_1 [29].

Oberkampf and Roy state in [2] that a significant merit of the Area Metric is that the units are that of the quantity in question, i.e. if the random variable X were an observation of stress in MPa then the units of the Area Metric are also MPa, since probability is dimensionless [2]. The distance therefore scales with the units of observed quantity.

5 Numerical Case Studies

In order to compare the statistical distances/divergences introduced in Sections 3 and 4 against the criteria in Section 2, several numerical examples are considered. These case studies are intended to demonstrate relative differences between the measures, in regard to the validation metric criteria, and not as a complete mathematical analysis of each equation's sensitivities.

The scenarios considered in this section are all comparisons of continuous distributions with known mathematical forms. In order to keep comparisons consistent, numerical integration is implemented to calculate each distance/divergence (however, it is noted that for certain distribution forms, the integrals in some distances/divergences can be solved in closed form e.g., the Hellinger distance between two Gaussian distributions).

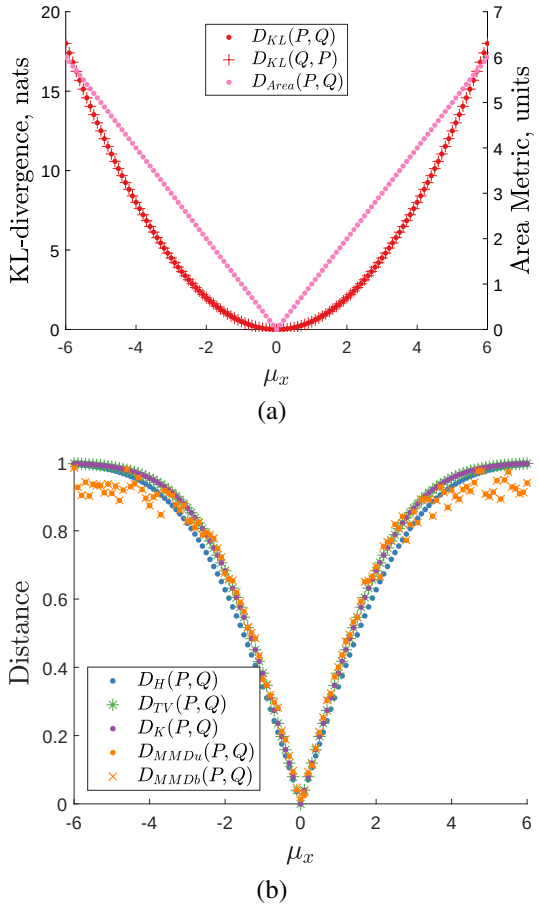


Fig. 5. A comparison of probabilistic distances/divergences for two Gaussian distributions, $\mathbb{P} \sim \mathcal{N}(0,1)$ and $\mathbb{Q} \sim \mathcal{N}(\mu_x,1)$; where the mean μ_x is varied from [-6 6] with a fixed variance. Panel (a) shows the KL-divergences and Area Metric, as these have units. Panel (b) presents the Hellinger, Total Variation, Kolmogorov and MMD distances. The MMD distance is calculated from 2000 samples with a radial basis kernel where σ is inferred from the median heuristic, and all other distances from numerical integration over the range [-30 30] in 0.01 steps.

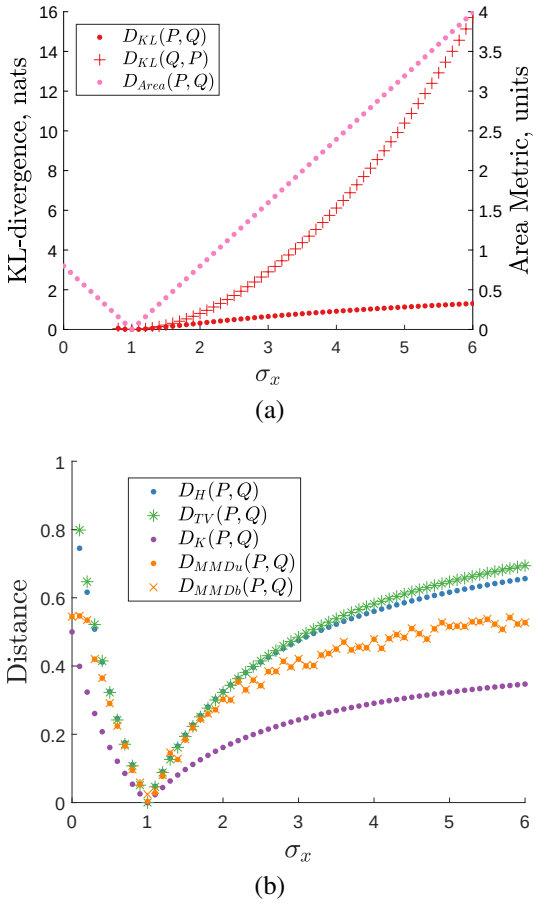


Fig. 6. A comparison of probabilistic distances/divergences for two Gaussian distributions, $\mathbb{P} \sim \mathcal{N}(0,1)$ and $\mathbb{Q} \sim \mathcal{N}\left(0,\sigma_x^2\right)$; where the standard deviation σ_x is varied from [0 6] with a fixed mean. Panel (a) shows the KL-divergences and Area Metric, as these have units. Panel (b) presents the Hellinger, Total Variation, Kolmogorov and MMD distances. The MMD distance is calculated from 2000 samples with a radial basis kernel where σ is inferred from the median heuristic, and all other distances from numerical integration over the range [-30 30] in 0.01 steps.

The first two scenarios explore the sensitivity of these distance/divergence measures to changes in lower order moments, specifically in the context of Gaussian distributions, $\mathbb{P} \sim \mathcal{N}(0,1)$ and $\mathbb{Q} \sim \mathcal{N}\left(\mu_x,\sigma_x^2\right)$. In the first case study the mean μ_x is varied and the variance σ_x^2 is fixed, the second case considers the mean μ_x fixed and the standard deviation σ_x variable. The third example quantifies each distance/divergence between several other distribution forms. As a result, comments are made about each measure's sensitivity to general changes in probability mass, such that the fourth validation metric criteria in Section 2 can be more widely assessed.

5.1 Sensitivity to Variation in the Mean — Gaussian Distribution Case

Figure 5 displays a comparison of the distances/divergences when the mean is varied (in a Gaussian distribution context). Figure 5a presents the KL-divergences and Area Metric, as these both have units, with Fig. 5b showing a comparison of the remaining dimensionless measures.

For this example the KL-divergence is symmetric (i.e. $D_{KL}(\mathbb{P},\mathbb{Q}) = D_{KL}(\mathbb{Q},\mathbb{P})$). It is also slow to increase and as a result may struggle to detect small variations in the mean. The unbounded nature of the KL-divergence also makes it a difficult measure to interpret, especially if used as a validation metric. In contrast, the Area Metric values are equal to the distance between the two distribution means i.e. when $\mu_x = 2$, $D_{Area}(\mathbb{P},\mathbb{Q}) = 2$. This result follows, as the Area Metric mathematically becomes the distance between the two distribution means, when the remaining statistical moments (in this case the variances) are the same.

Comparing the distance metrics bounded on [0 1] — the Hellinger, Total Variation and Kolmogorov distances — illustrates that Total Variation and Kolmogorov distances are equally more sensitive to the change in mean (based on these

measures gradients) between [-2 2], where outside of this interval the Hellinger distance is then more sensitive. With the knowledge that these have an upper bound of 1, the distances become quite large relatively quickly, i.e. when $\mu_x = 2$ Total Variation and Kolmogorov distances are 0.68 compared with 0.62 for the Hellinger distance. For this scenario the distances can be interpreted as not close and would lead to an acknowledgement of significant inadequacy in the relationship between the simulator and observations. It is argued that these distances give a better indication of the relative difference between the distributions providing a more objective comparison when compared with the KL-divergence and Area Metric. The MMD distances do not have an upper bound but track relatively consistently with the Total Variation, Kolmogorov and Hellinger distances. It is noted that the MMD's non-parametric, sample-based approximation of the distributions leads to oscillations in the metrics. Additionally, both bias and unbiased results are very similar and become less sensitive to changes in the mean ≥ 4 and ≤ -4 when compared with the Kolmogorov and Hellinger distances.

5.2 Sensitivity to Variation in the Standard Deviation — Gaussian Distribution Case

The second scenario, shown in Fig. 6, considers variations in the standard deviation with a fixed mean. Figure 6a presents the KL-divergences and Area metric. This example demonstrates the asymmetric nature of the KL-divergence where more nats of information are required in order to encode $\mathbb{Q}$ when $\mathbb{P}$ is the model distribution than in the opposing case. This is because there is a greater overlap in probability mass when $\mathbb{Q}$ approximates $\mathbb{P}$, and therefore less information required to encode $\mathbb{P}$, than in the alternative case for this example (however, in the scenario where the means are varied and the standard deviations are fixed the overlap in probability mass is the same for both cases). This means that the KL-divergence will often favour conservative model distributions, which can be useful for a validation setting. However, this can also be a negative attribute of the KL-divergence, as it could lead to a modeller over-inflating the predictive uncertainties from a simulator such that it produces a lower KL-divergence. Moreover, the units of the KL-divergence are difficult to intuitively interpret. The Area Metric, on the other hand, linearly scales with a change in variance and appears almost symmetric about the variance of $\mathbb{P}$. This suggests that the Area Metric struggles to differentiate between under- and over-estimations of the variance; an unhelpful property in validation. Nonetheless the Area Metric is valuable as the units are the same as the quantity of interest.

In comparison, Total Variation, Hellinger and Kolmogorov distances, displayed in Fig. 6b, appear more sensitive to underestimation of the variance, indicated by a steeper gradient of distances below a standard deviation of 1. In this case study, Total Variation is more sensitive to changes in the standard deviation than the Hellinger or Kolmogorov distances. Here, the Kolmogorov distance becomes less sensitive than the Hellinger distance, which is due to the fact that the Kolmogorov distance is less sensitive to changes in the tails, compared to difference in the central probability mass. Again, both MMD distances track in a similar manner to the Hellinger distance between standard deviations of 0.5 and 2, becoming less sensitive outside these values, but still penalising under-estimation of the variance more heavily than over-estimation.

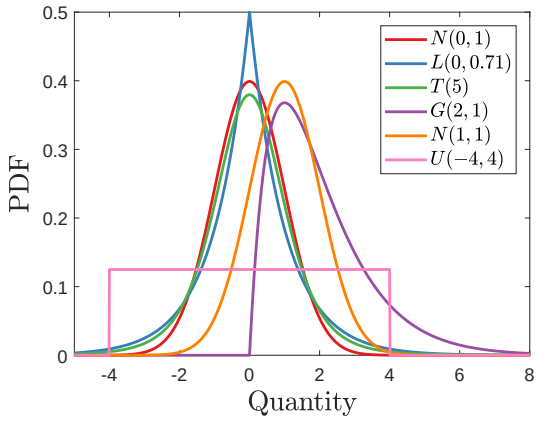


Fig. 7. Distributions used in the comparison of distance/divergences.

5.3 Different Distribution Forms

The next examples, presented in Tables 1 and 2, compare the statistical distances for different forms of distribution. The first two examples compare standard Gaussian and Laplace distributions (with the same mean and variance) — example one — as well as standard Gaussian and Student's t distributions — example two. These two comparisons have been chosen as the distribution forms in each case have small dissimilarities, as shown in Fig. 7. For these two examples, the KL-divergences (in both directions) indicate that relatively small amounts of information are required to encode the 'true' distribution, from the low KL-divergences given the log ratio relationship.

The Kolmogorov distance shows very small distances, which is expected given its insensitivity to differences away from the central probability mass. The MMD distances, both biased and unbiased, produce comparable results calculating larger distances for the Laplace than the Student's t distributions. The biased MMD produces almost equivalent distances to the Total Variation distance. The Hellinger distances also show that the standard Gaussian is closer to the Student's t distribution than the Laplace distribution, but by a relatively smaller amount. The two Area Metrics for these examples are equal. This demonstrates a failure to capture the knowledge that a Student's t is expected to be closer to the standard normal than a Laplace distribution.

Evaluating the KL-divergence for the next two examples — a comparison of Gamma and Gaussian distributions in example three, and of uniform and Gaussian distributions in example four — presents issues with using numerical integration, but provides informative results. The Gamma dis-

Example	$\mathbb{P}$	$\mathbb{Q}$	$D_{KL}(\mathbb{P},\mathbb{Q})$	$D_{KL}(\mathbb{Q},\mathbb{P})$	$D_{H}(\mathbb{P},\mathbb{Q})$	$D_{TV}(\mathbb{P},\mathbb{Q})$
1	$\mathcal{N}(0,1)$	$\mathcal{L}(0,0.71)$	0.07	0.23	0.16	0.12
2	$\mathcal{N}(0,1)$	$\mathcal{T}(5)$	0.03	0.12	0.11	0.06
3	$\mathcal{G}(2,1)$	$\mathcal{N}(1,1)$	-	∞	0.38	0.25
4	$\mathcal{U}(-4,4)$	$\mathcal{N}(0,1)$	-	∞	0.46	0.49

Table 1. Examples of f-divergences for different distributions. Numerically integrated over the range [-30 30] in 0.01 steps. KL-divergences are in nats.

Example	$\mathbb{P}$	$\mathbb{Q}$	$D_{K}(\mathbb{P},\mathbb{Q})$	$D_{MMDu}(\mathbb{P},\mathbb{Q})$	$D_{MMDb}(\mathbb{P},\mathbb{Q})$	$D_{Area}(\mathbb{P},\mathbb{Q})$
1	$\mathcal{N}(0,1)$	$\mathcal{L}(0,0.71)$	0.06	0.12	0.12	0.15
2	$\mathcal{N}(0,1)$	$\mathcal{T}(5)$	0.03	0.04	0.05	0.15
3	$\mathcal{G}(2,1)$	$\mathcal{N}(1,1)$	0.25	0.26	0.26	1.00
4	$\mathcal{U}(-4,4)$	$\mathcal{N}(0,1)$	0.25	0.44	0.44	1.20

Table 2. Examples of IPM distances for different distributions. Numerically integrated over the range [-30 30] in 0.01 steps apart from the MMD distances which are estimated from 2000 samples with a radial basis kernel where σ is inferred from the median heuristic.

tribution contains no probability mass below zero, as it is bounded at one end. It is therefore impossible for a Gaussian distribution that has symmetric probability mass over the $[-\infty\ \infty]$ range, to ever be able to replicate the Gamma distribution, given any amount of additional information; it will always have some probability mass beyond the bound. In contrast, a Gamma distribution would require an infinite amount of additional information below zero to replicate the Gaussian distribution. The KL-divergence, calculated in this manner, is extremely informative in diagnosing these issues, i.e. that it is not possible to model the observational distribution using the simulator distribution. Similar problems also exist in the comparison of a uniform and Gaussian distributions, given that the uniform distribution contains no probability mass outside of its range.

The Kolmogorov distances for these examples are the same, illustrating once again the insensitivity of this measure to deviations that are outside the central probability mass. Moreover, the Total Variation, Hellinger and MMD distances, including the Area Metric, all quantify that the uniform and Gaussian distribution distances are further than the Gamma and Gaussian distribution. Once more the Total Variation is almost equivalent to the MMD distances.

5.4 Discussion of Numerical Case Studies

The results from empirical numerical observations indicate the strengths and weaknesses of the distances/divergences considered. It can be summarised that the KL-divergence becomes very sensitive in scenarios where large amounts of extra information are required to replicate the 'true' distribution, and its convex nature makes it ideal for optimisation settings. This makes the divergence useful for scenarios when the question of whether to obtain more observations or simulator runs to solve issues of inadequacy are asked. The major drawback of the KL-divergence it is not easily interpretable.

The Kolmogorov distance is flawed as a general distribution validation metric for the aforementioned reasons. It is not recommended as the sole qualification of the distance between distributions as it fails to adequately meet the fourth validation metric criteria in Section 2. The Total Variation, Hellinger and Kolmogorov distances are arguably more objective in comparing two distributions given that 0 indicates they are the same and 1 that the distributions are as far as possible — criteria three from Section 2. Furthermore, the Total Variation and Hellinger distances provide better quantification of a wider variety of differences when compared to the Kolmogorov distance. These two distances are sensitive to a variety of differences in probability mass and would be appropriate for most engineering applications, and in the author's opinion are relatively interpretable from the results in Table 1.

Furthermore, the MMD distances for these numerical case studies tend to provide similar distances to both the Total Variation and Hellinger distances, and may be practical in a variety of settings due to its non-parametric formulation. However, for small sample sizes it will be more dependent on kernel and hyperparameter choices adding a level of modeller input that may be unwanted — although calculation of the median heuristic removes a level of subjectivity.

Lastly, the Area Metric, although in the units of the quantity of interest, is relatively hard to objectively interpret. The Area Metric also displayed difficulty in differentiating between under- and over-estimation of the variance for these numerical examples, often problematic when conservative results are required.

It is noted that all the examples considered here have been for univariate distributions. Different conclusions may be found with higher-dimensional distributions in line with the findings of Aggarwal et al. where fractional norms increase sensitivity for high-dimensional non-statistical distances [32]. This is left as further research, as this paper

Fig. 8. Test setup of the representative five storey building structure.

is focused on providing a framework for utilising statistical distances in the validation of probabilistic model outputs.

6 Case Study: Bayesian History Matching Example

An experimental case study is provided in order to demonstrate the applicability of the considered distance/divergence measures as validation metrics. The case study considers a five storey building structure displayed in Fig. 8 constructed from aluminium 6082. The objective of this analysis was to calibrate the three material properties $\theta = \{E, \nu, \rho\}$ of a finite element computer model, using Bayesian History Matching (BHM) in order to predict the first five bending natural frequencies $\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$ of the structure under varying levels of mass, $x = \{0, 0.1, \ldots, 0.5\}$kg, attached to the first floor.

Experimental data were obtained using experimental modal analysis, whereby the structure was excited laterally with a 409.6Hz bandwidth Gaussian excitation via an electrodynamic shaker and five accelerometers used to capture the response at each floor. The sample rate and sample time were chosen such that the frequency resolution was 0.05Hz. 40 averages were acquired for each measurement and for each level of mass, ten repeats were performed in order to obtain an understanding of the underlying modal frequency distribution.

The data used in the calibration process were the mean natural frequencies when the mass was $x_z = \{0, 0.3, 0.5\}$kg. The remaining full repeat data were used as an unseen validation set z_*. The prior bounds on the material properties were $\pm 10\%$ of the typical values for aluminium 6082; $E = 71$GPa, $\nu = 0.33$ and $\rho = 2770$kg/m^3.

6.1 Bayesian History Matching

Bayesian history matching is a methodology for calibrating statistical models of the form,

$$z_j(x) = \eta_j(x, \theta) + \delta_j + e_j \tag{15}$$

where $z_j(x)$ is the jth observational output given inputs x, $\eta_j(x, \theta)$ is the jth simulator given x and parameters θ. The model discrepancy and observational uncertainty are δ and e, respectively. The model assumes that the simulator, model discrepancy and observational uncertainty are independent and does not seek to define the model discrepancy's functional form.

The 'likelihood free' technique utilises an implausibility metric to iteratively discard parts of the input space that were unlikely to have generated the observational data, given a set of uncertainties, defined as,

$$I_j(x, \theta) = \frac{|z_j(x) - \mathbb{E}\left(\mathcal{GP}_j(x, \theta)\right)|}{\left(V_{o,j} + V_{m,j} + V_{c,j}(x, \theta)\right)^{1/2}} \tag{16}$$

where V_o, V_m, and $V_c(x, \theta)$ are variances associated with the observational, model discrepancy and code uncertainties (the variance of the Gaussian Process (GP) emulator) and $\mathbb{E}(\mathcal{GP}(x, \theta))$ is the mean of the GP emulator. Due to the focus of this paper being on the assessment of validation metrics the reader is referred to [33, 34] for a more detailed overview of BHM.

Once calibrated, the outputs from BHM can be used to infer the functional form of the model discrepancy term. Here an importance sampling approach is implemented, whereby a second GP model is inferred whilst marginalising out the posterior parameter distribution $p(\theta | Z)$. Again, due to the scope of this paper the reader is referred to [35,36] for a more detailed explanation of the analysis. The result of this approach is that calibrated and bias-corrected predictive distributions can be inferred across the input space.

The predictions from both the BHM and importance sampling approach for the five storey building structure are presented in Figs. 9 and 10.

6.2 Validation of output predictions

The proposed validation metrics outlined in the previous sections were applied to the BHM predictions shown in Figs. 9 and 10. It is noted that the Normalised Mean Squared Error (NMSE) for each natural frequency prediction were 157.60, 0.07, 0.01, 0.01 and 0.12 respectively. This deterministic metric would indicate that the mean predictions are adequate for the second to fifth natural frequencies with large errors in the first natural frequency (as visually intuitive from Figs. 9 and 10).

To analyse the predictions, further distance metrics were applied. The f-divergence measures were all compared to Kernel Density Estimates (KDEs) of the observational data and calculated via numerical integration, as presented in Fig. 11. The KL-divergence (where $\mathbb{P}$ is the observational data and $\mathbb{Q}$ the model predictions, Fig. 11a), clearly captures the large discrepancy for the first natural frequency predictions at 0.1 and 0.2kg. In general, the first natural frequency predictions all produce relatively large (> 2) KL-divergences. Apart from the third natural frequency predic-

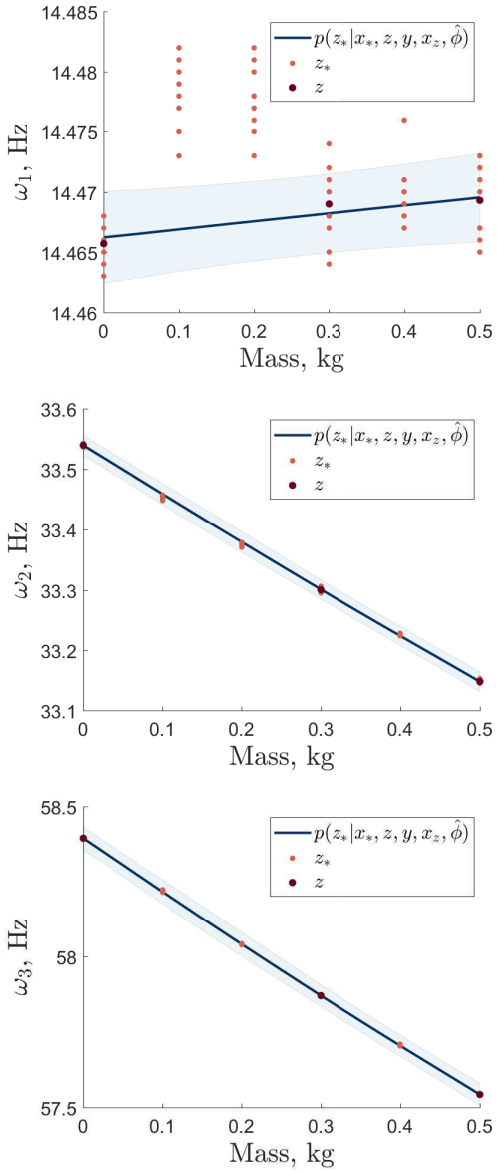


Fig. 9. BHM predictive outputs ($\omega_1, \omega_2, \omega_3$) with inference of model discrepancy via importance sampling trained GPs. The shaded regions indicate $\pm 3\sigma$.

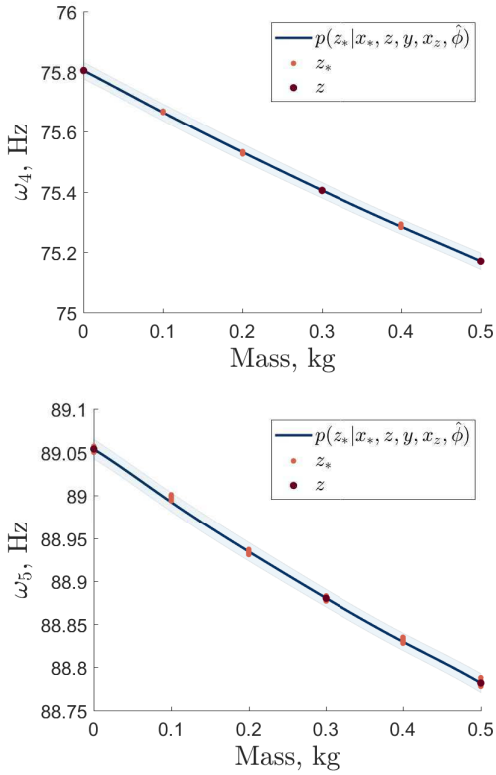


Fig. 10. BHM predictive outputs (ω_4, ω_5) with inference of model discrepancy via importance sampling trained GPs. The shaded regions indicate $\pm 3\sigma$.

tions at 0.2 and 0.3kg, the remaining predictions all have a KL-divergence < 1.5, with the majority being below 1, informing relatively 'good' agreement.

The Hellinger and Total Variation distances (Fig. 11b and Fig. 11c) also confirm that the first natural frequency predictions are 'far' from the observational data, especially at 0.1 and 0.2kg. Both of these distances show very similar distances and relative trends, e.g. that the fifth natural frequency is closest for the 0, 0.2, 0.4 and 0.5kg masses, and far at 0.1kg due the slight offset in mean. A difference between these two distances occurs for the first natural frequency at 0.1kg, where Total Variation quantifies a larger discrepancy.

The IPMs are displayed in Fig. 12. The Kolmogorov distance (Fig. 12a) and Area Metric (Fig. 12c) are compared to empirical CDFs of the observations. Both of these metrics indicate that the first natural frequency predictions at 0.1 and 0.2kg are the furthest away from the observations, with the Area Metric also stating that the 0.4kg prediction is close. In addition, both of these metrics better capture that the second natural frequency predictions at 0.1kg and 0.2kg have large discrepancies, due to an offset in the predictive mean. A challenge here is that the Area Metric magnitudes are all relatively low, at an order of magnitude of 10^{-3}Hz. This is caused by the close spacing of the observational points, leading to small areas between the empirical and predicted CDFs. At these magnitudes of frequency the Area Metric would therefore indicate that all predictions, even for the first natural frequency, are 'good', and may lead to the acceptance of an inadequate model. The biased MMD distance (Fig. 12b) is utilised in this case study and calculated from the average distance when 100 repeats of ten samples are drawn from the predictive distribution. In agreement with the Area Metric, the MMD distances follow a similar pattern for the first natural frequency, with it stating that the prediction at 0.4kg is close.

Finally, a key benefit of the MMD distance over the other distances/divergences is the ability to interrogate the differences between distributions via the witness function. This provides a potentially useful and powerful diagnostic tool for determining where modelling improvements may be made. Fig. 13 presents a comparison of the simulator and observational distributions against the witness function,

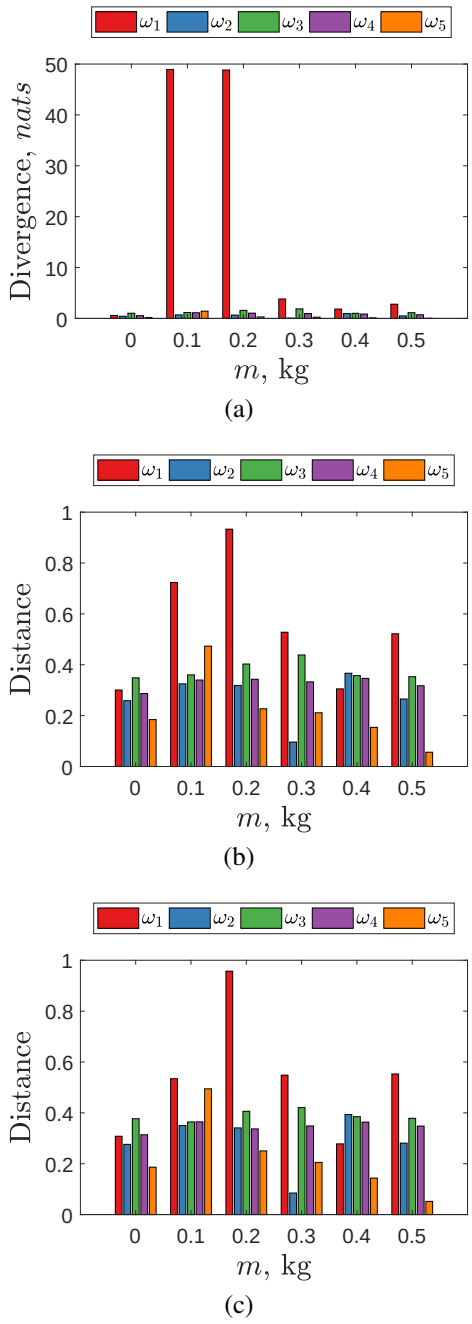


Fig. 11. f-divergence measures applied to the BHM and importance sampling predictions. Panel (a), (b) and (c) are the KL-divergence, Total Variation and Hellinger distance, when compared to KDEs of the observational data. These measures have been calculated via numerical integration.

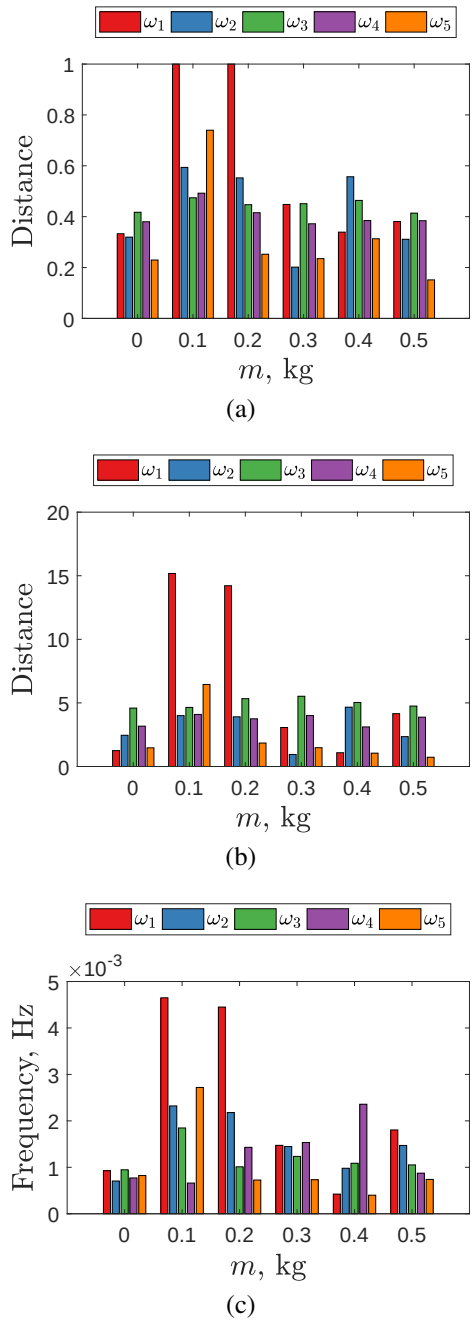


Fig. 12. IPM statistical distances applied to the BHM and importance sampling predictions. Panel (a) and (c) are the Kolmogorov distance and Area Metric when compared to empirical ten point observational CDFs. Panel (b) is the averaged MMD distance over 100 repeats of ten samples from the predictive distribution. A radial basis kernel where σ is inferred from the median heuristic is implemented.

demonstrating its diagnostic capabilities. Even though the fifth natural frequency has been 'adequately' captured by the simulator, the witness function clearly highlights several differences. The 0, 0.2 and 0.3kg predictions all over-estimate the variance with slight shifts in the mean values, indicated by the witness function being negative about the mean and asymmetric. These results can be interpreted as conservative, given the relatively small number of observations. For the 0.1kg case it can clearly be seen that there is an offset in the mean value, although the observation distribution is still within the majority of the simulators probability mass. The 0.4kg case shows an offset between the two distributions. Furthermore, although the simulator appears to have almost matched the observational data for the 0.5kg case, the witness function has highlighted that the simulator has a higher prediction of the mean with a larger variance than the observational distribution. This highlights the witness func-

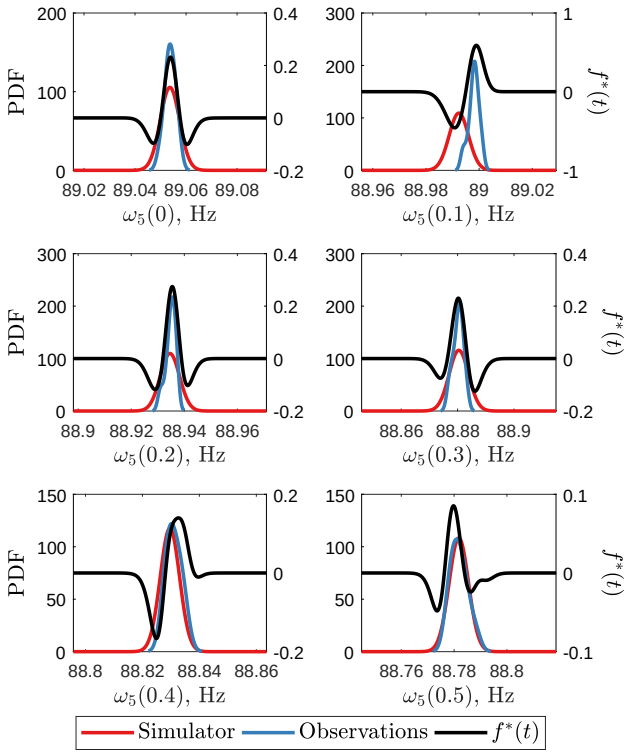


Fig. 13. Witness function ($f^*(t)$) for the fifth natural frequency compared with the Gaussian simulator distributions and KDEs of the observational data. The witness functions are constructed from a radial basis kernels where σ is inferred from the median heuristic.

tion's use in quantifying where the differences in probability mass occur, potentially aiding the correction of the simulator or leading an improved experiential test strategy.

7 Conclusion

Understanding and quantifying uncertainties in simulator predictions requires the development of validation metrics that can assess the differences between the simulator and observational distributions. This paper has categorised existing validation metrics within two families of statistical distances/divergences, namely f-divergences — KL-divergence, Hellinger distance, Total Variation distance — and IPMs — Total Variation distance, Kolmogorov distance, MMD distance and the Area Metric. This has shown that a wider variety of statistical distances/divergences exist that could be implemented as potential validation metrics.

It is noted that these measures all rely on multiple samples of the observations, which may be challenging to obtain in real world applications; although this paper assumes enough samples are obtainable. For this reason understanding the convergence rates of non-parametric estimators of these measures should be investigated as further research. Moreover, the distance/divergence values can be difficult to objectively interpret. As each of the measures outlined in this paper have an equivalent frequentist hypothesis test, these should be investigated such that their performances as validation metrics can be further scrutinised.

The measures discussed in this paper have been compared both in numerical examples and an experimental case study. The numerical case studies have led to the conclusion that the Kolmogorov distance is often insensitive to differences outside of the central probability mass, making it impractical for some validation contexts. The KL-divergence will often be difficult to interpret, but can provide useful information in diagnosing problems where significant differences (or impossibilities) in the probability mass are present. Both Total Variation and Hellinger distances show a good level of sensitivity to differences in distributions. The MMD distances produced similar distances to the Total Variation and Hellinger distance for these numerical example, meaning that it could be an informative and stable method for providing a non-parametric distance between samples. Finally, the Area Metric is useful in that it quantifies the distance in terms the quantity of interest units. Despite this, the Area Metric can be hard to objectively compare. Furthermore, it appears to fail to distinguish between under- and over-estimation of the variance for the case studies provided. It is therefore suggested that for most validation applications, a combination of the KL-divergence, Area Metric and either the Total Variation, Hellinger or MMD distances would be effective in assessing the simulator's adequacy.

The experimental case study again confirmed the difficulties in interpreting the KL-divergence, with it being most useful in situation where large differences are present. Both the Total Variation and Hellinger distances provide similar quantifications of the differences between distributions and are able to quantify a range of dissimilarities between two distribution's probability mass. In addition, the Total Variation and Hellinger distances, along with the Kolmogorov distance, are standardised across problems due to being bounded [0 1]. The Area Metric produced very small magnitudes in distance between the simulator predictions and the observations, which could lead to miss-identifying inadequacy. Furthermore, the MMD distance provides both a non-parametric method for assessing distance but also the ability to interrogate the differences in probability mass using the witness function. This can be a key tool in diagnosing areas of difference as part of a wider validation strategy.

Acknowledgements

The authors would like to acknowledge the support of the UK Engineering and Physical Sciences Research Council via grant number EP/R006768/1.

References

[1] ASME PTC Committee, 2009. Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer: ASME V&V 20.

[2] Oberkampf, W. L., and Roy, C. J., 2010. *Verification and Validation in Scientific Computing*. Cambridge University Press, Cambridge.

[3] 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), pp. 2408–2430.

[4] Liu, Y., Chen, W., Arendt, P., and Huang, H.-Z., 2011. "Toward a Better Understanding of Model Validation Metrics". *Journal of Mechanical Design,* ***133***(7), p. 71005.

[5] Xu, H., Jiang, Z., Apley, D. W., and Chen, W., 2016. "New Metrics for Validation of Data-Driven Random Process Models in Uncertainty Quantification". *Journal of Verification, Validation and Uncertainty Quantification,* ***1***(2), jun, pp. 21002–1.

[6] Wang, Z., Fu, Y., Yang, R.-J., Barbat, S., and Chen, W., 2016. "Validating Dynamic Engineering Models Under Uncertainty". *Journal of Mechanical Design,* ***138***(11), p. 111402.

[7] Sankararaman, S., and Mahadevan, S., 2013. "Assessing the Reliability of Computational Models under Uncertainty". In 54th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, American Institute of Aeronautics and Astronautics.

[8] Li, C., and Mahadevan, S., 2016. "Role of calibration, validation, and relevance in multi-level uncertainty integration". *Reliability Engineering & System Safety,* ***148***, pp. 32–43.

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

[10] Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G., 2009. "On integral probability metrics, phi-divergences and binary classification". *arXiv*(1), jan, pp. 1–18.

[11] Murphy, K. P., 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.

[12] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D., 2017. "Variational Inference: A Review for Statisticians". *Journal of the American Statistical Association,* ***112***(518), apr, pp. 859–877.

[13] Yang, P., and Chen, B., 2019. "Robust Kullback-Leibler Divergence and Universal Hypothesis Testing for Continuous Distributions". *IEEE Transactions on Information Theory,* ***65***(4), apr, pp. 2360–2373.

[14] Wang, Q., Kulkarni, S., and Verdu, S., 2005. "Divergence Estimation of Continuous Distributions Based on Data-Dependent Partitions". *IEEE Transactions on Information Theory,* ***51***(9), sep, pp. 3064–3074.

[15] Nguyen, X. L., Wainwright, M. J., and Jordan, M. I., 2007. "Nonparametric estimation of the likelihood ratio and divergence functionals". In IEEE International Symposium on Information Theory, pp. 2016–2020.

[16] Lin, J., 1991. "Divergence measures based on the Shannon entropy". *IEEE Transactions on Information Theory,* ***37***(1), pp. 145–151.

[17] Abou-Moustafa, K. T., and Ferrie, F. P., 2012. "A Note on Metric Properties for Some Divergence Measures : The Gaussian Case". In Proceedings of the Asian Conference on Machine Learning, pp. 1–15.

[18] Basu, A., Mandal, A., and Pardo, L., 2010. "Hypothesis testing for two discrete populations based on the Hellinger distance". *Statistics and Probability Letters,* ***80***(3-4), pp. 206–214.

[19] Zhixiang Wang, 2010. "An application of Hellinger distance: Hypothesis testing for two continuous populations". In IEEE International Conference on Intelligent Computing and Intelligent Systems, Vol. 1, pp. 340–342.

[20] Christofides, T. C., and Vaggelatou, E., 2009. "Bounds for the Distance Between the Distributions of Sums of Absolutely Continuous i.i.d. Convex-Ordered Random Variables with Applications". *Journal of Applied Probability,* ***46***(1), pp. 255–271.

[21] Devroye, L., Gyorfi, L., and Lugosi, G., 2002. "A note on robust hypothesis testing". *IEEE Transactions on Information Theory,* ***48***(7), jul, pp. 2111–2114.

[22] Massey, F. J., 1951. "The Kolmogorov-Smirnov Test for Goodness of Fit". *Journal of the American Statistical Association,* ***46***(253), pp. 68–78.

[23] Adell, J. A., and Jodrá, P., 2006. "Exact Kolmogorov and total variation distances between some familiar discrete distributions". *Journal of Inequalities and Applications,* ***2006***, pp. 1–8.

[24] Gretton, A., Borgwardt, K. M., Rasch, M. J., Scholkopf, B., and Smola, A., 2012. "A kernel two-Sample test". *Journal of Machine Learning Research,* ***13***(1), pp. 723–773.

[25] Lloyd, J. R., and Ghahramani, Z., 2015. "Statistical Model Criticism using Kernel Two Sample Tests". In Advances in Neural Information Processing Systems, pp. 829–837.

[26] Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J., 2008. "A kernel statistical test of independence". In Neural Information Processing Systems, pp. 585–592.

[27] Gretton, A., Sriperumbudur, B., Sejdinovic, D., Strathmann, H., and Pontil, M., 2012. "Optimal kernel choice for large-scale two-sample tests". In Neural Information Processing Systems, pp. 1205–1213.

[28] Chwialkowski, K., Strathmann, H., and Gretton, A., 2016. "A Kernel Test of Goodness of Fit". In International Conference on Machine Learning, Vol. 48.

[29] Barrio, E. d., Gine, E., and Matran, C., 1999. "Central Limit Theorems for the Wasserstein Distance between the Empirical and the True Distributions". *The Annals of Probability,* ***27***(2), pp. 1009–1071.

[30] Gibbs, A. L., and Su, F. E., 2002. "On choosing and bounding probability metrics". *International Statistical Review,* ***70***(3), pp. 419–435.

[31] Ramdas, A., Trillos, N., and Cuturi, M., 2017. "On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests". *Entropy,* ***19***(2), jan, p. 47.

[32] Aggarwal, C. C., Hinneburg, A., and Keim, D. A., 2001. "On the Surprising Behavior of Distance Metrics

in High Dimensional Space". In *International Conference on Database Theory*. pp. 420–434.

[33] Andrianakis, I., Vernon, I. R., McCreesh, N., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G., 2015. "Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda". *PLoS Computational Biology,* ***11***(1), p. e1003968.

[34] Gardner, P., Lord, C., and Barthorpe, R. J., 2019. "Bayesian History Matching for Forward Model-Driven Structural Health Monitoring". In *Proceedings of IMAC XXXVI*. pp. 175–183.

[35] Gardner, P., Rogers, T. J., Lord, C., and Barthorpe, R. J., 2019. "Learning of model discrepancy for structural dynamics applications using Bayesian history matching". *Journal of Physics: Conference Series,* ***1264***, p. 012052.

[36] Gardner, P., 2019. "On novel approaches to model-based structural health monitoring". PhD thesis, University of Sheffield.