Statistical validation of simulation models

Ramesh Rebba, Shuping Huang, Yongming Liu and Sankaran Mahadevan*

Department of Civil and Environmental Engineering,
Vanderbilt University, Nashville, USA
E-mail: sankaran.mahadevan@vanderbilt.edu
*Corresponding author

Abstract: This paper investigates various statistical methodologies for validating simulation models in automotive design. Validation metrics to compare model prediction with experimental observation, when there is uncertainty in both, are developed. Two types of metrics based on Bayesian analysis and principal components analysis are proposed. The validation results are also compared with those obtained from classical hypothesis testing. A fatigue life prediction model for composite materials and a residual stress prediction model for a spot-welded joint are validated, using the proposed methodology.

Keywords: Bayesian statistics; fatigue life; hypothesis testing; PCA; validation.

Biographical notes: Ramesh Rebba is a Doctoral student of Civil Engineering at Vanderbilt University. He received Masters Degree from the same university and a Bachelor's degree from Indian Institute of Technology (IIT), Madras, India. His research interests include uncertainty analysis, computational model verification and validation, Bayesian techniques and structural reliability methods. He is currently working on a research project being sponsored by NSF-Sandia National Laboratories life cycle engineering program.

Dr. Shuping Huang is a Post-Doctoral Research Associate in Civil Engineering department at Vanderbilt University. She received her PhD degree from National University of Singapore (NUS). Her research interests include stochastic process simulation, stochastic finite element analysis and error estimation in numerical analysis. She published several journal papers and is currently working on a research project being sponsored by NSF-Sandia National Laboratories life cycle engineering program.

Yongming Liu is a Doctoral student of Civil Engineering at Vanderbilt University. He received both Bachelors' and Masters' degree from Tongji University, China. His research interests include probabilistic mechanics, fatigue and damage analysis for composite materials. He is currently working on a research project being sponsored Union Pacific railroad, developing an inspection plan for rail-wheels subject to fatigue.

Sankaran Mahadevan is Professor of Civil and Environmental Engineering and Professor of Mechanical Engineering at Vanderbilt University. He is also the director of NSF-funded Interdisciplinary Graduate Education and Research

Training (IGERT) program in risk and reliability engineering and management at Vanderbilt University. His research contributions include model-based simulation, reliability and risk assessment, optimisation and model validation. He has applied these methods to engine structures, composite structures, automotive quality, aging aircraft and aging electronics, funded by NASA, DOD, DOE, NSF, GM, DaimlerChrysler and Sandia National Laboratories. In 2002, he received the Distinguished Probabilistic Methods Educator Award from the Society of Automotive Engineers.

1 Motivation

Mathematical and computational models are extensively used to approximate the behaviour of many engineering systems, in order to facilitate cost-effective analysis and design. However, such models need to be validated with data on actual behaviour, based on testing of the real-life product or system, before they can be used with confidence for reliability prediction and reliability-based design decisions. Due to the various uncertainties in model-based prediction, as well as in actual measurement, both sets of data (prediction and observation) have uncertainty. Therefore, model validation has to properly account for these uncertainties. This paper investigates various methodologies for this purpose. The use of simple graphical comparison between prediction and observation is inadequate in the presence of various uncertainties. Therefore, this paper focuses on numerical techniques drawn from classical and Bayesian statistics, and pattern recognition methods.

Several studies are investigating the fundamental concepts and terminology for validation of large-scale computational models, such as by the ASCI (Advanced Simulation and Computing Initiative) programme of the US Department of Energy (Ang et al., 1998; Trucano, 2000), American Institute of Aeronautics and Astronautics (AIAA, 1998), American Society of Mechanical Engineers Standards Committee (ASME PTC#60) on verification and validation of computational solid mechanics, and the Defense Modeling and Simulation Office (DMSO, 1996) of the US Department of Defense. Model validation has been pursued with respect to computational fluid dynamics codes for aerospace systems (Aeschliman and Oberkampf, 1998; Oberkampf and Trucano, 2002; Roache, 2002), environmental quality modelling (Beck, 1987; Tsang, 1989) etc. An explicit treatment of uncertainty is missing or treated cursorily in these studies. Standards pointing out clear and accepted procedures for model validation under uncertainty are yet to be developed (Thacker and Huyse, 2002; Thacker, 2003). A method to validate reliability prediction models based on Bayesian hypothesis testing has recently been proposed when reliability testing is feasible (Zhang and Mahadevan, 2003). A six-step procedure based upon a Bayesian statistical methodology was proposed (Bayyari et al., 2002) and applied to problems of interest in automotive industry. Computer model outputs were modelled as response surfaces, and the Bayesian calibration/tuning/updating was performed for the models to be used in future prediction in the range beyond validated domain. Sacks et al. (2000) studied the validation of simulation models of traffic operations. Validation studies are also found in the field of structural dynamics (Anderson et al., 2000; Hasselman and Chrostowski, 1997;

Hasselman and Wathugala, 2002). All these studies focused mainly on model calibration and an explicit metric is not proposed.

Three types of validation metrics under uncertainty are investigated in this paper, based on classical statistics, Bayesian statistics, and pattern recognition methods.

- considers whether the difference between the prediction and observation is statistically significant

- explicitly considers the variability in the experimental data and compares the prior and posterior densities of the model output or its distribution parameters

- uses similarity tests based on principal components analysis for comparing groups of data.

Two numerical examples, related to composite materials durability and spot-weld residual stress modelling, are discussed in Section 3 to illustrate and compare the various methods for model validation.

2   Model validation methods

A model may be a response surface equation or a computer code (e.g., finite element analysis code). The inputs to such models like material properties, geometry, loading, etc., may have random variability and hence the model output is also a random variable. When model predictions are made at various instances of time or loading conditions, response at each instance will be a random variable. Inadequate data leads to statistical uncertainty in the input and output variables. Various simplifications in mathematical modelling, followed by approximations in numerical solution, lead to model uncertainty and error. Also, the measurements made in the laboratory have some error in them. Thus, model validation has to compare two uncertain quantities (prediction and observation), in order to decide whether to accept or reject the model prediction.

The currently used method of 'graphical validation' (i.e., by visually comparing graphs of prediction and observation either by overlays or by side-by side comparisons) is inadequate. With a qualitative graphical comparison, one does not explicitly analyse the numerical error or quantitative uncertainties in the predictions and experimental results. A scatter plot of experimental values against predicted response with a linear fit has been used for validation purposes in a simple example (Hills and Trucano, 1999). A statistical inference of model validity, based on confidence bounds, was made in this reference to evaluate whether the line, so fit, has zero intercept and a slope of one. Also, various statistical tests can be performed on the residuals for normality. Thus, this paper focuses on quantitative estimation of validation metrics.

In order to reduce testing costs, historical data may be used to validate a newly developed simulation model. The model output may or may not follow a parametric statistical distribution. A large number of samples can be generated from the distribution of the model output and these samples may be treated as the population. The finite number of existing test data, spanning the range of possible model response values, can be treated as a sample set. Then model validation may be done through standard hypothesis testing where the samples belong to population are tested.

2.1 Classical hypothesis testing

Several preliminary ideas for model validation metrics that consider uncertainty in prediction and observation have been suggested (Oberkampf and Trucano, 2002). One such metric normalises the difference between model predictions and experimental values, computes a relative error norm for discrete and continuous domain problems. Another metric considers the uncertainty in the experimental value due to limited data, and uses classical hypothesis testing (Hills and Leslie, 2003; Urbina et al., 2003). A common practice is to conduct *t*-tests to decide whether or not the sample belongs to the population.

In standard statistical tests, null (H_0) and alternative hypotheses (H_a) are usually required. A test statistic, S, computed using the data set, is used to decide whether the available evidence rejects the null hypothesis H_0. S may refer to a t-statistic for the mean or F- statistic for the variance. H_0 is rejected if S lies in a rejection region C (region outside a particular interval, for example $[-S_{crit}, S_{crit}]$). The bounds of the critical region are selected, such that the probability of making a Type I error (i.e., rejecting a null hypothesis which is actually true) is, say, 0.05 or 0.01. Thus, S_{crit} is selected, such that d and such that $P(S \in C|H_0) = 1 - P(|S| > S_{crit}|H_0) = 0.05$ or 0.01.

Suppose that the model prediction follows a statistical distribution with mean μ_0 and variance σ^2_0. These values are treated as population parameters. The data mean and variance of *n* samples are denoted by μ and σ^2. The null hypothesis could be H_0: $\mu = \mu_0$ and $\sigma^2 = \sigma^2_0$ while the alternative hypothesis is H_a: $\mu \neq \mu_0$ or $\sigma^2 \neq \sigma^2_0$. A *t*-test statistic for the mean and *F*-test statistic for the variance are calculated as follows:

$$t = \frac{\sqrt{n}\,|\mu - \mu_0|}{\sigma} \text{ and } F = \frac{(n-1)\sigma^2}{\sigma_0^2} \tag{1}$$

The null hypothesis H_0 is accepted if $t < t_{\alpha,\ n-1}$ and $F < f_{\alpha,\ n-1}$, where $t_{\alpha,\ n-1}$ is the $(1-(\alpha/2))$ percentile value of *t*-distribution with $n - 1$ degrees of freedom and $f_{\alpha, n-1}$ is the $(1-(\alpha/2))$ percentile value of *F*-distribution with $n-1$ degrees of freedom.

Alternatively, the null hypothesis can be rejected if the difference between the observation and prediction is statistically significant. This can be done using the *p*-value of the observation. The *p*-value is defined as the probability of getting a test statistic as extreme or more extreme than that observed by chance alone i.e., $P(S \geq S_{obs}|H_0)$. A small *p*-value indicates that S_{obs} is too far from zero, i.e., the difference between prediction and observation is statistically significant, and hence H_0 is rejected. The *p*-value based hypothesis testing has been argued to have some logical flaws, leading to misleading conclusions (Hwang et al., 1992). Hence Bayesian hypothesis testing has been suggested as an alternative to the classical approach. The major distinction between classical and Bayesian hypothesis testing for model validation is discussed in the following section.

2.2 Bayesian methods

Consider two models M_i and M_j. Their prior probabilities of acceptance are denoted by $P(M_i)$ and $P(M_j)$. Using Bayes' theorem, when an event/data is observed, the relative posterior probabilities of the two models are obtained as (Berger and Pericchi, 1996; Leonard and Hsu, 1999):

$$\frac{P(M_i \mid \text{data})}{P(M_j \mid \text{data})} = \left[\frac{P(\text{data} \mid M_i)}{P(\text{data} \mid M_j)} \right] \left[\frac{P(M_i)}{P(M_j)} \right]. \tag{2}$$

The term in the first set of square brackets on the right hand side is called the 'Bayes factor' (Jeffreys, 1961). If the Bayes factor is > 1.0 then it can be inferred that the data favours model M_i more than model M_j. If only a single model is proposed, the model could be either accepted or rejected. When an observation is made, the Bayes factor can be used to estimate the ratio of relative likelihood of the null hypothesis (i.e., data support the proposed model) and the alternative hypothesis (i.e., data do not support the proposed model). For example, let x_o and x be the predicted failure and true response respectively of an engineering system. The value x_o is predicted by model M. This can be considered as a point null hypothesis ($H_o: x = x_o$). To estimate the Bayes factor in equation (2), an alternative hypothesis ($H_1: x \neq x_o$) needs to be constituted.

Suppose the corresponding observed value in an experiment is y. Then the probability of observing the data under the null hypothesis, $P(y|H_o: x = x_o)$, can be obtained from the likelihood function. For a discrete distribution, the likelihood function of the parameter is the probability of observing the data given the parameter. For a continuous distribution, however, the likelihood function is (Pawitan, 2001) proportional to the density of data y given the parameter x, and for the problem under consideration, $P(y|H_o : x = x_o) = L(x_o) = \varepsilon f(y|x_o)$. Similarly, the probability of observing the data under the alternative hypothesis $P(y|H_1: x \neq x_o)$ can be obtained from $\int L(x)g(x)\mathrm{d}x$ or $\int \varepsilon f(y|x)g(x)\mathrm{d}x$, where $g(x)$ is the prior density of x under the alternative hypothesis. Since no information on $g(x)$ is available, one possibility is to assume $g(x) = f(x)$. Then, using equation (2), the Bayes factor is computed using Bayes theorem as

$$B(x_0) = \frac{P(y \mid H_0 : x = x_o)}{P(y \mid H_1 : x \neq x_o)} = \frac{L(x_o)}{\int L(x)g(x)\mathrm{d}x} = \frac{f(y \mid x_o)}{\int f(y \mid x)f(x)\mathrm{d}x} = \frac{f(x \mid y)}{f(x)}\Big|_{x=x_0}. \tag{3}$$

Thus, the Bayes factor simply becomes the ratio of posterior to prior PDFs of the predicted response, when $g(x) = f(x)$. This result helps to validate model prediction with corresponding experimental measurement. If $g(x) \neq f(x)$, then the Bayes factor is computed using equation (3) with $g(x)$, instead of $f(x)$, in the denominator. Further, since x is the true solution, x_0 is the model output and y is the observed value, the following equations hold:

$$x = x_0 + \varepsilon_{\text{pred}} \tag{4a}$$

$$x = y + \varepsilon_{\text{exp}} \tag{4b}$$

where $\varepsilon_{\text{pred}}$ is the model prediction error and ε_{exp} is the measurement error. If there is no prediction error, i.e., ($H_0: x = x_0$), the observed value will simply be $y = x_0 - \varepsilon_{\text{exp}}$. From this relation and Gaussian experimental error assumption, for example, we obtain $f(y|x_0) \sim N(x_0, \sigma_{\varepsilon \text{exp}}^2)$ where $\sigma_{\varepsilon_{\text{exp}}}^2$ is the variance of measurement error. Non-Gaussian experimental errors may also be considered here.

In Bayesian hypothesis testing (Berger and Pericchi, 1996), the statistical parameters of the model output are treated as random variables and updated using test data. Validation metrics based on the ratio of posterior and prior densities of the statistical parameters can be used to infer whether the sample belongs to a population. The model is

said to be supported by the data if the Bayes factors for all the statistical parameters are larger than one.

From the discussion in Sections 2.1 and 2.2, one can notice some similarities between Bayesian and classical hypothesis testing. However, the Bayesian approach offers some alternatives (Carlin and Louis, 2000), preferable to the classical (often called frequentist, based on the idea of long-term frequency of outcomes in imagined repeats of experiments or samples) methods, for hypothesis testing as well as for estimation and decision-making (Box and Tiao, 1992; Berger, 1985; Johnson, 1999). In model validation, the value of a parameter (say mean or variance, or shape and scale factors) is predicted from theory, and it is more reasonable to test whether or not that value is consistent with the observed data, than to calculate a confidence interval (Berger and Delampady, 1987; Zellner, 1987). For testing such hypotheses, what is usually desired is $P(H_0 \mid \text{data})$, i.e., assessing how much evidence is there for the null hypothesis or for the model prediction being right. Bayesian hypothesis testing directly addresses this question. Classical hypothesis testing answers a different question, i.e., whether or not there is evidence to reject H_0. The *p*-value is used to infer whether or not the difference between observation and prediction is statistically significant. Not having enough evidence to reject H_0 is not the same as accepting H_0. The illustrative examples in Section 3 compare the validation inferences made, using both Bayes factor and *t*-test statistics.

The Bayes factor for validating the prediction of a single quantity can be extended for validating a vector of decision variables (e.g., multiple responses at one location or one response variable at different locations). However, computation of such a validation metric is difficult for realistic problems, since it involves evaluation of high-dimensional multivariate densities. On the other hand, computing individual Bayes factor values for each response quantity may give conflicting inference about overall model validity. Therefore, an aggregate metric for validating a vector of decision variables is proposed and investigated next, in which both model prediction and experimental observation can be viewed as multivariate data, and the two data groups can be compared to make an inference about the validity of model prediction. In this method, principal components analysis (PCA) (Jolliffe, 1986; Oja, 1983; Singhal and Seborg, 2001; Timm, 2002) is first used for dimension reduction and noise filtering. After PCA is completed, validation is performed in the space of principal components (also referred to as features). Two similarity metrics (weighted average of cosine similarity, and distance similarity based on the Mahalanobis distance) are employed in the feature space, to characterise the statistical equivalence of experimental data and model prediction.

2.3 Metrics based on probabilistic PCA

Consider an $N \times k$ data matrix, X, consisting of N samples (either experimental observations or predicted values) of k variables. PCA can be described as a decomposition of the matrix X, giving a set of scores in a $N \times q$ matrix Z describing the principal components, and a set of loadings in a $k \times q$ matrix W, describing the influence on the principal components. The decomposition of X can be described by $X = ZW^T + E$, where E is an $N \times k$ error matrix. The dimensionality of the data matrix can be reduced by retaining 'q' principal components ($q < k$) with the largest eigenvalues that capture most of the variation in the data, assuming that other components capture the contaminating noise. The values of the eigenvalues decide how many components should

be included in the analysis. Probabilistic PCA (PPCA) (Tipping and Bishop, 1999) is an analogue to PCA, which incorporates information about measurement errors to develop PCA models that are optimal in a maximum likelihood sense (therefore, also referred to as MLPCA) (Wentzell et al., 1997). PPCA not only does the same operation as PCA for dimension reduction, but also accounts for variance introduced by the error E.

After noise filtering using probabilistic PCA, two similarity measures can be employed in the feature space to characterise the statistical equivalence of the experimental data and model prediction. These are

- weighted average cosines (Baeza-Yates and Ribeiro-Neto, 1999)

- distance based on the Mahalanobis distance, as described below.

The cosine of the angle between two vectors x_r and x_s is defined as

$$\cos\theta = \frac{x_r \cdot x_s}{\|x_r\|\|x_s\|}. \tag{5}$$

Consider two data sets (from simulation and physical testing) that contain the same number of variables k but not necessarily the same number of measurements. Assume that the PCA model for each data set contains q principal components, where $(q < k)$. Denoting S and T as sub-spaces calculated from PCA for the two data sets, the similarity between the two data sets is then quantified by comparing their features (principal components in the subspace). The weighted average of cosine similarity is defined, to compare the spatial orientation of the feature space of the two data sets:

$$S_\theta = \sum_{i=1}^{q} c_i \cos\theta_i \tag{6}$$

where c_i are weights for each feature calculated using the eigenvalues.

However, the distance similarity factor distinguishes two datasets that might have the same spatial orientation but are located far apart. Distance similarity is particularly useful when two datasets have similar spatial orientation, but the centre of data cloud is very different. One of such distance measures is Mahalanobis distance

$$d_{rs}^2 = (x_r - x_s)C^{-1}(x_r - x_s)' \tag{7}$$

where C is the sample covariance matrix. The Mahalanobis distance normalises the features using the covariance matrix. In the context of this paper, the Mahalanobis distance is from the centre of the simulation data to the centre of the test data, and is computed on principal component scores, rather than actual datasets (PCA reduces the data the into a smaller set of representative numbers-scores):

$$d_{ST}^2 = (\overline{x}_S - \overline{x}_T)C_T^{-1}(\overline{x}_S - \overline{x}_T)' \tag{8}$$

where C^T is the transpose of the pooled covariance matrix of features of test data and simulated data. There are several possibilities for converting such a distance metric (in $[0, \infty]$) into a similarity measure (in $[0,1]$) by a monotonic decreasing function. The distance d and similarity measure S_d are related by using (Strehl, 2002)

$$S_d = e^{-d^2} \tag{9}$$

This form of distance similarity measure has been used by Hills and Leslie (2003), but PCA was not used in that study to reduce the size of data groups. The use of PCA makes the validation process computationally more efficient.

Distance similarity is translation invariant but scale sensitive, while cosine similarity is translation sensitive but scale invariant. The two metrics are complementary to each other, and together give insights on both spatial orientation and location of the two datasets under comparison.

3 Numerical examples

The three types of model validation, based on classical hypothesis testing, Bayesian statistics and pattern recognition techniques, are illustrated in this section with the material and mechanical behaviour models.

3.1 Example 1: validation of a fatigue life prediction model for composite materials

Composite materials are widely used in automotive, naval, and aerospace structures, where they are often subjected to cyclic fatigue loading (Van Paepegem and Degrieck, 2002). Fatigue is one of the most common failure modes in all structural materials, including composite materials. Accurate fatigue life prediction needs to include the random variation in material properties (Young's modulus, strength, etc.), loading history, component geometry, and environmental conditions. A damage accumulation model for fatigue life prediction of composite laminates is used in this example. The model is constructed at the ply level and uses a new multi-axial damage index to consider the damage caused by different stress components. The material properties and geometry properties are randomised, based on the experimental data or some assumptions.

The predictive model is basically an *S–N* curve-based fatigue progressive damage model and ignores the detailed analysis of local failure. It uses fatigue data from the family of S-N curves, and uses a special damage variable to account for multi-axial fatigue in each ply. Refer Liu and Mahadevan (2005) for details of these calculations. For the purpose of the discussion, the predictive model is treated as a 'black-box' and only the model output is used for validation purposes. Monte Carlo simulation is used to calculate the distribution of fatigue life under different load amplitudes.

The analytical model predicts fatigue life, *N* cycles or log(*N*), for each stress level with amplitude S_{max}. Since the material properties of the laminate are random, the life also follows a statistical distribution at each applied stress. Fatigue properties of a glass-fibre-based composite, available from the US Department of Energy/Montana State University (DOE/MSU) Composite Materials Fatigue Database (Mandell and Samborsky, 2003), are used in this section to validate the fatigue calculations at various stress levels.

Validation of the predictive model may be performed by comparing the statistical distributions of data and the predictions for different values of S_{max}. Since there are very few samples available, the validation in this case simply involves comparing the predicted and observed mean life. Table 1 shows the fatigue test data for three samples of laminates, conducted at four different stress levels.

Table 1   Fatigue test data

| S_{max} (MPa)	No. of tests	Life log(N)
41	3	5.887, 6.091, 6.11
48	3	5.139, 5.419, 5.507
55.2	3	4.586, 4.712, 4.799
69	3	3.661, 3.86, 3.802

Fatigue life is predicted at any stress level, expressed as log(N), followed Gaussian distribution and their statistics are provided in Table 2.

Table 2   Statistics of predicted and observed life – log(N)

S_{max} (MPa)	Data mean	Pred. mean	Pred. stdev
41	6.029	6.080	0.127
48	5.355	5.357	0.133
55.2	4.699	4.809	0.146
69	3.774	3.350	0.237

As a preliminary step in validation, the predicted mean-life and available test data are first plotted to observe any interesting trends.

Even with a visual inspection of Figure 1, it is easy to identify that model prediction is somewhat close to the observation. A more rigorous quantitative estimation of model acceptance or rejection may be possible through the proposed validation methodology.

Figure 1   Test vs. mean predicted life
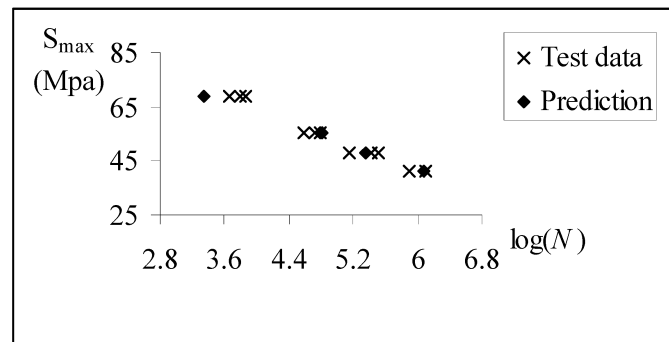
3.1.1 Classical hypothesis testing

It is a standard technique in classical hypothesis testing to make use of *t*-tests to compare the sample mean with the population mean for a known variance. Utilising that concept, test data is compared with mean predicted life for each stress level. A *t*-test conducted with 0.05 acceptance level proved that the model prediction is close to the observation at all stress levels as shown below in Table 3.

Table 3 *t*-test for mean life

S_{max} (MPa)	41	48	55.2	69
$t = \sqrt{n}|\bar{y}\text{-}\mu|/s$	0.709	0.019	1.793	3.088
$t_{0.05,\,n-1}$	4.3026	4.3026	4.3026	4.3026
Result	Pass	Pass	Pass	Pass

3.1.2 Bayesian hypothesis testing

Bayesian hypothesis testing can be conducted as proposed in Section 2.2 to validate the mean predicted life. In this demonstration problem, fatigue log-life predicted by the analytical model followed the Gaussian distribution at all stress levels. Thus, the Bayesian validation metric can be estimated at each stress level.

Suppose that the log-life for S_{max} follows normal density $N(\mu, \sigma)$, then the observed data $y = \{y_1, y_2, ..., y_n\} \sim N(\mu, \sigma)$. Since the aim is to validate the predicted mean life, μ is assumed to be an unknown random variable that follows a distribution $f(\mu) \sim N(\mu_\mu, \sigma_\mu)$. The likelihood function of μ is given by (Ang and Tang, 1975)

$$L(\mu)\alpha\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right]\alpha N_\mu\left(\bar{y},\frac{\sigma}{\sqrt{n}}\right). \tag{10}$$

Now the null hypothesis can be stated as $H_0: \mu = \mu_\mu$ and the alternate hypothesis as H_1: $\mu \neq \mu_\mu$. The posterior density for mean life can be determined analytically in this case as:

$$f(\mu \mid y) \sim N_\mu\left(\frac{\bar{y}\sigma_\mu^2+\mu_\mu(\sigma^2/n)}{\sigma_\mu^2+(\sigma^2/n)},\frac{\sigma_\mu^2(\sigma^2/n)}{\sigma_\mu^2+(\sigma^2/n)}\right). \tag{11}$$

Then the validation metric can be computed using equation (3) as:

$$B = \frac{f(\mu/y)}{f(\mu)}|_{\mu_\mu}. \tag{12}$$

The value for σ_μ is usually arbitrarily chosen and the Bayes factor is highly sensitive to the ratio σ_μ/σ. A small value of σ_μ would reject the null hypothesis (prior being too rigid) while a large value tends to accept null hypothesis (prior being too relaxing), irrespective of the fact that data is far from prediction. Hence a proper choice $\sigma_\mu = s$ (sample standard deviation) is made in this paper for most calculations to provide 'flexibility' to the mean value μ. When s^2 is smaller than σ^2/n, we can use σ^2/n for σ_μ^2.

A typical validation metric calculation is as follows: Suppose the life $\log(N)$ at a stress level 48 MPa follows $N(5.3575, 0.13323)$, according to Table 2. Validation data is available from Table 1 corresponding to $S_{max} = 48$ and the observed data is $y = \{5.1396, 5.4189, 5.5077\}$. In this case, $n = 3$, $\mu_\mu = 5.3575$, $\sigma = 0.13323$, and from Table 2, $y = 5.3554$ and $s = 0.192$. Thus, a value for the validation metric is computed as 2.688 as shown in Table 4:

Table 4   Validation inference for mean predicted life

S_{max} (MPa)	41	48	55.2	69
B	1.644	2.688	0.947	0.226

The validation metric values $B < 1$ for some stress levels in Table 4 indicate that the predictive model is not acceptable at all stress levels. Thus, the model is expected to have some prediction error (difference between the observation and prediction), which could be attributed to model form error or vast errors in the statistical distributions of the input parameters. This inference calls for model calibration or re-evaluation of the modelling process.

3.1.3 PCA-based similarity tests

PCA-based validation can be used to give an overall evaluation of the prediction model, instead of evaluating at each stress level. The life log(N) at the four stress levels are treated as four variables and 6,000 samples are simulated using the life prediction model. It is assumed that both test data and data from model simulation are contaminated by noise. The noise in the test data is due to measurement error, and is assumed to be zero-mean Gaussian, i.e., $\epsilon_x \sim N(0, Q_{\epsilon x})$. The error covariance matrix (Ψ) is assumed to be a diagonal matrix with variances of the errors. The signal-to-noise ratio is assumed to be 10. The comparison of test and prediction is done using the proposed PCA and similarity factors-based model validation procedure as described in Section 2.

The test set is a 3×4 data matrix consisting of three samples, of which four variables have been measured and serves as a template. The data set from simulation is divided into 2,000 sub-sets, which are the same size as the test data-set.

MLPCA is performed for the test data and each subset of the model simulation data. The similarity factors between each of 2,000 subsets of the model simulation data and the test data set are calculated in their feature spaces using equations (6) and (9). The approximate statistical distributions of the similarity factors (S_θ and S_d) are obtained from the 2,000 samples. The threshold for both S_θ and S_d is set to 0.9. If data from model simulation are viewed as response clouds, $S_\theta > 0.9$ implies that the orientation of the two clouds is similar, and $S_d > 0.9$ implies that the centres of gravity of the two clouds tend to overlap. Thus, the model is assumed to be valid, if $P(S_\theta > 0.9) \geq 0.95$ and $P(S_d > 0.9) \geq 0.95$. The metrics are computed and shown in Table 5. The model passed the angle similarity test but barely failed in distance test. However, the validation decision depends on the threshold. If the threshold is relaxed a little, the prediction model is able to pass both tests.

Table 5   Validation metrics using PCA techniques

Metric	P(S > 0.9)	Fail/pass
S_d	0.884	Fail
S_θ	0.9	Pass

Before closing the example, it is worth noting that the S-N approach could bring additional errors that affect the validation inference. Typically, the fatigue life tests are based on displacement control. The transformation from displacement to stress is usually

calculated using classical laminate theory and then used for fatigue life prediction. However, this transformation may introduce additional modelling error. Thus, one needs to account for errors in processing the validation test data before it is used to compare with model prediction.

3.2 Example 2: validation of a stress prediction model of a spot-welded joint

Resistance spot welding (RSW) is a widely used joining process for sheet metal assemblies, especially in the automotive and railroad industries. Due to the high stress concentration near the nugget and degradation of material properties caused by the welding process, spot welds are most critical fatigue locations of the structure. Several methods have been proposed to predict spot weld fatigue life, but few of those methods include welding residual stress, since welding stress analysis is quite complicated (Huh and Kang, 1997; Radaj et al., 1990). According to Bae et al. (2003), residual stress reduces fatigue life by as much as 25%.

The predictive model in this example uses finite element analysis (FEA) procedure to simulate the coupled multi-physics (electrical-thermal-structural) related to residual stresses. Since the geometrical properties and input electric current could be random, the residual stress would also be a random variable. Since Monte Carlo simulations over this complicated FE code could be computationally expensive, a response surface method (RSM) combined with design of experiments (DOE) is used to obtain a simplified empirical formula considering several sources of variation. For increased accuracy, a stochastic field expansion technique based on the Karhunen-Loeve method (Ghanem and Spanos, 1991) is applied to simulate the random field of the residual stress. The predictive capabilities of each of these two models need to be assessed so that a favourable model may be selected for future applications.

The FEA analysis is performed using the dimensions and materials properties reported by Henrysson (2001). Also experimental data for residual stress measured at three locations in a spot weld is used (Henrysson, 2001) to validate the predictive models. Table 6 shows the test data at locations whose distances from a reference point are normalised. The stresses are also normalised with respect to the yield stress of the material used for welding.

Table 6   Residual stress data

Normalised, distance	No. of tests	Observed normalised. Stress
0.4	4	0.75, 0.614, 0.477, 0.364
0.5	7	1.023, 0.704, 0.636, 0.545, 0.409, 0.273, 0
0.7	6	0.545, 0.5, 0.159, 0.091, –0.023, –0.432

The statistics of residual stresses obtained by two response surface models are presented for three different locations on the worksheet. The model response in all cases followed normal distribution, and hence only the mean and standard deviation are presented in Table 7.

Table 7   Statistics of predicted and observed life – log(N)

Normalised distance	Data mean	Regression model	K-L model
0.4	0.5511	0.8154	0.8177
0.5	0.5129	0.51635	0.5125
0.7	0.1402	0.1821	0.1897

A plot of test data over mean residual stress in Figure 2 qualitatively shows that the model predicts reasonably well in two of three locations far from the centre of nugget. A quantitative estimate of the model-data agreement may be possible through hypothesis testing.

Figure 2   Test vs. predicted stress at different locations
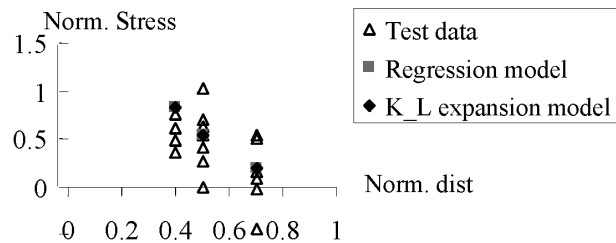
3.2.1   Classical hypothesis testing

A *t*-test conducted for the test data in Table 8 and predicted mean in Table 9 for both regression-based and KL expansion-based models showed that predictions match well with data in two out of three locations.

Table 8   *t*-test for mean residual stress using regression

Normalised distance	0.4	0.5	0.7
$t = \sqrt{n}\,|\bar{y} - \mu|\,/\,s$	3.158	0.0271	0.285
$t_{0.05,\,n-1}$	3.182	2.571	2.447
Result	Barely pass	Pass	Pass

Table 9   *t*-test for mean residual stress using KL expansion

Normalised distance	0.4	0.5	0.7
$t = \sqrt{n}\,|\bar{y} - \mu|\,/\,s$	3.1851	0.00361	0.3368
$t_{0.05,\,n-1}$	3.182	2.571	2.447
Result	Just fail	Pass	Pass

3.2.2   Bayesian hypothesis testing

Bayesian validation metrics are estimated at various locations using an analysis similar to the one described in the previous example and are presented in Table 10.

Table 10 Validation inference for residual stress

Normalised distance	0.4	0.5	0.7
B Reg	1.0E-55	1.37	2.7E-6
B KL	0.141	4.52	3.788

It is clear from Figure 2 and Tables 8 and 9 that the KL expansion based model prediction is more acceptable. It is also evident that the prediction at a location closer to the nugget edge is found to be unacceptable. Since each of these response surfaces are derived using finite element analysis, it is quite possible that the singularity in the vicinity of nugget is causing a large error in the prediction. This calls for mesh refinement in FE code and/or correcting measurements taken at the centre of spot weld during a validation test.

3.2.3 *PCA-based similarity tests*

In many real world applications, including this example, data sets with missing values are quite common. Missing values occur for several reasons and in various situations: Some measurements for a particular sample may not have been recorded or recorded wrongly so that they are deleted as erroneous, insufficient samples are available for some test region et al. In this paper, missing data is handled by using the EM (Expectation Maximisation) algorithm (Oba et al., 2003), which estimates missing values and fill them by the estimated values.

The residual stresses at the locations are treated as three variables and 7,000 samples are simulated using the prediction models (RSM and K-L). It is assumed that both test data and data from model simulation are contaminated by noise. The noise in the test data is due to measurement error, and is assumed to be zero-mean Gaussian, i.e., $\in_x \sim N(0, Q_{\in x})$. The error covariance matrix (Ψ) is assumed to be a diagonal matrix with variances of the errors. The signal-to-noise ratio is assumed to be 10. The test set is a 7×3 data matrix consisting of seven samples upon which three variables have been measured and serves as a template. The data set from simulation is divided into 1,000 sub-sets, which are the same size as the test data set.

The comparison of test and prediction is done for both RSM and K-L prediction models, using the proposed PCA and similarity factors-based model validation procedure as described in Section 2. The similarity between each of 1,000 sub-sets of the model simulation data and the test data set are calculated in their feature spaces using equations (6) and (9). The approximate statistical distributions of the similarity factors (S_θ and S_d) are obtained from the 1,000 samples. The threshold for both S_θ and S_d is set to 0.5. If data from model simulation are viewed as response clouds, $S_\theta > 0.5$ implies that the orientation of the two clouds is similar and $S_d > 0.5$ implies that the centres of gravity of the two clouds tend to overlap. Thus, the model is assumed to be valid, *if* $P(S_\theta > 0.5) \geq 0.95$ and $P(S_d > 0.5) \geq 0.95$. The threshold is set low to account for large variability in the residual stress measurement. The metrics for the K-L based model are calculated and shown in Table 11.

Table 11   Validation metrics using PCA

Metric	$P(S > 0.5)$	Fail/pass
S_d	0.953	Pass
S_θ	0.96	Pass

4   Conclusion

Statistics-based validation metrics to compare model prediction with experimental observation under uncertainty are developed in this paper. Treating model validation as equivalent to hypothesis testing, three types of metrics based on classical statistics, Bayesian statistics, and Principal Components Analysis (PCA) are investigated. Each validation approach answers a different question. Also, the first two approaches are useful in validating individual quantities or distribution parameters, whereas the third approach is useful for aggregate validation with multiple quantities. In the Bayesian approach, selecting proper priors for the distribution parameters is a challenging task. In the PCA-based approach, selection of threshold values for S_d and S_θ is also a challenge. In this paper, the validation metric based on hypothesis testing uses only model prediction and experimental data, and, thus, is only related to the overall difference between the two quantities. The difference may arise from many sources: model form error, numerical errors related to solution convergence and model resolution, stochastic analysis errors, and measurement errors in model input and system output. In some cases, the non-linear combination of these errors may result in an overall prediction error to be close to zero. The metric explicitly includes the measurement error, thus, acknowledging the validation inference only under a given purview of measurement uncertainty. Since the goal of a validation metric is to determine whether or not there is a significant overall error, we ignore the model bias issue. However, experimental bias is addressed indirectly in the metric. See the paper by Rebba et al. (2004) for more details on this topic.

A fatigue life prediction model for composite materials, and a residual stress prediction model for a spot-weld joint in an automotive structure were validated using the proposed methodology. The results of model validation could be used to improve or calibrate the model. Model improvement needs to include the quantification of various errors such as discretisation error, stochastic analysis error, model form error, etc. Methods have been developed to quantify finite element discretisation error and stochastic analysis error; methods to quantify model form error need considerable study.

Acknowledgement

References

Aeschliman, D.P. and Oberkampf, W.L. (1998) 'Experimental methodology for computational fluid dynamics code validation', *AIAA Journal*, Vol. 36, No. 5, pp.733–741.

American Institute of Aeronautics and Astronautics (1998) *Guide for the Verification and Validation of Computational Fluid Dynamics Simulations*, AIAA-G-077-1998, Reston, VA.

Anderson, M.C., Hasselman, T.K. and Crawford, J.E. (2000) 'A toolbox for validation of nonlinear finite element models', *Proceedings of 6th International LS-DYNA Users Conference, Simulations 2000*, Dearborn, MI, USA, 9–11 April.

Ang, A.H-S. and Tang, W.H. (1975) *Probability Concepts in Engineering Planning and Design, Volume I: Basic Principles*, John Wiley & Sons, New York.

Ang, J.A., Trucano, T.G. and Luginbuhl, D.R. (1998) *Confidence in ASCI Scientific Simulations*, Report No. SAND98-1525C; Sandia National Laboratories Albuquerque, NM.

Bae, D.H., Sohn, I.S. and Hong, J.K. (2003) 'Assessing the effects of residual stresses on the fatigue strength of spot welds', *Welding Journal*, pp.18–23, January.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999) *Modern Information Retrieval*, ACM Press, New York.

Bayarri, M.J., Berger, J.O., Higdon, D., Kennedy, M.C., Kottas, A., Paulo, R., Sacks, J., Cafeo, J.A., Cavendish, J., Lin, C.H. and Tu, J. (2002) *A Framework for Validation of Computer Models*, Technical Report 128, NISS.

Beck, M.B. (1987) 'Water quality modeling: a review of the analysis of uncertainty', *Water Resources Research*, Vol. 23, No. 8, pp.1393–1442.

Berger, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York.

Berger, J.O. and Delampady, M. (1987) 'Testing precise hypotheses', *Statistical Science*, Vol. 2, pp.317–352.

Berger, J.O. and Pericchi, L.R. (1996) 'The intrinsic Bayes factor for model selection and prediction', *J. Amer. Statist. Assoc.*, Vol. 91, pp.109–122.

Box, G.E.P. and Tiao, G.C. (1992) *Bayesian Inference in Statistical analysis*, John Wiley & Sons, New York.

Carlin, B.P. and Louis, T.A. (2000) *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall/ CRC, New York.

Defense Modeling and Simulation Office (1996) *Verification, Validation, and Accreditation* (*VV&A*) *Recommended Practices Guide*, Office of the Director of Defense Research and Engg, www.dmso.mil/docslib, Alexandria, VA, April.

Ghanem, R. and Spanos, P.D. (1991) *Stochastic Finite Elements: A Spectral Approach*, Springer-Verlag, New York.

Hasselman, T.K. and Chrostowski, J.D. (1997) 'Effects of product and experimental variability on model verification of automobile structures', *Proceedings of the 15th International Modal Analysis Conference*, Orlando, FL, February.

Hasselman, T.K. and Wathugala, G.W. (2002) 'A hierarchical approach for model validation and uncertainty quantification', *Proceedings of Fifth World Congress on Computational Mechanics* (WCCM V), Vienna, Austria, July.

Henrysson, H.F. (2001) *Fatigue Life Prediction of Spot Welded Joints: Experiment and Life Prediction*, PhD Thesis, Chalmers University of Technology, Sweden.

Hills, R.G. and Leslie, I.H. (2003) *Statistical Validation of Engineering and Scientific Models: Validation Experiments to Application*, Report No. SAND2003-0706, Sandia National Laboratories, Albuquerque, NM.

Hills, R.G. and Trucano, T.G. (1999) *Statistical Validation of Engineering and Scientific Models: Background*, Report No. SAND99-1256, Sandia National Laboratories, Albuquerque, NM.

Huh, H. and Kang, W.J. (1997) 'Eletro-thermal analysis of the electric resistance spot welding process by a 3D finite element method', *J. of Material Processing Technology*, Vol. 63, Nos. 1–3, pp.672–677.

Hwang, J.T., Casella, G., Robert, C., Wells, M.T. and Farrell, R.H. (1992) 'Estimation of accuracy in testing', *Ann. Statist.*, Vol. 20, No. 1, pp. 490–509.

Jeffreys, H. (1961) *Theory of Probability*, 3rd ed., Oxford University Press, London.

Johnson, D.H. (1999) 'On the insignificance of statistical hypothesis testing', *J. Wildlife Manage*, Vol. 63, pp.763–772.

Jolliffe, I.T. (1986) *Principal Component Analysis*, Springer-Verlag, New York.

Leonard, T. and Hsu, J.S.J. (1999) *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*, Cambridge University Press, Cambridge.

Liu, Y. and Mahadevan, S. (2005) 'Probabilistic fatigue life prediction of multidirectional composite laminates', *Composite Structures*, Vol. 69, No. 1, pp.11–19.

Mandell J.F. and Samborsky, D.D. (2003) '*DOE/MSU Composite Materials Fatigue Database: Test Methods, Materials, and Analysis*, Sandia Technical report, Sandia National Laboratories, Albuquerque, NM, February.

Oba, S., Sato, M.A., Takemasa, I., Monden, M., Matsubara, K.I. and Ishii, S. (2003) 'A bayesian missing value estimation method', *Bioinformatics*, Vol. 19, No. 6, pp.2088–2096.

Oberkampf, W.L. and Trucano T.G. (2002) 'Verification and validation in computational fluid dynamics', *Progress in Aerospace Sciences*, Vol. 38, No. 3, pp.209–272.

Oja, E. (1983) *Subspace Methods of Pattern Recognition*, New York, John Wiley.

Pawitan, Y. (2001) *In All Likelihood: Statistical Modeling and Inference using Likelihood*, Oxford Science Publications, New York.

Radaj, D., Zheng, Z. and Moehrmann, W. (1990) 'Local stress parameter at the spot weld of various specimens', *J. of Engg Fracture Mech.*, Vol. 37, No. 5, pp.993–951.

Rebba, R., Mahadevan, S. and Huang, S. (2004) 'Validation and error estimation of computational models', *Proceedings of Fourth International Conference on Sensitivity Analysis of Model Output* (*SAMO*), Santa Fe, NM, March.

Roache, P.J. (2002) 'Recent contributions to verification and validation methodology', *Proceedings of the Fifth World Congress on Computational Mechanics* (*WCCM V*), Vienna, Austria, July.

Sacks, J., Rouphail, N.M., Park, B.B. and Thakuriah, P. (2000) *Statistically-Based Validation of Computer Simulation Models in Traffic Operations and Management*, Technical Report, NISS, www.niss.org/downloadabletechreports.html.

Singhal, A. and Seborg, D.E. (2001) 'Matching patterns from historical data using PCA and distance similarity factors', *Proceedings of the American Control Conference*, Arlington, VA.

Strehl, A. (2002) *Relationship-Based Clustering and Cluster Ensembles for High-Dimensional Data Mining*, PhD Thesis, The University of Texas at Austin.

Thacker, B.H. (2003) 'The role of non-determinism in verification and validation of computational solid mechanics models', *Reliability & Robust Design in Automotive Engineering*, SP-1736, *Proc. SAE 2003 World Congress & Exposition*, SAE International Paper 2003-01-1353, Detroit, MI, March.

Thacker, B.H. and Huyse, L.J. (2002) 'Role of non-determinism in validation of computational mechanics models', *Proceedings of the Fifth World Congress on Computational Mechanics* (*WCCM V*), Vienna, Austria, July.

Timm, N.H. (2002) *Applied Multivariate Analysis*, Springer-Verlag, New York.

Tipping, M.E. and Bishop, C.M. (1999) 'Probabilistic principal component analysis', *J. of the Royal Statistical Society*, Series B 21, Vol. 3, pp.611–622.

Trucano, T.G. (2000) 'Aspects of ASCI code verification and validation', *Presented at Quality Fourm 2000*, Sponsored by the Department of Energy, Santa Fe, NM, April.

Tsang, C.F. (1989) 'A broad view of model validation', *Proceedings of the Symposium on Safety Assessment of Radioactive Waste Repositories*, Paris, France, pp.707–716.

Urbina, A., Paez, T.L., Hasselman, T.K., Wathugala, G.W. and Yap, K. (2003) 'Assessment of model accuracy relative to stochastic system behavior', *Proceedings of 44th AIAA Structures, Structural Dynamics, Materials Conference*, 7–10 April, Norfolk, VA.

Van Paepegem, W. and Degrieck, J. (2000) 'Numerical modeling of fatigue degradation of fibre-reinforced composite materials', *Proceedings of the Fifth International Conference on Computational Structures Technology*, Volume F: Computational Techniques for Materials, Composites and Composite Structures, Leuven, Belgium, September, Civil-Comp Press, pp.319–326.

Wentzell, P.D., Andrews, D.T., Hamilton, D.C., Faber, K. and Kowalski, B.R. (1997) 'Maximum likelihood principal component analysis', *J. Chemomet.*, Vol. 11, pp.339–366.

Zellner, A. (1987) 'Comment', *Statistical Science*, Vol. 2, pp.339–341.

Zhang, R. and Mahadevan, S. (2003) 'Bayesian methodology for reliability model acceptance', *Reliability Engg. and System Safety*, Vol. 80, No. 1, pp.95–103.