2012 NDIA GROUND VEHICLE SYSTEMS ENGINEERING AND TECHNOLOGY SYMPOSIUM
MODELING & SIMULATION, TESTING AND VALIDATION (MSTV) MINI-SYMPOSIUM
AUGUST 14-16, MICHIGAN

MODEL VALIDATION FOR SIMULATIONS OF VEHICLE SYSTEMS

Hao Pan, Michael Kokkolaras, Gregory Hulbert
Department of Mechanical Engineering
University of Michigan
Ann Arbor, MI

Matthew Castanier, David Lamb
US Army TARDEC
Warren, MI

ABSTRACT

This paper deals with model validation of dynamic systems (with vehicle systems being of particular interest) that have multiple time-dependent output. First, we review several validation methodologies that have been reported in the literature: graphical comparison, feature-based techniques, PDF/CDF based techniques, Bayesian posterior estimation, classical hypothesis testing and Bayesian hypothesis testing. We discuss their advantages and disadvantages in terms of several attributes: applicability to different types of models, need for assumptions, computational cost, subjectivity, propensity to type-I or II errors, and others. We then proceed with the most important attribute: can the validation method provide a quantitative measure of the goodness of the model? We conclude that Bayesian-based model validation frameworks answer this question positively. A bootstrap method is presented that obviates the need to assume a statistical distribution model. The features of the Bayesian validation framework are illustrated using a thermal benchmark problem developed by Sandia National Laboratories and a battery model developed in the Automotive Research Center, a US Army Center of Excellence for modeling and simulation of ground vehicle systems.

1. INTRODUCTION

Modeling and simulation are indispensable tools in engineering design and development, in general, and vehicle systems, in particular. However, the efficacy of this computer-aided engineering paradigm depends largely on the validity of the utilized models. Verification, validation and accreditation (VV&A) deal with various aspects of this challenging issue. In brief, verification asks the question of whether the mathematical model is being solved correctly; validation concerns the question of whether a model (assuming that it is being solved correctly) is an adequate representation of the "real" physical system at hand; accreditation provides certification for a model to be exercised within a well-defined scope.

In this paper, we consider the challenge of model validation. Typically, model validation entails the comparison of numerical predictions (CAE data) to experimental data (test data). Clearly, validation is a highly contextual process; e.g., a low-fidelity model may be adequate for a specific application, while even a high-fidelity model may fail to capture nuances of natural phenomena. In addition, the decision of whether a model is "good enough" is almost always subjective as it is based on human perceptions and knowledge that may be incomplete. Moreover, the nature of the system being modeled and the type of model output considered can vary significantly. In this regard, there does not seem to be a "silver bullet" approach to model validation.

This paper deals with model validation of dynamic systems (with vehicle systems being of particular interest) that have multiple time-dependent output. The remainder of this section provides a listing of attributes that are desirable for validation methodologies, followed by our classification of existing validation methodologies, along with their brief descriptions.

1.1 Attributes of Validation Techniques

We conducted an extensive literature review to identify attributes that validation techniques should possess. Over fifty of the most relevant publications are cited in this paper.

To our knowledge, this is the first such tabulation of desirable validation attributes.

Validation techniques may be applied across a wide range of engineering systems. We identify the following attributes that should be considered when assessing the utility of any validation technique:

Applicable to scalar data: the suitability of a validation technique to be applied for comparing scalars. A scalar is a single numerical quantity observed/calculated in one or multiple repeated experiments/computations.

Applicable to vector data: the suitability of a validation technique to be applied for comparing vectors. A vector is a finite collection of scalars.

Applicable to scalar time series: the suitability of a validation technique to be applied for comparing scalar time series, comprising a sequence of scalars recorded at successive time points. Unlike scalar and vector data, time series data often have serial dependence, in which there is statistical dependence between a value observed at time point t_i and the value observed at another time point t_j.

Applicable to vector time series: the suitability of a validation technique to be applied for comparing vector time series which are a sequence of vectors recorded at successive time points. Vector time series can be considered as a collection of multiple scalar time series; consequently, they too often have serial dependence.

Consider multivariate correlation: the ability of a validation technique to use the correlation information of multivariate data. Although a validation technique suitable only for univariate data could be applied to each response of the multivariate data, the validation results for each response might be in conflict.

Include objective criteria: the status of a validation technique to have objective criteria to accept/reject a model. An objective criterion is developed based on mathematical or statistical reasoning.

Quantify model confidence: the ability of a validation technique to provide a quantitative assessment of the validity of the model in terms of model confidence. For example, in hypothesis testing, the null hypothesis is set up to support the fact that the computer model is accurate. Model confidence is the probability of this null hypothesis being true.

Incorporate SME opinions: the ability of a validation technique to utilize information provided by Subject Matter Experts (SME) in the process of validating a computer model.

Normality assumption independence: the independence of a validation technique on the use of normality assumption for the distribution of either test data or CAE data. More generally, it is desirable that a validation technique does not require any particular distribution model.

Insensitivity to type-I error: the insensitivity of validation results to the type-I error level specified for classical hypothesis testing validation techniques. Type-I error level, or the rate of type-I error, is the probability of rejecting the null hypothesis when it is true. It is known that specifying the type-I error at different values can lead to different validation results (i.e. from accept to reject the model) [1].

Low computation cost: the time needed to execute the validation technique.

Sample size independence: the insensitivity of the validation results to the selection of sample size. Sample size is the number of observations in a sample which is a subset of the population. Validation results should be similar if data of different sample sizes are used.

1.2 Categorization of Validation Techniques

Figure 1 depicts the classification of validation techniques that we consider in this paper.

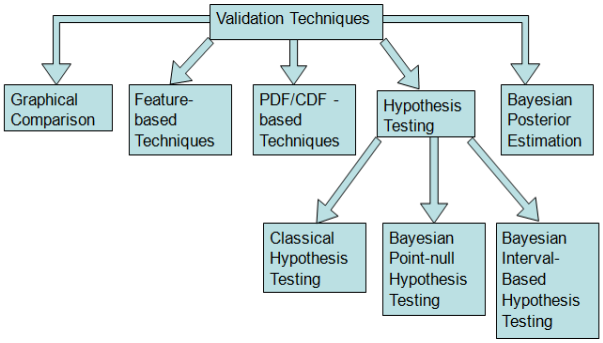

Figure 1: Categorization of validation techniques

Graphical comparison: validation techniques that generate validation results from the plot of test data and CAE data. An intuitive approach is to plot experimental measurements and simulation outputs on the same graph. One decides whether or not to accept the model by inspecting the difference between the two data. No quantitative measure of the difference between the two quantities compared is involved. In [2] the authors superimposed the computer-simulated deformation curve onto the experimental curve image taken by a high speed camera, qualitatively compared the shape of the curves and stated that the two curves have good correspondence. In [3] the authors plotted the test data as x-coordinates, and CAE data as y-coordinates. If the two data agree with each other, the collection of all the data points plotted should form a line of a unit slope (base line). Error bounds are formed by drawing two lines parallel to the base line. If two computer models are compared using this plot, one model would be preferred if considerably fewer points are outside the error bounds. Similar examples of graphical techniques can be found in [4, 5]. Graphical

comparison may be subjected to reader misinterpretation because of unknown underlying data structure [6], and can be biased and subjective [7].

The application of graphical comparison is limited to scalar and univariate time series as it cannot handle the correlation structure, although it can be applied to each response of multivariate data. Graphical comparison lacks objective rejection criteria as it is often based on subjective judgment (past experience, SME opinions, etc). Additionally, it does not quantify the model. SME opinions can be coupled with graphical comparison. For example, the acceptance region on the graph can be set up based on inputs from SME's. There are neither issues associated to type-I error nor to sample size since graphical comparison is not based on hypothesis testing. Computational cost is minimal. Graphical comparison is best used as a supplementary tool together with other validation techniques.

Feature-based techniques: validation techniques that draw validation conclusion based on the difference between features, e.g., magnitude, shape and phase of a scalar time series. Several magnitude-only error metrics such as mean absolute error (MAE) and the root mean square error (RMSE) are discussed in [6].

The Sprague and Geers metric (SG) [8], Knowles and Geer's metric (KG) [9] and Russell's metric (R) [10] are similar metrics that address the assessment of magnitude and phase error simultaneously. The EARTH metric [11] evaluates three features: phase, magnitude and topology, where topology is the shape or slope of a scalar time series. Discrepancy in phase (both global and local timing error) is removed by shifting the time history before analyzing the magnitude and topology errors. Local timing error is taken care of by the use of dynamic time warping (DTW). Unlike KG, SG and R metrics, there is no comprehensive form of the EARTH metric (i.e. a single number that summarizes all the validation results for different features). When evaluations from subject matter experts (SME) are available, a regression is performed to generate comparable ratings.

Feature-based techniques do not require a distribution assumption. Their application is limited to scalar and scalar time series as they cannot handle the correlation structure of a vector or vector time series. This limitation can be removed by the use of dimensionality and correlation reduction techniques. Feature-based techniques lack objective rejection criteria. Model confidence is not quantified. SME opinions can be incorporated (see [11] for an example of building regression-based validation models using SME opinions). There are neither issues associated to type-I error nor to sample size since feature-based techniques are not based on hypothesis testing. Computational cost is low.

PDF/CDF-based techniques: validation techniques that draw validation conclusions based on the distance between the probability density function/cumulative density function of test data and CAE data. Non-deterministic test data and CAE data are considered as random variables. In [12] the authors examined whether or not the deterministic scalar test data are within the highest density region (HDR) of the PDF of the CAE data. In [13] the authors developed a maximum horizontal distance between the two CDF's. The selection of a rejection criterion is subjective. Similarly, the Kolmogorov-Smirnov statistic measures the vertical distance between the two CDF's. If, however, the data have a very small variability (almost deterministic), the vertical distance could be very large even though the two CDFs are very close to each other horizontally.

Another measure of the distance between CDF's was developed in [14], where the area between the two CDF's was suggested as a validation metric. It was argued that the area metric enjoys several advantages such as ease of interpretation, objectiveness and ability to express validation results in terms of physical units. The CDF of the CAE data is assumed to be known. The authors suggested that this CDF be obtained by solving the mathematical model analytically or by propagating a large number of replicate samples via Monte-Carlo simulation. The test data, on the other hand, is usually provided as a collection of point values in a data set. The empirical cumulative distribution function (ECDF) was used to describe the distribution of the test data. The authors illustrated that this area metric is better than those based solely on the mean or/and variance of the data as it was able to detect the difference when the mean and variance of observations are matched but the distribution isn't. When applied to scalar time series data, the u-pooling method was developed to pool all the observations together and use statistical tests (e.g. Kolmogorov-Smirnov test) to evaluate the accuracy of the model since the pooled points should form a uniform distribution if test data match CAE data. The threshold value was not provided since the authors consider it as the task of decision makers. In the u-pooling method the CAE data distribution is assumed to be known but in practical this is often not the case.

In [15] the author proposed a discretized version of the area metric and gave the flexibility to reflect what portion of the ECDF to be emphasized for comparison. In [16] the authors used the Anderson-Darling test statistic as a measure of the discrepancy between two CDFs. The Anderson-Darling test uses a weighted quadratic ECDF statistic to measure the distance between the two CDF's and penalizes heavily deviations from the tail portion of the CDF. It was shown that the Anderson-Darling test has more statistical power than the Kolmogorov-Smirnov test [17].

PDF/CDF-based techniques do not require a distribution assumption. Their application is limited to scalars. The only implementation for scalar time series is the use of the u-pooling technique developed by [14]. PDF/CDF-based

techniques cannot handle the correlation structure of multivariate data. Some of the PDF/CDF-based techniques have objective rejection criteria but require the PDF/CDF of the experimental SRQ to be known. Model confidence is not quantified by PDF/CDF based techniques as only a measure of the distance between the two PDF's/CDF's is calculated. SME opinions can be incorporated to reveal the distribution of either test data or CAE data. Issues related to type-I error do not exist since PDF/CDF-based techniques are not based on hypothesis testing. Computational cost is negligible.

Bayesian posterior estimation techniques: validation techniques that estimate the posterior distribution of test data and CAE data using the Bayes theorem. Bayesian posterior estimation techniques can be considered as a combination of feature-based techniques and PDF/CDF-based techniques, in that a bias function is used to quantify the discrepancy in the magnitudes of test data and CAE data, and a Gaussian process is implemented to handle non-deterministic data. These techniques can be traced to [18], where a Gaussian Process was used to model the test data and CAE data (scalar time series) and the posterior parameters in the Gaussian process were inferred using Bayes' theorem. The authors suggested performing normality transformations if the data is not normal.

Bayarri et al. (see [19]) developed tolerance bounds for model predictions. Their perspective of validation is not simply to provide answer (yes/no) to the question whether to accept the computer model, but rather, to evaluate the accuracy of computer model prediction (CAE data) for the intended use.

Higdon (see [20]) developed posteriors based on non-normal priors of parameters of the Gaussian process model. Chen et al. [21, 22] developed posteriors for both model bias and output using a more flexible beta distribution prior. Tolerance bounds were developed for validation purposes. The traditional criterion for validation is that the model is accepted if the interval of the model bias contains zero or if the interval of the true value of the system response quantity contains the computer model output. This criterion can be problematic since it tends to reject the computer model at regions with many physical observations (and thus prediction intervals are narrow) but fails to reject the computer model at regions with few or no physical observations (and thus prediction intervals are wide).

Bayesian posterior estimation techniques are dependent on a normality assumption since a Gaussian process model is used. Sample size has a significant effect on the width of tolerance bounds. The technique is limited to scalar time series. Bayesian posterior estimation techniques do not have objective rejection criteria. Model confidence can be quantified. SME opinions are incorporated in terms of prior distributions of the parameters of the Gaussian process model. Bayesian posterior estimation techniques are not subject to issues related to type-I error since they are not based on hypothesis testing. Computational cost is high due to the use of the Gaussian process, MCMC and MLE.

Classical hypothesis testing techniques: validation techniques that employ a defined hypothesis to evaluate. For non-deterministic scalar data, the t-test is used to assess the similarity between the means of test data and CAE data [6, 23, 24], and the F-test to assess the similarity between the variances [6, 23, 24]. Extension to vector data can be achieved by using Hotelling's T^2-test for comparing multivariate means [25, 26], and Wilk's $\boldsymbol{\Lambda}$-distribution for comparing covariance matrices [26, 27]. Multivariate hypothesis tests (hypothesis test that is designed for vector) limit the inflation of type-I error present in multiple univariate tests (hypothesis test that is designed for scalars) [28]. Normality is assumed for both the test data and CAE data in all these hypothesis tests [23]. When this assumption is not valid, transformation to normality is suggested [24]. Alternatively, the bootstrap method was suggested to estimate the distribution of data [26]. In [24] the authors suggested to use univariate and multivariate tests collectively. The univariate tests can yield conflicting validation results but can identify which response in the multivariate data is most suspect. Multivariate tests, on the other hand, take into account the correlation structure.

A method closely related to Hotelling's T^2-test is the r^2 method developed by [29] (referred to as Mahalanobis distance later). The r^2 method assumes normality and the r^2 statistic follows a χ^2 distribution. The critical value is determined as the cumulative probability of a χ^2 random variable greater than the given significance level. The computer model is rejected if the probability of r^2 being greater than the critical value is less than the significance level. The r^2 method is applicable for both scalar and vector data and takes into account uncertainty in the model parameter. This method was further developed by formulating confidence intervals for the r^2 statistic [30]. It was extended to non-normal data by the use of the maximum likelihood estimation (MLE) [31]. The rejection criteria can be determined by Monte Carlo simulation.

Classical hypothesis testing techniques depend on a normality assumption except for the modified r^2 method in [31]. Classical hypothesis testing techniques are of the point-null hypothesis testing type and validation results are affected by sample size [28]. Application to time series is not appropriate because of the serial dependence. Classical hypothesis testing techniques have objective rejection criteria. Model confidence is not quantified because classical hypothesis testing techniques only judge whether a computer model is accurate. SME opinions are not currently incorporated but can be useful for determining the distribution used in the modified r^2 method [31]. Classical

univariate hypothesis testing is subject to accumulation of type-I error when applied to each response of multivariate data. The choice of significance level has a substantial effect on validation results. Computational cost is low except for the r^2 method. Classical hypothesis testing technique is best used for validating computer model generating non-deterministic scalar or vector outputs assuming normality.

Bayesian hypothesis testing techniques: validation techniques that combine classical hypothesis testing techniques and Bayes theorem to update validation results based on available data and SME opinions. Using Bayes factor [23, 32, 33], the authors set up hypothesis testing to examine whether the Bayes factor is above or below unity [34]. Normality is no longer required but can be used to provide an explicit expression of the posterior distribution. In [35] the authors treated the Bayes factor as a random variable to address the uncertainty in model parameters. In [24] the authors transformed non-normal data to normal and showed how the transformation helps reduce the type-I error. In [36, 37] multiple data sets were considered by assuming the data in each set are independent. The overall Bayes factor is calculated by multiplying together the individual Bayes factors for each data set. In [38] the authors derived model confidence based on Bayes factor and claimed to be the first to derive explicit expression of the model confidence for Bayesian point-null hypothesis testing.

A comparison between point-null and interval based hypothesis testing was made in [16, 39]; it was shown that the chance of rejecting a correct model increases as the sample size increases for point-null hypothesis testing.

To have more consistent results, a Bayesian interval-based hypothesis testing method (BIH) was proposed [38]. Bayesian hypothesis testing techniques were demonstrated to be superior to classical hypothesis testing because both hypotheses (null and alternative) are considered simultaneously [35]. Similarly, it was shown that the p-value used in classical hypothesis testing can engender misleading results [40].

Bayesian hypothesis testing techniques are not dependent on a normality assumption although the selection of a non-normal distribution may increase the computational cost. Sample size does not have a significant effect on Bayesian interval-based hypothesis testing. Bayesian hypothesis testing techniques have objective rejection criteria based on model confidence. SME opinions are incorporated to determine parameters used in the prior distribution of the test statistic. Bayesian hypothesis testing techniques are not subject to issues related to type-I error. Computational cost is modest, although not as low as the previously described methods.

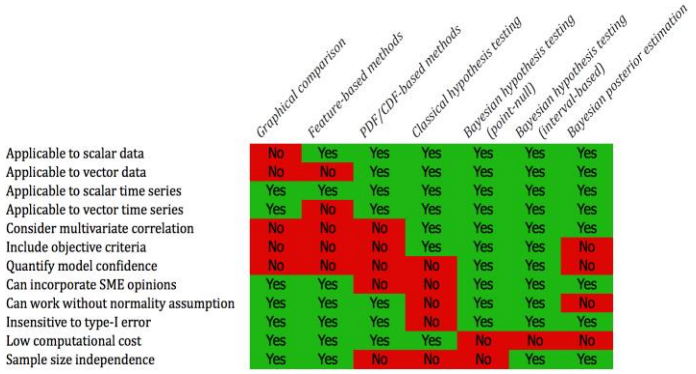

Figure 2: Attributes of validation techniques

Figure 2 summarizes the above validation techniques with respect to the validation attributes presented in Section 1.1, in which a "Yes" indicates the validation technique does possess the corresponding attribute.

2. METHODOLOGY

Dimensionality reduction techniques are used commonly for multivariate. In the context of validation, Principal Component Analysis (PCA) was coupled with the r^2 method [40], and with Hotelling's T^2-test [12]. However PCA lacks the ability to deal with non-deterministic data. BIH was coupled with Probabilistic Principle Component Analysis (PPCA) to remove correlation of data, reduce dimensionality and handle non-deterministic data [41-43]. This is the basis of the Bayesian validation framework whose process is shown in Figure 3.

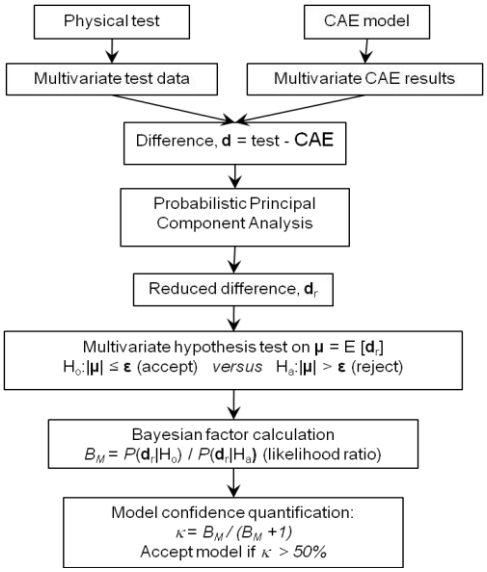

Figure 3: Bayesian interval-based hypothesis testing coupled with PPCA

First multivariate test and CAE data are obtained from experiments and simulations. PPCA is applied to the difference between test and CAE data to obtain the reduced difference. The PPCA transformation matrix is a function of the eigenvalues and eigenvectors of the covariance matrix of the difference data. A latent variable model is established to relate the difference data (observed) to a corresponding vector of latent (unobserved) variables. The reduced difference is the expectation of the latent variable. The dimensionality reduction is achieved by retaining only a few of the largest eigenvalues so that the resulting reduced difference data represent at least 95% of the variability information in the difference data.

After PPCA, the reduced difference data is uncorrelated. As a result, various validation techniques can be considered that are only suitable for univariate data (scalar or scalar time series). The Bayesian hypothesis testing technique is selected here as it is the only technique that produces model confidence which provides a quantitative assessment of the goodness of the model.

Bayesian interval-based hypothesis testing is performed on the reduced difference data. The test examines whether the expected value of the reduced difference is within the integration bounds of the integral of Eq. (2.1). The null hypothesis is that the expected reduced difference is within the integration bounds (accept the model). The prior distribution of the expected reduced difference is assumed to be Gaussian. Its posterior is obtained by applying Bayes' rule to update the prior using the observed data (reduced difference) and a Gaussian model with mean vector $\boldsymbol{\rho}_0$ and covariance $\boldsymbol{\Lambda}_0$. The model confidence is calculated as:

$$\kappa = \int_{-\boldsymbol{\varepsilon}}^{\boldsymbol{\varepsilon}} \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Lambda}_0|}} exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\rho}_0)^T \boldsymbol{\Lambda}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\rho}_0)\right) d\boldsymbol{\mu} \tag{2.1}$$

The model confidence is the probability that the expected reduced difference falls in the integration bounds with respect to its posterior probability density function.

2.1 Integration Bounds

Model confidence was shown to be sensitive to the selection of the integration bounds [41]. Here two methods of selecting the integration bounds will be explored.

Norm-based integration bounds: As illustrated in Figure 4, error bounds $[-\mathbf{e}, \mathbf{e}]$ are symmetrically set up around the test data defined as the maximum allowable deviation from the data:

$$\mathbf{e} = b\|\mathbf{t}\|_\infty \tag{2.2}$$

where $\|\cdot\|_\infty$ denotes the infinity norm or maximum norm of the test data and $\mathbf{e} \in \mathbb{R}^{m\times 1}$; $\mathbf{t}$ is the test data, $\mathbf{t} \in \mathbb{R}^{m\times n}$, and m is the number of responses and n the number of observations of each response.

The magnitude of $\mathbf{e}$ is chosen to be some fraction, b, of the L_∞ norm of the test data based on intended engineering applications or SME opinion.

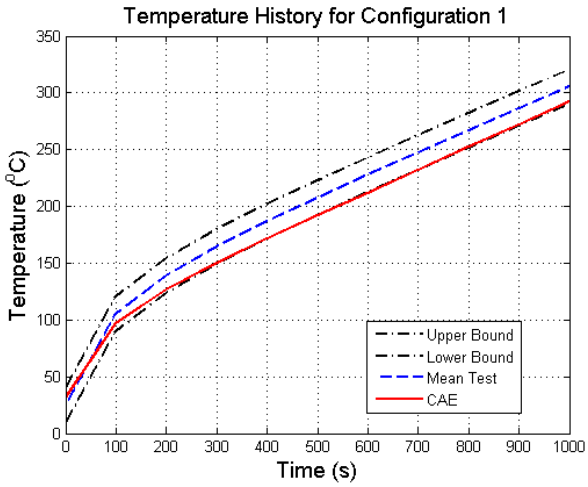

Figure 4: Example of norm-based integration bounds

The magnitude of the integration bounds used in the calculation of model confidence is calculated using:

$$\boldsymbol{\varepsilon} = abs(\mathbf{M}^{-1}\widehat{\mathbf{W}}^{\mathrm{T}}\mathbf{e}) \tag{2.3}$$

where $abs(\cdot)$ returns the absolute value. The matrix product $\mathbf{M}^{-1}\widehat{\mathbf{W}}^{\mathrm{T}}$ is the same transformation matrix applied to the difference data to obtain the reduced difference in the PPCA transformation.

Variability-based integration bounds: Following the procedure outlined in [41], the integration bounds magnitude is calculated as a fraction of the standard deviations of the reduced test data:

$$\boldsymbol{\varepsilon} = b\sqrt{diag(\boldsymbol{\Sigma}_t)} \tag{2.4}$$

where $diag(\cdot)$ returns the diagonal components of a matrix as a vector, and b is determined iteratively by considering only the covariance of the reduced test data in Eq. 2.1. $\boldsymbol{\Sigma}_t$ is the uncertainty associated with the test data.

2.2 Bootstrapping

In the methodology described in Section 2, it is assumed that the reduced difference follows a multivariate normal distribution. Oftentimes, this assumption may not

necessarily hold. To remedy this, a bootstrap-based technique was developed as an alternative approach to calculate model confidence without relying on distributions for the error model.

The bootstrap method was introduced by Bradley Efron [44] in the 1980s; the primary objective was to calculate confidence intervals for parameters in situations where standard methods were not applicable [45]. For example, asymptotic results are unacceptably inaccurate when the number of observations is small. Since its invention, the bootstrap method has been applied to many engineering fields such as geophysics, biomedical engineering, image processing, environmental engineering, artificial neural networks, etc.

The bootstrap-based technique developed for the research presented in this paper is illustrated in Figure 5.

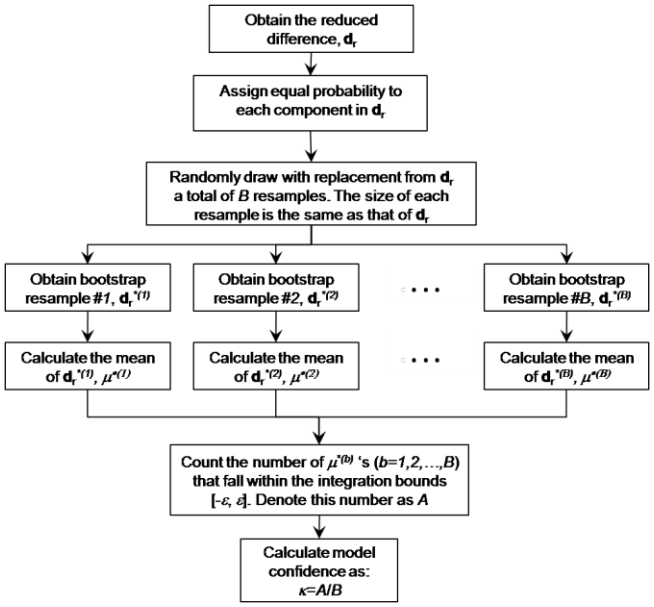

Figure 5: Bootstrapping technique

In most practical applications, the number of resamples B that need to be drawn should be of the order of a thousand [46]. More detailed guidelines on choosing B are provided in [47]. The bootstrap method employed here is of the non-parametric type; however, parametric bootstrapping will also be considered in future research since it is noted in [46] that parametric bootstrap methods can be more accurate than non-parametric ones when the sample size is small. In addition, the i.i.d. assumption for the samples is arguable; therefore, we will also consider bootstrap methods designed for dependent data, e.g., moving-block bootstrapping [46].

3. THERMAL BENCHMARK PROBLEM

In this section, we illustrate the presented Bayesian methodology for quantifying model confidence using a benchmark validation problem from the literature. Specifically, a thermal benchmark problem [48] was developed for a model validation challenge workshop held at Sandia National Laboratories in 2006. The computational model to be validated is a one-dimensional heat conduction model that predicts temperature for a material layer of thickness L subject to a specific heat flux q (Figure 6).

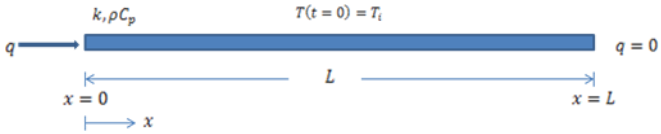

Figure 6: Schematic of the heat conduction problem [48]

The boundary conditions are specified flux q on the $x = 0$ face and adiabatic on the $x = L$ face. The computational model for temperature prediction is given by:

$$T(x,t) = T_i + \frac{qL}{k}\left[\left(\frac{k}{\rho C_p L^2}\right)t + \frac{1}{3} - \frac{x}{L} + \frac{1}{2}\left(\frac{x}{L}\right)^2 - \frac{2}{\pi^2}\sum_{n=1}^{6}\frac{1}{n^2}e^{-n^2\pi^2\left(\frac{k}{\rho C_p L^2}\right)t}\cos\left(\frac{n\pi x}{L}\right)\right] \quad (3.1)$$

The thermal properties k and ρC_p, and the initial condition for temperature T_i are prescribed constants.

Four replicate experiments were conducted for each of four configurations (combinations) of thickness L, and heat flux magnitude q, on the $x = L$ face (two levels for each variable) to obtain test data. The values of q and L in each configuration are given in Table 1.

Table 1: Values of q and L in each configuration [48]

Configuration	Heat flux, q (W/m^2)	Thickness, L (cm)
1	1000	1.27
2	1000	2.54
3	2000	1.27
4	2000	2.54

All the experimental data are provided in [48]. It is assumed that there is no measurement error. Graphical comparison of test data to CAE data is shown in Figures 7-10. The error bars indicate the maximum and minimum values of the four replicate experiments.

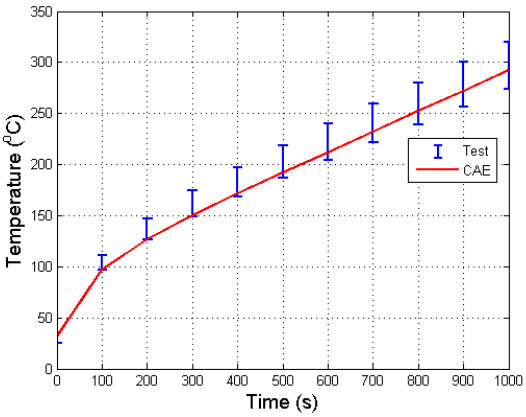

Figure 7: Graphical comparison of test and CAE data for configuration 1

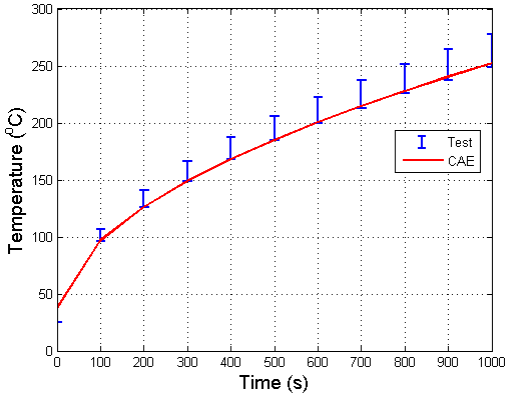

Figure 8: Graphical comparison of test and CAE data for configuration 2

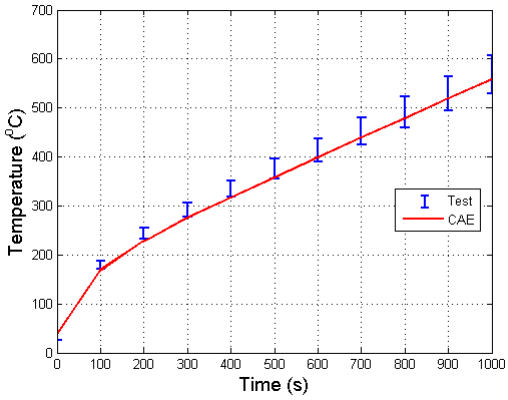

Figure 9: Graphical comparison of test and CAE data for configuration 3

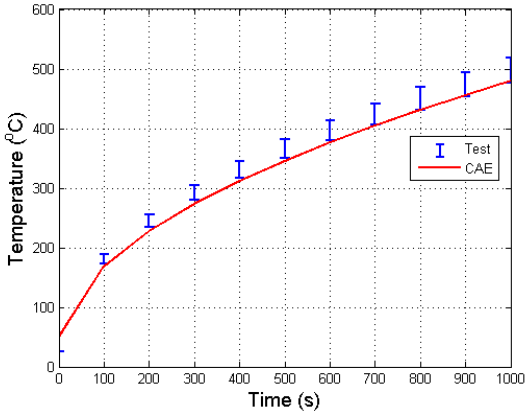

Figure 10: Graphical comparison of test and CAE data for configuration 4

3.1 Validation Results

There are 7 published solutions to the benchmark problem [49-55]. Each of these approaches fall under one of the categories presented in Section 1.2 (see Figure 11). All of these approaches yield qualitative assessments, as summarized in the authors' own words in Table 2.

We calculated model confidence for four variations of the presented Bayesian validation framework; results are presented in Figure 12. 'Norm-based Bayesian' refers to the method that employs the norm-based integration bounds introduced in Section 2.1, and calculates the model confidence using Bayes factor. 'Norm-based bootstrap' refers to the method that employs the same norm-based integration bounds, but calculates the model confidence using the bootstrap technique presented in Section 2.2. 'Variability-based Bayesian' and 'Variability-based bootstrap' differ from 'Norm-based Bayesian' and 'Norm-based bootstrap', respectively, only in the fact that the variability-based integration bounds were used instead of the norm-based integration bounds.

While there is a small variation in the results, it can be concluded that for this benchmark problem i) the normality assumption made in the Bayesian calculation does not have a significant impact on model confidence quantification; ii) validation results are relatively insensitive to the technique for determining integration bounds; and iii) the computational model can be accepted as adequate representation of reality since confidence is well above 50%.

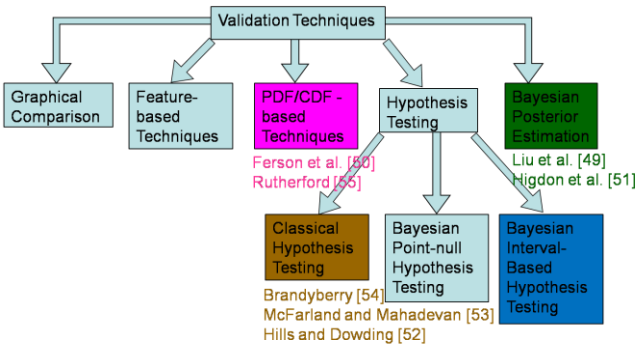

Figure 11: Categorization of solutions to the thermal benchmark problem

Table 2: Summary of validation results from the published solutions to the thermal benchmark problem

Liu *et al.* [49]	"Negligible bias"
Ferson *et al.* [50]	"Mismatch"
Higdon *et al.* [51]	"Small discrepancy"
Hills and Dowding [52]	"Poor"
McFarland and Mahadevan [53]	"Valid"
Brandyberry [54]	"Equivalent means"
Rutherford [55]	"Inadequate"

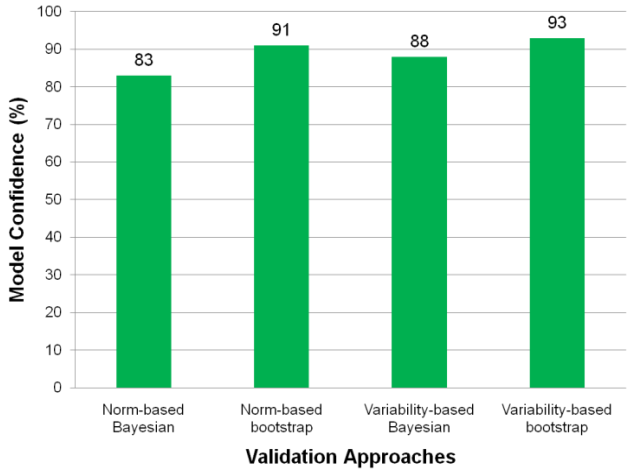

Figure 12: Comparison of validation results

3.2 Statistical Power

Statistical power is a useful tool to assess the robustness of the computed model confidence. Formally, it is defined as the probability that the hypothesis test procedure will reject the null hypothesis when it is false (i.e., the probability of not committing a type-II error) [56].

For the thermal benchmark problem, the statistical power of the Bayesian validation framework is significantly higher than that of classical hypothesis testing (79% vs. 11%). Bayesian hypothesis testing supports the null hypothesis directly by providing the probability of it being true, while the classical hypothesis testing does so by concluding that there is not sufficient evidence to reject the null hypothesis.

The factors that influence statistical power are the type of hypothesis testing used, sample size and the distance between the test statistic (the expected reduced difference) and the integration bounds bound. Statistical power is low if the integration bounds are set to be too narrow, or the sample size is not large enough. Guidelines can be established to choose the ideal sample size or the integration bounds to achieve a certain level of statistical power.

4. ONGOING AND FUTURE WORK

The research presented in this paper is supporting the activities of a tri-service Energy/Power Community of Interest (E/P CoI) for providing best practice guidelines for model sharing and verification and validation. Current members of this CoI include the Air Force Research Laboratory, the U.S. Army Tank Automotive Research, Development and Engineering Center, the Navy Surface Warfare Center, Carderock Division, the Automotive Research Center at the University of Michigan and the Electric Ship Research and Development Consortium at Florida State University. A straw-man model [57] has been developed by the Air Force Research Laboratory in order to be used as a testbed for the CoI's activities, and an electro-thermal battery model, developed at the Automotive Research Center at the University of Michigan [58] has been integrated in the straw-man model (Figure 13).

As a first step towards validating the straw-man model, the Bayesian validation framework was used to quantify model confidence for the electro-thermal battery model. The model confidence is 99%, indicating good match between test data and CAE data, which is consistent with the graphical comparison shown in Figures 14 and 15. Fragments of data are presented for clarity.

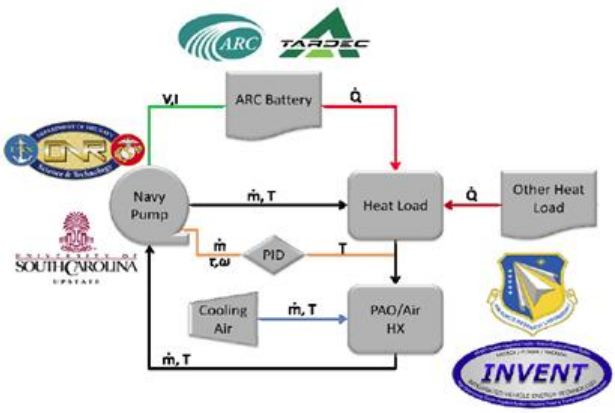

Figure 13: Straw-man model [59]

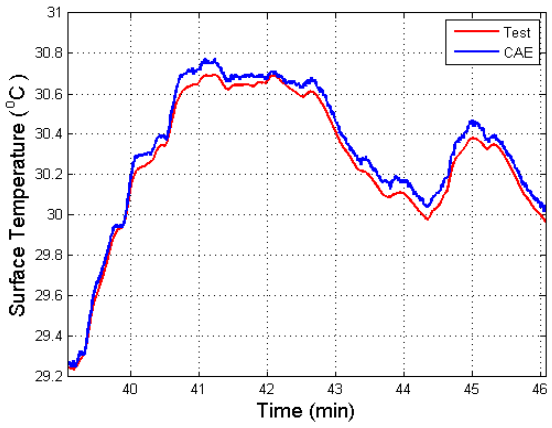

Figure 14: Graphical comparison of test and CAE data for battery surface temperature

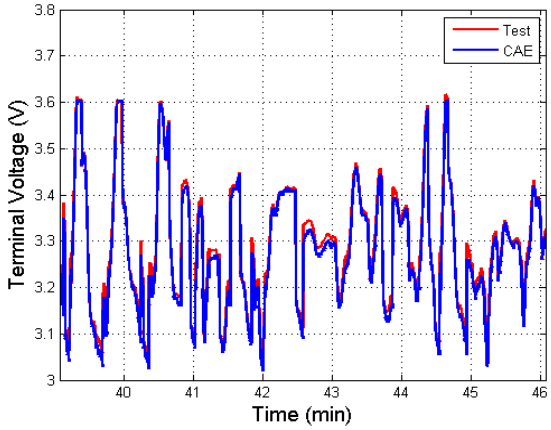

Figure 15: Graphical comparison of test and CAE data for battery terminal voltage

5. SUMMARY

There exist various validation techniques developed for different purposes and applications. However, there are no clear formal guidelines for using these techniques. Categorization of existing validation methods is thus essential to compare them systematically in order to establish suitable application domains for each category. We have presented such a categorization in this paper based on several attributes that highlight their advantages and disadvantages. The Bayesian validation framework was found to be the only validation technique that yields quantitative (as opposed to qualitative) assessment of the goodness of a model. Based on this finding, we implemented a Bayesian validation method and: i) developed alternative techniques for determining the integration bounds used for computing model confidence and ii) incorporated a bootstrap-based technique to eliminate the need to assume any distribution model for the data. We also used statistical power to assess the robustness of model confidence, showing that Bayesian hypothesis testing is superior to classical hypothesis testing.

ACKNOWLEDGEMENT

This work is supported partially by the Automotive Research Center (ARC), a U.S. Army Center of Excellence in Modeling and Simulation of Ground Vehicles led by the University of Michigan. Such support does not constitute an endorsement by the sponsors of the opinions expressed in this article.

DISCLAIMER

Reference herein to any specific commercial company, product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the Department of the Army (DoA). The opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government or the DoA, and shall not be used for advertising or product endorsement purposes.

REFERENCES

[1] T. Buranathiti, J. Cao, W. Chen, L. Baghdasaryan and Z. C. Xia, “Approaches for model validation: Methodology and illustration on a sheet metal flanging process”, AIAA Journal, 42, 2004.

[2] Y. Sugawara, K. Shinohara and N. Kobayashi, “Quantitative validation of dynamic stiffening represented by absolute nodal coordinate formulation”, in Proceedings of the ASME 2009 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, 2009.

[3] X. Qiu, D. Japikse, J. Zhao and M. R. Anderson, "Analysis and validation of a unified slip factor model for impellers at design and off-design conditions", Journal of Turbomachinery, 133, 2011.

[4] I. Arias, J. Knap, V. B. Chalivendra, S. Hong, M. Ortiz and A. J. Rosakis, "Numerical modeling and experimental validation of dynamic fracture events along weak planes", Computer Methods in Applied Mechanics and Engineering, 196:3933-3840, 2007.

[5] Y. Liu, W. Chen, P. Arendt, and H. Huang, "Experimental testing and validation of a magnetorheological (MR) damper model", Journal of Vibration and Acoustics, 131, 2009.

[6] D. G. Mayer and D. G. Butler, "Statistical validation", Ecological Modelling, 68, 1993.

[7] W. L. Oberkampf and T. G. Trucano, "Verification and validation benchmarks", Nuclear Engineering and Design, 238, 2008.

[8] M. A. Sprague and T. L. Geers, "Spectral elements and field separation for an acoustic fluid subject to cavitation", Journal of Computational Physics, 184:149-162, 2003.

[9] L. E. Schwer, "Validation metrics for response histories: perspectives and case studies", Engineering with Computers, 23:295-309, 2007.

[10] D. M. Russell, "Error measures for comparing transient data", in Proceedings of the 68th Shock and Vibration Symposium, 1997.

[11] H. Sarin, M. Kokkolaras, G. Hulbert, P. Papalambros, S. Barbat and R. J. Yang, "A comprehensive metric for comparing time histories in validation of simulation models with emphasis on vehicle safety applications", in Proceedings of the ASME 2008 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, 2009.

[12] S. Mahadevan and J. McFarland, "Error and variability characterization in structural dynamics modeling", Computer Methods in Applied Mechanics and Engineering, 197:2621-2631, 2008.

[13] I. Babuska, F. Nobile and R. Tempone, "A systematic approach to model validation based on Bayesian updates and prediction related rejection criteria", Computer Methods in Applied Mechanics and Engineering, 197:2517-2539, 2008.

[14] S. Ferson, W. L. Oberkampf and L. Ginzburg, "Model validation and predictive capability for the thermal challenge problem", Computer Methods in Applied Mechanics and Engineering, 197:2408-2430, 2008.

[15] B. M. Rutherford, "Computational modeling issues and methods for the regulatory problem in engineering - solution to the thermal problem", Computer Methods in Applied Mechanics and Engineering, 197:2480-2489, 2008.

[16] R. Rebba and S. Mahadevan, "Computational methods for model reliability assessment", Reliability Engineering & System Safety, 93:1197-1207, 2008.

[17] M. A. Stephens, "EDF statistics for goodness of fit and some comparisons", Journal of the American Statistical Association, 69:730-737, 1974.

[18] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models", Journal of the Royal Statistical Society Series B, 63, 2001.

[19] M. J. Bayarri, J. O. Berger, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C. H. Lin and J. Tu, "A framework for validation of computer models", Technometrics, 49:138-154, 2007.

[20] D. Higdon, C. Nakhleh, J. Gattiker and B. Williams, "A Bayesian calibration approach to the thermal problem", Computer Methods in Applied Mechanics and Engineering, 197:2431-2441, 2008.

[21] W. Chen, Y. Xiong, K. L. Tsui and S. A. Wang, "Design-driven validation approach using Bayesian prediction models", ASME Journal of Mechanical Design, 130, 2008.

[22] S. Wang and W. Chen, "Bayesian validation of computer models", Computer Methods in Applied Mechanics and Engineering, 198, 2009.

[23] R. Rebba and S. Mahadevan, "Model predictive capability assessment under uncertainty", AIAA Journal, 44:2376-2384, 2006.

[24] S. Mahadevan and R. Rebba, "Validation of models with multivariate output", Reliability Engineering & System Safety, 91:861-871, 2006.

[25] O. Balci and R. G. Sargent, "A methodology for cost-risk analysis in the statistical validation of simulation models", Communications of the ACM, 24:190-197, 1981.

[26] S. Mahadevan and J. McFarland, "Multivariate significance testing and model calibration under uncertainty", Computer Methods in Applied Mechanics and Engineering, 197:2467-2479, 2007.

[27] K. V. Mardia, J. T. Kent and R. F. Gunst, "Multivariate Analysis", Academic Press, London, 1979.

[28] R. Rebba, "Model validation and design under uncertainty", PhD thesis, Vanderbilt University, 2005.

[29] R. G. Hills and T. G. Trucano, "Statistical validation of engineering and scientific models: Background", Technical report, Sandia National Laboratories, 1999.

[30] J. B. Weathers, R. Luck and J. W. Weathers, "An exercise in model validation: Comparing univariate statistics and Monte Carlo-based multivariate statistics", Reliability Engineering and System Safety, 94:1695-1702, 2009.

[31] R. G. Hills and T. G. Trucano, "Statistical validation of engineering and scientific models: A maximum likelihood based metric", Technical report, Sandia National Laboratories, 2002.

[32] R. Rebba, S. Huang and S. Mahadevan, "Validation and error estimation of computational models", Reliability Engineering & System Safety, 91:1390-1397, 2006.

[33] S. Mahadevan and R. Rebba, "Validation of reliability computational models using Bayes networks", Reliability Engineering & System Safety, 87, 2005.

[34] R. E. Kass and A. Raftery, "Bayesian factors", Journal of the American Statistical Association, 90:773-795, 1995.

[35] R. Zhang and S. Mahadevan, "Bayesian methodology for reliability model acceptance", Reliability Engineering & System Safety, 80:95-103, 2003.

[36] S. Sankararaman, Y. Ling and S. Mahadevan, "Uncertainty quantification and model validation of fatigue crack growth prediction", Engineering Fracture Mechanics, 78:1487-1504, 2011.

[37] S. Sankararaman and S. Mahadevan, "Model validation under epistemic uncertainty", Reliability Engineering and System Safety, 96:1232-1241, 2011.

[38] X. Jiang and S. Mahadevan, "Bayesian validation assessment of multivariate computational models", Journal of Applied Statistics, 35:49-65, 2008.

[39] X. Jiang and S. Mahadevan, "Bayesian inference method for model validation and confidence extrapolation", Journal of Applied Statistics, 36:659-677, 2009.

[40] R. Rebba, S. Huang, Y. Liu and S. Mahadevan, "Statistical validation of simulation models", International Journal of Materials and Product Technology, 25:164-181, 2006.

[41] Y. Pai, "Investigation of Bayesian model validation framework for dynamic systems", Master's thesis, University of Michigan, 2009.

[42] X. Jiang, R.-J. Yang, S. Barbat and P. Weerappuli, " Bayesian probabilistic PCA approach for model validation of dynamic systems", SAE International 2009-01-1404.

[43] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis", Journal of the Royal Statistical Society Series B, 61:611-622, 1999.

[44] B. Efron, "Bootstrap methods, another look at the jackknife", Annals of Statistics, 7:1-26, 1979.

[45] B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation", The American Statistian, 37:36-48, 1983.

[46] A. M. Zoubir and D. R. Iskander, "Bootstrap techniques for signal processing", Cambridge University Press, 2004.

[47] P. Hall and D. M. Titterington, "The effect of simulation order on level accuracy and power of Monte Carlo tests", Journal of the Royal Statistical Society, Series B, 51:459-467, 1989.

[48] K. J. Dowding, M. Pilch and R. G. Hills, "Formulation of the Thermal Problem", Computer Methods in Applied Mechanics and Engineering 197:2385-2389, 2008.

[49] F. Liu, M. J. Bayarri, J. O. Berger, R. Paulo and J. Sacks, "Bayesian analysis of the thermal challenge problem", Computer Methods in Applied Mechanics and Engineering 197:2457-2466, 2008.

[50] S. Ferson, W. L. Oberkampf and L. Ginzburg, "Model validation and predictive capability for the thermal challenge problem", Computer Methods in Applied Mechanics and Engineering 197:2408-2430, 2008.

[51] D. Higdon, C. Nakhle, J. Gattiker, B. Williams, "A Bayesian calibration approach to the thermal problem", Computer Methods in Applied Mechanics and Engineering 197:2431-2441, 2008.

[52] R. G. Hills and K. J. Dowding, "Multivariate approach to the thermal challenge problem", Computer Methods in Applied Mechanics and Engineering 197:2442-2456, 2008.

[53] J. McFarland and S. Mahadevan, "Multivariate significance testing and model calibration under uncertainty", Computer Methods in Applied Mechanics and Engineering 197:2467-2479, 2008.

[54] M. D. Brandyberry, "Thermal problem solution using a surrogate model clustering technique", Computer Methods in Applied Mechanics and Engineering 197:2390-2407, 2008.

[55] B. M. Rutherford, "Computational modeling issues and methods for the ''regulatory problem" in engineering - Solution to the thermal problem", Computer Methods in Applied Mechanics and Engineering 197:2480-2489, 2008.

[56] P. D. Ellis, "The Essential Guide to Effect Sizes: An Introduction to Statistical Power, Meta-Analysis and the Interpretation of Research Results", Cambridge University Press, 2010.

[57] C. Miller, "Straw-man Design Description," Technical Communication, Air Force Research Laboratory, December 2012.

[58] X. Lin, H. E. Perez, J. B. Siegel, A. G. Stefanopoulou, Y. Li, R. D. Anderson, Y. Ding and M. P. Castanier, “Online Parameterization of Lumped Thermal Dynamics in Cylindrical Lithium Ion Batteries for Core Temperature Estimation and Health Monitoring”, IEEE Transactions on Control System Technology, under review