LA-UR-08- 08-7313

Approved for public release; distribution is unlimited.

Title:	PREDICTIVE MATURITY OF COMPUTER MODELS USING FUNCTIONAL AND MULTIVARIATE OUTPUT *(Manuscript)*
Author(s):	H. Sezer Atamturktur, Los Alamos National Laboratory, X-3 François M. Hemez, Los Alamos National Laboratory, X-3 Cetin Unal, Los Alamos National Laboratory, X-4 Brian Williams, Los Alamos National Laboratory, CCS-6
Intended for:	Proceedings of the 27th SEM International Modal Analysis Conference (IMAC-XXVII), Orlando, Florida, February 9-12, 2009

Los Alamos
NATIONAL LABORATORY
EST. 1943

Form 836 (7/06)

This page is left blank intentionally.

Predictive Maturity of Computer Models Using Functional and Multivariate Output

Sezer Atamturktur,[1] François Hemez,[2] Cetin Unal,[3] Brian Williams[4]

Los Alamos National Laboratory, Los Alamos, New Mexico 87545

ABSTRACT: Computer simulations are valued in science and engineering because they enable us to gain knowledge about phenomena that would otherwise be difficult to understand. Dependency on simulations primarily stems from our inability to conduct a sufficient number of experiments within the desired settings or with sufficient detail. However, if one were able to conduct a large-enough number of experiments, it is reasonable to envision that a simulation model could be calibrated to the point that its predictive uncertainty is reduced down to uncontrolled, natural variability. We inductively conclude that, as new experimental information is used for calibration, the calibrated parameters should stabilize, and thus, the disagreement between simulations and experiments should be reduced down to "true" bias. We propose to use the stabilization of the incremental improvement to assess the predictive maturity of a model. Accordingly, we develop a Prediction Convergence Index (PCI) that approximates the convergence of predictions to their "true" or stabilized values or, conversely, can be used to estimate the number of experimental tests that would be required to reach stabilization of predictions. Once the predictive maturity of a model has been assessed, we argue that it is acceptable to extrapolate its predictions away from settings or regimes where validation tests have been conducted as long as the physics involved and modeled by the code remains unchanged. The application of the PCI is illustrated using a Preston-Tonks-Wallace material model for Tantalum and six experimental datasets in the form of strain and stress cures. For the given model, the extent to which extrapolation and interpolation are acceptable is investigated. The results agree with our hypothesis and suggest that the approach proposed can prove useful for claiming completion of the calibration phase and providing insight into the predictive maturity of numerical models. *(Approved for unlimited, public release on November xxx, 2008, LA-UR-08-xxxx, Unclassified.)*

1. INTRODUCTION: WHY IS PREDICTIVE MATURITY IMPORTANT?

We firmly believe that the use of computer simulations to support high-consequence decision-making that affects environment, health, safety and security, will continue to grow. The demand to assess the accuracy and uncertainty of model predictions is the driving force behind activities collectively referred to as Verification and Validation (V&V). Verification, in simple terms, deals with the consistency between numerical solutions of a code and exact solutions of continuous partial differential equations; validation deals with fidelity of a code to real-world situations. An intrinsic component of the V&V process is model calibration that attempts,

[1] Graduate student in X-Division and Ph.D. candidate, The Pennsylvania State University. Mailing: Los Alamos National Laboratory, X-3, Mail Stop P365, Los Alamos, NM, 87545. Phone: 505-665-2613. E-mail: sezer@lanl.gov.

[2] Technical Staff Member in X-Division. Mailing: Los Alamos National Laboratory, X-3, Mail Stop B259, Los Alamos, NM 87545. Phone: 505-667-4631. E-mail: hemez@lanl.gov.

[3] Technical Staff Member in X-Division. Mailing: Los Alamos National Laboratory, X-4, Mail Stop T086, Los Alamos, NM 87545. Phone: 505-665-2539. E-mail: cu@lanl.gov.

[4] Technical Staff Member in CCS-Division. Mailing: Los Alamos National Laboratory, CCS-6, Mail Stop F600, Los Alamos, NM 87545. Phone: 505-667-2331. E-mail: brianw@lanl.gov.

through the comparison of simulation predictions (running a computer code) and physical observations (collecting measurements), to gain a better understanding of imprecise model parameters and inadequate physics in the simulation. For calibration to be meaningful, the quantification of uncertainty both in simulation predictions and physical observations must be an integral part of the process. Calibration is particularly demanding on resources as it requires that large numbers of computational solutions and experimental measurements be obtained.

In the conventional realm of V&V, calibration does not have a clear definition of completion other than *"the predictions must match the measurements perfectly."* It means that there is no clear definition of ***sufficiency*** for the number of calibration experiments crucial to reach a model of desired accuracy. The notion of completion is usually established according to other factors driven by non-scientific constraints. They include, for instance, budgetary and time constraints. In contrast with this conventional approach, a scientifically rigorous criterion capable to evaluate progress in predictive maturity throughout the phases of calibration would allow one to allocate resources more intelligently. Such a criterion would be particularly useful in the context of multi-scale, multi-physics models where needs in terms of experimentation to support calibration could rapidly overwhelm the resources available.

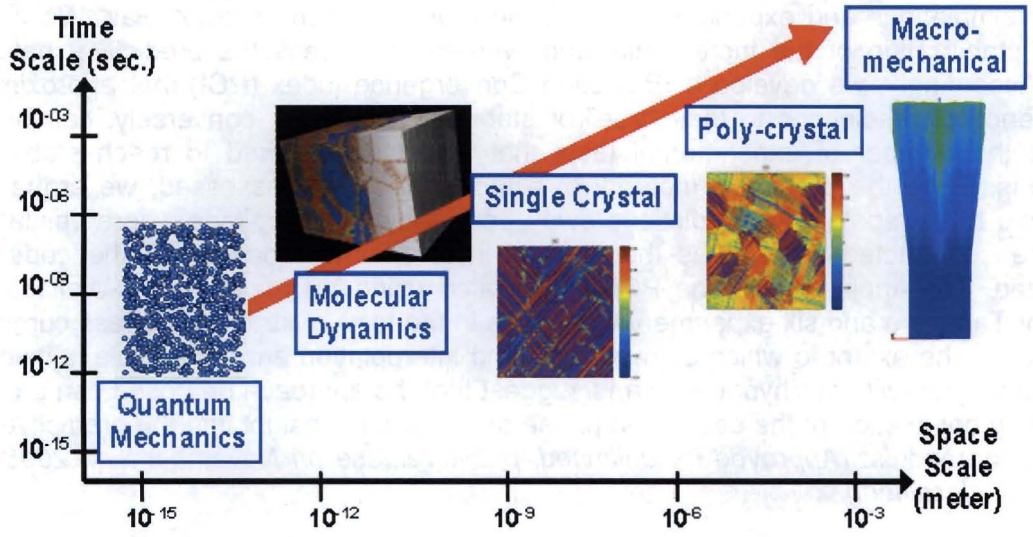


Figure 1. The typical space and time scales of various scientific modeling disciplines.

Figure 1 illustrates typical space and time scales needed to represent material behavior from quantum mechanics and molecular dynamics to the macroscopic, mechanical behavior. Such multi-scale models are needed to simulate, for example, the performance of new generations of nuclear fuels in support of the Global Nuclear Energy Partnership (GNEP) program of the U.S. Department of Energy.

For instance, molecular dynamics simulations require the resolution of non-linear systems of equations that may count up to 10^{+8} degrees-of-freedom over time with scales that range from 10^{-12} to 10^{-1} sec. Engineering descriptions of material behavior, on the other hand, must address millimeter to meter-size problems to simulate, for example, the performance of a nuclear fuel under various settings of temperature and irradiation in a reactor. Aside from adequate physics-based codes and sufficient computing resources, these applications demand large validation datasets collected over vastly different space and time scales to, among other things, calibrate many model parameters. Without criteria to select which physical experiments are the most useful in assessing the sufficiency of the number of experiments performed, the physics-based, multi-scale approach to modeling and simulation is doomed to failure.

Because of the unavoidable sparsity of experimentation, models are usually calibrated and validated using physical tests performed for a limited number of settings. Once the cycles of verification, calibration and validation are completed, the models are then applied to predict at settings or regimes other than those used for validation. Projection can be achieved by, first, constructing a function that closely fits the predictions of a validated model at discrete settings. The best-fitted function is then exercised to ***forecast***, that is, make predictions with quantified uncertainty bounds, at new settings that may not be located within the domain of validation. For our study, we fit a Gaussian Process Model (GPM) to discrete datasets. Interpolation generally refers to making predictions between, or within, the tested settings while extrapolation is a forecasting estimation outside the domain of validation. Figure 2 illustrates, for a 1D curve, the difference between extrapolation and interpolation.

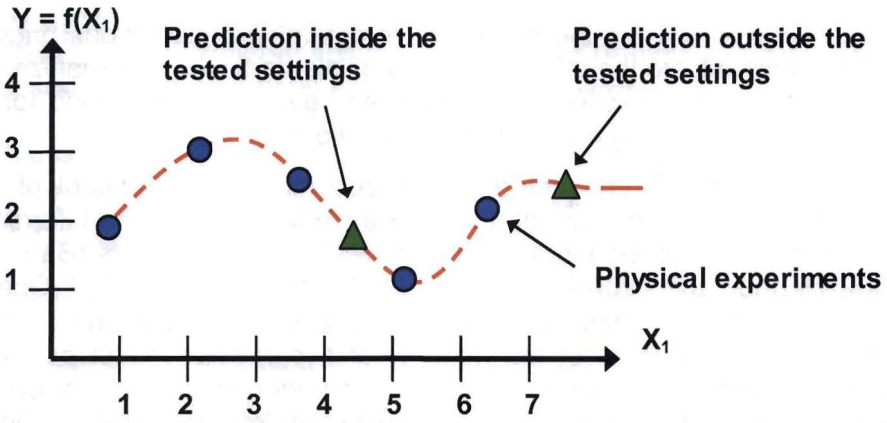


Figure 2. Illustration of interpolation and extrapolation predictions.
(In this illustration, interpolation means estimating an untested setting within the bounds of the tested domain of validation, while extrapolation means estimating an untested setting outside the bounds of the tested domain.)

The concepts of interpolation and extrapolation for science-based predictions have been a matter of heated debate in the scientific community for many decades. Although, mathematically speaking, there is no particular difference between interpolation and extrapolation, the idea of making predictions in an extrapolating mode is opposed by many. The primary reason for the opposition is that, while interpolation has clear upper and lower bounds, the bounds in which extrapolation is acceptable lack a clear definition. If extrapolation is stretched far enough from the domain of validation, then errors (bias and uncertainty) of the best-fitted model are likely to reach unacceptable levels. We argue that it is acceptable to extrapolate predictions (to a certain extent) away from the settings or regimes where validation tests have been conducted as long as the predictive maturity of the model has been assessed, prediction uncertainty is quantified, and the physics involved and modeled by the code remains unchanged. The latter requirement is equivalent to assuming that the physics modeled in the code are "smoothly varying" in that they do not change significantly from the validation regime to the extrapolation regime.

In this work, we investigate the extent to which extrapolation and interpolation are acceptable for a Preston-Tonks-Wallace (PTW) material model of plasticity for the Tantalum metal, using test data in the form of stress-strain curves at various settings of temperature and strain rate. The PTW model predicts the strain-stress behavior of Tantalum given seven material-dependent parameters. Typically these parameters are inferred from the experimental datasets through a calibration procedure. The experiments, although imperfect, are reflections of reality and, as more experimental datasets become available, the calibrated model should stabilize to an acceptable representation of reality.

As discussed in Section 4, we argue that predictive maturity achieved through calibration must be judged on the basis of stabilization of prediction errors with increased numbers of physical experiments. We define ***prediction error*** as the error that occurs when the model is used to predict at untested settings, in other words, when prediction is made in an extrapolating mode (or forecasting). Therefore, it is reasonable to expect that prediction error should be reduced and stabilized with increased coverage of the validation domain. Observing this trend would support the argument that extrapolation beyond untested regimes is acceptable for mature model. The proposed approach can prove useful for claiming completion of the calibration phase and providing insight into the predictive maturity of numerical models.

2. EARLIER CONTRIBUTIONS TO THE CONCEPT OF PREDICTIVE MATURITY

Methods developed to assess the quality of computational predictions originate from the need to effectively communicate the information to multiple parties in an organization, especially to decision makers. Several predictive maturity scales have been developed for this purpose, with independent efforts originating from different institutions.

An approach that is commonly encountered is to decompose the concept of credibility into several components. For each component, the qualitative and quantitative information available is subjectively evaluated and assigned scores by multiple team members (Balci, Adam, Myers and Nance, 2002; Harmon and Youngblood, 2003-2005; Blattnig, et al., 2007; Oberkampf, Pilch, and Trucano, 2007). Although valuable in organizing and communicating information, these maturity scales are not a measure of the maturity of a particular model ***per se***; rather their intended purpose is to measure the maturity of activities undertaken for model development and V&V. Such an approach is based on the basic premise that improved rigor in the V&V process will directly correlate to improved predictive maturity in the simulation model. However in real-life applications, there is no guarantee that increasing these efforts will necessarily yield increased confidence in model predictions. Also, since credibility is subjective, there is the risk of different decision makers reaching various conclusions when provided with the same data.

This paper takes a step towards removing the subjective content from the evaluation of model maturity. A quantitative method is missing in the literature, yet, it is of high importance because it would enable us to allocate time and budgetary resources more effectively.

3. METHODOLOGY TO USE EXPERIMENTAL DATA TO IMPROVE A MODEL

In model calibration, we attempt to improve the predictive capability of an initially inaccurate computer simulation through comparisons of a single (univariate) or multiple (multivariate) of its solutions with incomplete and imprecise physical measurements. Traditionally, model calibration strategies have two types that differ in the formulation through which they improve the models. The first type is the ***parameter calibration approach*** that captures the inaccuracy of the model parameters. Most finite element updating techniques, for example, are parameter calibration approaches that optimize the values of model parameters to minimize a cost function of test-analysis correlation.

The second type is the ***bias correction approach*** that captures the inadequacy of the physics model. If a model is missing a component that is significant in terms of representing the physics involved, then its predictions tend to exhibit a systematic bias from the physical measurements even at the "true" values of model parameters. In this case, calibration activities cannot reduce the disagreement between predictions and measurements any further. The systematic bias caused by the missing or inaccurate implementation of physics in the model is dealt with the bias correction approach.

These two fundamental concepts are combined together in the context of Bayesian inference in the landmark study of Kennedy and O'Hagan (2000, 2001). Their approach can simultaneously calibrate model parameters and correct bias. The method not only provides a statistically meaningful comparison of computational predictions and experimental measurements, but also incorporates the uncertainty associated with each of them. We adopt a later implementation of the method by Higdon et al. (2008), which is deeply rooted in the following relation:

$$y_{obs}(x) = y_{sim}(x,\theta) + \delta(x) + \varepsilon(x). \quad (1)$$

In the above relation, $y_{obs}(x)$ and $y_{sim}(x,\theta)$ are the experimental and numerical predictions, $\delta(x)$ corresponds to a discrepancy term that represents the systematic bias, and $\varepsilon(x)$ represents the random experimental error.

The parameter x of equation (1) denotes the controlled variable, which defines the validation domain. Figure 3 shows an illustration in 2D where $x = (x_1,x_2)$. The most important distinction between control parameters and calibration parameters is the experimentalist's lack of control over the latter during physical testing. Calibration parameters, also referred to as ancillary variables in statistical sciences, are either introduced by specific choices of assumptions or models, or represent parameters that cannot be measured or controlled experimentally.

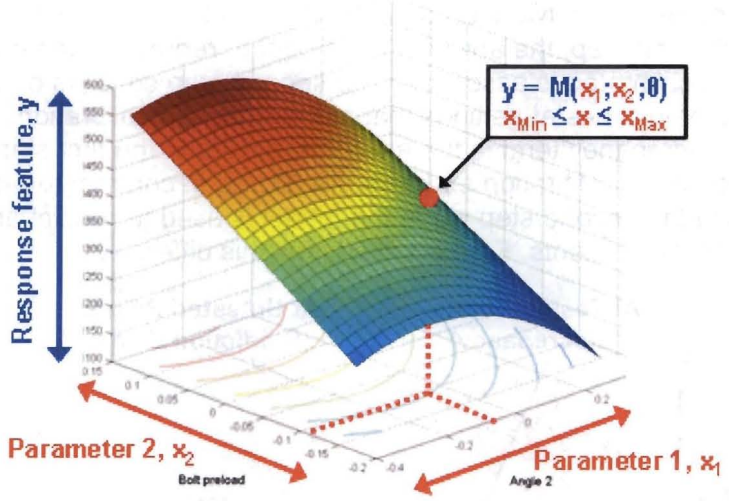


Figure 3. Model predictions, y, plotted against two control parameters, x_1 and x_2.

The control parameters define the validation domain, that is, the domain of configurations, settings or operational conditions within which the model is developed to make predictions. A subset of the validation domain is the domain defined by the settings at which experiments are conducted. Because calibration may be pursued only at tested settings, we refer to this domain as calibration (tested) domain. For example, if the performance of a material at varying strain rates and temperature is of concern, then the strain rate and temperature become our control variables. Other components of the model or simulation, such as the boundary condition or other material properties that may be poorly known by the analyst, would constitute the calibration parameters. In this case, physical experiments would be conducted at varying settings of strain rate and temperature in the validation domain. It is good practice to define the validation domain from the lower and upper bounds of these control parameters. Based on the available set of experiments, the other (ancillary) parameters would be calibrated. In practice, physical experiments cannot always be conducted to optimally cover the validation domain. However, as demonstrated in Section 6, coverage of the validation domain by experiments is of crucial importance for successful calibration.

With the formulation given in equation (1), we seek the values of parameters θ that, according to the Bayesian approach, represent the true but unknown values of calibration parameters. The unique contribution of Bayesian inference lies in the fact that it interprets probability as a degree of subjective opinion about the occurrence of an event. Bayes' perspective in that regard plays a central role in debates around the very foundation of the theory of probability. What this implies for our problem is that Bayesian inference requires an analyst's ***a priori*** opinion about the values of calibration parameters θ. Initial knowledge about these values is typically imprecise and, thus, defining an a priori distribution is remarkably difficult, if not impossible. To rectify this obstacle, a non-informative prior distribution can be assigned. The non-informative prior assigns equal probability of occurrence for all parameter values between upper and lower bounds. The bounds are usually more easily known than the parameter distribution. Once these priors have been assigned, the Bayesian inference machinery can be deployed to combine them with a likelihood function that describes the agreement between model predictions and measurements. The result is a posterior distribution of calibration parameters θ that can be used for forecasting and uncertainty quantification.

4. QUANTITATIVE APPROACH TO PREDICTIVE MATURITY

The proposed approach is a two step process that combines a fidelity to data metric with a stability criterion. In the first step, the ability of the model to reproduce the physical observations with which it has been calibrated is assessed. The second step consists of assessing the ability of the model to make predictions at settings that are not used in calibration. Strictly speaking, in this formulation, the discrepancy term is the error that occurs in the first step when the model is used to predict at settings of calibration experiments. This is in contrast to prediction error that is the error that occurs in the second step when the model is used to predict at settings other than those of the calibration experiments. Figure 4 illustrates this difference.

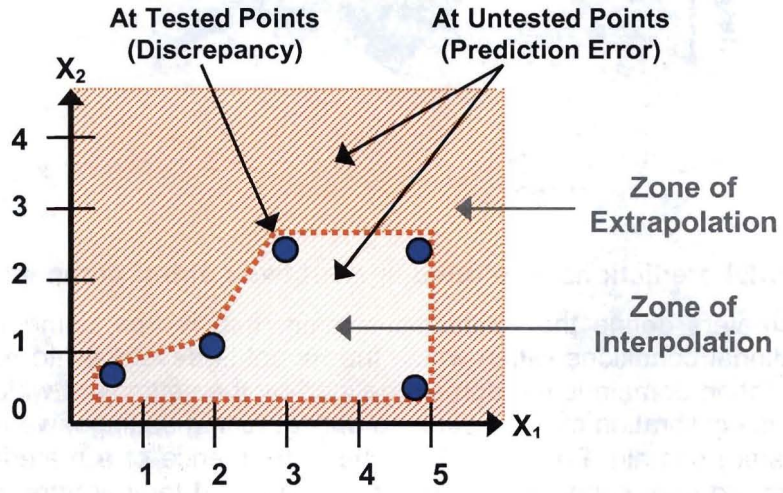


Figure 4. Illustration of discrepancy at tested settings vs. prediction error.
(Prediction error at settings of the calibration experiments is defined by the discrepancy term, while prediction error at untested settings is defined by the prediction error term.)

In the first step, the fidelity to data metric is defined by the discrepancy term. As coverage of the validation domain increases when new physical experiments are added, the calibrated model is expected to "converge" to a representation of reality and, likewise, the discrepancy term is expected to reach ***stabilization***. We propose to calibrate the model multiple times sequentially by increasing the number of experimental datasets and monitor changes in the discrepancy

term and its statistics. The expected behavior of the discrepancy term is illustrated conceptually in Figure 5. Once the discrepancy has been stabilized, which corresponds to the 5th experiment in the figure, it is clear that allocating resources to conduct additional experiments would provide only marginal improvement in predictive maturity of the model.

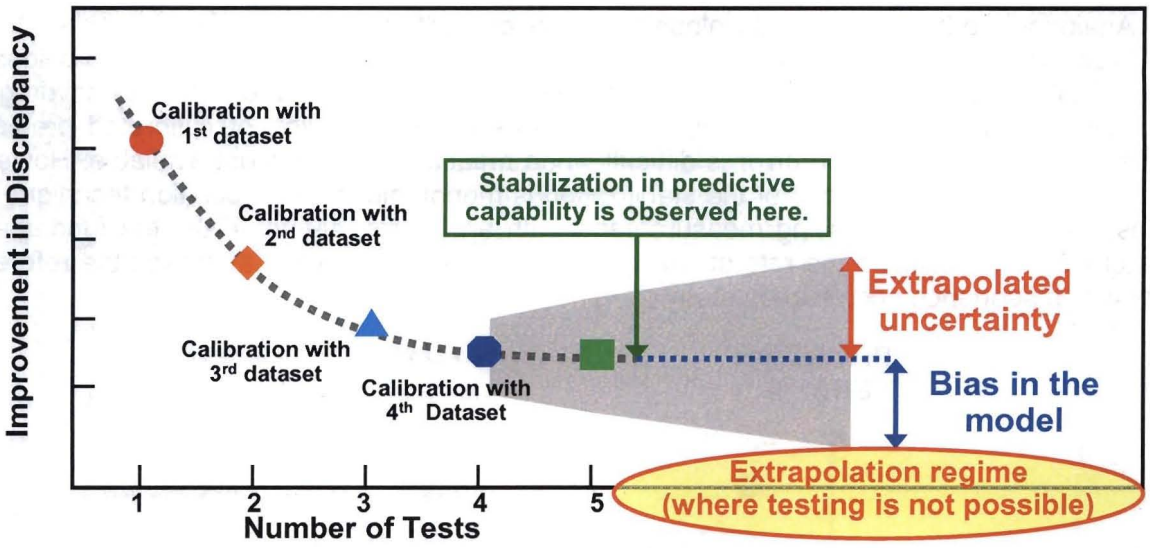


Figure 5. Behavior of the discrepancy term vs. coverage of the validation domain.

Once the stabilization of calibration is observed, predictive maturity can then be investigated. The model is calibrated with a select few experiments and used to predict the hold-out datasets. The number of experimental datasets fed to the procedure is successively increased to observe the behavior of prediction error as a function of coverage of the validation domain. According to our working hypothesis, prediction error should be reduced as coverage increases.

As explained in Section 5, coverage of the validation domain is quantified by the nearest neighbor metric, which also incorporates the sensitivity of control parameters. By combining the incremental improvement of model predictions with a sequence acceleration method, we define a Prediction Convergence Index (PCI) to approximate the "***reference prediction***." The PCI is used to support the inference of prediction error depending on the nature of predictions, that is, whether the model is used to interpolate or extrapolate.

Going back to equation (1), the following holds when the model is used to make predictions:

$$y_{obs}(x^{prediction}) = y_{sim}(x^{prediction}, \theta) + \delta\ (x^{prediction}) + E^{p}(x^{prediction}). \qquad (2)$$

The surrogate model $y_{sim}(x^{prediction})$ and error model $\delta(x^{prediction})$ estimate the prediction and its bias error. Although the calibrated model yields our best estimate, it usually cannot identically match measurements outside its calibrated settings. What is left between the calibrated and bias-corrected model predictions and hold-out experiments is the prediction error. In an attempt to predict a hold-out experiment, the only unknown in equation (2) is prediction error, denoted by $E^{p}(x^{prediction})$. Similar to the behavior of error in the calibration domain (or discrepancy term), the error at untested settings (or prediction error) is also expected to be reduced with increased coverage of the validation domain. Thus, the notional representation of discrepancy is similar to the expected behavior of prediction error (see Figure 5).

The maturity of a model is assessed quantitatively through the ability that the model exhibits to stabilize the prediction error as additional datasets are fed to the sequence of calibration and forecasting steps. Once it is deemed mature, the model is used to make predictions at settings

that have not been tested experimentally. How far one may safely venture away from settings that have been tested is determined by the PCI metric proposed in the next section.

5. THE PREDICTION CONVERGENCE INDEX

Analogous to the paradigm developed for solution verification proposed by Roache (1994), our Prediction Convergence Index (PCI) postulates that the calibrated model that provides an approximate solution approaches the exact but unknown value at a rate of "p," as coverage of the validation domain is refined. Figure 5 illustrates the stabilization of calibrated prediction errors. Quantifying prediction error is difficult when measurements are not available. However, by assuming that convergence of the stabilization is monotonic, an extrapolation technique can be used to replace the missing measurements with an approximation. Thus, the fundamental idea of PCI is to combine the rate at which calibrated prediction errors approach the reference error with a sequence acceleration method.

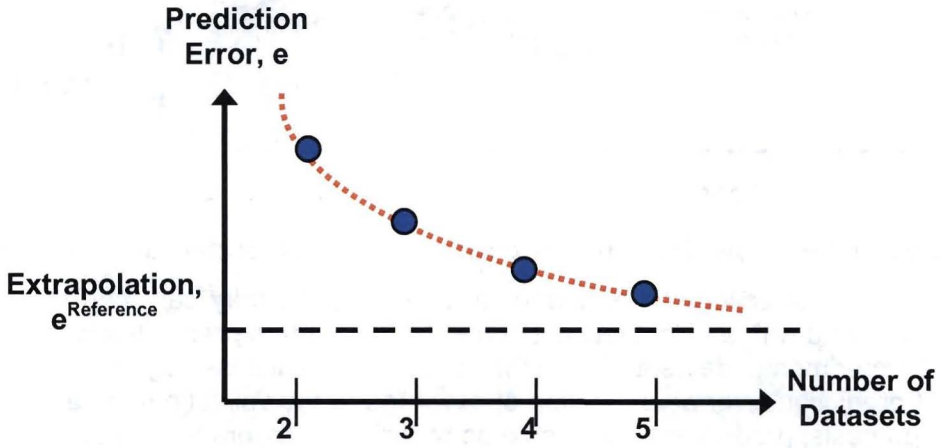


Figure 6. Monotonic convergence as a function of the coverage of validation domain.

To quantify the coverage of the validation domain by experiments, we propose a sensitivity adjusted nearest-neighbor metric. The dimensions of the validation domain, initially normalized between 0 and 1, are scaled according to the effect that each control parameter exercises on the output. These effects can be obtained through sensitivity analysis. Next, the dimensions of the validation domain are discretized into a sufficient number of grid points. These points are then categorized based on experimental datasets they are nearest to. The distance of these points to their corresponding datasets are calculated and the total distance of all points is scaled with respect to the total number of points in the validation domain. The normalization ensures that the final result is not affected by the initial grid size. This calculation provides a metric of coverage of the validation domain by experiments.

As coverage is improved, the successive nearest neighbor metrics are reduced. The ratios of these coverage metrics obtained for successive iterations are used to calculate the refinement ratio obtained by adding experiments. The refinement ratio is defined as $R = C^{(n)}/C^{(n+1)}$ where $C^{(n)}$ denotes the coverage metric obtained with "n" experiments and $C^{(n+1)}$ is the next value that corresponds to adding a $(n+1)^{th}$ experiment to the previous "n" ones. A value $R \leq 1$ indicates that the $(n+1)^{th}$ dataset provides better coverage than the previous, n^{th} dataset. It is analogous to the refinement ratio $R = \Delta x_C/\Delta x_F$ of a mesh refinement study, where Δx_C and Δx_F denote coarse-grid and fine-grid levels of resolution, respectively.

If the refinement ratio happens to be constant as additional datasets are provided for calibration, then the rate of convergence, p, can be estimated analytically as:

$$p = \log\left(\frac{e^{(n+1)} - e^{(n)}}{e^{(n+2)} - e^{(n+1)}}\right) \Big/ \log(R), \tag{3}$$

where $e^{(n)}$, $e^{(n+1)}$ and $e^{(n+2)}$ denote the prediction errors obtained from models calibrated with datasets $C^{(n)}$, $C^{(n+1)}$ and $C^{(n+2)}$, respectively. If the refinement ratio is not constant, then the rate is obtained by solving a non-linear equation. This procedure is analogous to the calculation of a reference solution from a mesh refinement study (Roache, 1994).

The error herein refers to the prediction error of a particular hold-out experiment. The selection of this hold-out experiment can be guided by the desired "direction" of prediction that is either an interpolation within the validation domain or an extrapolation outside the domain. We assume, without loss of generality, that such error should be scalar-valued. In the case of multivariate or functional analysis, we suggest reducing the error vector down to a scalar value by calculating, for example, its root mean square. Once the refinement ratio $R = C^{(n)}/C^{(n+1)}$ and corresponding rate of convergence, p, have been estimated for two sets of experiments, the reference error is estimated simply as:

$$e^{\text{Reference}} \approx \frac{R^p e^{(n+1)} - e^{(n)}}{R^p - 1}, \tag{4}$$

where the rate "p" results from equation (3) or solving a similar, non-linear equation. Following the idea of Roache (1994), the PCI estimates an error bound between model predictions and the reference prediction:

$$PCI = \frac{1}{R^p - 1}\left|\frac{e^{(n+1)} - e^{(n)}}{e^{(n+1)}}\right|, \tag{5}$$

and its application is illustrated with the Tantalum datasets in Section 7. The PCI estimates a bound that indicates what the true but unknown prediction error may be equal to, relative to the estimated error: $|e^{\text{Truth}} - e^{(n+1)}| \leq PCI \times |e^{(n+1)}|$. The PCI can, therefore, be used to derive an uncertainty (or bias) bound, where uncertainty originates from not possessing an infinite number of datasets to calibrate the model.

6. PREDICTIVE MATURITY OF THE PTW MODEL OF MATERIAL BEHAVIOR

The Preston-Tonks-Wallace (PTW) model of plastic deformation represents plastic stress in a material as a function of strain, strain rate, material temperature and seven dimensionless, calibration parameters. Even though the PTW model is appropriate to make predictions over a wide range of triplets of control parameters (strain, strain rate, temperature), no single setting of calibration parameters can identically reproduce multiple experimental datasets. Discrepancy is due to several factors, among which we mention the variation between experimental samples, measurement error and potentially erroneous assumptions and inadequacy of the PTW model itself. For instance, the model ignores non-isotropic plasticity and material texture effects. It also assumes that the plastic stress is independent of the history of material loading.

In this section, we briefly explain how we attempt to better characterize the seven calibration parameters needed in the definition of the PTW model. An application is discussed in Section 7 using experimental measurements for the Tantalum metal.

From an experimental standpoint, the plastic behavior of a metal is characterized through quasi-static compression tests along with Hopkinson-bar experiments. The quasi-static compression tests consist of small cylinders of material being compressed at constant, relatively slow rates. Rates typically used in quasi-static compression tests are about 1 $sec.^{-1}$ or less. The change in

nominal length of the cylinder is measured in relation to the load applied on its cross-section. The unit change in length is converted to strain and the unit load applied to the cross-section is converted to stress. To collect measurements in regimes of higher strain rages, typically around 10^{+3} sec.$^{-1}$, Hopkinson-bar experiments are conducted. An elastic wave is transmitted through a thin cylinder of material while the change in dimension of the cylinder is measured. Assuming that a constant strain rate is kept, it is straightforward to convert Hopkinson-bar measurements to stress-strain curves. An experimental apparatus is illustrated in Figure 7.

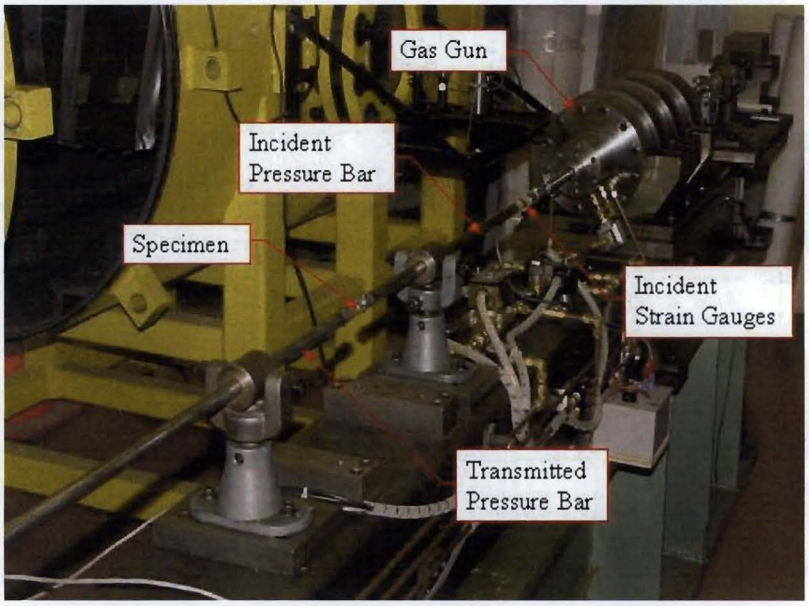


Figure 7. An experimental apparatus for Hopkinson bar testing.

We apply the Bayesian calibration procedure described briefly in Section 3 to calibrate the prior distribution of the seven dimensionless parameters (θ, κ, γ, y_0, y_∞, s_0, s_∞) of the PTW model. The Bayesian calibration is completed with 10,000 Markov Chain Monte Carlo (MCMC) iterations along with 500 burning runs at seven different levels. To train the GPM meta-models needed in the analysis, a 100-run Latin Hypercube maxi-min design-of-computer-experiments is analyzed. This design varies the seven parameters between their minimum and maximum values as given in Table 1. Prior distributions are defined as uniform, non-informative priors that assign an equal probability within the lower and upper bounds of each calibration parameter. We further assume that there is no correlation between priors of the seven calibration parameters.

Table 1. Lower and upper bounds of PTW calibration parameters for Tantalum.

Symbol	Description	Minimum	Maximum
θ	Initial strain hardening rate	2.78×10^{-5}	0.0336
κ	Material constant in thermal activation energy term (relates to the temperature dependence)	0.438	1.110
γ	Material constant in thermal activation energy term (relates to the strain rate dependence)	6.96×10^{-8}	6.76×10^{-4}
y_0	Minimum yield stress (at T = 0 K)	0.00686	0.0126
y_∞	Maximum yield stress (at T ≈ melting)	7.17×10^{-4}	0.00192
s_0	Minimum saturation stress (at T = 0 K)	0.0126	0.0564
s_∞	Maximum saturation stress (at T ≈ melting)	0.00192	0.00616

Since the physical experiments are conducted at varying settings of strain, strain rate and temperature, we can investigate the behavior of the discrepancy term at different points in the validation domain. Once the model is confirmed to reproduce the calibration experiments with acceptable accuracy, we can then investigate the behavior of prediction error by holding out some of the available experiments. It is emphasized that ***we define prediction error as the ability of the calibrated and bias-corrected model to predict an experiment that is not used for calibration***. As coverage of the validation domain is increased by adding physical experiments, we expect to observe the stabilization of prediction error.

7. APPLICATION TO MODELING THE PLASTICITY OF TANTALUM METAL

We analyze a total of six Tantalum datasets, each of which has been conducted at varying strain, strain rate and temperature setting. The example follows a framework first proposed by Hanson and Hemez (2004). The measurements of strain-stress curves collected from multiple Hopkinson bar tests are illustrated in Figure 8.

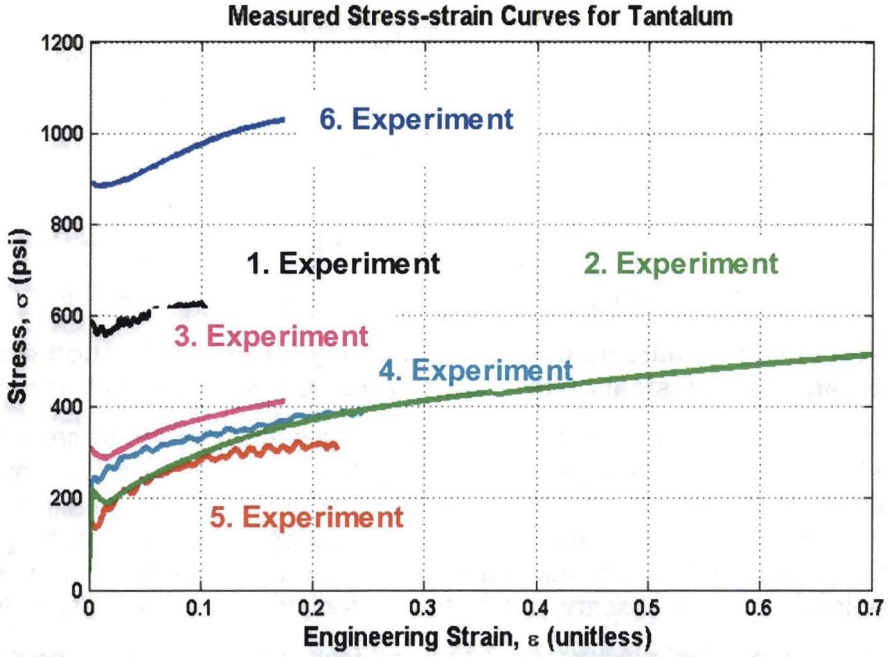


Figure 8. Stress-strain curves of Tantalum measured from Hopkinson bar testing.

All measurement pairs of (strain, stress) values shown in Figure 8 for these six experiments are combined in a pool. Five different calibration datasets are then defined as explained below. The first dataset uses only four data points from the pool, that is, four pairs of (strain, stress) values. A second dataset is created by selecting eight data points. The number of data points that form the next dataset is, each time, multiplied by two. The 5th and last dataset used for calibration counts a total of 64 data points. These pairs of (strain, stress) values are selected from an interleaving sequence of data points of the overall pool. It ensures that each calibration dataset provides an adequate coverage of the three-dimensional validation domain defined in terms of strain, strain rate and temperature settings.

The procedure based on Bayesian inference and described previously is applied to each one of these five datasets. As noted previously, Table 1 defines the nominal ranges within which the MCMC algorithm is allowed to sample the calibration parameters (θ, κ, γ, y_0, y_∞, s_0, s_∞). Model calibration is repeated five times, with these 4, 8, 16, 32 and 64 data points, to observe whether

stabilization of prediction error can be observed. Prediction error is assessed from calculations of stresses for hold-out settings of strain, strain rate and temperature that have not been used for calibration. Prediction error datasets are therefore independent from calibration datasets.

The maximum discrepancy estimated by the error model is found to be around 5% of the mean stress predictions, as shown in Figure 9. Although stabilization of the discrepancy term is not observed, the overall level of 5% discrepancy is below experimental uncertainty for Hopkinson bar tests of Tantalum. Thus, it is reasonable to consider that the calibration has been stabilized from the perspective of reproducing these experiments.

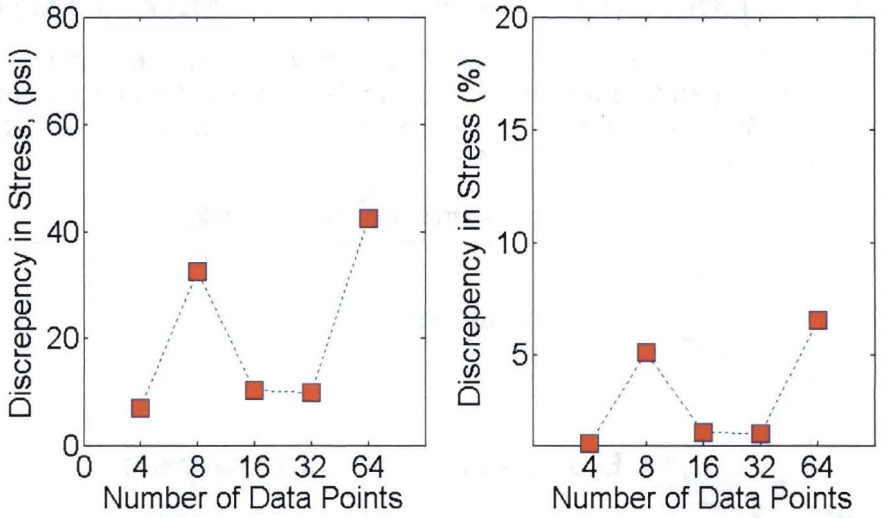


Figure 9. Root mean square values of discrepancy for five calibration datasets. (Left: RMS values in physical units; Right: Percentages of mean stress values.)

It is our belief that the concept of maturity goes beyond the stability of the discrepancy term and extends to the stability of calibration parameters provided, of course, that the coverage of the validation domain is satisfactory. By stability, we mean that when new information is provided to the calibration method, the parameters remain stable as opposed to fluctuating within their allowed ranges. Stabilization implies that calibrated parameters have reached maturity and further experimentation is not necessary, unless further reduction in uncertainty is desired).

Table 2. Statistics of PTW parameters obtained from the five calibration datasets.

Mean	**θ**	**κ**	**γ**	**y_0**	**y_∞**	**s_0**	**s_∞**
4 Data Points	0.3315	0.4984	0.5330	0.4782	0.4924	0.5644	0.4526
8 Data Points	0.3462	0.5383	0.5794	0.5072	0.5050	0.8223	0.4760
16 Data Points	0.4332	0.5566	0.6204	0.5140	0.4822	0.8402	0.3543
32 Data Points	0.4731	0.5585	0.6447	0.5044	0.4872	0.8367	0.2873
Overall Range	**0.1416**	**0.0659**	**0.1116**	**0.0358**	**0.0268**	**0.2758**	**0.1887**
Std. Deviation	**θ**	**κ**	**γ**	**y_0**	**y_∞**	**s_0**	**s_∞**
4 Data Points	0.1309	0.2822	0.2832	0.2879	0.2882	0.2866	0.2761
8 Data Points	0.1287	0.2870	0.2848	0.2945	0.2870	0.1519	0.2792
16 Data Points	0.0982	0.2825	0.2649	0.2878	0.2877	0.1184	0.2340
32 Data Points	0.0904	0.2828	0.2596	0.2855	0.2841	0.1089	0.1914

Table 2 lists the mean and standard deviation values obtained when parameters are calibrated with 4, 8, 16, 32 or 64 data points. It can be observed that variation in mean values is minimal, except for the initial strain hardening rate (θ), minimum saturation stress (s_0) and maximum saturation stress (s_∞). Although the three parameters show relatively large variations as different datasets are used for calibration, it is important to note that their standard deviation values are reduced with the addition of more experiments. The overall trend indicated by Table 2 is that the statistics of standard deviation estimated from the posterior distribution of calibration parameters are either stable or reduced as more experimental data points are added. Because the standard deviation represents variability of a distribution of values, this finding confirms that the calibrated parameters are, although slowly, converging to stable values.

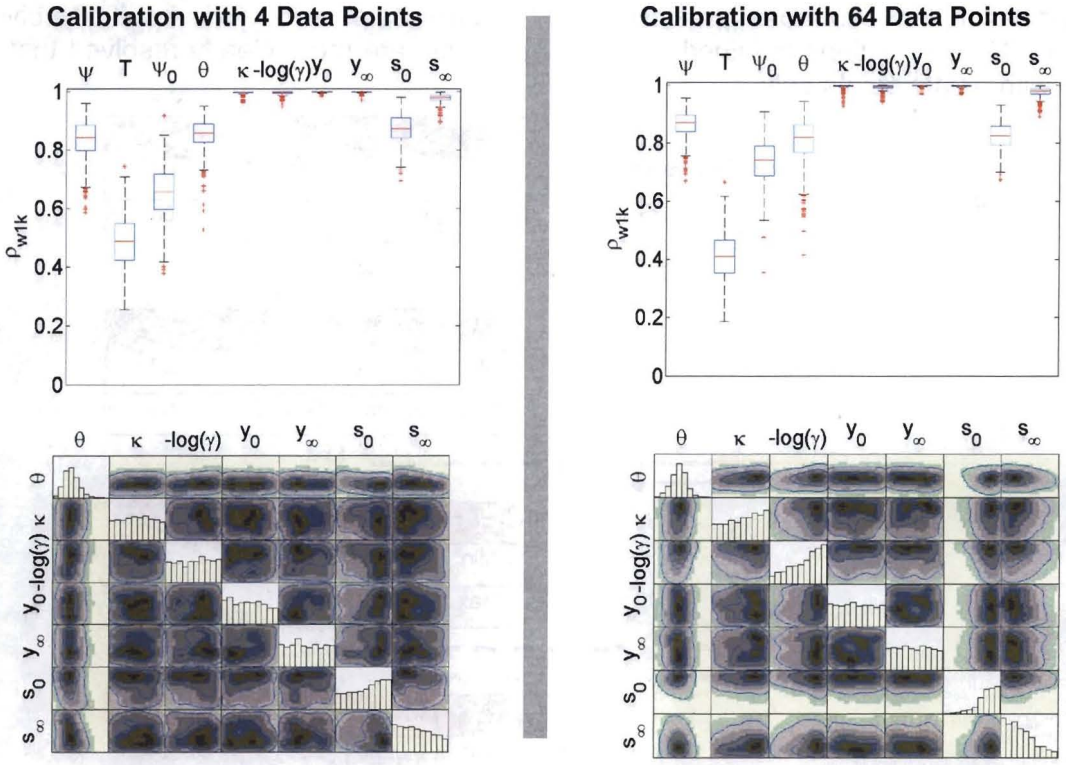
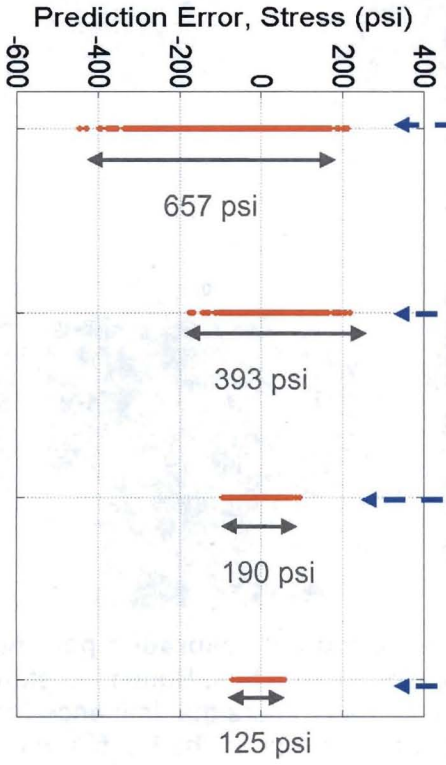


Figure 10. Sensitivity and posterior distribution of calibration parameters.
(Top: Sensitivity of control parameters and calibration parameters; Bottom: posterior distribution of calibration parameters. Significant effects, that is, parameters that influence the prediction of stress values, are indicated by statistics that deviate from one in the top figures. Histograms of the bottom figures show marginal distributions on the main diagonal while contour plots indicate joint distributions for pairs of calibration parameters.)

Figure 10 illustrates the sensitivity analysis obtained with model calibration against 4 and 64 data points. No significant change is observed between the two cases, indicating that the effects that parameters exercise on the prediction of stress are properly identified at the inception of the calibration studies. The three control parameters denoted by Ψ (strain), T (temperature) and Ψ_0 (strain rate) are, as expected, statistically significant in terms of influencing the stress prediction. It is also observed that only three of the seven calibration parameters are statistically significant. They are the initial strain hardening rate (θ), minimum saturation stress (s_0) and, to a lesser extent, maximum saturation stress (s_∞). The fact that four parameters are not significant could be used advantageously to keep them constant and equal to their nominal values, and eliminate them from the analysis.

Figure 10 also illustrates the marginal posterior distributions of single parameters and bivariate, joint posterior distributions of pairs of parameters. This information needs to be examined in light of sensitivities summarized in the top figures because calibrating a parameter that has little-to-no sensitivity does not yield a meaningful result. This can be observed in Figure 10. The four calibration parameters that are not statistically significant get assigned "flat," or non-informative, marginal distributions and these marginal distributions remain unchanged as the number of data points used for calibration is increased from 4 to 64.

For the triplet of significant parameters (θ, s_0, s_∞), on the other hand, it can be observed that feeding more information to the calibration tends to "narrow" the marginal and joint distributions. It means that the additional data points help to constraint the parameters to values that lead to better correlation between model predictions and test measurements. As a result, the bivariate, joint probability distributions obtained with 64 data points are more clearly resolved that those obtained with 4 data points only.



4 Dp	Strain	Temp.	Strain Rate
Min	0.0228	77	0.001
Max	0.8057	1074.67	3900

8 Dp	Strain	Temp.	Strain Rate
Min	0.0228	77	0.001
Max	0.8057	1085.08	3900

16 Dp	Strain	Temp.	Strain Rate
Min	0.0228	77	0.001
Max	0.9692	1091.04	3900

32 Dp	Strain	Temp.	Strain Rate
Min	0.0228	77	0.001
Max	0.9696	1093.98	3900

64 Dp	Strain	Temp.	Strain Rate
Min	0.0228	77	0.001
Max	0.9696	1095.57	3900

Figure 11. Statistics of prediction error obtained with the five calibrated models.

The next step is to calculate prediction errors at settings of strain, strain rate and temperature of the subsequent sequence added to the calibration. The first dataset, for instance, counts 4 data points for calibration while the second dataset counts 8 data points. The model obtained after calibration with the first dataset is used to predict stresses at the four settings (or data points) added to define the second dataset. This process is repeated until all 64 data points are used for calibration. The prediction errors of stress values are obtained in the form of random draws from the 10,000 MCMC runs. A total of 225 runs are selected for estimating the statistics.

Figure 11 is a simplified representation of the posterior distributions of prediction errors. The first model, for example, that predicts four new settings of strain, strain rate and temperature not used for calibration, provides a stress prediction error that ranges up to 657 psi based on 900 random draws. The last model, that is used to predict 32 data points not used for calibration, offers a prediction error that ranges up to 125 psi based on 7,200 draws. It clearly indicates that the overall prediction error of stress values is reduced as model calibration is completed with an increasing number of experiments. Figure 11 also indicates the levels of coverage of the three-dimensional, validation domain that these six datasets provide. Coverage is simply estimated by reporting the minimum and maximum values of strain, strain rate and temperature. These tables indicate that there is no significant difference in terms of coverage between the six datasets. It means that the reduction of prediction error comes from including more data points as opposed to better covering the validation domain.

The behavior of prediction error observed in this application, where it gets reduced as additional datasets are provided to the calibration procedure, comforts our hypothesis of stabilization. To investigate the rate at which this stabilization occurs, an experiment is randomly selected, held out during calibration and predicted with three different models. The held-out experiment is defined by the settings (strain = 0.235, Temperature = 703.35 K, strain rate = 2,600 s^{-1}). The analysis is focused on the three calibration studies performed with 8, 16 and 32 data points. Coverage of the validation domain is estimated using the nearest neighbor index, as discussed in Section 5. To estimate coverage, each control parameter is normalized between 0 and 1 and scaled with the sensitivity that the stress prediction exhibits to this parameter. The effect of this scaling is to "dilate" the dimension for a control parameter that does not exhibit a sensitivity that would be as important as sensitivities of the other parameters.

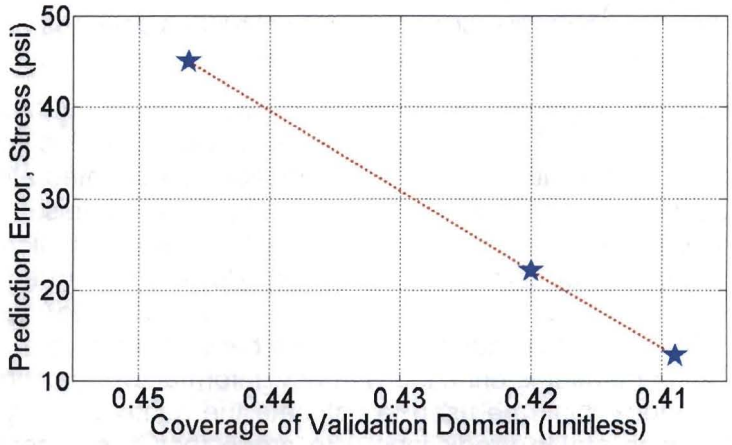


Figure 12. Values of the PCI metric as a function of coverage of the validation domain.

The coverage and prediction error metrics obtained with this procedure are shown graphically in Figure 12 and listed in Table 3. Pairs of (coverage; error) metrics are identified with blue star symbols in Figure 12, while the red, dashed line that connects them represents a least-squares, best-fit. The slope of this line represents the rate of convergence of equation (3), see Section 5.

Table 3. Coverage and prediction error metrics for three calibration datasets.

Metrics	8 Data Points	16 Data Points	32 Data Points
Coverage (unit-less)	0.446	0.420	0.409
Prediction Error (psi)	44.959 psi	22.147 psi	12.853 psi

Using the data from Table 3, the rate of convergence is estimated as p = 15.4 and the reference error of equation (4) is calculated as $y^{Reference}$ = 6.457 psi. This reference represents the best possible estimate of the "true" prediction error for the held-out experiment. In other words, the model calibrated with 32 data points makes a prediction that is 6.396 psi away from the true but unknown value of stress. This error corresponds to 1% error, approximately, which is a very low level of discrepancy between model prediction and physical measurement.

For Tantalum datasets described herein, the PCI is calculated to be equal to 2.1, or 210%, using the prediction errors from the 16-point and 32-point datasets of Table 3. The PCI converts the RMS of prediction error into a rough estimate of how far predictions of the model may be from the true but unknown value. One may, therefore, write $|e^{Truth} - e^{(32)}| \leq 210\% \times e^{(32)}$, where $e^{(32)}$ denotes the prediction error obtained with the model that uses 32 data points for calibration. If this "factor of safety" of 210% is deemed too large, then one could decide to either include more datasets in the calibration procedure or reduce the extrapolation to settings that are closer to those that have been tested experimentally (and used for calibration). We acknowledge that further studies of the PCI metric are needed to develop an empirical and intuitive understanding of which values of the PCI may indicate a mature predictive capability.

8. CONCLUSION

This paper starts from the assertion that simulation models will continue to be called upon by decision makers. Although the discipline of calibrating models to better match measurements from physical experiments is mature, when to bring closure to a calibration activity remains, to a great extent, judgment-based. We take a preliminary step towards formulating an approach to assess the predictive maturity of models and thereby increase confidence in the predictions of simulation codes. Defining an objective completion of calibration activities has practical value for all scientific fields where experiments are conducted to calibrate models.

We suggest assessing predictive maturity from a recursive strategy that builds on the strengths of Bayesian statistical inference. These advantageous properties include the quantification of prediction uncertainty and ease with which new information can be integrated to the procedure. In our approach, measurements are used not only to improve the models through calibration but also to assess the maturity of predictions. As new experimental information becomes available for calibration, the consistency of calibrated parameter values is monitored. Stabilization of the disagreement between predictions and measurements defines how the predictive maturity of a model is assessed. A convergence index is proposed based on the rate at which predictions stabilize as the calibration dataset is enriched with new information. Although not demonstrated here, the convergence index could be used to estimate the number of experimental tests that would be required to reach stable predictions. We argue that it is acceptable to extrapolate predictions away from settings or regimes where validation tests have been conducted as long as the predictive maturity has been assessed (which is what we attempt to do here); prediction uncertainty is quantified rigorously; and the physics involved in the application of the code remains unchanged as one ventures from the validation domain to the extrapolation regime.

An example is discussed for the predictive maturity of a non-linear model of plasticity that uses Hopkinson bar experiments performed with samples of Tantalum metal. It indicates that our hypothesis, namely, that stabilization of prediction discrepancy can be reached if a sufficient number of experimental datasets are used for calibration, is verified. This application outlines, first, the importance of collecting datasets that cover the entirely of the validation domain and, second, the point of diminishing return beyond which adding new experiments for calibration will not significantly improve prediction accuracy. Even though this example deals with a material

model, nothing prevents the presented approach from being applied to other types of simulation models in computational sciences.

ACKNOWLEDGEMENTS

This work is performed under the auspices of the Modeling and Simulation (M&S) program element of the Global Nuclear Energy Partnership (GNEP) program at Los Alamos National Laboratory (LANL). The first author expresses her gratitude to Michael Fugate, CCS-6, for his patient assistance during the course of this research. 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 U.S. Department of Energy under contract DE-AC52-06NA25396.

BIBLIOGRAPHICAL REFERENCES

D.L. Preston, D.L. Tonks, D.C. Wallace, (2003), "Model of plastic deformation for extreme loading conditions," Journal *of Applied Physics*, Vol. 220, pp. 93-211.

M. Fugate, B. Williams, D. Higdon, K.M. Hanson, J. Gattiker, S.-R. Chen, C. Unal, (2005), "Hierarchical Bayesian analysis and the Preston-Tonks-Wallace model," *Technical Report LA-UR-05-3935*, Los Alamos National Laboratory, Los Alamos, New Mexico.

O. Balci, R.J. Adams, D.S. Myers, R.E. Nance, (2002), "Credibility assessment: a collaborative evaluation environment for credibility assessment of modeling and simulation applications," *34th Winter Conference on Simulation: Exploring New Frontiers*, San Diego, California, 2002, pp. 214-220.

S.R. Blattnig, L.L. Green, J.M. Luckring, J.H. Morrison, R.K. Tripathi, T.A. Zang, (2007), "Towards a credibility assessment of models and simulations," *9th AIAA Non-Deterministic Approaches Conference*, Oahu, Hawai'i, April 23-26, 2007.

S.Y. Harmon, S.M. Youngblood, (2005), "A proposed model for simulation validation process maturity," *Journal of Defense Modeling and Simulation*, Vol. 2, No. 4, pp. 179-190.

W.L. Oberkampf, M. Pilch, T.G. Trucano, "Predictive capability maturity model for computational modeling and simulation," *Technical Report SAND-2007-5948*, Sandia National Laboratories, Albuquerque, New Mexico, October 2007.

D. Sornette, A.B. Davis, K. Ide, K.R. Vixie, V. Pisarenko, J.R. Kamm, (2007), "Algorithm for model validation: Theory and applications," *Proceedings of the National Academy of Sciences*, Vol. 104, No. 16, pp. 6562-6567.

D. Higdon, J. Gattiker, B. Williams, M. Rightley, (2008), "Computer model calibration using high-dimensional output," *Journal of the American Statistical Association*, Vol. 103, No. 482, pp. 570-583.

M. Kennedy, A. O'Hagan, (2000), "Predicting the output from a complex computer code when fast approximations are available," *Biometrika*, Vol. 87, pp. 1-13.

M.A. Christie, J. Glimm, J.W. Grove, D.M. Higdon, D.H. Sharp, M.M. Wood-Schultz, (2005), "Error analysis and simulations of complex phenomena," *Los Alamos Science*, No. 29.

K.M. Hanson, F.M. Hemez, (2004), "Inference about the plastic behavior of materials from experimental data," *4th International Conference on Sensitivity Analysis of Model Output*, Santa Fe, New Mexico, March 8-11, 2004, pp.126-136.