A Probabilistic Metric for the Validation of Computational Models

Ksenija Dvurecenska[1], Steve Graham[2], Edoardo Patelli[1] & Eann A. Patterson[1]

1. School of Engineering, University of Liverpool, Liverpool, L69 3GH, UK

2. National Nuclear Laboratory, Chadwick House, Warrington, WA3 6AE, UK

ABSTRACT

A new validation metric is proposed that combines the use of a threshold based on the uncertainty in the measurement data with a normalised relative error, and that is robust in the presence of large variations in the data. The outcome from the metric is the probability that a model's predictions are representative of the real world based on the specific conditions and confidence level pertaining to the experiment from which the measurements were acquired. Relative error metrics are traditionally designed for use with series of data values but orthogonal decomposition has been employed to reduce the dimensionality of data matrices to feature vectors so that the metric can be applied to fields of data. Three previously published case studies are employed to demonstrate the efficacy of this quantitative approach to the validation process in the discipline of structural analysis, for which historical data was available; however, the concept could be applied to a wide range of disciplines and sectors where modelling and simulation plays a pivotal role.

Keywords: Model validation, relative error, computational modelling, orthogonal decomposition.

INTRODUCTION

Computational models are widely used to evaluate and predict the future behaviour of engineering systems. Recent increases in computational capabilities have made it possible to simulate a large variety of processes. For instance, simulations are used to understand the mechanical behaviour of novel materials and to develop and optimise sustainable designs for engineering structures. The results from a simulation are nearly always used to inform decisions that are likely to have socio-economic and, or human consequences. In most cases the modeller will not and, it has been argued philosophically [1], should not be the decision-maker which implies that the credibility of the results or predictions from the model becomes vital and can be enhanced through a Verification and Validation (V&V) process [2]. Verification[♣] can be summarised as ensuring that the mathematics of the model are being solved correctly whereas validation[♠] is establishing a level of confidence in a model as an accurate and reliable representation of the reality of interest.

From these definitions, it can be seen that verification of the model should precede validation and usually it is a process undertaken by the purveyors of commercial and academic software packages using verification benchmarks [3,4]. In this study, the focus has been on the validation process which is usually undertaken by a modeller who is using a verified software package. Initial discussions about computational model validation appeared in the literature during the second half of 20th century and coincided with the advent of simulation and modelling techniques that were enabled by the availability of computing power. Fishman and Kiviat [5] and Van Horn [6] were amongst the first to consider the idea of validation, and related questions, in the context of models in economics science; but their ideas are relevant to simulations in many areas of science and technology. They identified that a computational model is usually developed with particular objectives that reflect the intended use; and consequently, the simulation results have to be evaluated against these objectives. Sargent [7] added further specificity by including the term 'for the intended use' in the definition of model validation. The concept of model validation emerged during the 1980s [8–10] as being the comparison of model behaviour with the behaviour of a real system when both the simulation and observations are conducted under identical conditions; and it was consolidated into two guides for engineers, namely the AIAA guide for computational fluid dynamics simulations [11] and the ASME guide for computational solid mechanics models [12] in 1998 and 2006 respectively. These guides provide concise definitions and a generalised

[♣] The ASME guide [12] defines verification formally as 'the process of determining that a computational model accurately represents the underlying mathematical model and solution'.
[♠] The ASME guide [12] defines validation formally as 'the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model'

methodology for performing verification and validation, but neither include definitive step-by-step procedures.

A common approach to validation is to divide the available empirical data set into a calibration or training subset and a validation subset; then to 'tune' the model parameters using the calibration subset only before testing the model predictions with the validation subset and then repeating the entire process with a different division of the data set. There are some concerns about the dual use of data, or double-counting, involved in such an approach. However, Steele and Werndl [13] have argued this practice of double-counting is legitimate within a Bayesian framework, such as used recently for a linear regression model of the strength of composite laminates containing manufacturing defects [14]. In this example, the prediction uncertainty of the model was estimated using Leave One Out Cross Validation [15]. When large data sets are unavailable for calibration and validation and, or the model has multiple input and, or output parameters, such as when modelling the spatial distribution of mechanical strain in an engineering structure over time, then a different approach is required. Recently, a CEN workshop agreement [16] has provided a detailed methodology for performing validations of computational solid mechanics models.

In solid mechanics, it has been common practice to validate numerical models using single data points, for example the maximum and minimum values of a response measured by a strain gauge. However, recent work has extended this approach to using fields of data acquired from optical measurement techniques [17], e.g. stereoscopic digital image correlation. In these circumstances, the measured and predicted data fields are rarely in the same coordinate system or have the same data pitch, orientation or perspective and this renders direct comparisons problematic. Patterson and his co-workers have represented both measured and predicted data fields as images in order to apply orthogonal decomposition techniques [18] and enable straightforward comparison for the purpose of validation [19] as well as model updating [20]. In image decomposition, a set of polynomials are used to represent the image such that only the moments or coefficients of the polynomials are required to describe the image [21]. Image decomposition using orthogonal moments not only reduces the dimensionality of the data from an image matrix to a feature vector but is also invariant to rotation, scale and translation of the images [19]. Clearly, it is important to ensure that a feature vector is good representation of the original data before utilising it in a validation procedure and the CEN guide [16] recommends that the reconstructed image must satisfy the criteria that the uncertainty introduced by the decomposition process, u_{deco} must be less than the minimum measurement uncertainty in the measured data set, u_{cal} and that there should be no cluster of points in the image where the residual is greater than three times the decomposition uncertainty, u_{deco}. A cluster was defined as a region of adjacent pixels representing more than 0.3%

of the data and the minimum measurement uncertainty was obtained using a calibration procedure for the measurement apparatus [23], while the decomposition uncertainty was defined as:

$$u_{deco}^2 = \frac{1}{N}\sum_{i,j=1}^{N}\left(\hat{I}(i,j) - I(i,j)\right)^2 \quad (1)$$

where $I(i,j)$ and $\hat{I}(i,j)$ are the original and reconstructed images respectively, and N corresponds to the number of orthogonal moments used in decomposition. When the images of the measured and predicted data fields have been decomposed using the same process, the resultant moments can be plotted against one another to provide a simple comparison, as illustrated in Figure 1. The CEN guide recommends that the model can be considered valid if all of the moment pairs fall within the zone described by

$$S_P = S_M \pm 2u_{exp} \quad (2)$$

where S_P and S_M are the moment values representing predicted and measured data fields, respectively, and u_{exp} is the total uncertainty in the measured data, which is given by:

$$u_{exp} = \sqrt{u_{cal}^2 + u_{deco}^2} \quad (3)$$

Although the CEN guide [16] was prepared from the perspective of solid mechanics and hence the predicted and measured data are in the form of displacement and strain fields, because the decomposition process is applied to images of the data fields, it could be used for any application in which predicted and measured data fields can be treated as images. The approach results in a statement about the adequacy of the representation of reality by the model but does not provide information about the degree to which the predictions represent the measurements.

VALIDATION METRICS

Although the validation of simulation results is often referred to as a single process, at a more detailed scale it can be divided into two activities [24]: first, the difference between the predicted and measured results is computed with the aid of a statistical comparison; and second, the outcome is evaluated in the context of the adequacy requirements. The statistical comparison is usually expressed in the form of a validation metric, i.e. a function representing the distance between the two results in the appropriate domain [25]. An ideal validation metric should be quantitative, objective and include a measure of the uncertainty in the measured and predicted results [12, 24, 26, 27]. Berger and Bayarri [28] have suggested that validation methodologies can be classified as either frequentist or Bayesian; however, the approach recommended in the CEN guide is a form of hypothesis testing that provides a Boolean result, i.e. the model is either acceptable or

unacceptable, without any indication of the quality of the results from the model. In some instances, particularly when the model has been found to be unacceptable, without information about the quality of the predictions, decision-makers will be unable to identify an efficient trade-off for the next set of actions, apart from a general decision to refine the model [12]. This information gap has been closed in this study by integrating a Boolean decision, based on the CEN approach, with a quantitative validation metric that is frequentist in nature.

In the frequentist approach, the measured data are assumed to be true and used to compute a relative error in the predicted data, i.e. the difference between the response of the model and the experiment. In reality, the measured data cannot be considered as absolutely true because there will be uncertainties and errors associated with the measurements and these should be accounted for when evaluating the discrepancy between the measured and predicted data sets [26]. Oberkampf and Barone [24] calculated both an average and a maximum relative error and then estimated confidence intervals for both relative errors, which allowed the degree of validity to be expressed. Their work is often cited, e.g. [29–31], both for its summary of the validation procedure and its definition of a validation metric; however, the metric is often simplified [32,33] because it is not robust when a system response cannot be time-averaged or is close to zero-valued. Kat and Els [34] avoided these issues by evaluating the absolute percentage relative error of each pair of data values and comparing it to a specified threshold set by the accuracy requirements, which allowed them to provide a probability of the predictions from the model being within the specified threshold. They assumed that the data were deterministic quantities and did not include an uncertainty analysis. Bayesian analysis permits uncertainty to be considered but does not appear to have been used to produce a statement about the validity of a model, i.e. to quantify the degree to which predictions are a reliable representation of reality. Instead, most reports in the literature on this topic are associated with model calibration or updating [30,31,35], which is the process of adjusting model parameters to reduce the discrepancy between predictions and a specified benchmark. In Bayesian analysis, initial information about the quantity of interest is described by a probability distribution, known as a prior distribution, and is updated using additional information described in a probability distribution, known as a likelihood, to produce a new or updated probability distribution describing the quantity of interest, known as the posterior distribution.

The ratio of the prior and posterior distributions is known as the Bayes factor, which both Rebba & Mahadevan [36] and Liu et al [27] have identified as a possible validation metric together with an associated confidence index. While it might be possible to use an uninformed, naïve prior derived from theory, in general, the choice of the prior distribution and the data to be included in the likelihood is subjective [37] which is inappropriate for an objective validation metric.

DEVELOPING A NEW VALIDATION METRIC

The motivation for developing a new validation metric was to advance the approach recommended in the CEN guide [16]. A new probabilistic metric, which is applicable to data fields, is proposed to include a measure of the extent to which predicted data is representative of reality as described by measured data. The predicted and measured data are represented by a pair of feature vectors, S_P and S_M respectively obtained by orthogonal image decomposition following the process described by CEN [16]. The proposed validation metric is evaluated in four steps: (i) compute a normalised relative error, e_k for each pair of vector components; (ii) compute a weight for each error, w_k; (iii) define an error threshold, e_{th}; and (iv) calculate the validation metric, VM as the sum of those weighted errors, w_i less than the error threshold, e_{th}.

The normalised relative error is defined as

$$e_k = \left| \frac{S_{P_k} - S_{M_k}}{\max\limits_{m \in S_M} |S_{M_m}|} \right| \qquad (4)$$

where S_{P_k} and S_{M_k} are the k^{th} vector components representing the predicted and measured results respectively and $\max\limits_{m \in S_M} |S_{M_m}|$ is the magnitude of the measurement vector component with the largest absolute value. A bar chart of a typical set of normalised relative errors, e_k is shown in figure 2 for the longitudinal strain field in an I-beam subject to three-point bending. The weight, w_k of each error is defined as its percentage of the sum of the errors, i.e.

$$w_k = \frac{e_k}{\sum_{k=1}^{n} e_k} \times 100 \qquad (5)$$

where n is the number of components in each vector. The error threshold, e_{th} is calculated by combining the approaches employed by Kat & Els [34] and Sebastian et al [19] and normalising the expanded uncertainty in the measurement data, i.e.

$$e_{th} = \frac{2u_{exp}}{\max\limits_{m \in S_M} |S_{M_m}|} \times 100 \qquad (6)$$

This error threshold has been evaluated for the data in figure 2 and shown as a dashed line. Once these three steps were completed, the weighted errors, w_k, were compared to the error threshold, e_{th}, and the sum of those errors less than the threshold computed to yield the validation metric, VM, i.e.

$$VM = \sum_i w_i \mathbb{I}_{w_k < e_{th}} \qquad (7)$$

where $\mathbb{I}$ is an indicator function which takes the value 1 when $w_k < e_{th}$ and otherwise has a value zero. This process is represented graphically in the bottom graph of figure 2 by ranking the values in

the top graph and then calculating their cumulative weighted value. Following the interpretation of Kat and Els [34], this sum corresponds to the probability of the normalised errors being equal to or less than the experimental uncertainty. From the validation perspective, VM represents the probability that model is representative of reality for a specified intended use.

A minimum number of points are required to define the cumulative distribution, shown in the bottom half of figure 2, in order for the validation metric to yield reliable results. It is impossible to define this minimum number of points for an unknown distribution; however, for the simplest non-linear curve, i.e. a conic, at least five points are required assuming there is no uncertainty associated with the location of the points, according to Pascal's theorem [38]. Hence, it is reasonable to assume that $n_{min} \geq 5$. In addition, the number of points in the cumulative distribution corresponds to the number of components in each of the feature vectors, S_P and S_M, or the number of moments used in the orthogonal decomposition of the images of the predicted and measured data fields. For data fields in which the variable is a non-linear function of both spatial coordinates, the orthogonal polynomials recommended in the CEN guide [16], i.e. Chebyshev and Zernike, will require at least six moments or shape descriptors to describe the data field. Data fields that are linear functions of the spatial coordinates can be compared using simpler approaches than proposed here, so that practically, $n_{min} = 6$.

CASE STUDIES

The application of this new validation metric has been demonstrated for three case studies, which are described in this section, using previously published predictions and measurements, including two from a recent inter-laboratory study (or round-robin exercise) on validation [39]. In part, these case studies were chosen because the data were available and the minimum measurement uncertainties had been established following methodologies similar to that recommended by the CEN guide [16] and were relatively small. In each case, data fields from computational models and physical experiments were treated as images and post-processed using an identical orthogonal decomposition methodology, following Sebastian et al [19], to produce feature vectors, S_p and S_M.

I - I-beam subject to three-point bending

The data for this case study was taken from an earlier study [40] of the efficacy of the validation methodology described in the CEN guide [16] and; hence, only brief details of the model and experiments are included here. A half-metre length of aluminium I-section with overall cross-section dimensions 42x65mm was subject to static bending by a central load while supported symmetrically by two 50mm diameter solid rods of circular cross-section that were 450mm apart. The thickness of

the web and flange was 2.5mm and a series of four 35mm diameter circular holes penetrated the web at 100mm intervals along its length, as shown in figure 3. In the experiment, a stereoscopic digital image correlation system was used to acquire displacement data and the minimum measurement uncertainty was established as $10\mu m$ for displacement and $30\mu m$ for strain measurements using the calibration procedure described in [41]. A finite element model was created using 23,135 shell elements with the Ansys software package and employing an elasto-plastic material model with kinematic hardening. The predicted and measured data fields were decomposed using 400 Zernike moments, but only significant coefficients, i.e. lower terms of the polynomial that represent main features of the data field within the specified threshold [40], were included in the validation.

In this case study, the extent to which the predictions represent the measurements of the transverse displacement of the flange and the longitudinal strain in regions 1 and 2 in Figure 3 were evaluated. The probability that the predictions are acceptable was found to be 100% and 48% for the transverse displacement and longitudinal strain respectively in region 1 while the corresponding probability in region 2 for the strain was 100%. These results are summarised in Table 1 together with corresponding values of measurement uncertainty, u_{exp} (from [40]) and error threshold, e_{th} computed using expression (6). No value of the validation metric was calculated for the displacement in region 2 because less than six shape descriptors were required to represent the displacement field due to its simple shape, as shown in figure 3.

These outcomes correlate well with those in Figure 1 obtained by following the CEN guidelines. For example, a relatively low probability was found for the longitudinal strain, ε_x in region 1, which corresponds to the widely scattered data points in the bottom left graph in Figure 1. Hence, it can be concluded that implementation of the relative error metric improves upon the binary outcome of the CEN methodology by quantifying the quality of the predictions.

II - Rubber block subject to indentation

The indentation of a 60x60x25mm rubber block by a rigid wedge has been investigated previously by experiment and modelled analytically [42] and computationally [39]. Consequently, only a brief outline is provided here. Deformation data for the rubber block was acquired using a stereoscopic digital image correlation system when a compressive displacement load of 2mm was applied across the entire 25mm thickness of the block by an aluminium alloy wedge of external angle 73.45 degrees and tip radius 1.68mm (see figure 4). The stereoscopic digital image correlation system was calibrated and found to have minimum measurement uncertainties of $3.2\mu m$ and $23.8\mu m$ for the in-plane [23] and out-of-plane [43] displacements respectively. Predictions of the x-, y- and z-direction

displacements were obtained from a finite element model simulated in the Abaqus 6.11 software package using 49,920 three-dimensional eight-noded linear elements for the block and 2,870 three-dimensional four-noded bilinear quadrilateral elements for the wedge. The material of the wedge was assumed to be rigid while the rubber was modelled as a hyperelastic material defined by the Mooney-Rivlin relationship with the constants taking the following values: C_{10}=0.9 and C_{01}=0.3 with a bulk modulus, J=20.

The measured and predicted displacement fields are shown in figure 5 and were decomposed using Chebyshev polynomials. In order to achieve average reconstruction residuals that were just below the appropriate minimum measurement uncertainties, as recommended by the CEN guide [16], 170, 210 and 15 moments were employed to describe the surface displacement in x-, y- and z- directions respectively. The values for validation metric, VM, for the x-, y- and z- direction displacements were found to be 82.48%, 62.42% and 34.3% respectively based on error thresholds of 9.95%, 1.19% and 24.63% for the x-, y- and z-direction displacements.

These results correlate well with outcomes observed in figure 5, which were obtained by following CEN methodology. As was expected from the visual comparison of the fields in figure 5, the model is quite poor at predicting displacement in the z-direction and, even given the high uncertainty, the value of VM is very low. At the same time, the probabilities for the predictions of displacements in x- and y-directions have been reflected successfully and the validation metric quantified the differences.

III - Bonnet liner impact

Burguete et al [44] have described the analysis of the displacement fields for an automotive composite liner for a bonnet or hood subject to an impact; and so only an outline of the data acquisition and processing is given here. The composite liner, which had overall dimensions of approximately 1.5x0.65x0.03m, was subject to a high velocity (70m/s), low energy (<300J) impact by a 50mm diameter projectile with a hemi-spherical head. A high-speed stereoscopic digital image correlation system was used to obtained maps of out-of-plane displacements at 0.2ms increments for 100ms. The minimum measurement uncertainty was established to be 14$\mu\varepsilon$ at 290$\mu\varepsilon$ rising to 29$\mu\varepsilon$ at 2110$\mu\varepsilon$ [44]. The finite element code Ansys-LS-Dyna was employed to model the bonnet liner following impact using an elastic-plastic material model with isotropic damage and four-noded elements based on a Belytschko-Tsay formulation. Typical fields of predicted and measured fields of out-of-plane displacements are shown in figure 6 and were decomposed using adaptive geometric moment descriptors (AGMD) specifically tailored for the complex geometry of the liner. Burguete et al [44] compared the data fields from the model and experiment for 100ms following impact by

plotting the absolute difference between pairs of corresponding AGMDs as shown in figure 7. They concluded that when any of the absolute differences were greater than the uncertainty in the experiments, indicated by the broken lines in figure 7, then the model was not valid. In this study, the probability of the model being acceptable was assessed using the validation metric in equation (7) for each increment of time for which a displacement field was measured. The result is shown in figure 7 together with the result obtained by Burguete et al [44]. The trends in acceptability implied by both plots in figure 7 are similar with the predictions being a reasonable representation of the experiment for about 0.035s after impact. Burguete et al observed that, after this time instant, a crack developed unexpectedly in the test specimen which was not permitted to develop in the model and this accounts for the poor performance of the predictions.

DISCUSSION

The proposed validation metric is based on a relative error metric but, through the application of appropriate normalising of the relative error and the error threshold, the drawbacks of the previous frequentist approaches are avoided. This means that, unlike previous metrics, the proposed metric is capable of evaluating data with a naturally high variance between the individual values in the data set, including very small values close to zero. It also takes into account uncertainties in the measurement data. In part, these advantages are a result of the choice of mean absolute percentage error as a basis for calculating the validation metric following the work of Kat and Els [34] and which allows the measurement uncertainty to be directly employed as an error threshold. This ease of interpretation and the direct proportionality of the influence of each contribution to the absolute value of error were additional reasons for the choice of mean absolute percentage error instead of a root mean square error. The result is a value for the probability that predictions from a model are a reliable representation of the measurements based on the uncertainty in the measurements used in the comparison. This allows the outcome of the validation process to be expressed in a clear quantitative statement that reflects the complete definition of the validation process. Such a statement should include the following three components:

- the probability of the model's predictions being representative of reality
- for the stated intended use and conditions considered, and
- based on the quality of the measured data defined by its relative uncertainty.

For example, one of the validation outcomes for the rubber block case study can be expressed as: *'there is an 83% probability that the model is representative of reality, when simulating x-direction*

displacements induced by a 2mm indentation, based on experimental data with 10% relative uncertainty'. The implementation of this type of statement would represent a significant advance on current practice and could be interpreted relatively straightforwardly by decision-makers. The outcome of this type of modified validation process allows the decision-maker, e.g. customer or stakeholder, to make the final judgment based on the evidence from the validation and their required or desired level of quality. When the level of agreement between predicted and measured data is inadequate for the intended purposes of the model, then both ASME [12] and CEN [16] guides recommend that both the model and the experiment should be reviewed before repeating the validation process. The use of model updating techniques [20] might be appropriate at this stage.

Brynjarsdottir and O'Hagan [45] have discussed the issue that experiments and simulations both mimic reality so that both have a certain level of approximation which has to be accounted for during a validation process. In particular, analysis that does not account for the discrepancies arising from these approximations may lead to biased and over-confident predictions. Hence, it is not enough to compare a simulation with an experiment, but also it is necessary to consider the relationship of the experiment to reality [46]. In other words, to recognise that the process of experiment design results in a representation of the real-life situation based on our current understanding; and that the resultant measurements should not be regarded as the absolute truth. Of course, measurements made directly in the real-life situation are likely to be closer to the truth than those made using physical models; but the measurement process will always influence the measurement data leading to uncertainty about the truth. Hence, the last component of the statement above, would ideally include information about the discrepancy between the truth and the measurements used in the validation process. However, this information is usually unavailable and, as a consequence, some caution, and awareness of context, needs to be exercised in employing the type of statement expressed above in italics; nevertheless, it represents a potential improvement on current practice in terms of its specificity.

The new validation metric, VM in equation (7), has been described in generic terms and the case studies illustrate its application to information-rich spatial data fields using feature vectors; however, the vectors, S_P and S_M describing the predicted and measured data could be constructed from many types of data, including time-series data. There is an implicit assumption in the use of the orthogonal image decomposition process to compare data fields, which is of one-to-one correspondence between the components of the feature vectors representing the data fields. This could be viewed as a potential limitation of the approach because this correspondence might not be present when some decomposition processes are used; however, the decomposition process

employed here and recommended in the CEN guide [16] was designed to deliver this correspondence. The measurement data in each of the case studies were displacement fields obtained using digital image correlation and were chosen based on the availability of both predicted and measured data fields and of measurement uncertainties. Digital image correlation has become almost ubiquitous in experimental mechanics and hence its use here; nevertheless, the decomposition technique is widely applicable and has been used for data fields from thermoelastic stress analysis and projection moiré [47]. The generic nature of the approach should allow its application in a wide variety of disciplines, for instance computational fluid dynamics, computational electromagnetics or landscape topography evolution modelling, and sectors, including civil, electrical and mechanical engineering whenever the predicted and measured data are available as maps that can be treated as images. In some applications, it is not possible to generate measurement data at all points in the region of interest, such as when optical access is obstructed or only a small number of point sensors can be employed or when the system is inaccessible, for example in a nuclear power plant. In these circumstances, when there is a sparcity of data, the relative error cannot be calculated for all of the predictions and this shortfall should be reflected in the statement about the outcome of the validation process, i.e. it would be appropriate to state what percentage of the predictions were used in the constructing the validation metric and how well the position of these data values covered the region of interest. The interpretation of this additional information will be specific to the intended use of the model and hence, no prescription is provided here.

The three case studies are a progression from a linear elastic planar static analysis, through a large deformation elastic static analysis to a non-linear elasto-plastic time-varying analysis. Although this progression provides increasing challenges to both modellers and experimentalists, all of these cases are mechanical systems that can be represented by deterministic models and for which it is possible to design and conduct repeatable experiments with relatively low levels of measurement uncertainty. Many analyses in engineering will fall within the same classification; however, in its current form, the validation metric cannot be applied to probabilistic models or to non-linear dynamic models with solutions in state space that lie outside this classification.

The approach to the validation process described in the ASME V&V guide [12] implies that it should be an interactive effort between those responsible for the model and those developing and conducting the experiments required to generate measurement data. However, it is unlikely that either group will be responsible for making decisions based on the predictions from the model and hence the credibility of the model becomes a critical factor. Model credibility is the willingness of others to make decisions supported by the predictions from the model [48]. Thus, it is important to present the outcomes from the validation process in a manner that can be readily appreciated by

decision-makers who may not be familiar with principles embedded in the model or the approach taken to validation, including the techniques used to acquire the measurement data used in the validation process. Patterson and Whelan [49], in the context of computational biology, have discussed strategies for establishing model credibility, including incorporating a high degree of transparency and traceability in the validation process, recognising the inadequacy of experiments as representations of the real-world, stating the uncertainties associated with the data in the outputs from the validation process, and expressing the accuracy of the representation of the real-world in terms of probabilities. The new validation metric combined with the proposed statement about the outcome of the validation process addresses these last two issues.

CONCLUSIONS

A new validation metric based on a frequentist approach has been proposed. The advantages of the metric are that it can handle data sets with large amplitude variations in data values as well as close-to-zero values and that the uncertainty in the measured data is also included in the metric. When it is combined with an appropriate orthogonal decomposition technique, then the dimensionality of large matrices of data can be reduced to feature vectors that enable data-rich maps of measurements to be used in the validation of corresponding predictions. The new validation metric allows a statement to be constructed about the probability that the predictions from a model represent reality based on experimental data with a given relative uncertainty for a specified intended purpose.

Three case studies have demonstrated the use of the new metric in computational mechanics for a linear elastic planar static analysis, for a large deformation elastic static analysis, and for a non-linear elasto-plastic time-varying analysis. The outcomes obtained with the new validation metric were more quantitative and informative than the previous validation procedures but qualitatively equivalent. Although these case studies relate to structural analysis, the principles illustrated are applicable to analysis in a wide range of fields including bioengineering, earth sciences and nuclear engineering.

Finally, it is proposed that the new metric can be used to construct a clear quantitative validation statement about a model that contains three core components: (i) the probability that model's predictions are representative of reality; (ii) for the intended use and conditions for which the comparison with measurements was performed and (iii) the uncertainty in the measurement data.

Data availability. Our data are deposited at Dryad: https://doi.org/10.5061/dryad.2qp305p

Competing Interests. We have no competing interests.

Authors' contributions. KD performed the research and prepared the first draft of the paper. EAP conceived and supervised the study and prepared the final draft of the paper. SG and EP supervised the study and contributed to the interpretation of the results and refinement of the methodology, including the interpretation of the ranked weighted errors as a cumulative distribution function (EP). All authors gave final approval for publication.

Funding. KD was supported by a studentship funded by the UK Engineering and Physical Sciences Research Council and by the National Nuclear Laboratory. EAP was in receipt of a Royal Society Wolfson Research Merit Award.

Ethics. No ethics approvals were required for this research.

Permission to carry out fieldwork. No fieldwork was performed in this research.

Acknowledgements. The authors are grateful to George Lampeas, Vasilis Pasialis, Xiaoshan Lin, Luis Felipe-Sese, Xiaohua Tan and Weizhong Wang for access to their data for the case studies.

REFERENCES

1. Jeffrey RC, 1956, Valuation and Acceptance of Scientific Hypotheses. *Philosophy of Science,* 23(3), 237–246.
2. Patterson EA, 2015, On the credibility of engineering models and meta-models. *J. Strain Analysis,* 50(4), 218–220.
3. Oberkampf WL, & Trucano TG, 2008, Verification and validation benchmarks. *Nuclear Engineering and Design,* 238(3), 716–743.
4. NAFEMS, 2014, Quality Assurance. *Laboratory Investigation*. 94, 500–517. See http://www.nature.com/doifinder/10.1038/labinvest.2014.37.
5. Fishman GS, & Kiviat PJ, 1968, The statistics of discrete-event simulation. *Simulation,* 10(4), 185–195.
6. Van Horn RL, 1971, Validation of Simulation Results. *Management Science* 17(5), 247–258.
7. Sargent RG, 1979, Validation of simulation models. In *Proceedings of the 11th conference on Winter Simulation-Volume 2*, pp. 479–503. IEEE Press.
8. Gass SI, 1977, Evaluation of complex models. *Computers and Operations Research,* 4(1), 27–

35.

9. Shannon RE, 1981, Tests for the verification and validation of computer simulation models. In *Proceedings of the 13th conference on Winter Simulation -Volume 2*, pp. 573–577. IEEE Press.

10. Balci O, 1989, How to assess the acceptability and credibility of simulation results. In *Proceedings of the 21st conference on Winter simulation*, pp. 62–71. IEEE Press.

11. Oberkampf WL, Sindir M, & Conlisk AT, 1998, Guide for the verification and validation of computational fluid dynamics simulations. *American Institute of Aeronautics and Astronautics*, AIAA G-077-1998.

12. ASME, 2006, Guide for verification and validation in computational solid mechanics. *American Society of Mechanical Engineers*, ASME V&V 10-2006.

13. Steele K, & Werndl C, 2016, Model tuning in engineering: uncovering the logic, *J. Strain Analysis for Engineering Design*, 51(1):63-71.

14. Christian WJR, DiazDelaO FA, Atherton K, & Patterson EA, 2018, An experimental study on the manufacture and characterisation of in-plane fibre-waviness defects in composites. *R. Soc. open sci.* 5:180082.

15. Witten IH, Frank E, & Hall MA. 2011 *Data mining: practicalmachine learning tools and techniques*.Boston, MA: Morgan Kaufmann.

16. CEN, 2014, Validation of computational solid mechanics models. *Brussels: Comite Europeen de Normalisation (CEN)*, CWA16799(2014).

17. Rastogi PK & Hack E, 2012, *Optical Methods for Solid Mechanics: A Full-Field Approach*. John Wiley & Sons.

18. Patki AS, & Patterson EA, 2012, Decomposing Strain Maps Using Fourier-Zernike Shape Descriptors. *Experimental Mechanics,* 52(8), 1137–1149.

19. Sebastian C, Hack E, & Patterson EA, 2013, An approach to the validation of computational solid mechanics models for strain analysis. *J. Strain Analysis,* 48(1), 36–47.

20. Wang W, Mottershead JE, Sebastian CM, & Patterson EA, 2011, Shape features and finite element model updating from full-field strain data. *IJ Solids and Structures,* 48(11–12), 1644–1657.

21. Huazhong S, Limin L, & Coatrieux J-L, 2007, Moment-Based Approaches in Imaging. Part 1. Basic Features. *IEEE Engineering in Medicine and Biology Magazine,* 26(5), 70–74.

22. Mukundan R, Ong SH, & Lee PA, 2001, Image analysis by Tchebichef moments. *IEEE Transactions on Image Processing,* 10(9), 1357–1364.

23. Hack E, Lin X, Patterson EA, & Sebastian CM, 2015, A reference material for establishing uncertainties in full-field displacement measurements. *Measurement Science and Technology,* 26(7), 75004.

24. Oberkampf WL, & Barone MF, 2006, Measures of agreement between computation and experiment: Validation metrics. *J. Computational Physics,* 217(1), 5–36.

25. Upton G, & Cook I, 2008, *A Dictionary of Statistics*. 2nd edn. Oxford: Oxford University Press.

26. Rutherford BM, 2008, Computational modeling issues and methods for the 'regulatory problem' in engineering - Solution to the thermal problem. *Computer Methods in Applied Mechanics and Engineering* 197(29–32), 2480–2489.

27. Liu Y, Chen W, Arendt P, & Huang H-Z, 2011, Toward a better understanding of model validation metrics. *J. Mechanical Design* 133(7), 71005.

28. Berger JO, & Bayarri MJ, 2004, The interplay of bayesian and frequentist analysis. *Statistical Science,* 19(1), 58–80.

29. Bayarri MJ, Paulo R, Berger JO, Sacks J, Cafeo JA, Cavendish J, Lin CH, & Tu J, 2007, A framework for validation of computer models. *Technometrics,* 49(2), 138–154.

30. Wang S, Tsui KL, & Chen W, 2009, Bayesian validation of computer models. *Technometrics,* 51(4), 439–451.

31. Hills RG, Dowding KJ, & Swiler L, 2008, Thermal challenge problem: Summary. *Computer Methods in Applied Mechanics and Engineering,* 197(29–32), 2490–2495.

32. Fortunato V, Galletti C, Tognotti L, & Parente A, 2015, Influence of modelling and scenario uncertainties on the numerical simulation of a semi-industrial flameless furnace. *Applied Thermal Engineering,* 76, 324–334.

33. Slaba TC, Blattnig SR, Reddell B, Bahadori A, Norman RB, & Badavi FF, 2013, Pion and electromagnetic contribution to dose: Comparisons of HZETRN to Monte Carlo results and ISS data. *Advances in Space Research,* 52(1), 62–78.

34. Kat CJ, & Els PS, 2012, Validation metric based on relative error. *Mathematical and Computer Modelling of Dynamical Systems,* 18(5), 487–520.

35. Patelli E, Govers Y, Broggi M, Gomes HM, Link M, & Mottershead JE, 2017, Sensitivity or Bayesian model updating: a comparison of techniques using the DLR AIRMOD test data.

Archive of Applied Mechanics, 87(5), 905–925.

36. Rebba R, & Mahadevan S, 2008, Computational methods for model reliability assessment. *Reliability Engineering and System Safety,* 93(8), 1197–1207.

37. Kass RE, Wasserman L, 1996, The selection of prior distributions by formal rules. *J. Am. Statistical Association,* 91(435), 1343–1370.

38. Hamilton WR, 1847, On a proof of Pascal's theorem by means of quaternions; and on some other connected subjects. *Proc. Royal Irish Academy,* 3, 273–292.

39. Hack E, Lampeas G, & Patterson EA, 2016, An evaluation of a protocol for the validation of computational solid mechanics models. *J. Strain Analysis,* 51(1), 5–13.

40. Lampeas G, Pasialis V, Lin X, & Patterson EA, 2015, On the validation of solid mechanics models using optical measurements and data decomposition. *Simulation Modelling Practice and Theory* 52, 92–107.

41. Sebastian C, & Patterson EA, 2015, Calibration of a digital image correlation system. *Experimental Techniques,* 39(1), 21–29.

42. Tan X, Kang Y, & Patterson E, 2014, An experimental study of the contact of a rounded rigid indenter with a soft material block. *J. Strain Analysis,* 49(2), 112–121.

43. Felipe-Sesé L, Siegmann P, Díaz FA, & Patterson EA, 2014, Integrating fringe projection and digital image correlation for high-quality measurements of shape changes. *Optical Engineering,* 53(4), 44106.

44. Burguete G, Lampeas G, Mottershead JE, Patterson EA, Pipino A, Siebert T, & Wang WJ, 2014, Analysis of displacement fields from a high-speed impact using shape descriptors. *J. Strain Analysis ,* 49(4), 212–223.

45. Brynjarsdóttir J, & O'Hagan A, 2014, Learning about physical parameters: The importance of model discrepancy. *Inverse Problems,* 30(11), 801812.

46. Henninger HB, Reese SP, Anderson AE, & Weiss JA, 2010, Validation of computational models in biomechanics. *J. Engineering in Medicine,* 224(7), 801–812.

47. Felipe-Sese L, Diaz-Garrido FA & Patterson EA, Exploiting measurement-based validation for a high-fidelity model of dynamic indentation of a hyperelastic material, *IJ Solids & Structures*, 97-98:520-529, 2016.

48. Schruben LW, 1980, Establishing the credibility of simulations. *Simulation,* 34, 101–105.

49. Patterson EA, & Whelan MP, 2017, A framework to establish credibility of computational models in biology. *Progress in Biophysics and Molecular Biology* 129, 13–19.

Table 1: Case study 1: I-beam subject to three-point bending

		u_{exp}	e_{th}	Validation metric, *VM*
Region 1	u_y	2.69%	24.15%	100%
	ε_x	3.57%	15.11%	48%
Region 2	ε_x	3.97%	11.53%	100%

Figures

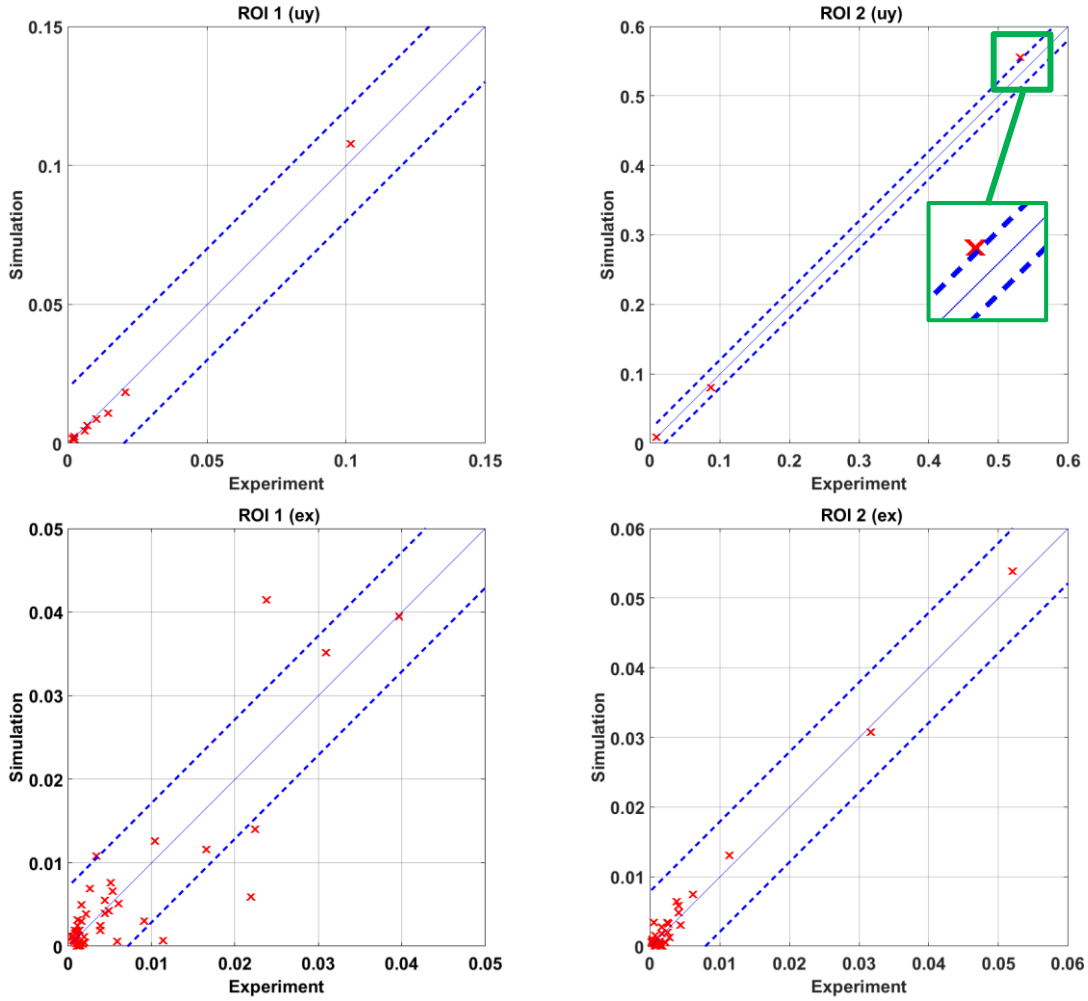

Figure 1 - Graphical comparisons, using the approach recommended by the CEN guideline [16] for evaluating the acceptability of model predictions, of the Zernike moments representing the predicted (y-axis) and measured (x-axis) transverse displacement (top) and longitudinal strain (bottom) in regions 1 (left) and 2 (right) of the I-beam subject to three-point bending shown in figure 3 (based on Lampeas et al [40]). The predictions can be considered acceptable when all of the data falls within the zone bounded by the broken lines that are defined by equation (2) based on the measurement uncertainty.

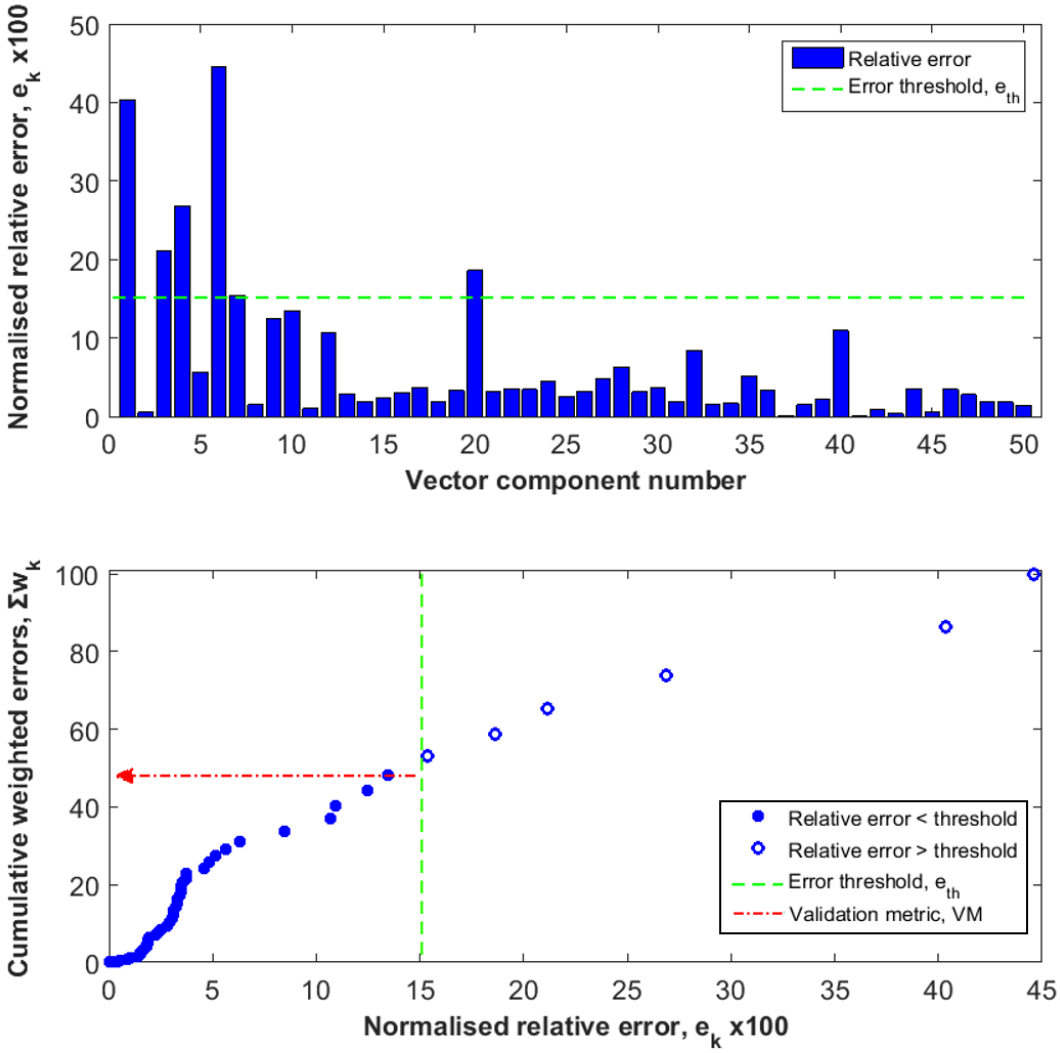

Figure 2 – A bar chart of normalized relative errors (top) based on equation (4) and multiplied by 100 to allow the error threshold from equation (6) to be shown; and the cumulative distribution (bottom) of ranked weighted errors computed using equation (5) for the predicted longitudinal strain field in region 1 of the of the I-beam subject to three-point bending shown in figure 3; based on equation (7) the validation metric is the sum of those errors below the threshold, i.e. the filled symbols in the bottom graph.

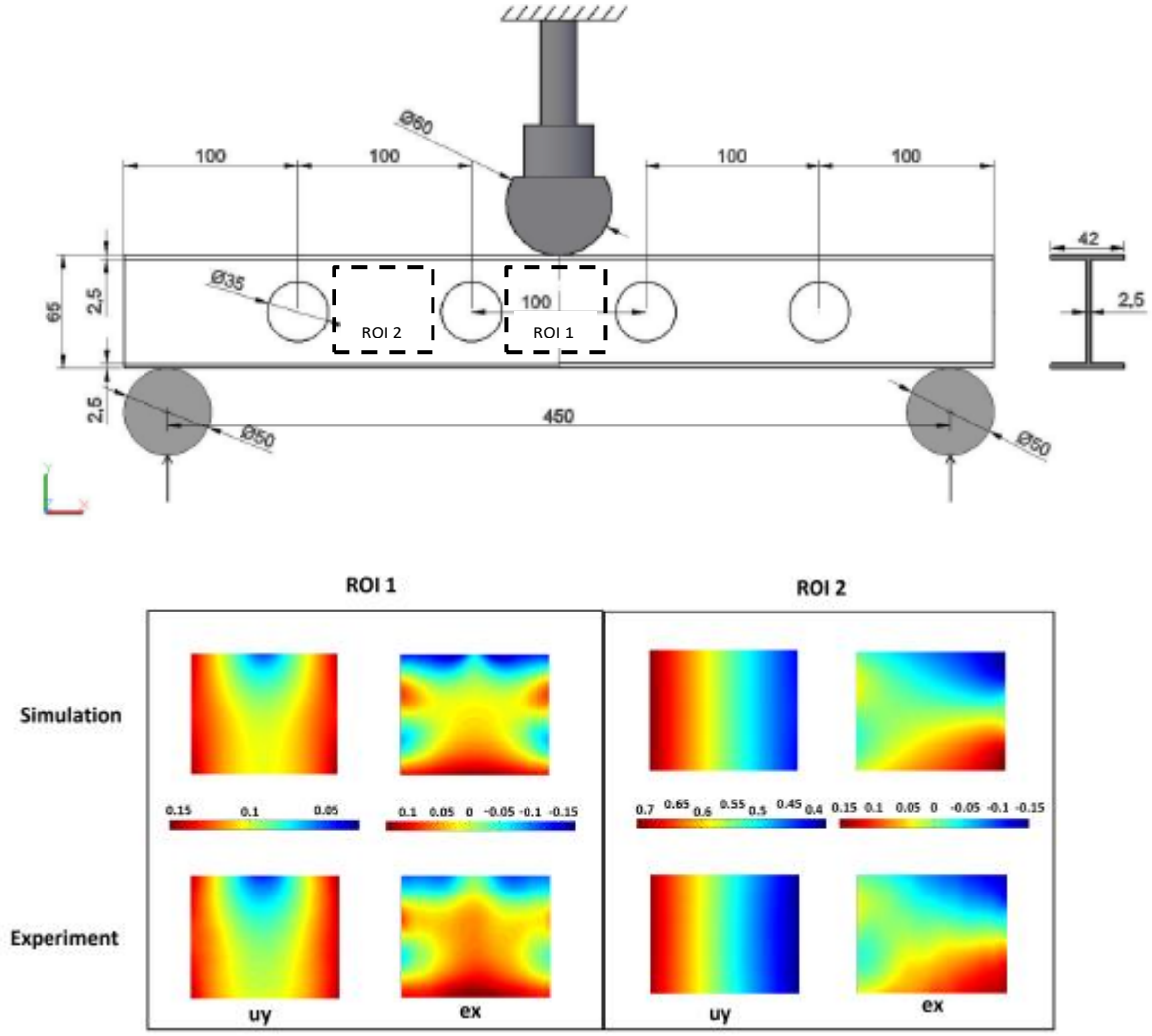

Figure 3 - A diagram (top) of the I-beam subject to three-point bending showing the regions of data used in case study 1 together with the predicted and measured fields of transverse displacement and longitudinal strain (bottom) (reproduced with permission from Lampeas et al [40]).

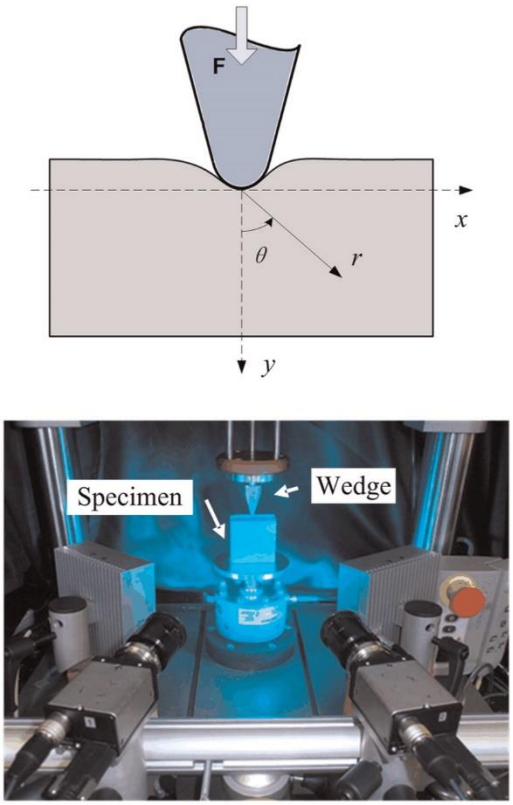

Figure 4 - Schematic diagram (top) and photograph (bottom) of the indentation of a rubber block (60x60x30mm) by a rigid indenter (reproduced with permission from Tan et al [42]).

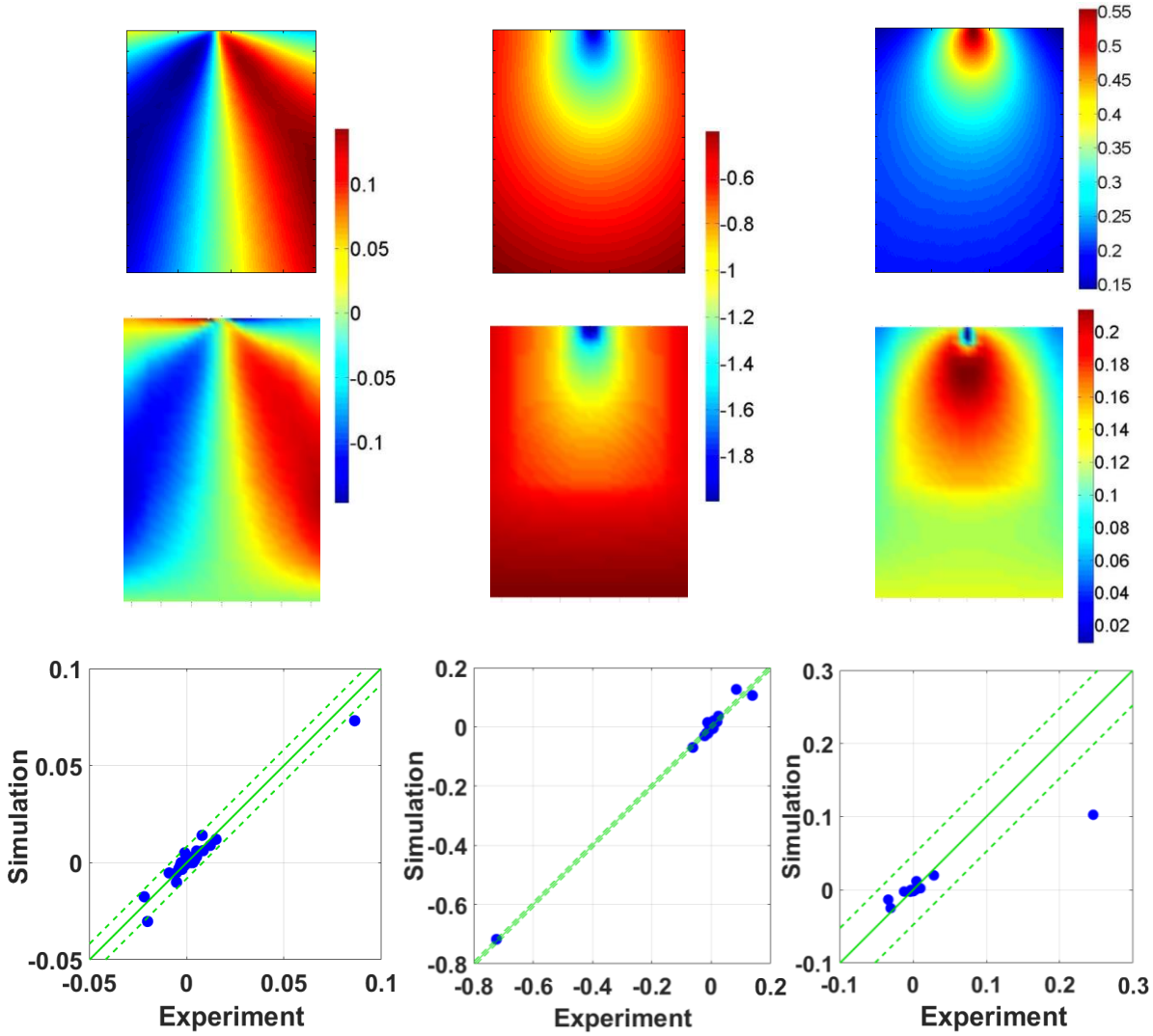

Figure 5 - Measured (top) and predicted (middle) x-direction (left), y-direction (centre) and z-direction (right) displacement fields for a 28.5x23mm area of the rubber block shown in figure 4 when it was subject to 2mm displacement load by the wedge in the y-direction; and plots obtained using the CEN methodology [16] (bottom). The centre of the top edge of each data area corresponds the location of contact by the wedge and the units are millimetres.

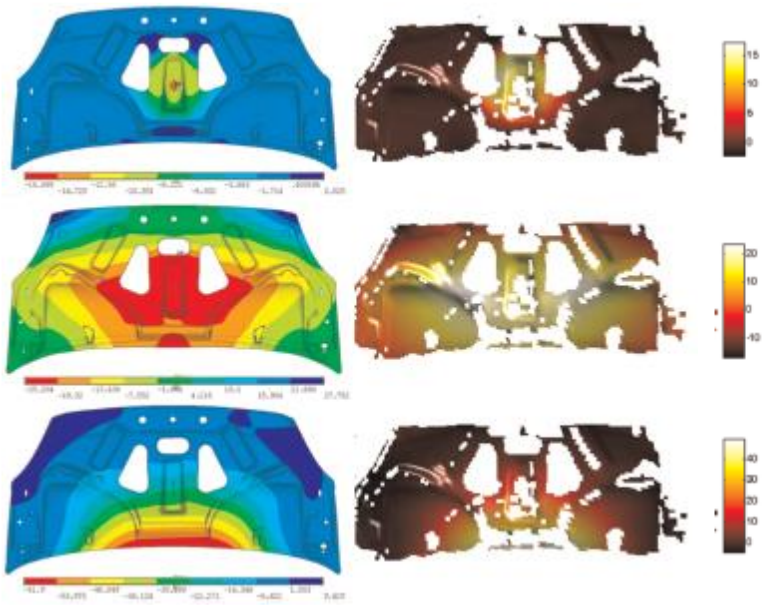

Figure 6 - Predicted (left) and measured (right) out-of-plane displacement fields for the automotive bonnet (hood) liner (approx. 1.5x0.65x0.03m) at 40, 50 and 60ms (from top to bottom) after a high-speed, low-energy impact by a projectile in the centre of the liner (reproduced with permission from Burguete et al [44]).

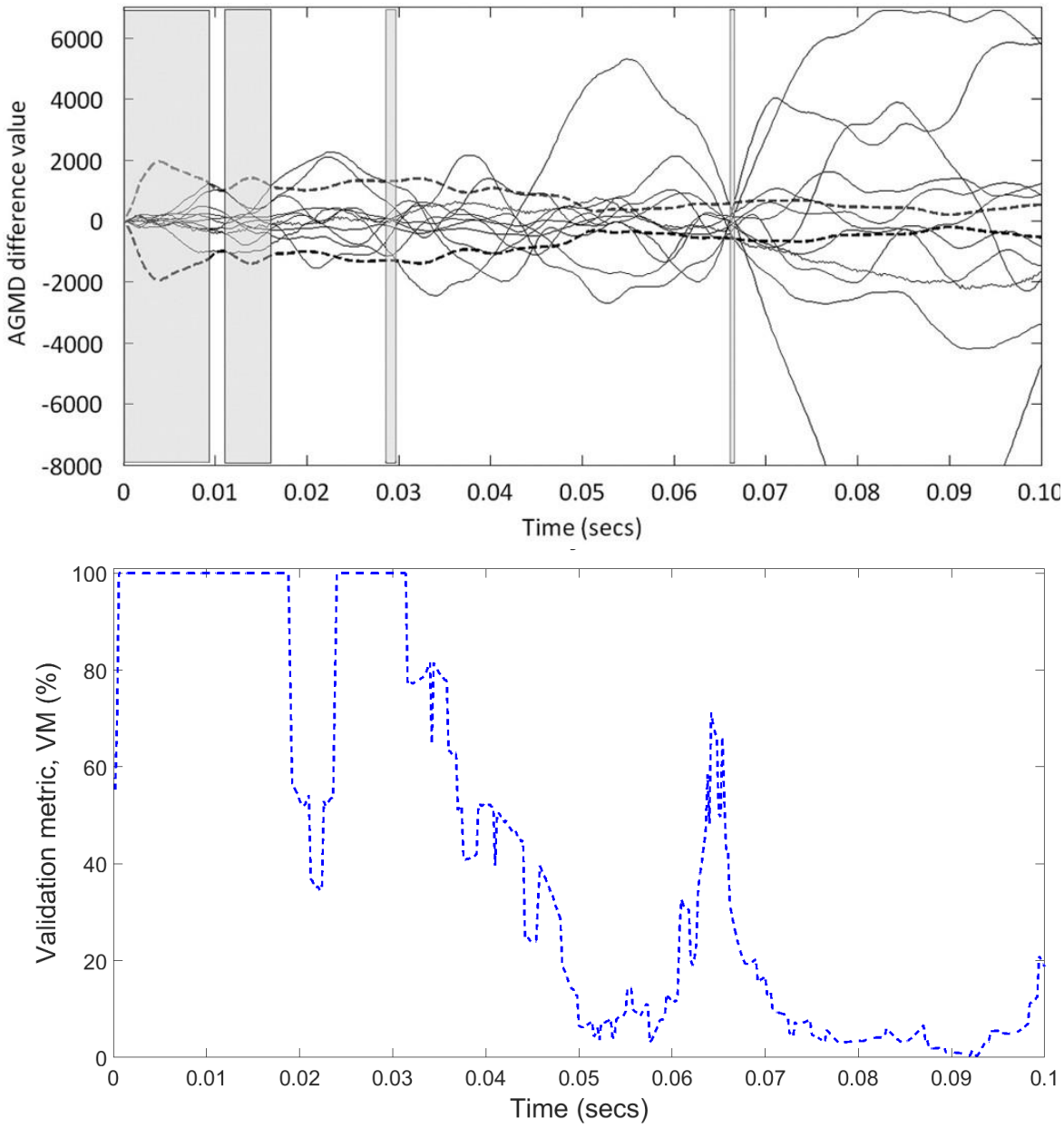

Figure 7 - Absolute difference between corresponding adaptive geometric moment descriptors (top) describing the predicted and measured out-of-plane displacement field of the automotive bonnet liner during the 0.1 seconds following impact and the corresponding probability of the predictions being a reliable representation of the measurements based on incorporating the weighted relative error and error threshold into the validation metric, VM, using equations (7) (the top graph is reproduced with permission from Burguete et al [44]).