Computers and Structures 88 (2010) 497–505

Contents lists available at ScienceDirect

Computers and Structures

journal homepage: www.elsevier.com/locate/compstruc

Defining predictive maturity for validated numerical simulations

François Hemez, H. Sezer Atamturktur *, Cetin Unal

Applied Physics Division (X-Division), Los Alamos National Laboratory, Los Alamos, NM 87545, USA

ARTICLE INFO

Article history:
Received 14 July 2009
Accepted 7 January 2010

Keywords:
Modeling and simulation
Model calibration
Predictive maturity metric
Decision-making

ABSTRACT

The increasing reliance on computer simulations in decision-making motivates the need to formulate a commonly accepted definition for "*predictive maturity*." The concept of predictive maturity involves quantitative metrics that could prove useful while allocating resources for physical testing and code development. Such metrics should be able to track progress (or lack thereof) as additional knowledge becomes available and is integrated into the simulations for example, through the addition of new experimental datasets during model calibration, and/or through the implementation of better physics models in the codes. This publication contributes to a discussion of attributes that a metric of predictive maturity should exhibit. It is contended that the assessment of predictive maturity must go beyond the goodness-of-fit of the model to the available test data. We firmly believe that predictive maturity must also consider the "*knobs*," or ancillary variables, used to calibrate the model and the degree to which physical experiments cover the domain of applicability. The emphasis herein is placed on translating the proposed attributes into mathematical properties, such as the degree of regularity and asymptotic limits of the maturity function. Altogether these mathematical properties define a set of constraints that the predictive maturity function must satisfy. Based on these constraints, we propose a Predictive Maturity Index (PMI). Physical datasets are used to illustrate how the PMI quantifies the maturity of the non-linear, Preston–Tonks–Wallace model of plastic deformation applied to beryllium, a light-weight, high-strength metal. The question "*does collecting additional data improve predictive power?*" is answered by computing the PMI iteratively as additional experimental datasets become available. The results obtained reflect that coverage of the validation domain is as important to predictive maturity as goodness-of-fit. The example treated also indicates that the stabilization of predictive maturity can be observed, provided that enough physical experiments are available.

1. Introduction

The increasing reliance on computer simulations in decision-making motivates the need to formulate a commonly accepted definition for "*predictive maturity*." Predictive maturity metrics are needed to better quantify the degree of confidence that decision-makers can place in decisions that are supported by numerical simulations. The concept of predictive maturity involves quantitative metrics that could guide the decision makers while allocating resources for physical testing and code development. Such metrics should be able to track progress (or lack thereof) as additional knowledge becomes available and is integrated into the simulations. Examples of knowledge integration include using collecting experimental datasets for model calibration and implementing better physics-based models in the codes.

This work addresses the definition of a quantitative metric for assessing predictive maturity. Rather than proposing a rigid definition, the emphasis is on formulating a number of axioms that such a metric should satisfy. An example of a predictive maturity metric is then proposed and its performance is showcased using the non-linear, Preston–Tonks–Wallace model of plasticity for the beryllium metal. This research contributes to the Verification and Validation (V&V) component of the fuels Modeling and Simulation (M&S) campaign of the Global Nuclear Energy Partnership (GNEP) at Los Alamos [1]. Further details and additional application examples can be found in Ref. [2].

2. A brief review of available literature

There is no track of research in the fields of computational sciences and engineering that addresses how to define or implement predictive maturity. Likewise there is no history of publishing on this topic in statistical sciences, at least not to the extent that statisticians have investigated, for example, how to correlate predictions and measurements through model calibration. One reason may be that the concept of maturity starts to matter only after one has demonstrated through the activities of V&V that a

* Corresponding author.
E-mail address: sez@clemson.edu (H.S. Atamturktur).

0045-7949/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compstruc.2010.01.005

predictive M&S capability is being reached for a specific application. One may argue that modeling is a 4000-year-old activity that, for example, finds one of its origins in the work of Pythagoras to predict the cycles of planets in 1900 BC. Seeking to achieve a science-based predictive capability, on the other hand, is a much younger endeavor that can be traced back to the convergence of high-performance computing platforms, advanced scientific languages and graphics capabilities of the mid-1980s.

Among the few contributions available in the literature, the following studies are found to be useful as a starting point to the present investigation.

Refs. [3–5] discuss the definition and implementation of validation metrics, also known as test-analysis correlation metrics. It is the starting point because quantitative metrics are needed to assess prediction accuracy and correlate physical measurements. These metrics, however, are only a part of the puzzle because they do not address the domain of applicability of the modeling capability. In particular, one must be careful not to confuse "*goodness-of-fit*," or the extent to which predictions match measurements and predictive maturity, that should include a statement about the ability of models to deliver accurate predictions over a range of settings.

Ref. [6] proposes online credibility assessment software with the purpose of being simultaneously and collaboratively used by multiple team members at different geographic locations. This methodology decomposes the concept of credibility into several components. For each of these components, evaluators integrate the available qualitative and quantitative information according to the given score set. Although this method has the merit of reducing subjectivity by combining multiple inputs from multiple evaluators, the role of these weightings is limited to representing the judgment of the evaluators about the model and not necessarily its quality.

In Refs. [7,8], the authors focus on the development of a goodness-of-fit metric to quantify the agreement, or correlation, between predictions and measurements. Their proposal includes the concept of "*novelty*" of a piece of information or experimental measurement. The novelty term quantifies whether the experimental data are new and independent in terms of exploring a physics regime that previous experiments have avoided. It also quantifies whether the data used for test-analysis comparison are relevant to the maturity of the model.

Informative discussions of goodness-of-fit, V&V, credibility, and confidence are offered in a collection of technical reports and presentations among which Refs. [9–11] are cited herein. The authors propose a metric called the Quantitative Reliability at Confidence (QRC) that attempts to go beyond the conventional goodness-of-fit metric to include a statement about the level of "*confidence*" that analysts can place in the predictive power of their models. The QRC comes in the form of a statistical test that relates goodness-of-fit to statistical confidence. One weakness is that strong, yet, potentially unwarranted, assumptions need to be formulated about the type of underlying statistical distribution. Although this contribution addresses some of the same points made here, to define predictive maturity, its emphasis is still too heavily weighted towards model calibration.

More recently, efforts have been made to define predictive maturity through the combination of qualitative assessment tables and quantitative scoring schemes. Research organizations of the US Department of Energy (specifically, Sandia National Laboratories in Ref. [12]) and National Aeronautics and Space Administration (specifically, NASA Langley Research Center in Refs. [13,14]) are leading these efforts. Refs. [12,14] include descriptions of the main elements of V&V, how they can be categorized and how to define levels of maturity for each category. The levels of maturity are defined on simple scales, for example, from 0 to 3. At this point, it is up to the analysts involved in assessing predictive maturity to agree upon the specific scores justified by the levels of coverage and rigor of their V&V activities. These scores can be color-coded or aggregated into a single, high-level score for judgment by the decision-makers.

What is observed from the literature available on the subject is that very few investigations venture away from the conventional concept of goodness-of-fit, with the exception of the utility metric of Refs. [7,8]. Predictive maturity, as described in the above paragraph, comes with the drawback of relying on qualitative statements and expert judgment, even if the underlying V&V activities are quantitative in nature. Taking a small step away from goodness-of-fit to include other aspects of predictive maturity is what is attempted in this work.

3. The concept of domain of applicability

The concept of domain of applicability is essential to the discussion of predictive maturity. This is because generally models are not developed, and codes are not written, to simulate a single reality of interest. Models and codes are, instead, applied to multiple regimes or settings that collectively define what is called the domain of applicability.

3.1. Terminology and definitions

In this paper, the terms "model," "code" and "simulation" are used inter-changeably to denote an analytical or numerical model. They are collectively referred to as the "model" that, in short, builds a functional relation between inputs and outputs:

$$y = M(p;\theta), \tag{1}$$

where the pair $(p;\theta)$ denotes inputs to the model and y is the output. Variables $(p;\theta)$ and y can be scalar-valued or multi-dimensional. The symbol p refers to inputs that define the domain of applicability while θ denotes other inputs such as ancillary variables or calibration variables.

The objective of this work is to assess the predictive maturity of the model or simulation defined in Eq. (1), that is, its ability to deliver accurate predictions over a domain of applicability. The terms "data" or "dataset" are used refer to physical measurements or observations collected by performing one or several experimental tests. The measurements are denoted by the symbol y^{Test}, a quantity that, again, can be either scalar-valued or multi-dimensional and that an analyst can compare to the predictions of Eq. (1). A V&V study assesses how close (or correlated) the predictions y and measurements y^{Test} are, together with a quantification of sensitivity and uncertainty for these predictions [15–17].

3.2. The domain of applicability

Fig. 1 illustrates what could be the domain of applicability of a code that simulates the aero-elastic behavior of an aircraft wing. It defines two dimensions: flow velocity and angle of attack. Defining the domain of applicability is the starting point of V&V because it represents the collection of settings or regimes where numerical simulations need to be performed. The model or code, therefore, needs to be verified and validated over the entire domain.

The domain of applicability is referred to with the symbol $\{\Omega_{\mathrm{V}}\}$. In general, it is an N-dimensional subspace, or $\Omega_{\mathrm{V}} \subset \Re^{\mathrm{N}}$. Referring back to Eq. (1), the control parameters $p \in \Re^{\mathrm{N}}$ define the dimensions of $\{\Omega_{\mathrm{V}}\}$. Control parameters p_k $(k = 1, \ldots, N)$ are quantities that usually have a clear and unambiguous physical meaning. They are measured or controlled during experimentation. This is in contrast to calibration variables, denoted by the symbol θ_k

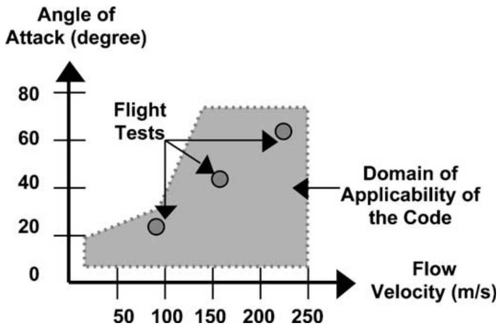

Fig. 1. A 2D domain of applicability defined for an analysis of aero-elastic flutter.

($k = 1, \ldots, m$). Calibration variables are introduced by the choices of physical models, numerical methods and solvers. They may have a physical meaning or be purely numerical in nature. The most important point is that calibration variables do not define the domain of applicability of the model.

4. Attributes of predictive maturity

The objective is to derive metrics that can quantitatively measure the predictive maturity of a model, that is, its ability to provide accurate predictions of a phenomenon of interest over a range of regimes or settings where the model is exercised. In Section 4, first the essential attributes of prediction maturity are discussed, then a metric to quantify the predictive maturity is proposed.

4.1. The essential attributes of predictive maturity

What are the essential attributes that make a model more or less predictive? To answer this question, the following three attributes, labeled (A-1) through (A-3) for convenience, are tabulated: (A-1) the extent to which available datasets "cover" the domain of applicability; (A-2) the "complexity" of the model; and (A-3) the level of accuracy to which model predictions match the available datasets.

The attributes (A-1)–(A-3) are consistent with those considered to define predictive maturity in Refs. [12–14]. Clearly one could think of many other attributes of a model that contribute to its predictive power. Examples include the robustness of model predictions to assumptions and time-to-solution. Time-to-solution is clearly important because getting the answer from a computer code that runs in 1 min as opposed to another code that runs in 1 h, with everything else being equal, matters greatly.

Robustness is essential to predictability. This is a point that has repeatedly been stressed in previous work; see Refs. [18,19]. Robustness refers to the extent to which predictions of a model vary or, to the contrary, are insensitive or "*robust*" when the assumptions upon which the model is based are modified. Achieving robustness means that predictions do not change too much and can therefore be trusted (whether they are accurate or not is another debate entirely), even if some of the assumptions that went into the model or numerical simulation are incorrect. Robustness can be dealt with by investigating the extent to which the predictive maturity metric is sensitive to assumptions that define the numerical simulation. This approach follows the application discussed in Ref. [19] in spirit. Nevertheless, it is recognized that this question remains somewhat unresolved and should be addressed in future research.

4.2. Measuring coverage of the domain of applicability

The coverage attribute (A-1) refers to the location (or position) of physical experiments performed in the domain of applicability $\{\Omega_V\}$. Fig. 2 illustrates this concept for a 2D domain. The figure shows that three tests are performed for various settings of control parameters $(X_1; X_2)$.

A simple approach to quantitatively measure coverage of the domain of applicability by assessing the extent to which the physical experiments fill the space can be achieved with two steps. One first computes the convex hull $\{\Omega_{CH}\}$ of the domain defined by the physical experiments. The convex hull is by definition, the smallest convex domain that includes all physical experiments. The second step is to calculate the ratio between the volume of the convex hull $\{\Omega_{CH}\}$ and that of the domain of applicability $\{\Omega_V\}$. The two steps result in the metric:

$$\eta_C = \frac{\text{Volume}(\Omega_{CH})}{\text{Volume}(\Omega_V)}, \qquad (2)$$

where Volume(•) denotes a function that calculates the N-dimensional volume of an arbitrary domain in $\Re^N$. Mathematical algorithms are available to calculate the convex hull of a domain defined by a set of points in N dimensions, as well as its volume (see, for example, function **convhulln.m** of MATLAB™). Fig. 3 illustrates the convex hull $\{\Omega_{CH}\}$ defined from the three experiments shown in Fig. 2. For a domain of applicability defined as $\{\Omega_V\} = [1; 5] \times [1; 4]$, the coverage metric of Eq. (2) would be equal to the ratio of highlighted area to the total area of the domain, that is, $\eta_C = \text{Area}(\Omega_{CH})/\text{Area}(\Omega_V)$.

Note that definition (2) has the advantage of simplicity and can be used to distinguish between "interpolation," or experiments located inside the convex hull, and "extrapolation" (experiments located outside). Yet, it can be misleading because experiments located inside the convex hull add no value to the coverage metric. In situations where many experiments are located inside $\{\Omega_{CH}\}$, the authors recommend a more suitable metric which accounts for the interior points.

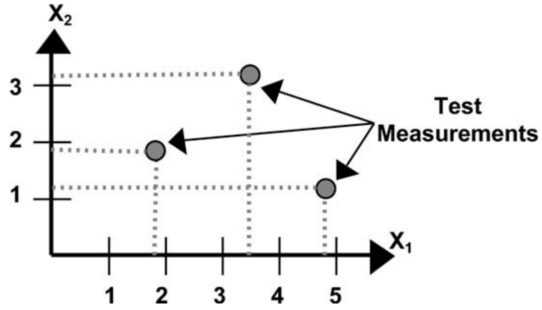

Fig. 2. Definition of a 2D domain of applicability.

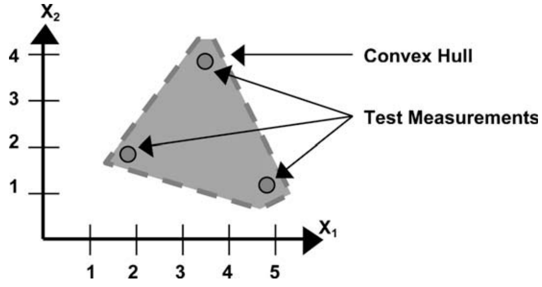

Fig. 3. Illustration of coverage of a 2D domain by the test datasets.

4.3. Measuring complexity of a model or numerical simulation

Defining the complexity of a model can be an extremely difficult task. It may involve making a statement about the sophistication of physical principles that are modeled; the complexity of mathematical spaces where the continuous or discrete solutions are constructed; the degree to which different physics are coupled; how many sub-models, algorithms or numerical methods are implemented; how many lines of codes are written; etc. The number of calibration "*knobs*," or ancillary variables more-or-less cuts across all of these aspects and it is chosen as a metric to address the complexity attribute (A-2).

While not necessarily the best option this choice offers is the undeniable advantage of simplicity. It is guided by the general principle that more sophisticated models possess larger numbers of calibration variables. A non-linear model of strain rate-dependent and temperature-dependent plasticity, for example, is more complicated and exhibits a few more free-parameters than an elastic, perfectly plastic constitutive law. An equation-of-state, on the other hand, can be more sophisticated than a rate and temperature dependent model.

In the remainder, the number of calibration knobs is denoted by the symbol N_K. Commonly encountered values of N_K that track complexity for our material modeling example would be $N_K = 2$ for an elastic, perfectly plastic constitutive law; $N_K \approx 5$–10 for a macroscopic, constitutive model; and $N_K \approx 10$–50 for an equation-of-state.

4.4. Measuring the overall level of accuracy of a model

The accuracy attribute (A-3) can be assessed using a goodness-of-fit or correlation metric defined for the purpose of test-analysis comparison. Examples include the absolute difference between a prediction and measurement, root mean square (RMS) error, correlation coefficient, or statistical test of consistency between two distributions of features. While calculating the predictive maturity metric nothing prevents one from using these metrics, the present study adapts the "*discrepancy*" term, as defined in Ref. [20], to quantify the accuracy of the model. The discrepancy term is derived from the following formulation:

$$y^{\mathrm{Test}} = y(p;\theta) + \delta(p) + \varepsilon^{\mathrm{Test}}, \tag{3}$$

where symbols $y^{\mathrm{Test}}(p)$, $y(p;\theta)$, $\delta(p)$ and $\varepsilon^{\mathrm{Test}}$ denote the physical measurements, model predictions, discrepancy term, and measurement (random) error. The measurement error usually takes the form of a zero-mean, Gaussian distribution, $\varepsilon^{\mathrm{Test}} \sim N(0;\sigma^{\mathrm{Test}})$, derived from repeated experiments (or replicates).

With the formulation of Eq. (3), one seeks the joint distribution of calibration parameters θ by comparing model predictions to experimental measurements. This is achieved through Bayesian inference in Ref. [20]. The role played by the discrepancy term is to capture residual differences between predictions and measurements that cannot be accounted for when calibration parameters are varied. Discrepancy, therefore, is a statistical process that represents model form error as opposed to the parametric uncertainty captured by parameters θ. The implementation of this procedure comes from Ref. [20] and has been perfected into a software package called Gaussian Process Modeling and Sensitivity Analysis (GPM/SA) at Los Alamos National Laboratory (see Refs. [21,22]).

Once the statistics of discrepancy $\delta(p)$ are computed, a metric for the accuracy attribute (A-3) that calculates the ratio of norms of discrepancy relative to prediction is proposed:

$$\delta_S = \frac{\|\delta(p)\|}{\|y(p;\theta)\|}, \tag{4}$$

where $||\delta(p)||$ denotes a user-defined norm of discrepancy and $||y(p;\theta)||$ denotes a similar norm of model predictions. Both norms must be evaluated over the domain of applicability. Examples of potential norms include using a range of values, variance of values, or the definition proposed in Ref. [2] that is based on a Principal Component Analysis (PCA) of datasets.

5. Functional and asymptotic properties of predictive maturity

So far it has been established that a metric of predictive maturity, that is referred to as the Predictive Maturity Index (PMI), should at a minimum be a function of coverage η_C, number of knobs N_K and goodness-of-fit δ_S:

$$\mathrm{PMI} = F(\eta_C; N_K; \delta_S), \tag{5}$$

where, without loss of generality, the value of PMI is bounded in the interval $0 \leqslant \mathrm{PMI} \leqslant 1$. The interpretation is intuitive: a value PMI = 0 means that the model has no predictive maturity. Likewise a value PMI = 1 means perfect maturity over the entire domain of applicability. Clearly, these two cases are asymptotes that cannot possibly be reached given a finite amount of simulation runs and test data.

Next, the mathematical properties of the PMI function are defined in Section 5.1 and limiting cases for instance, discrepancy reaching infinity, are discussed in Section 5.2.

5.1. Mathematical properties of predictive maturity

The mathematical properties defined here form a set of conditions that, any metric of predictive maturity should satisfy. In the following, these functional properties are labeled (C-1)–(C-4) for reference. Other conditions could, of course, be added but the main ones are captured in this list.

(C-1): The first property is that the value of a predictive maturity metric should increase when coverage increases. This can be written mathematically as a derivative with positive sign:

$$\frac{\partial \mathrm{PMI}}{\partial \eta_C} \geqslant 0. \tag{6}$$

The sign of the derivative is positive or zero to express the fact that one should take credit of adding more physical tests with better coverage of the overall domain of applicability, all other aspects of predictive accuracy being equal. The case "equal to zero" must be included because a test could be added that does not improve predictive maturity, for example, if it happens to be redundant with the tests already available.

(C-2): The second property is that the value of a predictive maturity metric should decrease when the number of knobs is increased. The sign of the partial derivative is, this time, negative:

$$\frac{\partial \mathrm{PMI}}{\partial N_K} \leqslant 0. \tag{7}$$

Adding more knobs, with everything else being equal, decreases predictive maturity. It means that simpler models, that tend to have fewer knobs, are considered more mature if they offer the same overall coverage and discrepancy as more complex models. The case of a derivative equal to zero must be included because knobs could be added that have no effect in terms of improving the goodness-of-fit and decreasing the discrepancy term.

(C-3): The third property is that the value of a predictive maturity metric should decrease when discrepancy increases. The sign of the partial derivative is strictly negative in this case:

$$\frac{\partial \mathrm{PMI}}{\partial \delta_S} < 0. \tag{8}$$

This constraint means that observing more discrepancy, with everything else being equal, results in a degradation of predictive maturity. The case "equal to zero" is, not allowed here because worsening prediction accuracy should always translate into less predictive maturity.

(C-4): Finally, a few purely mathematical properties may be considered. The function should be monotonic in its three arguments (δ_S; N_K; η_C). It should be sufficiently regular but this is not essential. (A discontinuous function, for example, could be defined.) For simplicity, a minimum of C^1 regularity is deemed to be necessary. These properties are included simply because it is easier to work with functions that obey these properties.

5.2. Asymptotic limits of predictive maturity

It is straightforward to define a mathematical function that satisfies properties (C-1)–(C-4). What is somewhat more challenging, however, is to satisfy the asymptotic limits. Once the PMI is defined as a generic function that is scaled in the interval [0; 1], the bounds "0" and "1" become asymptotic limits that should not be attained given finite amounts of data because there is no such thing as a model with either no or total predictive power. This reasoning leads to the four asymptotic cases discussed next and labeled (AL-1)–(AL-4) for convenience.

(AL-1): The first limit state examines the case when discrepancy tends to "infinity." As predictions of the model become increasingly more inaccurate, reaching a maximum of discrepancy relative to the overall range of predictions, or $\delta_S \to 1$, the predictive maturity deteriorates:

$$\lim_{\delta_S \to 1} \mathrm{PMI} = 0. \tag{9}$$

Eq. (9) states that maturity asymptotically reaches zero when accuracy decreases. This limit is independent of coverage of the domain of applicability.

(AL-2): The second limit state examines the case when the number of knobs tends to infinity. Predictive maturity should decrease because it becomes easier to match the test data with more calibration knobs. An extreme is reached when an infinite number of knobs are available, $N_K \to \infty$, because it is then possible to match any measurement. The limit case is:

$$\lim_{N_K \to \infty} \mathrm{PMI} = 0. \tag{10}$$

(AL-3): The third limit state addresses the case when coverage of the domain of applicability tends to zero, meaning that fewer and fewer experiments are available to assess the accuracy of predictions. Clearly, there cannot be any concept of predictability without test data, which translates into the following equation:

$$\lim_{\eta_C \to 0} \mathrm{PMI} = 0. \tag{11}$$

One could define a companion case to (AL-3) where $\eta_C \to 0$ to illustrate what happens if the coverage becomes "infinite." The case $\eta_C \to 1$ means that more and more experiments are available to fill the domain. However, it should not suffice to guarantee predictability because one still needs the model to be accurate. The asymptotic limit when $\eta_C \to 1$ must therefore be combined with the asymptotic limit of perfect accuracy, or $\delta_S \to 0$, in case (AL-4).

(AL-4): It is postulated that "perfect" predictive maturity can be reached with the combination of "perfect" accuracy, or $\delta_S \to 0$, and "infinite" coverage of the domain of applicability, or $\eta_C \to 1$:

$$\lim_{\delta_S \to 0, \eta_C \to 1} \mathrm{PMI} = 1. \tag{12}$$

Eq. (12) states that the limits $\delta_S \to 0$ and $\eta_C \to 1$ must occur simultaneously to get a value of maturity that tends to one. Reaching one limit but not the other one is not sufficient.

6. Proposal for a predictive maturity index

After having derived what is believed to be essential attributes and properties of a metric for predictive maturity, now it is necessary to find a mathematical function that satisfies the constraints (C-1)–(C-4) and asymptotic limits (AL-1)–(AL-4). There clearly are an infinite number of functions that satisfy these constraints. The point of this exercise is not to define "the" metric of maturity, but rather to provide an example for the mathematical implementation of the axioms enounced in Section 5.

The simplest possible function that verifies the afore-mentioned constraints is sought. Therefore, it is necessary to decouple the three attributes δ_S, N_K and η_C in the definition of the PMI. Remembering the constraints of monotonic behavior, a simple proposal is:

$$\mathrm{PMI} = (\eta_C) \times (N_K)^{-1} \times (\delta_S)^{-1}. \tag{13}$$

Eq. (13) verifies (C-1)–(C-4) but fails the asymptotic limits (AL-1) and (AL-4). After going through a few more steps, not reported here for brevity but explained in Ref. [2], we settle on the following definition of the PMI:

$$\mathrm{PMI}(\delta_S; N_K; \eta_C) = \eta_C \times \left(\frac{N_R}{N_K}\right)^{\gamma_1} \times (1 - \delta_S)^{\gamma_2} \times e^{(1-\eta_C^2)^{\gamma_3} - \delta_s^2}, \tag{14}$$

where γ_1, γ_2 and γ_3 are strictly positive, user-defined coefficients used to weight the effects of various contributions relative to the first one. Coefficient γ_1 weights the effect of the number of calibration knobs relative to coverage. Likewise coefficient γ_2 weights the effect of discrepancy relative to coverage. Coefficient γ_3 controls the coupling between discrepancy and coverage. Since it results from the multiplication of four components that are each scaled between zero and one, the PMI is automatically valued in $0 \leqslant \mathrm{PMI} \leqslant 1$. The introduction of a triplet of coefficients (γ_1; γ_2; γ_3) adds the flexibility to account for expert judgment in the definition of the PMI metric.

The symbol N_R of Eq. (14) represents a "reference" number of knobs. It is a characteristic number of calibration variables that one would expect to encounter in a class of similar models or codes. The ratio (N_R/N_K) helps define a non-dimensional variable for the number of knobs. It is needed because the concept of "complexity" of a model should always be relative to the current state-of-the-art. Hence, variable N_R represents the complexity of the state-of-the-art while N_K represents the complexity of the model that one is assessing.

Fig. 4 illustrates a "2D slice" of the PMI function obtained when the number of knobs is kept constant. Coefficients $\gamma_1 = 0$, $\gamma_2 = 1/4$ and $\gamma_3 = 2$ are used to produce the figure that shows the combined effect of discrepancy δ_S and coverage η_C. The PMI reaches "perfect" predictability, or PMI = 1, only when $(\delta_S; \eta_C) \to (0; 1)$. Predictive maturity then decreases either as coverage reduces or discrepancy increases. The function can be optimized though a judicious choice of coefficient γ_2 to lower the predictive maturity slowly-but-steadily as more discrepancy between measurements and predictions is observed. This should be true unless coverage of the domain of applicability is significantly reduced, at which point maturity rapidly decreases to zero.

7. Application to the maturity of a material model

An application of the previously derived PMI metric is presented. The goal is to assess the maturity of a material model that can be applied to physical tests in low-to-high regimes of strain rates and temperatures encountered when simulating the performance of novel nuclear fuels of the US Global Nuclear Energy Partnership (GNEP) program. The material model of interest is the

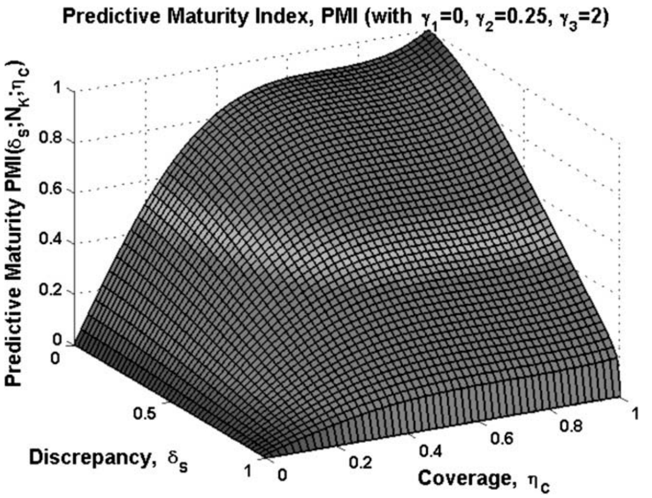

Fig. 4. Predictive maturity PMI(δ_S; N_K; η_C) with coefficients $\gamma_1 = 0$, $\gamma_2 = ¼$ and $\gamma_2 = 2$.

Preston–Tonks–Wallace (PTW) model of plastic deformation documented in Ref. [23]. The PTW model is applied to the light-weight, high-strength beryllium (Be) metal.

7.1. The Preston–Tonks–Wallace model of plastic deformation

The PTW model of plasticity describes strain–stress curves obtained at various regimes of strain rate and temperature. It models plastic flows for metals and is suitable to simulate material response to fast transients such as those from explosive loading or high velocity impacts. The main equations of the PTW model of plasticity from Ref. [23] are:

$$\sigma_S = s_0 - \frac{2(s_0 - s_\infty)}{\sqrt{\pi}} \int_0^{\kappa T \log(\gamma\dot{\varepsilon}/\theta)} e^{-\lambda^2}\, d\lambda \quad \text{and}$$

$$\sigma_Y = y_0 - \frac{2(y_0 - y_\infty)}{\sqrt{\pi}} \int_0^{\kappa T \log(\gamma\dot{\varepsilon}/\theta)} e^{-\lambda^2}\, d\lambda, \tag{15}$$

where symbols σ_S and σ_Y denote the dimensionless work hardening saturation stress and yield stress, respectively. The material specific coefficient λ determines the hardening behavior. Control parameters that define the two-dimensional domain of applicability for this application are the strain rate ($d\varepsilon/dt$) and temperature (T) of Eq. (15). Symbols θ, κ, γ, s_0, s_∞, y_0 and y_∞ are seven, also dimensionless, calibration variables defined in Table 1 for the Be metal. The ranges listed in the table are large, which justifies the need to resort to statistical inference from test datasets to reduce the variability.

Experimental datasets collected from Hopkinson bar tests performed on samples of Be material at varying strain-rates and temperatures are used. The goal of this exercise is to obtain a PTW model of plasticity that accurately describes datasets collected over the two-dimensional domain of control parameters. For this, the joint probability distribution of seven ancillary variables (θ; κ; γ; s_0; s_∞; y_0; y_∞) must be inferred from the measurements of strain–stress curves. The GPM/SA software used in this study not only achieves this task, but also the delivers an independent estimate of the discrepancy term $\delta(p)$ of Eq. (3).

7.2. Description of experimental datasets

The Hopkinson bar experiments performed on samples of Be metal yield the strain–stress curves illustrated in Fig. 5. Measurement error is modeled as a Gaussian process with zero mean and 2.5% variance of the measured stress, as described by the experimentalists. Table 2 lists the values of control parameter pairs ($d\varepsilon/dt$; T) that correspond to each Hopkinson bar test. The table also lists the maximum, measured strain ε_{Max}. Fig. 5 illustrates the range of stresses and variety of shapes that the PTW model is expected to reproduce over the domain of applicability.

Fig. 6 illustrates the location of physical experiments in the domain of applicability ($d\varepsilon/dt$; T). The blue, square symbols illustrate the location of Hopkinson bar experiments performed within the domain of applicability $\{\Omega_V\}$. The domain is a two-dimensional, hyper-cube defined as $\{\Omega_V\} = [10^{-4}; 4.10^{+3}] \times [70; 600]$ K s^{-1}. The convex hull $\{\Omega_{CH}\}$ is shown with the dashed line.

Fig. 6 shows that most physical tests are performed at the "edge" of the domain $\{\Omega_V\}$ and a few fall "inside," such as tests 10 and 11. The logic of this design of experiments is to locate Hopkinson bar tests, as much as possible, at the "edges" of the domain with tests 10 and 11 added at the "center." It offers a good balance between overall coverage and number of tests.

7.3. Statistical inference of calibration parameters and model discrepancy

To characterize uncertainty, the software GPM/SA explores the joint probability distribution of calibration variables (θ; κ; γ; s_0; s_∞; y_0; y_∞) with a Markov-chain random walk that is based on a simple but effective principle: predictions that better match the measurements originate from combinations of calibration variables that tend to be visited more frequently by the random walk. After performing a sufficient number of MCMC iterations, selected to be 10,000 here, statistics of calibration variables visited are computed to estimate the (unknown) joint probability distribution. Such posterior distribution characterizes the uncertainty of the PTW material model.

The procedure is initiated by performing a design of computer experiments. An orthogonal Latin Hypercube Sample (LHS) using a number of runs equal to 10 times the number of control parameters and calibration variables is analyzed. It leads to $10 \times (2 + 7) = 90$ runs, which is a small computational overhead cost to pay prior to initiating the statistical analysis. For each one of these runs, the PTW model is evaluated using a given combination of control parameters ($d\varepsilon/dt$; T) and calibration variables (θ; κ; γ; s_0; s_∞; y_0; y_∞), and the corresponding strain–stress curve is predicted. The collection of these 90 curves, together with the

Table 1
Definition of free dimensionless parameters of the PTW model for the Be metal.

Symbol	Description	Minimum	Maximum
θ	Initial strain hardening rate	0.009979	0.0480590
κ	Material constant in thermal activation energy term (relates to the temperature dependence)	0.013516	0.4901500
γ	Material constant in thermal activation energy term (relates to the strain rate dependence)	−22.15299	7.4708000
y_0	Minimum yield stress (at $T = 0$ K)	0.001054	0.0021643
y_∞	Maximum yield stress (at $T \approx$ melting)	0.000194	0.0016100
s_0	Minimum saturation stress (at $T = 0$ K)	0.002493	0.0480680
s_∞	Maximum saturation stress (at $T \approx$ melting)	0.000599	0.0080031

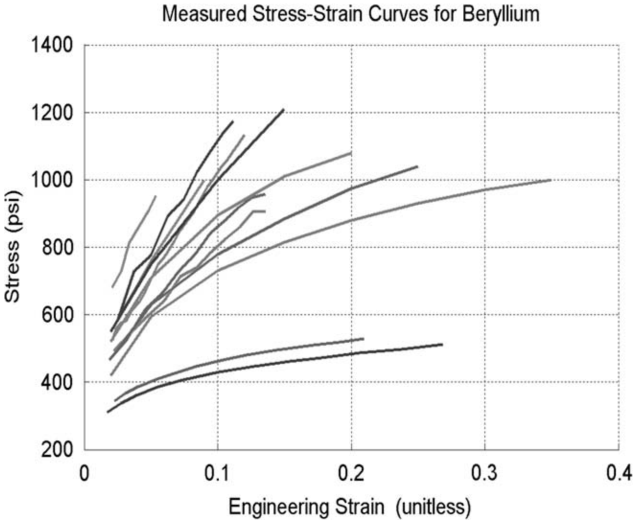

Fig. 5. Strain–stress curves measured from Hopkinson bar tests of Be metal.

Table 2
Definition of settings for experiments performed on Be samples.

Dataset	Maximum strain, ε_{Max}	Temperature, T (K)	Strain-rate, $d\varepsilon/dt$ (s^{-1})
1	0.0539	77.0	3000.0
2	0.1118	223.0	3500.0
3	0.1202	298.0	3500.0
4	0.1355	473.0	3700.0
5	0.1360	573.0	3900.0
6	0.2100	573.0	1.0
7	0.2689	473.0	0.001
8	0.090	293.0	2000
9	0.150	293.0	950.0
10	0.200	293.0	2.0
11	0.250	293.0	0.02
12	0.350	293.0	0.0001

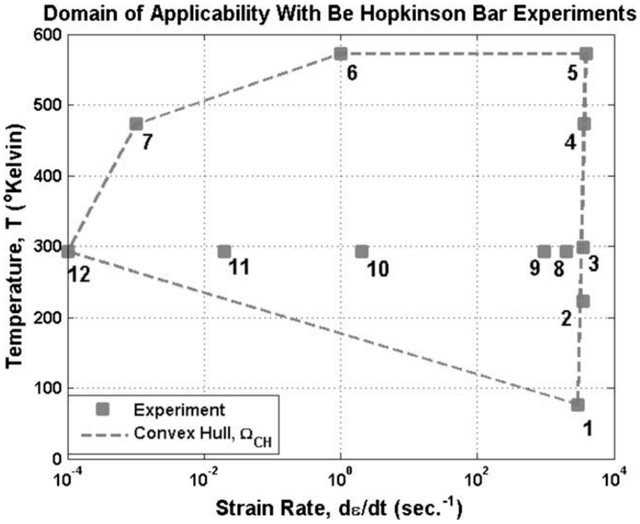

Fig. 6. Domain of applicability and coverage of experiments for the Be metal.

12 measurements of Fig. 5 and statistical model of measurement error, form the datasets fed to GPM/SA.

For computational efficiency, GPM/SA does not operate directly on strain–stress curves. Instead it uses their Principal Component Decomposition (PCD). The PCD is a mathematical transform that finds an orthogonal basis to describe the sub-space defined by multiple curves. The PCD is computed numerically from a Singular Value Decomposition (SVD). Implementation details can be found in Refs. [21,22]. For this application, it is deemed sufficient to operate on the first three PCD modes that capture over 95% of the strain–stress information. This projection reduces the computational complexity from the $N = 150$ samples of strain–stress values shown in Fig. 5 to only three principal mode components.

Even though this procedure is reminiscent of parametric calibration, such as finite element model updating [24–25], it is different in two major ways. First, instead of optimizing the calibration variables such that predictions match measurements, the aim is to infer from test-analysis correlation a description of the uncertainty with which the Be metal can be described. Such philosophy departs from calibration or ‘‘knob tuning,” even though some of the tools used are similar. The second significant departure from the calibration paradigm is that prediction accuracy is, here, assessed via the discrepancy term as opposed to test-analysis correlation error.

7.4. The hypothesis of improvement of predictive maturity

Even though it could be performed once using the 12 datasets defined in Table 2, the entire procedure is repeated nine times by progressively adding one experiment to datasets fed to the GPM/SA software. Table 3 defines these nine cases. Tests 8 and 9 are not used because they are located near test 3 and, therefore, do not improve coverage, as seen in Fig. 6.

The reason to define nine separate cases, as opposed to performing a single analysis using all 12 experiments, is to assess the effect that increasing the number of physical tests has on predictive maturity. *Our working hypothesis is that increasing the number of physical tests generally improves predictive maturity*. Of course, this assertion should be true as long as additional tests increase coverage of the domain while not deteriorating prediction accuracy too much. Results discussed in the remainder illustrates that the above hypothesis is reasonable. Specifically, they illustrate the working hypothesis that predictive maturity can eventually be reached.

‘‘Maturity,” does not necessarily imply that predictions of the model match exactly the measurements throughout the domain of applicability. Instead, it means that the statistical description of discrepancy eventually converges to a ‘‘*stable asymptote*” as more and more testing datasets are fed to the analysis.

7.5. Quantitative analysis of predictive maturity of the PTW model

In this section, results obtained by applying the previously described procedure, are described. The first objective is to quantify the extent to which the PTW model proposes a mature capability to simulate the strength behavior of Be metal. The second objective is to illustrate the hypothesis of Section 7.4 according to which

Table 3
Definition of the nine sets of Hopkinson bar experiments for the Be metal.

Case	List of experiments	Coverage, η_C (%)
1	1, 2	3.11
2	1, 2, 3	4.62
3	1, 2, 3, 4	8.17
4	1, 2, 3, 4, 5	9.68
5	1, 2, 3, 4, 5, 6	34.16
6	1, 2, 3, 4, 5, 6, 7	55.55
7	1, 2, 3, 4, 5, 6, 7 and 10	55.55
8	1, 2, 3, 4, 5, 6, 7 and 10, 11	62.74
9	1, 2, 3, 4, 5, 6, 7 and 10, 11, 12	74.25

Table 4
Predictive maturity for the nine sets of experiments (PTW model of Be metal). PMI values are obtained from Eq. (14) where $\gamma_1 = ½$, $\gamma_2 = ¼$, $\gamma_3 = 2$ and $N_R = 5$.

Case	Coverage, η_C (%)	Number of knobs, N_K	Discrepancy, δ_S (%)	PMI metric (%)
1	3.11	7	7.11	6.96
2	4.62	7	7.21	10.32
3	8.17	7	1.55	18.44
4	9.68	7	0.63	21.80
5	34.16	7	9.72	60.83
6	55.55	7	11.82	72.36
7	55.55	7	11.23	72.58
8	62.74	7	12.01	73.12
9	74.25	7	16.85	71.24

additional experimental datasets translate into an improvement of maturity.

Table 4 lists the values of coverage η_C, number of calibration variables N_K, scaled discrepancy δ_S and PMI for each one of the nine cases. Values of the PMI are computed from Eq. (14) with $\gamma_1 = ½$, $\gamma_2 = ¼$, $\gamma_3 = 2$ and $N_R = 5$. A reference of five calibration variables, that is, $N_R = 5$, is used because this number is typical of material models of plasticity in solid mechanics such as, for example, Johnson–Cooke or Zerilli–Amstrong. The scaled discrepancy metric δ_S is defined as a ratio of maximum singular values:

$$\delta_S = \frac{\sigma_{\text{Max}}(\delta(p))}{\sigma_{\text{Max}}(y(p;\theta))}, \tag{16}$$

where the operator $\sigma(\bullet)$ denotes the maximum singular value computed when all strain–stress curves of either discrepancy, that is, $\delta(p)$, or prediction, that is, $y(p;\theta)$, are decomposed with the SVD algorithm (see Ref. [2] for details). The maximum singular value is a convenient metric because it possesses the physical units of stress and it is independent of the number of curves analyzed for each case defined in Table 3. It is emphasized that the discrepancy $\delta(p)$ is estimated only for those experiments that define each case in Table 3. For example, in case 1 that uses experiments 1 and 2, $\delta(p)$ is obtained for experiments 1 and 2 only.

For this application, values for the triplet $(\gamma_1; \gamma_2; \gamma_3)$ are chosen to achieve equal "weight" for the individual effects. Since the number of calibration variables is the same for all cases considered, the selection of $\gamma_1 = ½$, or any other value, scales the PMI without changing the overall trend. We started with a nominal value of $\gamma_2 = ¼$ for the discrepancy term. To achieve equal importance between discrepancy and coverage, the value of γ_3 had to be no less than 2. Other experts may chose to weight these attributes differently, in which case the triplet $(\gamma_1; \gamma_2; \gamma_3)$ would change. This was intended to let subject matter experts incorporate their judgment.

It can be seen in Table 4 that coverage increases from 3.11% to nearly 75% of the domain of applicability. Discrepancy starts at 7.11% and decreases to 0.63% when test 5 is added to the datasets. Tests 4 and 5 are experiments performed at high temperatures; they are essential in terms of predictability of the PTW model, which explains the improvement of accuracy obtained when they are included in the datasets analyzed. Then it becomes more difficult to match the variety of settings obtained when new experiments are added at low strain-rates. It results that discrepancy increases again from 0.63% (case 4) to nearly 17% (case 9). The overall level of accuracy and discrepancy remain, however, acceptable for this application. Overall the maturity of the PTW model is deemed excellent to predict the strength behavior of beryllium.

Fig. 7 presents the same results as Table 4 but it better conveys the "stabilization" of the PMI metric as coverage is increased. The overall trend is that adding tests increases coverage while making it somewhat more difficult for the model to maintain its accuracy. The maturity metric nevertheless stabilizes, which illustrates that our working hypothesis ("*predictive maturity can be reached if enough experiments are analyzed*") is reasonable.

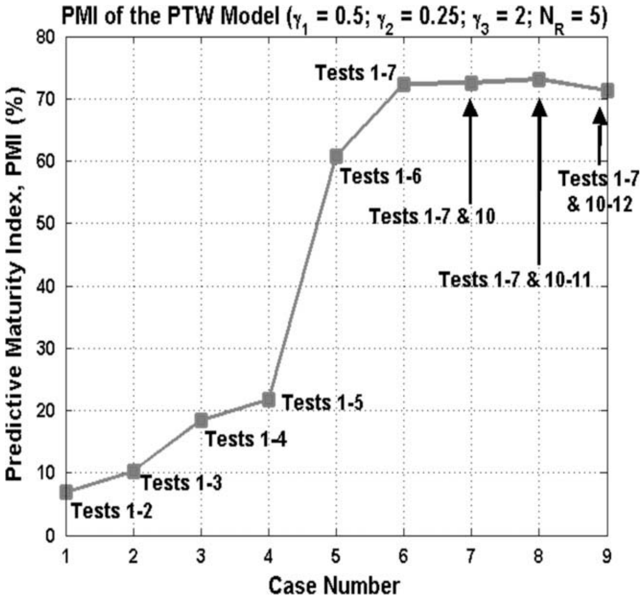

Fig. 7. Predictive maturity index of the PTW model with Be material datasets.

Even though this application indicates that the PMI is (almost monotonically) increasing as more physical tests are analyzed, it should not necessarily be the case for all applications. Predictive maturity could be stationary or decrease if the physical experiments added do not improve coverage; if they happen to be non-informative in terms of constraining the values of calibration variables; or if predictions of the model become inaccurate for the new test. Rather than defining a single metric, this work has attempted to identify these mechanisms and propose a framework that could take them into account. One has the flexibility, for example, of adjusting coefficients $(\gamma_1; \gamma_2; \gamma_3; N_R)$ of Eq. (14) to enhance, or reduce, the sensitivity of these effects.

8. Conclusion

This report documents an effort to define a quantitative metric to assess predictive maturity. The need to formulate a commonly accepted definition of predictive maturity comes from the ever increasing reliance on computer simulations in decision-making. The concept of predictive maturity involves the definition of quantitative metrics that could prove useful to allocate resources for physical testing and code development. Such metrics should be able to track progress (or lack thereof) as additional knowledge becomes available. Predictive maturity metrics are also needed to quantify the degree of confidence that decision-makers can place in decisions that are supported by numerical simulations.

It is argued herein that the assessment of predictive maturity must go beyond the goodness-of-fit of the model to the available test data. Maturity must consider the "knobs," or ancillary variables, used for calibration and, also, the degree to which physical tests cover the domain of applicability of the model or code. Metrics are proposed for three attributes of predictive maturity: goodness-of-fit, complexity and coverage. These attributes are used to define the mathematical properties and asymptotic limits of predictive maturity. Based on these constraints, a proposal is made for a PMI. Performance of the PMI is studied to quantify the maturity of a material model applied to Hopkinson bar tests performed on samples of beryllium metal.

The application illustrates that the working hypothesis, that is, *increasing the number of tests generally improves predictive maturity*, is founded. Sufficient maturity can be reached if a numerical model captures the "right" physics and remains capable of yielding accurate predictions throughout the domain of applicability. These findings open the door to a decision-making strategy that is based on quantifying prediction discrepancy over the domain of applicability and demonstrating that the discrepancy is *stable*. Stability is essential because it demonstrates that predictions of the numerical model are robust to assumptions that go into the calculation.

Acknowledgments

This work is performed under the auspices of the Validation and Uncertainty project of the Nuclear Energy Advanced Modeling and Simulation (NEAMS) program at Los Alamos National Laboratory (LANL). at Los Alamos National Laboratory (LANL). The first two authors are grateful to Cetin Unal, M&S project leader, for his support and technical leadership. LANL is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the US Department of Energy under Contract DE-AC52-06NA25396.

References

[1] DOE, Fuels modeling and simulation roadmap, Planning document GNEP-FUEL-SYSE-TD-PL-2008-073, Office of Nuclear Energy of the US Department of Energy, Global Nuclear Energy Partnership, Transmutation Fuel Campaign, Washington, DC; February 2008.

[2] Hemez FM, Atamturktur S, Unal C. Defining predictive maturity for validated numerical simulations. Technical report LA-UR-08-6741 of the Global Nuclear Energy Partnership (GNEP) Program, Los Alamos National Laboratory, Los Alamos, New Mexico; September 2008.

[3] Trucano TG, Easterling RG, Dowding KJ, Paez TL, Urbina A, Romero VJ et al. Description of the Sandia validation metrics project. Technical report SAND-2001-1339, Sandia National Laboratories, Albuquerque, New Mexico; July 2001.

[4] Oberkampf WL, Barone MF. Measures of agreement between computation and experiment: validation metrics. J Comput Phys 2006;217:5–36.

[5] Iuzzolino HJ, Oberkampf WL, Barone MF, Gilkey AP. User's manual for VALMET: validation metric estimator program. Technical report SAND-2007-6641, Sandia National Laboratories, Albuquerque, New Mexico; November 2007.

[6] Balci O, Adams RJ, Myers DS, Nance RE. Credibility assessment: a collaborative evaluation environment for credibility assessment of modeling and simulation applications. In: Proceedings of the 34th winter simulation conference: exploring new frontiers, San Diego, California; 2002. p. 214–20.

[7] Sornette D, Davis AB, Kamm JR, Ide K. A general strategy for physics-based model validation illustrated with earthquake phenomenology, atmospheric radiative transfer and computational fluid dynamics. In: Proceedings of the Lawrence Livermore National Laboratory workshop on computational methods in radiation and particle transport, Lake Tahoe, California, September 9–14. Berlin: Springer; 2006.

[8] Sornette D, Davis AB, Ide K, Kamm JM. Theory and examples of a new approach to constructive model validation. In: NATO RTA AVT-147 symposium on computational uncertainty, Athens, Greece, December 3–6, 2007. (Also, Technical report LA-UR-07-7013, Los Alamos National Laboratory, Los Alamos, New Mexico; 2007.)

[9] Logan RW, Nitta CK, Chidester SK. Verification and validation: process and levels leading to qualitative or quantitative validation statements. Technical report UCRL-TR-2001-31, Lawrence Livermore National Laboratory, Livermore, California; October 2003.

[10] Logan RW, Nitta CK. Verification and validation: goals, methods, levels and metrics. Technical report UCRL-PRES-153252, Lawrence Livermore National Laboratory, Livermore, California; July 2003.

[11] Logan RW, Nitta CK. Verification and validation methodology and quantitative reliability at confidence (QRC): basis for an investment strategy. Technical report UCRL-ID-150874, Lawrence Livermore National Laboratory, Livermore, California; November 2002.

[12] Oberkampf WL, Pilch M, Trucano TG. Predictive capability maturity model for computational modeling and simulation. Technical report SAND-2007-5948, Sandia National Laboratories, Albuquerque, New Mexico; October 2007.

[13] Zang TA. Perspectives on uncertainties (and margins) in NASA engineering decisions. In: Proceedings of the 10th AIAA non-deterministic approaches conference, Schaumburg, Illinois; April 7–10, 2008.

[14] NASA, Standard for models and simulation. Technical standard NASA-STD-7009, National Aeronautics and Space Administration, Washington, DC; November 2007.

[15] Roache PJ. Verification in computational science and engineering. Albuquerque (NM): Hermosa Publishers; 1998.

[16] Hemez FM, Doebling SW, Anderson MC. A brief tutorial on verification and validation. In: Proceedings of the 22nd SEM international modal analysis conference, Dearborn, Michigan; January 26–29, 2004.

[17] ASME, Guide for verification and validation in computational solid mechanics, Publication V&V-10-2006. American Society of Mechanical Engineers; 2006.

[18] Hemez FM. Answering the question of sufficiency: how much uncertainty is enough? In: Proceedings of the first international conference on uncertainty in structural dynamics, University of Sheffield, United Kingdom; June 11–13, 2007. (Also, Technical report LA-UR-07-3575, Los Alamos National Laboratory, Los Alamos, New Mexico; 2007.)

[19] Hemez FM, Ben-Haim Y. Info-gap robustness for the correlation of tests and simulations of a non-linear transient. In: Mechanical systems and signal processing, vol. 18; March 2004. p. 1443–67. (Also, Technical report LA-UR-02-3538, Los Alamos National Laboratory, Los Alamos, New Mexico; 2002.)

[20] Kennedy M, O'Hagan A. Predicting the output from a complex computer code when fast approximations are available. Biometrika 2000;87:1–13.

[21] Higdon D, Gattiker J, Williams B, Rightley M. Computer model calibration using high-dimensional output. J Am Stat Assoc 2008;103(482):570–83.

[22] Higdon D, Nakhleh C, Gattiker J, Williams B. A Bayesian calibration approach to the thermal problem. Comput Meth Appl Mech 2008;197:2431–41.

[23] Preston DL, Tonks DL, Wallace DC. Model of plastic deformation for extreme loading conditions. J Appl Phys 2003;93(1):220–1.

[24] Mottershead JE, Friswell MI. Model updating in structural dynamics: a survey. J Sound Vib 1993;167(2):347–75.

[25] Hemez FM, Doebling SW. Review and assessment of model updating for nonlinear, transient dynamics. Mech Syst Signal Process 2001;15(1):45–74 [LA-UR-00-0091].