SAND2021-14785C

AIAA Non-Deterministic Approaches Conference, SciTech 2022, Jan. 3-7, San Diego, CA

Arguments for the Generality and Effectiveness of "Discrete Direct" Model Calibration and Uncertainty Propagation vs. Other Calibration-UQ Approaches[1]

Vicente Romero
V&V, UQ, and Credibility Processes Dept.
Sandia National Laboratories,[2] Albuquerque, NM

Abstract

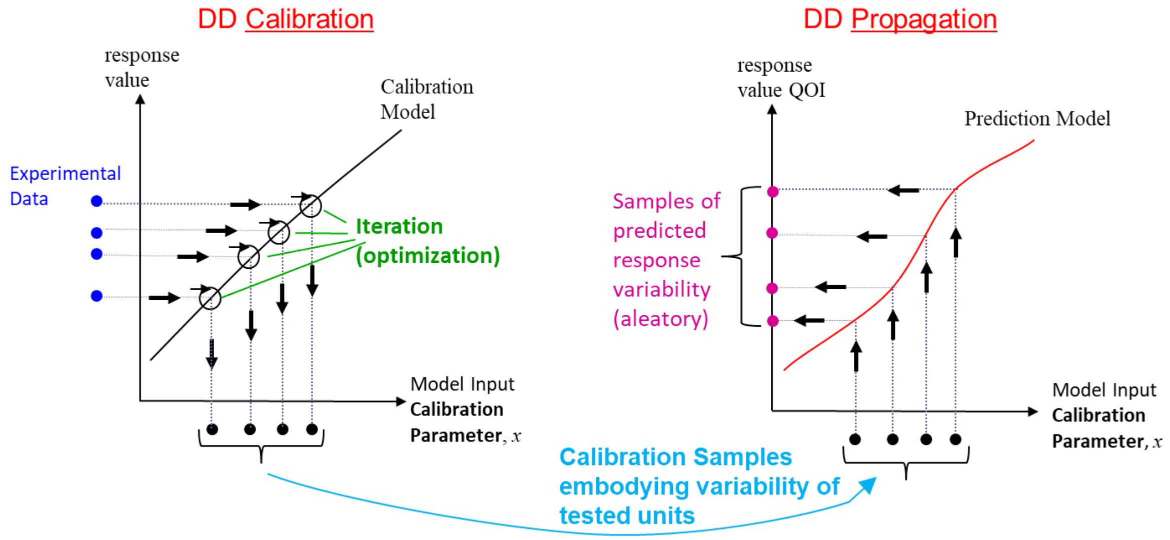
This paper describes and analyzes the Discrete Direct (DD) model calibration and uncertainty propagation approach for computational models calibrated to data from sparse replicate tests of stochastically varying phenomena. The DD approach consists of generating and propagating discrete realizations of possible calibration parameter values corresponding to possible realizations of the uncertain inputs and outputs of the experiments. This is in contrast to model calibration methods that attempt to assign or infer continuous probability density functions for the calibration parameters. The DD approach straightforwardly accommodates aleatory variabilities and epistemic uncertainties (interval and/or probabilistically represented) in system properties and behaviors, in input initial and boundary conditions, and in measurement uncertainties of experimental inputs and outputs. In particular, the approach has several advantages over Bayesian and other calibration techniques in capturing and utilizing the information obtained from the typically small number of replicate experiments in model calibration situations, especially when sparse realizations of random function data like force-displacement curves from replicate material tests are used for calibration. The DD approach better preserves the fundamental information from the experimental data in a way that enables model predictions to be more directly tied to the supporting experimental data. The DD methodology is also simpler and typically less expensive than other established calibration-UQ approaches, is straightforward to implement, and is plausibly more reliably conservative and accurate for sparse-data calibration-UQ problems. The methodology is explained and analyzed in this paper under several regimes of model calibration and uncertainty propagation circumstances.

1. Introduction

The "Discrete Direct" (DD) model calibration and uncertainty propagation approach was developed to calibrate otherwise deterministic models to reflect unit-to-unit variability of systems being modeled that have underlying random/stochastic/aleatory variability in material properties, geometries, surface interface behaviors, etc. The underlying variability causes nominally identical units to have varying behaviors and response outputs in replicate tests that provide the calibration data (after normalizing for any variations of input boundary and initial conditions from test to test). A model calibration process is used to map the experimental variability in important measured output responses into variability of one or more parameters of the model.

[1] Sandia National Laboratories document SAND2021-yyyy (unlimited release). This paper is a work of the United States Government and is not subject to copyright protection in the U.S. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.
[2] Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Figure 1 presents a simplified representation of the DD calibration and uncertainty propagation ("calibration-UQ") paradigm applied to an illustrative 1-D (one calibration variable) problem. The four deterministic experimental results (data points) in the figure come from different but nominally identical test units in four replicate tests. For now, let each test be performed at exactly the same input boundary conditions. A separate calibration is performed for each test/data set. This yields the four calibration parameter values indicated in the figure. These four values are then propagated to predictions in other modeling applications. For instance, the parameter values could be for a material model to be used under different geometry and/or initial and/or boundary conditions. (Model and parameter suitability for use at conditions different from the calibration conditions is a matter of expert judgment, best informed by model validation assessment of predictive accuracy at related extrapolation conditions.) The four prediction results in the figure are processed with specialized sparse-sample 1D UQ techniques (explained later) to obtain reliably conservative and efficient estimates of response variability and related statistical quantities.


DD Calibration
response value
Calibration Model
Experimental Data
Iteration (optimization)
Model Input Calibration Parameter, x
DD Propagation
response value QOI
Prediction Model
Samples of predicted response variability (aleatory)
Model Input Calibration Parameter, x
Calibration Samples embodying variability of tested units

- Use optimization to find N calibration parameter values that best match results from N individual experiments (N = 4 in the graphic)
- Proceed to prediction with the N discrete values of the calibration parameter(s)
- N runs of prediction model to propagate N parameter sets from N calibration experiments
- Preserves direct correspondence to the experimental data underlying the calibrations
- Use appropriate sparse-sample 1-D UQ techniques to process N discrete prediction results into reliably conservative distributions or statistics of response variability
- Simple to update with new experiments/data and calibrations if/when more data becomes available (without Bayes' rule & machinery)

Figure 1. Simplified representation of the Discrete-Direct model calibration and uncertainty propagation paradigm.

DD calibration straightforwardly accommodates multiple realizations of scalar experimental data (as indicated on the ordinate at left in Figure 1) or functional experimental data such as spatial and/or temporal response data. Problems with multiple calibration parameters are also easily accommodated. For

a calibration problem involving multiple calibration parameters, say (*x1, x2, x3*), a calibration parameter set (*x1, x2, x3*)$_i$ would exist from calibration to each of the i=1,I replicate experiments.
DD calibration and uncertainty propagation has been applied at Sandia to problems involving sparse realizations of functional experimental data and 1 to 11 calibration parameters in applications in solid mechanics ([1] - [4]); structural dynamics ([5], [6]); and radiation-damaged electronics ([7], [8]). References [2]-[4] and [6]-[9] have been used as test problems with synthetically generated data for known truth statistics to confirm the sparse-data DD calibration-UQ methodology over hundreds of random trials.

Beyond the deterministic model function and experimental data points indicated on the ordinate at left in Figure 1, many potential sources of error and uncertainty in the calibration model and experimental data can arise. These are itemized below. They result in uncertainty in the calibration parameter set determinable from a given experiment.

- random and/or systematic uncertainties on measured experimental inputs and outputs
- model discretization and solver related numerical solution uncertainties
- uncertainties associated with approximation errors in any statistical and probability models used in the processing of the experimental data

All these uncertainties can be straightforwardly accommodated in the DD methodology as demonstrated in [9] and discussed in Section 5 of the present paper. The DD calibration and uncertainty propagation methodology also straightforwardly reduces to handle the case of one test with no replicates and experimental uncertainties present ([6]) or not.

Treatments for the following other sources of error and uncertainty in the model calibration problem are less well established for the DD calibration-UQ approach. Plausible treatments are proposed and discussed in Sections 4 and 5.

- Model-form related error and uncertainty in the calibration problem and results occur when the model being calibrated does not have the necessary structural/mechanistic form and/or when active calibration parameters and/or allowed parameter ranges do not enable the model output(s) to completely match the various uncertainty realizations of the experimental output(s). This is particularly prevalent when functional experimental data and/or multiple output quantities are desired to be matched in the calibration. Model-form related error/uncertainty in the context of a DD calibration-UQ approach is discussed in Section 5.

- Uncertainty and/or non-uniqueness of calibration parameter values can stem from the following.
    - Precision error/uncertainty on calibration parameter values results from incomplete convergence of the optimizer due to finite convergence tolerances used for affordability, and/or when response-surface approximations are used as surrogate models for the computational physics model in the optimization procedure. Some progress on this front is summarized in Sections 4 and 5.
    - More fundamentally, non-uniqueness of calibration parameter values often exists. This is due to incomplete identifiability of the parameter(s) because of insufficient calibration data from the test or experiment. Prediction uncertainty due to parameter value non-uniqueness is ventured to be adequately quantifiable with multi-start optimization procedures [10], but this remains to be established. This is discussed further in Sections 4 and 5.

Section 3 outlines the *Simultaneous* DD (SDD) method for efficient calibration, propagation, and prediction when multiple material, phenomena, and/or component models of a system model are each calibrated to sparse replicate scalar or functional data. Before proceeding to the DD advances in sections 3 – 5, an assessment is conducted in Section 2 on the strengths of DD versus other optimization-based calibration-UQ approaches in the literature (e.g. [11]-[16]) and Bayesian calibration-UQ approaches (e.g. [11], [17]-[21]).

Optimization-based calibration approaches other than DD have not yet demonstrated provisions for all the types and combinations of heterogeneous uncertainties in the bulleted items above, nor for epistemic lack-of-knowledge uncertainty due to sparseness of experimental data replicates used for calibration. The latter is usually the largest source of epistemic uncertainty by far in stochastic model calibration in engineering practice (in the author's experience). The non-DD optimization based calibration-UQ approaches are also argued in the next section to be more complicated and less accurate than the DD approach under most realistic engineering circumstances. Bayesian approaches for calibration of stochastic models are even more complicated and so can be subject to significant analyst-to-analyst variability, as evidenced in this paper. Most of the bulleted uncertainty sources above have been addressed to some degree in Bayesian formulations, but the conventional zero-mean Normal distribution "noise" model in Bayesian formulations does not appear to fully address heterogeneous experimental data uncertainties (random and/or systematic, interval and/or probabilistic uncertainties on experimental inputs and outputs—all of which often arise in experiments). Current implementations of Bayesian approaches are also argued to not address sparseness of replicate experimental data as accurately and efficiently as the DD approach does. Some evidence of this is presented in the Appendix of this paper.

2. Structural advantages of the Discrete Direct approach over other calibration-UQ approaches for calibrating models of stochastic phenomena to sparse replicate tests

It is common that time and resource limitations allow only a few units from a large population to be experimentally tested to supply calibration data. With only a few replicate tests to sample the stochastic variability of behaviors in the population, it is often desired to appropriately calibrate and use a model to predict response variability for the whole population of units (under various use conditions). In this and the next section we consider epistemic uncertainty associated with sparseness of replicate experiments and corresponding model calibration and prediction. We assume the ideal case of full identifiability of calibration parameters such that a 1:1 correspondence exists between experiments and calibration parameter sets. In Sections 4 and 5 we relax this assumption and consider other epistemic uncertainties in calibration-UQ problems in the bulleted uncertainty sources in Section 1.

2.1 DD circumvents the difficulties of modeling and propagating calibration parameter distributions and correlations

One structural advantage of the Discrete-Direct method is that it determines and propagates discrete sets of calibration parameter values rather than modeling and propagating variability distributions and correlations of the calibration parameters. The latter approach if applied to the problem in Figure 1 would amount to inferring the parameters of a probability density model, such as the mean and standard deviation of a proposed Normal distribution, for variability of the calibration parameter x on the abscissa as illustrated in Figure 2. Such inference can be approached in a number of ways to be discussed in this subsection. The immediate point is the distinction between DD propagation of just the four discrete values x_i of the calibration parameter forward to predictions, versus inferring a probability density for variability of x and then propagating it.

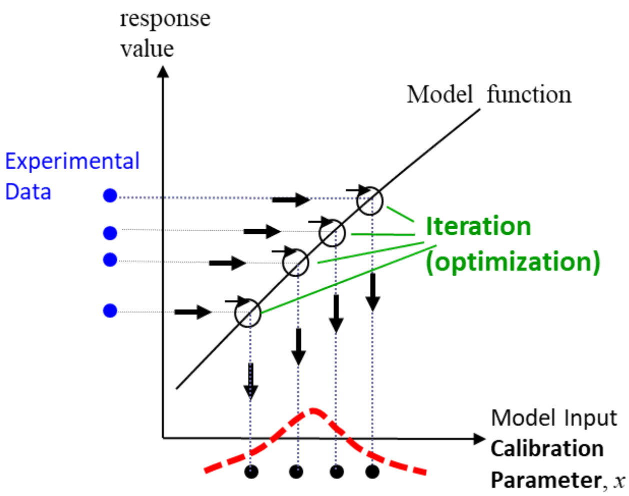


Figure 2. One approach to inference of a probability density (PD) for stochastic variability of a calibration parameter by fitting a PD model to realizations of calibration parameter values (on abscissa) from discrete calibrations.

The simplicity of the DD method carries over to problems with multiple calibration parameters, essentially mapping the problem into a one-dimensional (1D) uncertainty estimation problem as explained later, whereas inference approaches as in Figure 2 substantially increase in complexity for multi-dimension (multi-D) calibration-UQ problems. A first insight into this comes from considering a calibration-UQ method [15] that is “half way” to the DD approach. Figure 3 shows the results of 18 calibrations of a material plasticity model to 18 load-displacement curves from 18 uniaxial tension tests of a steel material. The material model has three parameters that are determined in each calibration; 18 values are shown for each calibration parameter. The results are from a real application at Sandia National Laboratories. The project’s discrete calibration approach follows [15] and coincides with that of the DD approach. However, the *propagation* of calibration parameter stochastic variability differs from DD propagation as explained next.

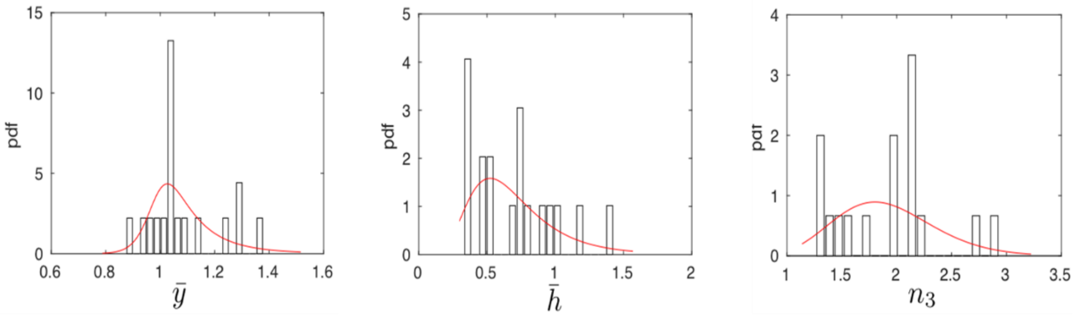


Figure 3. Histograms of calibration values for three parameters of a material plasticity model calibrated to each of 18 material tests individually, with probability densities (PDs, in red) from PD models fit to each parameter’s 18 values.

Following [15], 2-parameter Beta-family probability density models are fit to the 18 values of each parameter in Figure 3. Maximum likelihood estimation (further discussed in Section 2.2) is used to obtain the parameter values. The PD fits do not appear to accurately reflect the actual histogram densities in some locations. For instance, the population bump at the right end of the *y* parameter's histogram is not reflected in its PD. Even gross trends are off in some cases; the trend of increasing histogram density from right to left in the *h* parameter is not reflected in the trend at the left end of its PD. The particular forms of the PD models and their selection and fitting processes in this example are not important here. What is important is that *any* PD model or formulation and fitting process (parametric or non-parametric) will struggle to be adequately representative, even with the fairly rich data set of 18 sample values of each of the calibration parameters. Shape constraints of proposed PD models and lack of a very large number of data samples (calibration parameter values) to fit the PD models to, will in most realistic physical engineering situations preclude arriving at an accurate model of the true PD that would exist with a large number (hundreds to thousands) of tests, calibrations, and corresponding parameter value samples.

Error will also exist in correlation models needed to approximately enforce experimentally based sampling dependencies between the parameter values. That is, considering in this section the ideal case of perfect parameter identifiability from each experiment, the three calibration parameter values are uniquely determined by, and associated with, each experiment. Then DD propagation of the 18 3-parameter sets through a next-level analysis model (here a structural model) directly ties the variability in the 18 material-level tests to predicted variability in structural response (18 realizations for each output quantity of interest). Alternatively, it would be important to recognize and preserve the parameter relationships/dependencies in the 18 sets as accurately as possible in propagating marginal PDs fit to the calibration parameter values. The parameter dependencies could potentially be approximately preserved by the common mechanism of assigned linear correlations between the fitted PDs. Then in a next-level analysis, the parameter sets produced from sampling the PDs with assigned correlations could potentially have approximately the same parameter dependency relationships as the parameter sets from the 18 individual calibrations.

For scalar quantities of response, including scalar quantities of interest (QOIs) obtained from functional output quantities (e.g., high or low value of a field quantity, or value at a particular time and/or spatial location), the 18 samples of the scalar QOI set up a 1-D uncertainty estimation problem. We know that fitting a PD to sparse data samples normally implies substantial error and uncertainty, so we do not pursue this. Instead we pursue bounding estimation of statistical quantities commonly of prediction interest for design, analysis, and decision making, such as central 95% of response variability between the 2.5 and 97.5 percentiles; desired percentiles of response; and probability or proportion of response above or below a prescribed threshold value ("*tail probabilities*").

The estimates are intended to bound the true statistical quantities that would result from DD propagation of an asymptotically large population of calibration parameter sets from a corresponding large number of tests of the stochastically varying phenomena. The estimated bounds are obtained with specialized sparse-sample 1-D UQ methods ([22]-[26]) derived from well-established statistical methods based on sparse sampling of Normal populations. Extensive numerical studies have recently established the sparse-sample UQ methods to be robust and efficient bounds estimators for the mentioned statistical quantities and diverse varieties of highly non-Normal distributions.[3] At this point, substantial empirical evidence exists

[3] For example, statistical tolerance intervals (TIs) bound the central 95% of response with 90% confidence (95/90 TIs) when based on sparse samples drawn from a Normal distribution. Lesser but still reasonable and useful success rates occur with sparse samples from a diverse variety of other distributions. For instance, 89% of 144 PDFs of physics model outputs (including highly non-Normal multi-modal and/or highly skewed and even one-tailed distributions) in studies [22] and [23] had empirical confidence levels of ≥75% with 95/90 TIs and N = 2 to 4 random samples. (Empirical confidence levels were based on 10,000 random-sample trials per each PDF and number of samples N.) The average or expected confidence levels decline slowly as the number of samples N increases, for reasons explained in [22], [23]. In [2] with 90/90 TIs and N = 2, empirical confidence levels of ≥75% were obtained on 15 of 16 similarly challenging distributions (94% of PDFs). In [4] with 99/95 TIs and N = 4, empirical confidence levels of ≥75% were obtained on 35 of 40 similarly challenging PDFs (88%). As reviewed and

to provide a basis for reasonable confidence and credibility that conservative and efficient (not overly conservative) bounds are obtained by the methods.

Alternatively, consider a response QOI PD obtained from a propagated joint probability density (JPD) consisting of the calibration parameter marginal PDs and their correlations. The calibration parameter PDs and their correlations each have substantial error. The multiple uncharacterized error sources would make it very difficult to judge whether an estimated statistic from the resultant probability distribution is appropriately conservative, overly conservative, or unconservative.

Another source of uncharacterized error often comes from the propagation procedure itself. A Monte Carlo (MC) sampling procedure would require orders of magnitude more runs of the analysis model than the 18 simulations needed for DD propagation in the present example. A more computationally efficient alternative to pure MC propagation is to build a response-surface surrogate model and sample it in the MC propagation procedure (e.g., [27], [28]). Stochastic expansion or optimization-based reliability methods or other uncertainty propagation approaches could alternatively be used (see e.g. [29]-[31] for reviews). Depending on the nonlinearity of QOI response over the variability ranges of the calibration parameter PDs, it could easily take more than 18 simulations of the analysis model to obtain suitably accurate response-surface MC or alternative propagation methods for the three-variable example. Indeed, accuracy assessment in the Sandia project recommended 20 model evaluations for production work with the employed Stochastic Reduced-Order Model (SROM) propagation method ([32]), but even 80 model evaluations did not accurately capture the upper tail of the test QOI's response beyond the 90th percentile.

Propagation difficulty and cost (in terms of number of model simulations needed by alternatives to pure MC) quickly scale with the number of uncertainty sources (equal to the number of calibration parameters in the present context). The number of error sources from modeling the PDs and correlations to be propagated also scale with the number of calibration parameters.

For a given number of calibration parameters, if the number of affordable experiments/calibrations/ parameter sets is, say, just three instead of 18 (as would be realistic in many engineering settings), then much larger errors will exist in the modeled PDs and correlations propagated to QOI PDs. DD results will also reflect more uncertainty, but in a manner that explicitly controls for the number of experiments/calibrations/parameter sets to roughly yield a desired confidence in the estimated statistical bounds. DD propagation cost is just three analysis model simulations in this case, regardless of the number of calibration parameters, while non-DD propagation cost will scale with the number of calibration parameters, regardless of the number of experiments/calibrations/parameter sets.

Thus, the DD approach sets up a simple and inexpensive 1-D uncertainty bounding estimation problem for each output QOI and avoids a multi-D JPD inference and propagation problem. The DD approach is

referenced in those documents, the other methods tried or critically evaluated include Bootstrapping, optimized four-parameter Johnson-family distribution fit to the response samples, non-parametric kernel density estimation specifically designed for sparse data, non-parametric cubic-spline probability density functions fit to the data based on maximum likelihood, and Bayesian sparse-data approaches.

If the model response samples are instead to be used for bounding estimates of tail probabilities of response for robust/reliable design or safety/risk analysis, the response samples would be processed in a different way. This is demonstrated in recent investigations ([22], [24]-[26]) for N= 2 to 20 samples and 16 diversely shaped distributions (including multi-modal and/or highly skewed and even one-tailed distributions) and tail probability magnitudes from 10^{-5} to 10^{-1}. Reliably conservative (at 80% to 99% confidence for 15 of the 16 PDFs) and efficient estimates of small tail probabilities are obtained with the recommended methods itemized in [26]. Similar success rates are obtained in [4] and [8] on left and right tails of 41 similarly challenging PDFs (for tail probability magnitude $5x10^{-3}$ and N=4 samples per random trial and 100 trials). Further results are presented at the ends of Section 3 and of the Appendix of this paper.

substantially simpler, typically less costly, and generally more accurate in terms of reasonably controllable and knowable statistical confidence on calculated statistics of response quantities relevant to design, analysis, and decision making.

One could go about trying to account for errors and uncertainties associated with modeling and propagating JPDs of calibration parameter variability in multi-D problems, but this would incur substantial difficulty, cost, and complexity compared to DD propagation and UQ. Alternatively, an attempt to simply model the calibration parameter variabilities with conservatively wide marginal PDs such as tolerance-interval equivalent Normals ([33]) would not be desirable compared to the DD approach. Investigations in [33] and [34] show that the TI related conservatism of the individual PDs compounds in propagation when QOI response functions are monotonic over the joint PD parameter space, yielding over-conservatism in the resultant PD of response. This effect increases with the number of calibration parameters. On the other hand, if QOI response is non-monotonic over the JPD space, then propagating conservatively wide PDs will not yield a conservatively wide response QOI PD. Thus, a more careful and involved approach would be required to appropriately account for sparse-sample related error and uncertainty in modeling and propagating calibration parameter variability PDs and correlations. It does not appear possible to be competitive with the simplicity, economy, and reliability of the sparse-data DD propagation-UQ approach.

Outside the context of model calibration, one can ask whether the DD propagation-UQ approach is superior to alternative propagation-UQ approaches. It appears so if the input data to the propagation-UQ problem comes in discrete form from experiments, but not for UQ with specified continuous input PDs and/or interval uncertainties (unless low-cost sparse-sample conservative bounding estimates of statistics are desired).

Figure 4 shows 10 random realizations of data variability for each of six random-variable inputs to a structural model in a calibration-validation-UQ methods challenge problem (Sandia storage tank failure problem [51]). The figure also shows two potential sets of PD representations obtained in [20] from different methods applied to the sparse sample data of the six variables. One method used a stabilized Pearson method to fit a 4-parameter PD to each 10-sample data set. The authors cite the following advantages for this choice (paraphrasing): “because of its general better accuracy for the high and low probability levels compared to other PDF estimation methods, such as the saddlepoint approximation, maximum entropy principle, and [4-parameter] Johnson system. Many popular distributions (e.g., normal, beta, gamma, lognormal, etc.) are special cases of the Pearson system. Hence, the proposed approach can reduce the statistical uncertainty by eliminating improper distribution assumptions.” References for these methods are given in [20].

The second sparse-sample fitting approach was a Bayesian partial hyperparameter approach. A Normal PD was assumed for all six variables. Each Normal PD’s mean parameter was assigned a Normal distribution of uncertainty as a prior, with the prior mean and variance set at values consistent with the 10 data samples and then updated with them. The variance of each input variable’s Normal distribution was set to the variance from the 10 data samples and was not updated. The authors observed: “the approximate PDFs from the Pearson system match the test data better than the Bayes approach. However, it is worth noting that the test data are not sufficient and the Pearson system could create significant amount of statistical error by modeling all PDFs as irreducible uncertainty.” The latter is taken by the present author to mean that the parameters of the Pearson distribution arrived at are point values, with no expressed uncertainty in them. This contrasts with their Bayesian approach that included uncertainty in the mean parameter of the Normal distributions. The authors also noted that the Bayesian approach “produces conservative modeling which is more desirable when data are insufficient.” The authors went on to propagate the Bayesian PD results for their predictions.

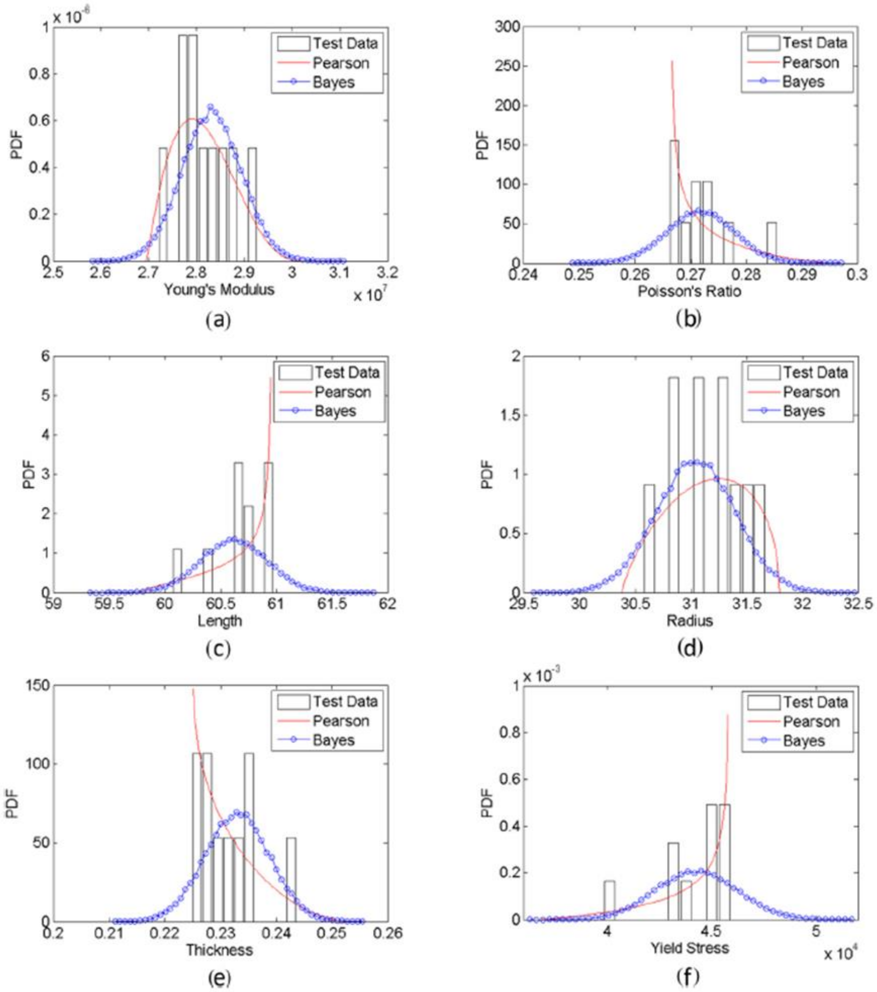


Figure 4. Estimated probability densities from Pearson and Bayesian method fits to 10 data samples of each random-variable quantity. (Figure reproduced from [20] with permission from the publisher.)

The results in Figure 4 show the very large differences that can occur in fitting PDs to sparse data with two different sophisticated and plausible approaches. The more conservative Bayesian results in the figure are less conservative than they would be if variance were also considered uncertain in the Bayesian updating procedure. Sparse realizations of random data tend to under-represent the variance of the true distribution the realizations are drawn from; the variance calculated from a small number of samples will typically be smaller than the variance calculated from a large number of samples, for many common PDF types. Significant error in the mean calculated from sparse sample data will of course also usually exist. Ref. [22] shows these error dynamics for Normally distributed data. The Bayesian PDs in the figure may not actually be conservative (e.g., at the right tail for Poisson's Ratio, the left tail for Length, and the left tail for Yield Stress) because uncertainty in variance was not accounted for (and the assumed Normal PD form for the variables may not be correct). If the PD fits all happen to be conservative, then their propagation together will create compounded conservatism (over-conservatism) in the outputs vs. their true variability.

Joint propagation in these situations will often be free of significant dependence between the input variables and therefore free of correlation modeling and propagation difficulties, but the other error/uncertainty issues remain. Indeed, if the Pearson-derived PDs are propagated it appears they would result in compounded under-conservatism. DD propagation to an equivalent 1-D UQ problem coupled with use of appropriate sparse-sample 1-D UQ methods features more controllable and reasonably predictably conservatism on calculated statistics. (See Section 3 for simultaneous DD propagation for these situations.) A caveat is that the DD approach is currently configured for the same number of experimental replicates for each input variable. Relaxation of this constraint is currently being studied.

For UQ analysis involving specified *continuous* input PDs (and correlations if applicable), DD propagation is equivalent to the Monte Carlo method. It is well known that the cost (number of model simulations) of MC can be improved upon for the same or better accuracy with stratified MC approaches such as Latin Hypercube Sampling [35]. Even more cost effectiveness is attainable in many cases with surrogate methods, expansion methods, or optimization-based reliability methods previously referenced. However, it should be kept in mind that the propagation problem in this realm would usually not involve asymptotically large data sets that would closely define the continuous input PDs. Instead, they would be roughly estimated from fairly sparse data as already discussed, or would involve even more epistemic (lack of knowledge) type uncertainty if based on questionable or only loosely relevant data from a somewhat similar problem or if based on complete speculation because of a dearth of relevant data. The latter circumstances are very common in engineering settings.

Care should be taken to interpret propagation results for output QOIs according to the degree of information and rigor with which the input uncertainties are defined and propagated. UQ results from propagating highly speculative or imagined PDs should only be interpreted as rough scoping indications of an uncertainty range. It would not be justified to claim, for example, that tail probabilities could be estimated with any reliability or accuracy. On the other hand, it is possible that a bounding approach could be taken if QOI response is monotonic over conservatively wide input PDs as discussed above, but substantial lack of knowledge would call for highly conservative estimates for the input PDs to be relatively sure that results are indeed conservative. The response QOI PD would have related compounded conservatism.

Thus, response QOI statistical quantities are most legitimately and accurately estimated from propagating actual sample data. In these cases, DD propagation and sparse-sample 1-D UQ methods are arguably most accurate and efficient if sparse sample data are involved. This includes PD propagation when relatively few samples can be afforded. The sparse-sample 1D UQ methods can give relatively inexpensive reliably conservative bounds on response statistics that would be much more expense to precisely resolve with fuller propagation, by whatever method.

Returning to the context of model calibration, two other significant advantages of a DD approach exist. Concerning the calibration parameter space, it is sometimes questionable whether it is legitimate to predict with other than actually determined parameter sets from calibration. Sometimes, models do not successfully run at interpolated or extrapolated sets of calibration parameters, or analysts do not trust model prediction results obtained with synthetic parameter sets. The DD approach avoids these issues and questions by propagating actual calibration parameter solution sets. On the other hand, calibration-UQ approaches built on a paradigm of propagating an inferred or modeled JPD of calibration parameter variability must interpolate and extrapolate about the actual calibration parameter solution sets that inform the inference process.

Furthermore, the addition of new replicate experiments to the calibration data set is easier and less computationally intensive with the DD approach than with other calibration-UQ approaches. One or more added experimental data points in Fig. 1 would simply require that number of new calibrations and parameter sets propagated to simulations of the prediction model. The augmented group of QOI realizations would then be processed with 1-D sparse-sample UQ techniques as before, in a relatively simple post-processing step. In contrast, calibration approaches built on a paradigm of an inferred or modeled joint density of calibration parameter variability must conduct the inference or modeling process completely anew. The computational machinery for updating the JPD is significantly more involved for the Bayesian and other calibration-UQ approaches (discussed next) than for the SROM calibration-UQ approach discussed above. Moreover, a complete new propagation of the updated joint density of the calibration variables is required. This could add significant cost if an adequate surrogate model does not already exist from a prior propagation of the JPD from a prior calibration before the added experimental data.

2.2 Additional difficulties of Bayesian and Optimization-Based calibration approaches that indirectly infer calibration parameter JPDs consistent with the experimental response realizations

The SROM calibration-UQ approach can be considered half-way to the DD approach because it performs a separate calibration to each experiment like DD, but then uses the calibration parameters' realizations to construct and propagate a JPD. In this subsection we consider methods that infer JPDs by various *indirect* approaches that iterate parameters of candidate JPDs and propagate the JPDs through the *calibration* model to obtain predictions of experimental response variability that are most consistent with the discrete experimental realizations according to some measure of agreement.

Referring to Figure 5, one could seek to suitably estimate a distribution of the scalar experimental response from the sample data on the ordinate (ignoring the difficulties associated with sparse data as explained previously). Then one could attempt to best match the response distribution as follows. Propose one or more candidate PD forms such as Normal, log-Normal, etc. for each calibration parameter and optimize the parameters of the candidate PD(s) (the parameters are called "*hyper*parameters" in a calibration context) such that propagation of the PD(s) through the calibration model yields an output response PD that best matches the estimated experimental PD (Figure 5) according to some maximized measure of agreement or minimized measure of disagreement between the two distributions. This type of hyperparameter function optimization (HFO) approach is pursued by several methods surveyed in [14] and [36]. Other HFO methods (e.g. [16]) attempt to best match calculated statistical moments such as mean and standard deviation of the raw experimental data samples and of the calibration model's predicted response distribution. Yet other HFO methods (e.g. [11]-[14]) attempt to get results most consistent with the experimental data points by maximizing a likelihood function as explained next.

A typical likelihood function in the scalar experimental output case ([14]) is the product of M probability density values obtained by evaluating the predicted PD of response (or usually a Normal approximation of it for convenience of mathematical formulation) at M experimentally measured values of response (e.g., at the M=4 points in Figure 5). The more consistent the experimental data points and the predicted PD or Normal approximation, the higher the value of the likelihood. The likelihood and thus the consistency are maximized by optimizing the hyperparameters of the proposed variability PDs for the calibration parameters.

Moment-matching, PD-matching, and likelihood maximization mechanisms for optimizing calibration hyperparameters all involve directly working with the experimental data realizations. However, it was previously pointed out that the variance calculated from relatively few samples will often be less than the variance of the true PDF for many common PDF types. Significant error in the mean calculated from sparse sample data will also usually exist. These concerns compound with likelihood functions involving sparse experimental replicates with *functional* data such as time-series data or load-displacement or stress-strain curve data as discussed below. The concerns also transfer to Bayesian calibration methods discussed below which perform sampling based hyperparameter optimization and use likelihood functions also not explicitly developed for sparse replicate experimental data. Prior and posterior uncertainty distributions on calibration hyperparameters in Bayesian methods could potentially account for some or all of the sparse-sample epistemic uncertainty implied by use of typical likelihood formulation and PD model-form approximations, but it isn't clear if or how this is effectively done, especially for functional data or multi-parameter calibration problems. Optimization based HFO approaches do not have this additional uncertainty mechanism available to address the said error/uncertainty because they arrive at point-values of the hyperparameters.

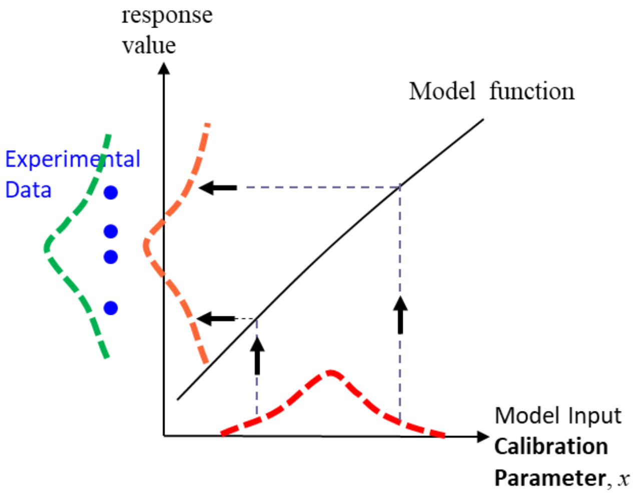


Figure 5. Illustrative indirect inference approach for a probability density of a stochastic calibration parameter. Propose one or more candidate PD forms for the calibration parameter. Then optimize the parameters of each candidate PD such that its propagation through the calibration model yields an output response PD that best matches the raw experimental data or estimated experimental PD according to some maximized measure of agreement or minimized measure of disagreement.

DD model calibration in the presence of functional data is relatively straightforward: calibrate to each test's output data field individually and propagate the resulting calibration parameter sets, much the same as in Figure 1. Additional measures are sometimes needed if physics model-form error is involved, see Section 5. The DD methodology has recently been confirmed to perform as expected (per Footnote 3) on several test problems ([2]-[4], [6]-[9]) involving functional experimental data, 1 to 11 calibration parameters, > 70 non-Normal response QOIs, and 20 to hundreds of random-data trials per stochastic calibration case.

Stochastic calibration problems with multiple experimental realizations of functional data are not as straightforward for the optimization-based HFO approaches or for Bayesian approaches to be discussed below. We cannot pursue moment-matching or PD-matching approaches because we do not have simple scalar results from each experiment to compute statistical moments or fit a distribution to. A likelihood metric is usually used to quantify agreement between experimental and model-predicted data fields. This presents some significant challenges. When scalar experimental data and predicted response are involved it is clear what the likelihood formulation for stochastic calibration problems means physically and how it supports optimal values of hyperparameters (e.g., [14]). However, things are much more complicated with likelihood functions for functional data.

Using an example of time-series data, one could consider N time-response curves from N tests, and their comparison to calibration model predicted curves (given a JPD of the calibration parameters as defined by proposed marginal PD forms or formulations, hyperparameters, and correlations). A likelihood function can be fashioned from selected time points of comparison. A likelihood function for stochastic calibration should reflect how well the population of experimental data points conform to predicted response variability at selected times over the duration of the event, such that maximizing this agreement will yield the best JPD for the calibration parameters. Though response variability at each time point will generally not be Normal, a Normal approximation is usually made in likelihood functions for mathematical and computational convenience.

The Normal distribution at each time point is formed from predicted mean response there and a variance determined in one of several ways. Some likelihood formulations for Bayesian calibration (e.g. [19], [21] and one form in [18]) tie the variance to random measurement noise in the experimental output data. This can in some cases be estimated from measurement instrumentation considerations, and perhaps as an implicit function of time if noise variance is a fixed percentage of response magnitude which evolves in time. However, this does not account for output variability due to variability of test units in the calibration experiments. This is the primary variability quantity of importance in stochastic calibration problems, i.e., mapping this variability to the calibration parameters through determination of an appropriate JPD by maximizing the likelihood function. Therefore, unit-unit variability effects should be included in the variance term in order for the likelihood function to measure how well the population of experimental data points conform to the predicted response variability at those times. It is said in [18] that the variance can be estimated if experimental output measurement random noise information is available (again, not sufficient without including estimated unit-unit variability), or an optimally effective variance value can be inferred (optimized) along with the other hyperparameters in the calibration problem, although with no added information to infer this additional degree of freedom. A second formulation in [18] uses the model-predicted response variance in the likelihood function.

A variance in the likelihood function that is significantly different from the actual data variance can cause suboptimal results in efforts to maximize the likelihood. Even if the variance used does not sensitively affect the optimal hyperparameter values arrived at, it was found with Bayesian calibration in [1] that the value of variance used can sensitively impact the posterior uncertainty about the optimal values and thus the propagated uncertainty to next-level predictions. These concerns also make it important to account (per Section 5) for effects on observed experimental output variability due to random measurement noise in experimental inputs and outputs when sparse experimental replicates are involved, because it is fairly likely that partial variance cancellation will make the observed unit-unit variability less than the actual variability when sparse replicate tests are involved ([37]).

The independent and identically distributed (IID) data assumption in the commonly used simple product form of the likelihood function requires that the comparison time-points used must be far enough apart to avoid time-correlation between them. The field data must therefore be thinned in time. Data thinning can eliminate useful experimental information for calibration.

Moreover, experimental data realizations for highly stochastic phenomena often have differences that are important to capture. For instance, time curves or stress-strain curves may not end at the same abscissa value because material samples fail at different strain values, or thermal-chemical phenomena experience exothermic runaway at different times, etc. In more mundane cases the experimental realizations may peak at substantially different strains, times, spatial locations, etc., or just generally be out of phase in behavioral trends. Capturing these individual differences can be very important for predicting variations in stochastic behavior. The DD approach produces calibration parameter sets that reflect individual behaviors. On the other hand, the likelihood function is a measure of global fit over all the experimental realizations. Maximizing the likelihood function rewards calibration hyperparameter values that yield a best global compromise fit to all the data. This does not best reflect individuality of stochastic behaviors in calibration or prediction settings.

The likelihood function's IID assumption arguably applies across experiments at a given point in time, but identically distributed results would not typically exist through time even if the responses at the thinned retained times are effectively independent. Experimental responses at one time would typically have different variance than responses at a substantially different time. The preceding paragraph presents some examples why. Other examples are if the responses of the tested systems flare out and/or come together (and even cross) at various times. Physics model-form error may also cause irreconcilable variance differences over time between experimental and calibrated model results. The freedom exists to have variance in the likelihood function be different for each comparison time-point, but this violates the IID assumption in the commonly used simple form of the likelihood function. In all, the global comparison metric and global compromise optimization associated with likelihood functions present complexity, difficulty, and uncertainty relative to DD's individual experiment best-possible calibrations and parameter propagations, especially for functional data.

Another concern is the need for response-surface surrogate models for temporal response as a function of the calibration variables' hyperparameters being optimized. Each original calibration parameter will have at least two hyperparameters in the PD model used to represent stochastic variability of that calibration parameter. This doubles the number of variables that must be inferred and propagated, adding substantial cost due to added dimensionality in both procedures. Therefore, response-surface surrogate models are usually needed for affordability of the optimization. These bring complexity and surrogate-related error/uncertainty that should be estimated and accounted for. (Surrogate related error can be substantial if the physics model is expensive and/or the parameter space is high-dimensional and/or nonlinear enough that an affordable budget of physics model simulations to construct the surrogate limits its accuracy substantially.) The DD approach is normally affordable without creating or using surrogate models.

Bayesian model calibration approaches also involve the complications of likelihood functions and surrogate models. They involve further difficulties of indirectly inferring appropriate uncertainty distributions for calibration parameters or hyperparameters, and in some formulations, also for model-form related bias/discrepancy and/or experimental "noise" terms. Bayesian approaches with hyperparameter formulations sample the hyperparameter uncertainty space and identify the more successful realizations that yield PD model realizations of calibration parameter variability whose propagated results produce high values of the likelihood function. These approaches yield uncertainty distributions on the hyperparameters instead of point values that the optimization-based HFO approaches do. Alternatively, some Bayesian formulations for the stochastic calibration problem may not solve for hyperparameters of PD models for calibration parameter variability. Instead, posterior uncertainty distributions are sought for the calibration parameters themselves (each treated as having a fixed value, but uncertain), with the stochastic variability effects mostly mapped into model-form related bias/discrepancy and/or experimental "noise" terms (see e.g. [11]).

In general for Bayesian methods, several pivotal choices must be made in terms of what formulation will be used, and within a formulation, what sampling procedures, surrogate model types, and calibration parameter priors will be used. Hence there is considerable variability in formulations, and procedures, and results interpretation in practice.

For example, for the stochastic calibration problem it is argued in [18] that a discrepancy term should not be used because variability mapped to it instead of to the calibration parameter variables would not be physically reasonable and would not extrapolate well to next-level predictions. The authors go on to demonstrate that a hyperparameter formulation with a noise term performs more physically accurate and consistent in prediction than a formulation for posteriors on the calibration parameters themselves and a noise term that most of the stochastic variability gets mapped into. Meanwhile, on the Sandia challenge [51], two hyperparameter Bayesian methods with noise term were applied by two teams [19], [20]), and another team ([21]) applied a non-hyperparameter formulation with noise and bias terms that map most of the experimental variability into the bias term. In the two hyperparameter approaches there were significant differences in priors, PD modeling/parameterization schemes, sampling procedures, surrogate modeling approaches, and measures of prediction and experimental data agreement (likelihood function vs. U-pooling metric—the latter also does not explicitly account for epistemic uncertainty due to sparseness of experimental replicates). Between the various hyperparameter formulations [18]-[20], different methods are used for calibration variable PD representation and determination (respectively a polynomial chaos expansion method with variable number of terms/hyperparameters to model each calibration variable PD vs. a Johnson family 4-hyperparameter model per PD vs. a Pearson system 4-hyperparameter model). The three hyperparameter methods [18]-[20] used different likelihood formulations.

Bayesian methods require proposed prior distributions for each of the calibration parameters or hyperparameters. The parameter values are said to be fixed but unknown to within uncertainty proposed by the specified initial or prior information. Bayes' rule and associated computational machinery are used to update the uncertainty (the prior) into a posterior distribution that is narrower than the prior if the results from calibrating to the experimental data are highly consistent with the prior. This is a double-edged sword, however. A prior that conflicts with or is non-representative of calibration parameter values indicated by the experimental data can increase the variance (though not the width) of the posterior distribution relative to the prior distribution. When the experimental data is relatively sparse, the prior (if not a sufficiently wide non-informative uniform distribution) will have substantial influence on the posterior distribution arrived at. Real engineering projects often have very sparse experimental replicates and little to no information for good priors. In such cases the required involvement of priors would appear to be more of a liability than an advantage. See the Appendix of this paper for some evidence of this. On the other hand, non-informative uniform priors do not accelerate arrival to accurate posteriors. So, numerous replicate tests are required for accurate posteriors whether inaccurate non-uniform priors or non-informative uniform priors are involved. In [18], for example, an impractically high 100 material tests were required to stabilize the hyperparameter posteriors starting from non-informative uniform priors.

2.3 Synopsis

In summary, substantial craft is involved in using Bayesian calibration methods for stochastic problems. Analysts must make consequential choices of: formulations; measures of agreement or disagreement with experimental data; sampling procedures; surrogate models; and priors. It is observed that substantially different choices are made by different analysis teams even on the same problem (as described above for the Sandia Storage Tank Problem, and summarized in [11] for the Sandia Thermal Challenge Problem [52], and summarized in the Appendix of the present paper in connection with the Sandia end-to-end UQ test problem [46]). Such variability in analysis choices and implementations would be expected to bring considerable variability to prediction results and interpretations as well. This occurred with the two Bayesian variants in [18], three variants on the Sandia tank problem [51], and three variants on the UQ test problem [46] (see Appendix). In many realistic engineering circumstances, the vagaries associated with specifying priors ([53]) are a disadvantage. Epistemic uncertainty due to sparse experimental replicates is not explicitly accounted for in calibration results and next-level predictions. Also, conceptual and implementational complexity is much higher than the alternative approaches.

Optimization-based HFO methods and Bayesian methods for stochastic calibration problems involve substantial approximations in likelihood functions, surrogate models, and calibration parameter variability distributions. These approximations may incur substantial calibration error. Moreover, epistemic uncertainty that accompanies sparse experimental replicates is not explicitly accounted for in calibration results and next-level predictions. The optimization-based HFO methods are conceptually and implementationally simpler than Bayesian methods, but only yield point values of the hyperparameters—

uncertainties from the approximations involved are not reflected. It is not apparent how calibration parameter or hyperparameter uncertainties obtained with Bayesian methods can be made to appropriately reflect the many approximations involved.

The optimization-based SROM calibration-UQ method is less complex than those above. It does not involve likelihood functions, hyperparameters or indirect inference of calibration parameter variability distributions, or construction of surrogates for the calibration procedure. However, it does involve modeling and propagation of calibration variable PDs and correlations. Modeling of the JPD will involve substantial unaccounted-for error/uncertainty, especially when sparse experimental replicates are involved. Additionally, propagation of the JPD to the next-level application may involve non-negligible propagation error. This could arise from a computer simulation budget that significantly limits the accuracy of the propagation results or of a surrogate model constructed for the propagation. These difficulties are also true for Bayesian and optimization-based HFO calibration approaches, but compounded by the higher dimensionality of the calibration and propagation problems when hyperparameters are involved (along with the added difficulties cited in the prior two paragraphs). All these sources of error/uncertainty, complexity, and cost are avoided with the DD method, whether for scalar or functional experimental data.

The DD approach is the simplest and usually least expensive approach. It directly reflects the aleatory variation of the experimental results without needing to explicitly construct variation and correlation structures of the calibration parameters. It also avoids the cost and uncertainty of propagating a JPD of the calibration parameters. DD propagates N calibration sets from N individual deterministic calibrations to N experiments. This provides a simple and directly traceable tie between experiments, associated calibration parameter sets, and predictions. The N samples of response for scalar QOIs are processed with relatively simple 1-D UQ methods that yield relatively robust and efficient bounds on statistical quantities commonly of engineering interest—even with very low numbers of calibration experiments. Relevant evidence of this is presented in ([2]-[4], [6]-[9]) and the Appendix of the present paper.

For these types of useful prediction quantities (statistics of response) and for the full-identifiability conditions specified at the start of this Section 2, the DD approach is concluded to generally be the simplest, least expensive, and most accurate of the calibration-UQ approaches the author is aware of. It is projected that DD retains a strong advantage when considering the other sources of epistemic uncertainty itemized in Section 1 and addressed in Sections 4 and 5 of this paper. DD brings these advantages to an efficient and effective "simultaneous" DD approach to calibration, propagation, and prediction where multiple component models of a system model are each calibrated to sparse replicate scalar or functional experimental data. This is explained in the next section.

3. Simultaneous Discrete-Direct Method for calibration of multiple component models of a system model

This section briefly explains the Simultaneous Discrete-Direct (SDD) method for efficient calibration, propagation, and prediction when multiple material, phenomena, and/or component models of a system model are each calibrated to sparse replicate scalar or functional data. Figure 6 shows an SDD calibration example where two calibration tests and DD calibrations are conducted for each of three submodels to be used in a system model.

Figure 7 shows two possible combinations of DD calibrations of submodels used in system-level predictions where each test and calibration result is used once and only once. (As explained in [2], the once-and-only-once usage has an analogue to stratified sampling approaches like Latin Hypercube Sampling for efficient Monte Carlo propagation of probabilistic uncertainty. An advantage is empirically

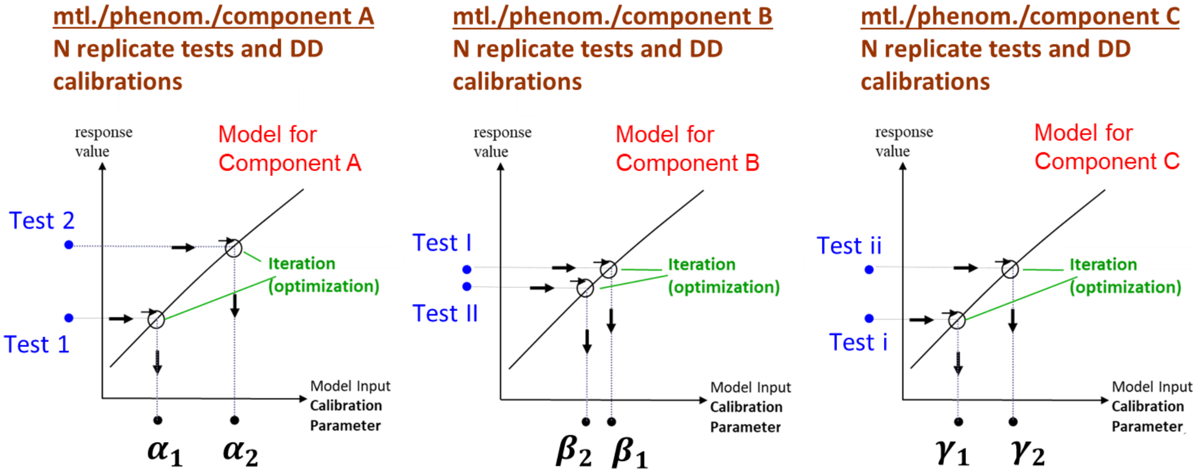


Figure 6. Simultaneous DD calibration example problem with N=2 calibration tests and discrete-direct calibrations for each of M=3 submodels to be used in a system model.

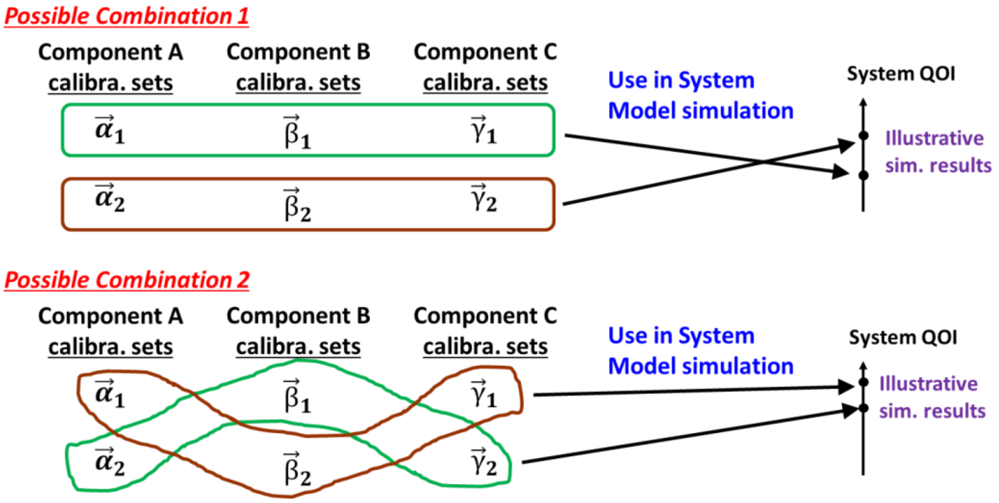

Possible Combination 1
Component A calibra. sets
Component B calibra. sets
Component C calibra. sets
$\vec{\alpha}_1$
$\vec{\beta}_1$
$\vec{\gamma}_1$
$\vec{\alpha}_2$
$\vec{\beta}_2$
$\vec{\gamma}_2$
Use in System Model simulation
System QOI
Illustrative sim. results
Possible Combination 2
Component A calibra. sets
Component B calibra. sets
Component C calibra. sets
$\vec{\alpha}_1$
$\vec{\beta}_1$
$\vec{\gamma}_1$
$\vec{\alpha}_2$
$\vec{\beta}_2$
$\vec{\gamma}_2$
Use in System Model simulation
System QOI
Illustrative sim. results

Figure 7. Two possible combinations where N=2 experiments and DD calibrations for each submodel are used once and only once in N=2 system-level simulations to get N=2 predictions (realizations) of response variability. Calibration parameter sets are designated as vectors in this figure to signify that each calibration problem may have multiple calibration parameters.

established vs. non-once-and-only-once usage schemes also tested and characterized in [2].) Each combination produces a number N of system-level simulations and predictions (realizations) of QOI response variability corresponding to the N tests and calibrations per submodel. Each set of N realizations is processed into one or more decision-relevant statistical measures of response with 1-D sparse-sample

UQ techniques per Footnote 3 in Section 2.1. The estimated statistic(s) from combinations 1 and 2 would be different but equally legitimate; there is no reason to favor results from either combination over the other. Recent research in [2], [4], and [24]-[26] has found that an average of such equally legitimate results is typically more accurate than individual results. For this N=2 M=3 SDD calibration and uncertainty propagation example, four equally legitimate combinations exist where each of N submodel tests and calibrations is used once and only once in the system-level simulations. Figure 8 shows the four combinations. It would take eight system model runs if one wanted to average results of all combinations to get the best result. For N>2 and/or M>3 there are more equally legitimate combinations available. Because one is averaging results, there is little reason to pick more than 5 combinations to average. This coincides with a strong knee in the curve of cost vs. accuracy of a mean calculated from random samples from a Normal distribution (see Figure 9). (It is anticipated that estimates (of statistical quantities) from equally legitimate combinations as discussed here are Normally distributed, or approximately so, although this has not yet been empirically examined.) Thus, the number of system model runs with SDD would scale at ≤5N, where N is the number of experiments and calibrations per each submodel of the system model. The number of calibration experiments per submodel is often just a few, so the system model propagation cost is on the order of 10 to 20 and therefore no system-level surrogate model is necessary. The SDD approach is currently configured for experimental designs where the same number N of replicate experiments and DD calibrations are performed for each submodel. Relaxation of this constraint is currently being studied.

The SDD methodology has recently been successfully applied in UQ test problems [4] and [7]. The test problem [7], for example, involved SDD calibration of models of three electronic devices in a system-level circuit model. For each submodel, 10,000 (10K) calibration parameter sets that match synthetic "experimental" test data for that device (time-response curves of realistic shape and random curve-to-curve variations) were generated. The 10K calibration parameter sets of each device were used in system-level circuit simulations to produce 10K realizations of time-dependent response for a scalar output quantity. Response values at a time A of interest were extracted to form a scalar QOI of 10K samples. Three threshold levels (L1, L2, L3) were determined such that proportions of the QOI's 10K samples above the thresholds were respectively 0.1, 0.01, and 0.001.

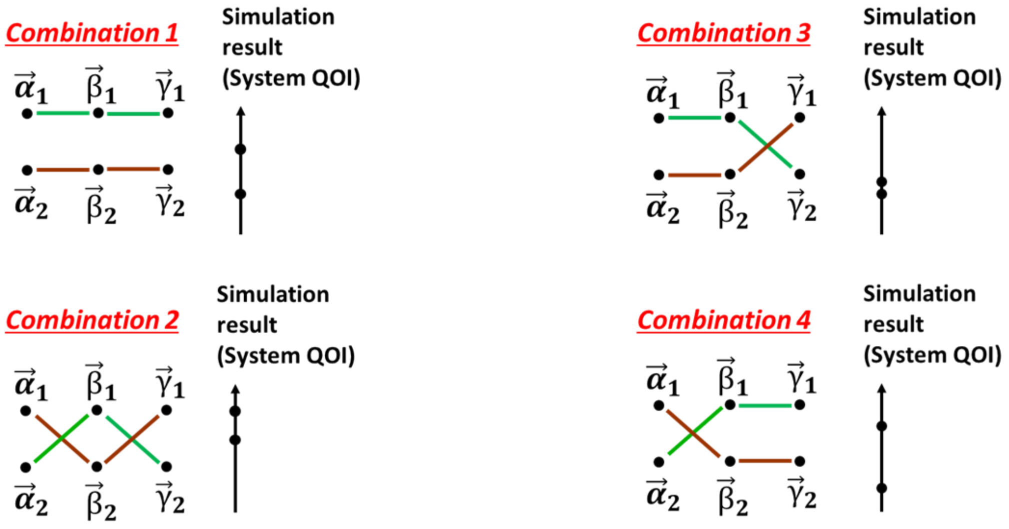

Combination 1
Simulation result (System QOI)
$\vec{\alpha}_1$ $\vec{\beta}_1$ $\vec{\gamma}_1$
$\vec{\alpha}_2$ $\vec{\beta}_2$ $\vec{\gamma}_2$
Combination 3
Simulation result (System QOI)
$\vec{\alpha}_1$ $\vec{\beta}_1$ $\vec{\gamma}_1$
$\vec{\alpha}_2$ $\vec{\beta}_2$ $\vec{\gamma}_2$
Combination 2
Simulation result (System QOI)
$\vec{\alpha}_1$ $\vec{\beta}_1$ $\vec{\gamma}_1$
$\vec{\alpha}_2$ $\vec{\beta}_2$ $\vec{\gamma}_2$
Combination 4
Simulation result (System QOI)
$\vec{\alpha}_1$ $\vec{\beta}_1$ $\vec{\gamma}_1$
$\vec{\alpha}_2$ $\vec{\beta}_2$ $\vec{\gamma}_2$

Figure 8. All four possible combinations for the N=2 M=3 SDD calibration and uncertainty propagation problem where each submodel test and calibration is used once and only once in system-level simulations to obtain N predictions (realizations) of response variability.

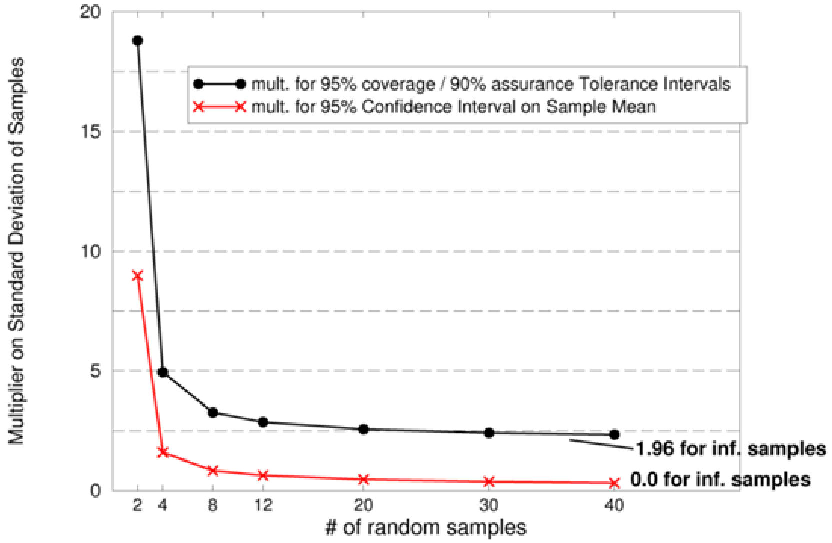


Figure 9. Multiplier on calculated standard deviation of Normally distributed samples to create a 95% confidence interval on the mean calculated from the samples: 95% CI = calculated mean ± [multiplier times calculated standard deviation]. (The multiplier curve for tolerance intervals is to be discussed in Section 5.)

These three “tail probabilities” of QOI response were then estimated from more realistic circumstances of each device model being calibrated to N “experimental” data curves for that device (where N =2, 3, or 4), as per Figure 6 for N=2. Then the N calibration parameter sets per device were used in N system-level circuit simulations to produce N response samples for the QOI. From the N response samples, tail probabilities were estimated given the three threshold levels L1, L2, L3. Estimates for the tail probabilities were obtained using the recommended sparse-sample methods in [26]. Estimates were obtained and averaged for equally legitimate combinations of the calibration parameter sets, as per Figure 8.

Averaged estimates were obtained in this way for 21 cases total: three random trials were performed for N=2 tests and calibrations per device, two trials for N=3, and two trials for N=4, for a total of seven trials for each of the three tail probability magnitudes. All 21 tail probability estimates were conservative, yielding larger tail probabilities than the true values, as expected. This is preferable to the other way around, which would incur far greater program risk when designing and assessing by model predictions.

A conservative bias is taken in the sparse-data UQ methodology because accurate estimates of tail probabilities are impossible with very sparse samples of response. The methods are designed to give high assurance of attaining conservative tail probability estimates (see Footnote 3 in Section 2.1). In the test problem [7] the results were very conservative for the 0.001 tail probability magnitude: estimates in the seven trials were 1 to 2 orders of magnitude higher probability than actual. This is the tradeoff for high assurance of not under-estimating tail probabilities. Nonetheless, the sparse-sample estimation methods recommended in [26] are the best known to the author for lowest average over-estimation error while maintaining high assurance of not under-estimating. The recommendations come from an extensive study involving over 300 million random trials over a test matrix of N=2 to 20 samples; tail probability magnitudes of 10^{-5} to 10^{-1}; 16 diversely shaped response PDs; and about 20 sparse-sample 1-D tail-probability estimation methods and combinations of methods.

The traditional alternative to the SDD paradigm would be to separately calibrate the M submodels to arrive at M PDs or joint PDs of calibration parameter values. These would be propagated in system-level simulations to estimate variability in response QOIs. This would be more complicated and typically more computationally expensive than the SDD approach, both in the calibrations (per Section 2) and in system-level UQ propagation. Another disadvantage is that any over-conservatism or under-conservatism of stochastic variability approximated in the M PDs or joint PDs of calibration parameter values is unlikely to offset completely when all the PDs are propagated to system level. Compounding of over-conservatism or under-conservatism can yield highly over-conservative or under-conservative stochastic variability in response QOIs (see discussion and references near mid-Section 2.1). Thus, there is far less control and predictability of estimation error/uncertainty than with the SDD approach.

4. DD method progress on addressing calibration parameter non-uniqueness

This section summarizes initial investigations regarding important sources of error and uncertainty in DD calibration related to fundamental lack of identifiability of calibration parameter values and/or incomplete convergence of parameter values due to finite numerical precision in the optimization. This section considers a multi-start calibration-UQ methodology to address these sources of error and uncertainty when calibration, propagation, and prediction are based on a single test. Section 5 describes how the methodology is adapted to DD calibration-UQ involving several replicate tests.

Figure 10 shows results of 100 calibrations of a Johnson-Cook (JC) strain-rate-dependent material strength model calibrated in DD fashion to synthetic load-displacement data curves from 100 simulated tension tests at each of two different strain rates [2]. The figure shows %deviations of the JC model's four calibration parameter values relative to their known true values in the calibration-UQ test problem. In the test problem the 100 "truth" calibration parameter sets were generated by uniform random sampling over a 4-D calibration parameter space defined by interval ranges of the four calibration parameters. The parameter ranges reflect realistic levels of stochastic variability of a material being modeled. An investigation into calibration parameter convergence precision error and non-uniqueness (not undertaken in [2]) is described next.

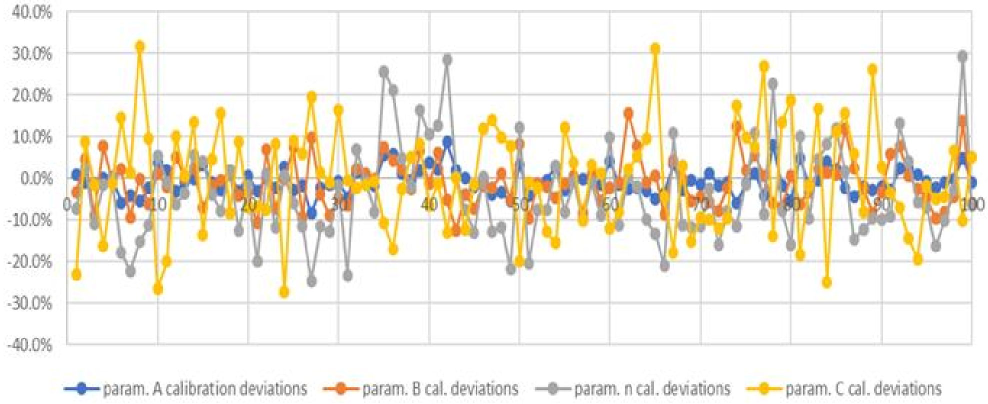


Figure 10. Approximately symmetrically distributed %errors (about zero) in calibrated material model parameters due to objective function numerical noise and parameter non-uniqueness.

The 100 4-parameter truth sets along with the JC stress-strain constitutive model were used in finite-element simulations of material tension tests at two different strain rates. For each strain rate, the simulation yields a load-displacement (L-D) curve given a 4-parameter set of JC model parameters. Thus, for each of the 100 4-parameter sets, two synthetic experimental L-D curves were generated for the two different strain-rate "tests". The corresponding pair of synthetic experimental L-D curves were matched as well as possible at 17 selected points along each curve by calibrating the 4 JC parameters in simulations of the two strain-rate tests. The calibrations minimized the sum of squared errors at the combined 34 comparison points of the two different strain-rate curves.

The errors in the final parameter values from the 100 calibrations are shown in Figure 10. There is no model-form error here; the same JC model and simulated tests were used to produce the L-D "experimental" data curves and in the calibrations to best match them. The precision errors or deviations in the calibration parameter values are attributed to incomplete optimizer convergence due to objective function numerical noise in the optimization process, and/or to flatness in the parameter space which is symptomatic of fundamental non-identifiability/non-uniqueness where multiple different realizations of the four parameters yield similarly good matches to the experimental data.

Some of the 100 cases have almost perfect match between the "experimental" and model-predicted L-D curves at the final optimum set of calibration parameters, yet errors relative to the exact calibration parameter values are substantial. The optimal {A, B, n, C} values for three such cases are {-0.5%, -2.5%, -2.0%, 5.1%}; {-0.5%, 0.6%, -0.8%, 3.7%}; and {0.9%, 1.5%, 1.8%, -8.3%}. This reveals substantial parameter value non-uniqueness where both the exact parameter sets and the different (calibrated) ones yield essentially the same L-D curves. These cases have lower average error in the data curve agreement than the cases with first and second lowest combined parameter value error magnitudes of {0.3%, -0.9%, 0.0%, -1.3%} and {-0.2%, 1.0%, -0.6%, -1.4%}. These cases also closely match the experimental L-D data curves. There are also many cases with neither close convergence to the exact parameter values nor great agreement with the calibration data curves. The worst case on both accounts has parameter value errors of {-8.5%, 9.7%, -24.7%, 19.6%} and average absolute deviation of 1.1% at the 34 data points of the L-D data curves to be matched as closely as possible in the calibrations. Some cases have even higher individual errors in A and/or B and/or n and/or C, but with lower combined errors than the worst case just cited.

Thus, the JC calibration study has substantial deviations from the known exact calibration parameter values that are attributable both to fundamental non-uniqueness and to numerical noise and finite convergence tolerances in the optimization. An approach to account for these effects in DD calibration and uncertainty propagation with multi-start optimization is investigated next.

We begin by noting that the calibrations in the study were performed with a hybrid global-local optimization procedure. First, a non-gradient based simultaneous global-to-local optimization procedure was used: the method of divided rectangles, DIRECT [38]. DIRECT has a good reputation for efficiency and robustness for small to moderate numbers of calibration parameters, simple bound constraints, and potentially noisy objective functions from optimizer evaluations of computational physics models. Once the prescribed stopping tolerance is reached in DIRECT, a handoff is made to a second, local optimizer (NL2SOL gradient based trust-region method with adaptive selection of Hessian approximations as explained in [1]) for fast convergence to the final local (hopefully global) optimum.

The optimization procedure starts from a definition of a hypercube over which the global-to-local optimization is to be conducted. In the present study this was the JC 4-parameter space discussed above. Even though the 100 calibration test points (to hopefully be converged to) were at different locations in the parameter space, the starting region for each calibration procedure was the same—the 4-D hypercube that was uniformly sampled to generate the 100 tests points. Given this setup, each calibration test point's calibration procedure was started from a hypercube that was not centered about the test point, so each test point is converged-to from a different direction and starting region as depicted notionally at left in Figure 11 for two different test points. This observation will be leveraged shortly.

Analysis of each parameter's deviations in Figure 10 reveals that they are approximately symmetrically distributed random quantities with ~zero mean error. More importantly, effects of the combined calibration parameter errors $\overrightarrow{\Delta p}$ have been examined on predictions of 29 highly non-linear stress and strain output QOIs from "Can Crush" structural simulations ([2]). For each QOI, 100 reference values are determined when the Can Crush model is run with the 100 truth parameter sets. The Can Crush prediction model is also run with the 100 calibration parameter sets from the 100 calibrations. The calibration related prediction errors Δ_iQOI in the 100 predictions of each QOI are also approximately symmetrically distributed about the exact results. This is notionally illustrated at far left in Figure 11 and is a very useful finding if the notion of reciprocity in the figure holds for the application problem [2].

The current findings are for 100 different calibration truth points converged upon starting from the same starting hypercube region. Per the notions presented in Figure 11 and its caption it is proposed as plausible that the 100 results are characteristically similar to results that would occur in this problem if we had a single truth calibration end-point converged upon by calibrations started from 100 random starting regions and directions about the true calibration end-point. If such reciprocity holds, then the observed calibration-related prediction error data Δ_iQOI in the test problem [2] can be used to assess whether a multi-start optimization procedure with sparse random starting points could support reliable confidence intervals for QOI prediction error/uncertainty due to calibration parameter non-identifiability, optimization numerical noise, and finite convergence tolerances.

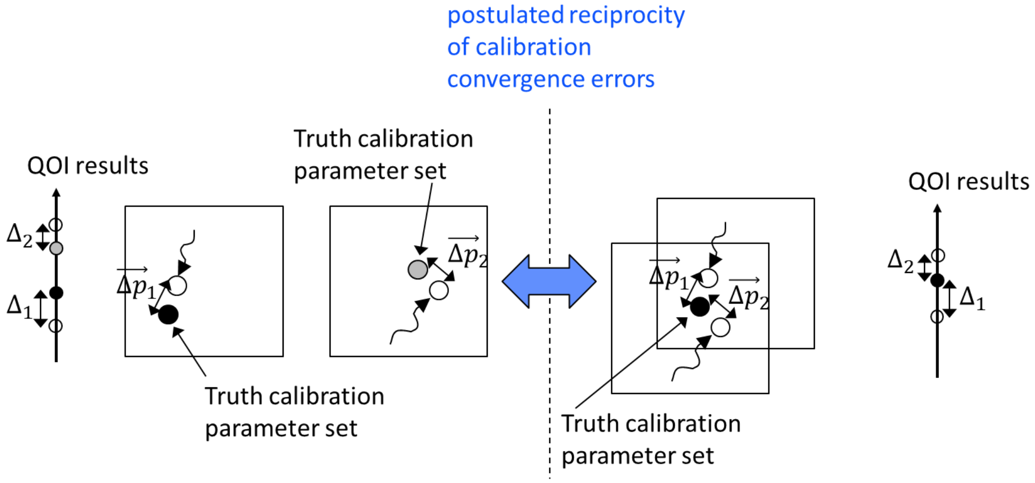

postulated reciprocity of calibration convergence errors
QOI results
Δ_2
Δ_1
$\overrightarrow{\Delta p_1}$
Truth calibration parameter set
Truth calibration parameter set
$\overrightarrow{\Delta p_2}$
$\overrightarrow{\Delta p_1}$
$\overrightarrow{\Delta p_2}$
Truth calibration parameter set
QOI results
Δ_2
Δ_1

Figure 11. Postulated reciprocity of calibration convergence errors in 2-D illustrative analogue of Johnson-Cook/Can-Crush test problem. The two calibration cases at left start from the same global hypercube region for the DIRECT optimization procedure but the region is not centered about either truth solution set of calibration parameters and the end points of each calibration differ from the truth parameter sets by the indicated convergence errors $\overrightarrow{\Delta p}_i$. Statistically similar convergence errors are postulated at right in a proposed notional "reciprocity" of calibration errors from a related calibration problem with the indicated truth parameter set and two calibrations starting from different hypercube regions and converging toward the truth set but terminating at the indicated points with the said convergence errors. The proposed reciprocity in the figure postulates that the starting regions, calibration problems, and convergence errors are relationally and statistically similar relative to their truth end-points in the two cases at left and at right.

For illustration, consider the QOI results in Figure 11. If reciprocity exists, then the prediction errors Δ_1QOI and Δ_2QOI at left in the figure from propagating the two calibration parameter sets would be representative of calibration related prediction errors Δ_1QOI and Δ_2QOI about a truth response value at right in the figure corresponding to propagation of two parameter sets obtained from different optimization starting regions. As said above, the 100 prediction errors Δ_iQOI for each of the 29 QOIs were observed to be approximately symmetrically distributed about zero. A statistical confidence interval (CI) can be constructed from a random set of error samples Δ_iQOI under an assumption that they come from a Normal population. Then the CI would be expected to contain the mean value, which effectively represents zero prediction error.

Under this possibility, 95% CIs were calculated from each consecutive set of five error samples from the 100 samples per QOI. This yielded 20 CIs for each of 29 QOIs. It was found that 577 of the 580 95%CIs (29 QOIs X 20 random trials each) successfully contained the mean zero-error values. This presents an empirical success or confidence rate of 99.5%, much higher than the nominal advertised 95% rate with the 95% CIs. If sets of 10 of the 100 results per QOI are used to construct 95% CIs, an empirical success rate of 100% is obtained in the 290 cases (= 29 QOIs X 10 random trials each).

Hence, it is plausible that DD calibration with multi-start optimization started at five random starting points (or hypercube regions for DIRECT) within a domain of reasonable uncertainty on the calibration parameters may enable effective sampling and UQ treatment of calibration parameter value non-uniqueness and optimizer precision error effects.

A more direct investigation is discussed next. Figure 12 shows errors in response QOI predictions with calibrated parameters vs. true parameter values. As part of the work [7], 20 random values of a calibration parameter of a stochastic electrical device response model were generated. For each of these 20values a simulation was performed for time-dependent device response under an applied excitation. Then a calibration was performed to attempt to recover each parameter value by best matching the associated time-dependent response curve. A DIRECT + NL2SOL optimization method as described previously was used for the 1-D calibration problem. Optimizations/calibrations were performed with both "loose" and "tight" convergence tolerances and starting with two different interval ranges centered at 0.9 and 1.1 of the true parameter value. This was done for all 20 cases. This is a "2-point" multi-start optimization approach for the 1-D calibration problem. Ultimately, three to five different starting points or hypercube regions are recommended even for a 1-D calibration problem (see Section 5.1).

Figure 12 shows % deviations of peak predicted responses (over time) using the calibrated parameter values vs. predicted peak responses with the exact parameter values in the 20 cases. The results using calibration parameters from optimizations with tight tolerances plot essentially on top of each other. However, noticeable separation exists between pairs of loose-tolerance results obtained with calibration starting regions centered at 0.9 and 1.1 about the true parameter values. For these cases, 95% CIs were formed from each pair of results. In all 20 cases the 95% CIs contained the zero-error value which corresponds to predictions with the true calibration parameter values. This supports the proposition that DD calibration with even a 2-point multi-start optimization can enable effective sampling and treatment of calibration parameter value non-uniqueness and optimizer precision error effects.

Very different results occur with the tight tolerances, but end conclusions are similar. The tight-tolerance prediction results are closely co-located, so their 95% CIs are extremely small. None of the 20 CIs contain the zero-error value. However, this is not a significant miss for 17 of the 20 cases where the prediction deviations are < 0.4% and this would often be considered acceptably small calibration/optimization related precision error or noise. The remaining three cases have significantly larger (but not large) optimization-related deviations of about 1.5% to 2%. It is notable that these optima were repeatable, arrived at from starting regions on different sides of the true parameter values. This may have something to do with the global-to-local tradeoffs in searching the space with the non-gradient based DIRECT optimizer and/or a premature handoff to the local gradient-based NL2SOL optimizer with its Hessian approximation method.

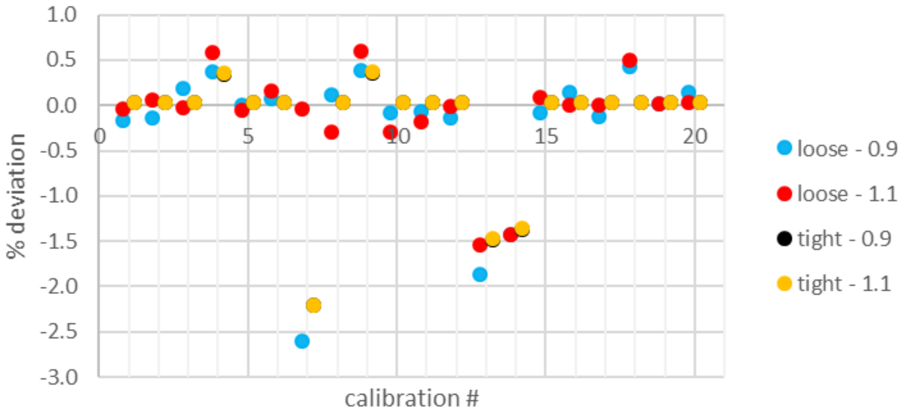


Figure 12. Errors in response QOI predictions with calibrated parameters vs. true parameters, for 20 electronic device calibrations at loose and tight convergence tolerances, each starting from two substantially different regions in the parameter space. For each of the 20 cases, predictions with calibration parameters from loose optimization tolerances are slightly jittered leftward relative to results with parameters from tight tolerances. The tight-tolerance results lie essentially on top of each other.

In summary, the tight-tolerance calibration cases yielded QOI prediction errors that were not identified or accounted-for with the applied 95% CI 2-point multi-start calibration-UQ approach, but only 15% of the 20 cases involved calibration precision related prediction errors of a couple %. The other 85% of cases involved inconsequential failures of the CI error estimates if prediction error < 0.4% due to calibration precision error can be considered reasonable and acceptable. For the less expensive loose-tolerance calibrations, the DD method with 95% CIs effectively captured the true response predictions in all 20 cases, including 6 cases with QOI prediction errors as large as 0.5% to 2.5%. On balance, this is favorable for the proposed multi-start calibration-UQ approach, although much more testing and characterization of the approach is needed on a large diversity of calibration problems. Another favorable indicator, though less direct, comes from very high >99% success rates of 95% CIs in the first investigation in this section if the conjectured reciprocity holds.

5. Incorporating experimental, non-uniqueness, and model related uncertainties into the DD calibration-UQ method

This section describes combined treatment of stochastic/aleatory and epistemic/lack-of-knowledge type uncertainties in DD calibration, propagation, and prediction. The emphasis is other sources of epistemic uncertainty beyond that from sparseness of experimental data addressed by the 1-D sparse-sample UQ techniques already discussed. Most of these other sources of epistemic uncertainty were addressed in [9] except for the calibration/optimization related uncertainties considered in Section 4. All are addressed in the UQ treatment outlined in this section.

5.1 Calibration optimization uncertainties from Section 4 extended to multiple replicate tests

We start by explaining how multi-start treatment of calibration/optimization related uncertainties in Section 4 is adapted for DD calibration-UQ methodology involving multiple replicate tests. (Section 4 considered multi-start calibration-UQ methodology for a single test.) Figure 13a shows illustrative results for 5-point multi-start DD calibration methodology applied for each of N=3 tests. For each test there are K=5 randomized starting points for K calibrations (for a total of K*N calibrations in all). The three horizontal lines in Figure 13a signify response QOI predictions from ideal best-possible calibration parameter values for each test. These would yield the best-possible match between the experimental data and model-predicted response in the calibration if the optimizer does not get trapped in local optima from computational-noise or physical origins and so finds the best global optimum (or one of many equally best optima if non-unique global optima exist in the calibration problem). Thus, the ideal best-possible parameter values are most reflective of the experimental results, given the model being calibrated. Any model-form related calibration and prediction errors/uncertainty are normally well addressed in a context of next-level predictions through model validation, bias correction, and extrapolation approaches.[4]

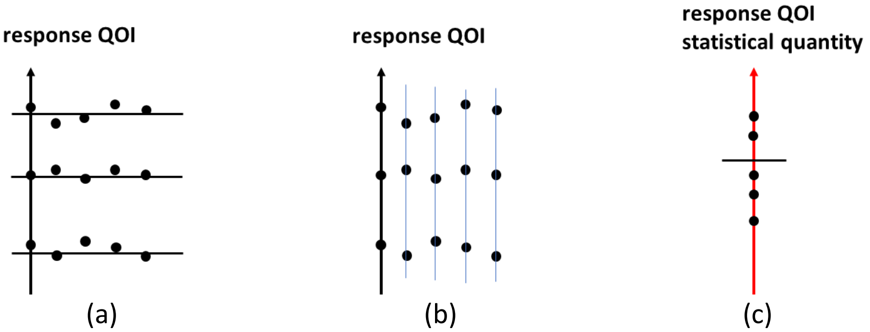


Figure 13. (a) Illustrative QOI estimates from calibration parameter sets of 5-point multi-start DD calibrations applied for N=3 tests. (b) K groupings of N=3 response QOI estimates. (c) K estimates of response QOI statistical quantity from the K groups of N=3 samples. Axis is a different color (red) to signify that it is different from the QOI.

From the investigations in Section 4 it is apparent that the ideal best-possible calibration parameter values may not be exactly obtained in practice. The multi-start calibration approach provisionally addresses this as investigated in Section 4. The approach also caters to any issues of fundamental non-uniqueness of best parameter sets by seeding a diverse sampling of the parameter space to identify multiple approximately equally good parameter sets if they exist. Multiple parameter sets that match the calibration data equally well may yield somewhat different QOI results when evaluated in a next-level prediction model. Because each is an equally legitimate and good parameter set, it would be important to sample their possibly

[4] In some cases a model may have sufficient calibrated and non-calibrated degrees of freedom to fully match the experimental data. However, models are always imperfect abstractions of reality so will contain model-form error whether or not they fully match the experimental data upon calibration. For example, full matching was achieved in [3] with a simplified isotropic constitutive model fit to synthetic tension-test data generated from a higher-fidelity anisotropic model. Even though the isotropic model perfectly matched the calibration data, next-level structural predictions with the simplified calibrated model had differing accuracy for different QOIs when compared against results with the anisotropic truth model used in the structural predictions. In general, model-form error exists whether the calibration data is fit perfectly or not. The model-form error may be consequential or inconsequential for a given next-level prediction, depending on a number of factors including: model discretization effects; degree of extrapolation in boundary conditions, geometry, and state conditions like temperature and strain rate; the particular output response quantity of interest; and the accuracy needed for the engineering decision or task. The most trustworthy way to quantify the effects of model-form related prediction error is through well designed and executed model validation assessments that are sufficiently representative of the eventual predictions to be made with the model. The quantitative validation results can then be used for prediction bias correction, if needed, in a manner that can be extrapolated on a rational basis with associated uncertainty, e.g. [40].

different prediction effects. This can be done at no added cost in the course of sampling and propagating other sources of uncertainty in the calibration problem (see Section 5.2).

Figure 13a shows illustrative results from predictions with K=5 parameter sets obtained from 5-point multi-start calibrations to each of the N=3 test data sets (for a total of K*N evaluations of the next-level analysis model to get K*N response values per output QOI). If any of the final parameter sets correspond to significantly inferior fits to the calibration data, the optimization run likely got trapped in an inferior physical or numerical local minimum. The inferior set(s) could be replaced with one or more sets having better match to the data. This would be akin to a selective bootstrap sample from the K parameter sets. Alternatively, inferior set(s) could be replaced by more optimizations started from new random starting points until all K parameter sets yield ~equally good matches to the experimental data.

In Figure 13a the QOI predictions with the final K=5 parameter sets corresponding to each test are shown to vary about the ideal best-possible calibration values for each test. This reflects the error deviation properties found in Section 4. The last paragraph of the present subsection 5.1 comments on robustness of the following prediction-UQ approach when the K prediction results per each test are all systematically offset by reasonably small magnitudes from the QOI prediction that would occur with ideal best-possible calibration parameters from the test.

Figure 13b shows K=5 vertical groupings of QOI values. For each grouping of N=3 results, the statistical quantity/ies of interest in the analysis (like tail probability) are estimated with the appropriate sparse-sample 1-D UQ method. The estimates will come with a high level of confidence that they are conservative bounds on the desired statistic(s). K=5 estimates are obtained as exemplified in Figure 13c. The variance in the estimates reflects sampled uncertainty from calibration/optimization effects; no such effects would correspond to no variance in the estimates. More extensive sampling (K >> 5) would result in a wider range of estimates. A reasonable approach to bound the small-K effect is to form, say, 95/90 tolerance intervals (see Footnote 3 in subsection 2.1) based on the sparse K=5 estimates.

The choice K=5 is very cost effective because it marks a relatively distinct knee in the TI cost-accuracy curve of uncertainty magnitude vs. number of samples (see Figure 9). Evidence of the effectiveness of the choice K=5 is cited in the next footnote. One can go as low as K=2, corresponding to 2N calibrations and evaluations of the next-level analysis model to propagate the uncertainty, but 95/90 TIs on predicted statistics will be considerably larger, with results commensurately more conservative in most cases.

Realistic circumstances often limit to just a few replicate experiments; 2 to 5 times “a few” implies order-4 to 20 calibrations and next-level model evaluations per calibrated model. The relatively low numbers of calibrations and analysis model simulations will typically be affordable without constructing surrogates for either the calibration model or the next-level analysis model. This portends less cost and complexity with the DD approach relative to other calibration-UQ approaches.

If each test’s associated K predictions of the response QOI bound that test’s associated ideal best-possible prediction (as illustrated in Figure 13a where calibrated model response predictions are above and below the horizontal prediction line from that test’s idealized best-possible calibration parameters), then the K statistical estimates in Figure 13c will likely bound the statistical estimate from the ideal best-possible predictions. The latter estimate is depicted by the horizontal cross-line in Figure 13c. This claim is made for K=5 in Figure 13b because it is very likely that at least one of the 5 vertical point sets in Figure 13b will have smaller variance and at least one will have larger variance than the ideal best-possible prediction results in Figure 13a.[5] Then the QOI statistical estimate from the ideal best-possible prediction results

[5] A quick investigation found this to be the case in 92.5% or 37 of 40 random trials in an example problem with N=3 ideal best-possible QOI prediction results set at (-1, 0, +1) and K=5 estimates with random Normal variations about

would usually be bounded above and below by estimates in Figure 13c from the K groups. (Variance differences in the K groups are amplified by the sparse-sample UQ methods for statistics other than mean response, and drive the spacing of the K notional statistical estimates in Figure 13c.)

Even if the statistical estimate from the ideal best-possible prediction results lay outside the K estimates contrary to what is shown in Figure 13c, a 95/90 TI formed from the K estimates may well contain the ideal best-possible statistical estimate. An analogue is the capture of the zero-deviation exact result by 95% CIs for loose-tolerance cases in Figure 12 where both results lie to one side or another of the exact result and do not straddle it (95/90 TIs would be significantly wider per Figure 9.) Thus, significant margin often exists for significant systematic offsets of the K random estimates in Figure 13a relative to their corresponding best-possible QOI results. The methodology is fairly robust to these situations.

5.2 Extension to uncertainties in the model and experimental inputs and outputs

Here we extend the DD calibration-UQ methodology to other sources of epistemic uncertainty. Figure 14 shows a situation where the multiple calibration experiments occur at slightly different experimental input conditions due to control variability of initial and boundary conditions. These differing conditions input to the model being calibrated result in a slightly different model response function in the calibration to each experiment. This is indicated in the figure by the slightly offset and differently sloped response function regions corresponding to the different experiments. We point this out on the way to generalizing the DD calibration-UQ methodology in the following, but there are no meaningful procedural differences vs. the situation in Figure 1 where all the replicate tests were said to occur at the exact same input conditions and just the tested units varied.

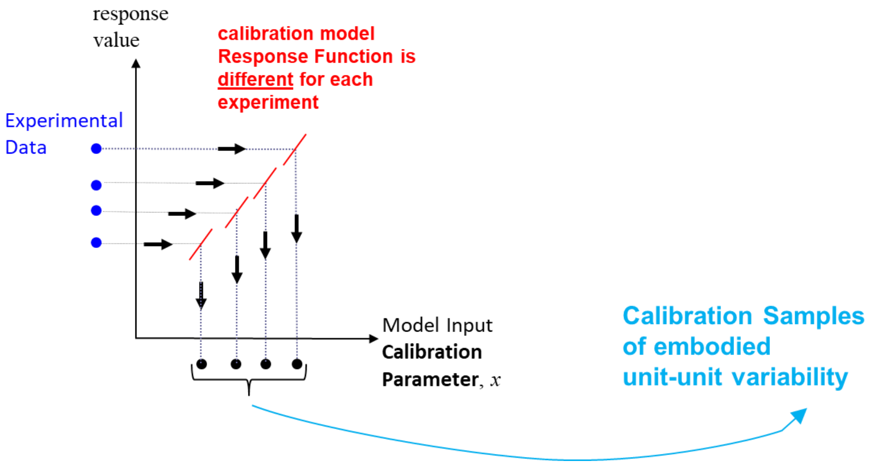


Figure 14. DD calibration problem where the multiple experiments for calibration data occur at slightly different experimental initial and boundary conditions. The calibration model response function is slightly different for calibration to each experiment as shown.

each best-possible result on a scale of about ±2% (standard deviation = 0.01). The proportion was 87.5% (35 of 40 trials) for the same problem but with non-equally spaced ideal best-possible QOI prediction results at (-1, 0, +0.5). The lower proportion in the latter case is attributed to sampling error; there is no other obvious reason that a systematic difference in proportions would exist between these two circumstances.

Other potential uncertainties that do not have to be explicitly accounted for in the calibration problem are "traveling" uncertainties of model inputs that are approximated reasonably well through measurement or estimation. A traveling input is one that appears consistently in the calibration problem and in the application problem where the calibrated model is to be used. For example, in [6] the material properties of the steel bolt and the object bolted to the test shaker are known to within a few % of nominal estimates. The six calibration parameters of the bolted-joint behavioral model depend on the material properties of the bolt and bolted object, so are coupled to their material property values used. However, if the bolt and object remain as a pair (say the object is bolted in the same fashion to an applied hardware system), then the material property values of the bolt and object are the same in the calibration and application/prediction uses of the model. It makes little difference if the nominal property values used in the model calibrations are correct or in error by a small amount because there is compensation by the calibration parameters, which come coupled with the (approximate but incorrect) nominal values of the material properties. Such calibration parameter compensation does not exist, however, for uncertainties of non-traveling inputs to the calibration problem, so these if significant must be explicitly treated in the calibration problem.

Figure 15 concerns model response-function uncertainties related to model input uncertainties (non-traveling) from experimental initial and boundary conditions (ICs/BCs), geometry, material and surface properties, etc. Some of these inputs may be estimated and others measured. In both cases, some of the uncertainties may be mathematically described by probability distributions and others by intervals. The measurement uncertainties can have both a random/uncorrelated component that varies from test-to-test and a systematic component for measurement errors that are uncertain but effectively the same (fully correlated) from test to test.

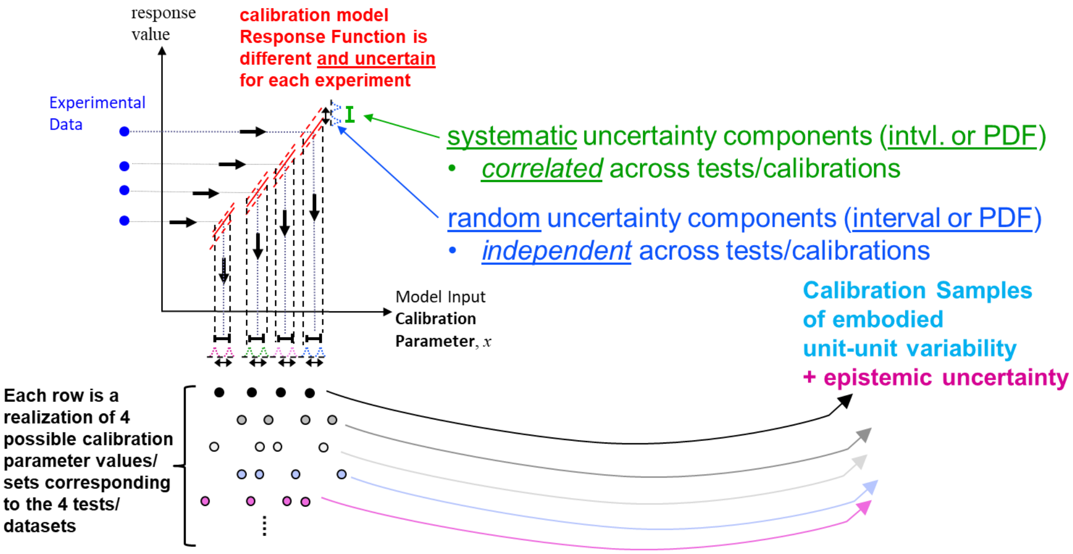

response value
calibration model Response Function is different and uncertain for each experiment
Experimental Data
I
systematic uncertainty components (intvl. or PDF)
• correlated across tests/calibrations
random uncertainty components (interval or PDF)
• independent across tests/calibrations
Model Input Calibration Parameter, x
Calibration Samples of embodied unit-unit variability + epistemic uncertainty
Each row is a realization of 4 possible calibration parameter values/ sets corresponding to the 4 tests/ datasets

Figure 15. Calibration model response-function uncertainties related to epistemically uncertain model inputs and precision errors from model-discretization, the optimization process, and any use of surrogate models.

Figure 15 also includes uncertainties related to model discretization and solution. These can be straightforwardly incorporated as uncertainty in the calibration model response function. Mesh discretization related solution error/uncertainty may be treatable as systematic across the calibrations (extrapolating from investigations in [41], [42]). The uncertainty is estimated by solution or calculation verification methods, e.g., [43], [44]. Model solver related errors/uncertainties are usually random/uncorrelated across the calibrations and are typically controlled to be much smaller in magnitude than mesh discretization related uncertainties. Model solver related random errors/uncertainties are presumed to be sampled in the multi-start calibration-UQ procedure described in Section 4 and subsection 5.1.

The same is presumed for any random calibration/optimization errors from use of response-surface surrogates in the optimization procedure. Any systematic surrogate-related error/uncertainty effects on the model response function across the calibrations would have to be separately estimated. It can then be treated like the other sources of systematic uncertainty in the model response function.

The DD calibration-UQ approach straightforwardly handles all these heterogeneous uncertainty possibilities, as well as random and systematic uncertainties on measured *outputs* of the tests as indicated in Figure 16 and demonstrated in [6] and [9].

When the heterogeneous uncertainties are sampled in a Monte Carlo approach, each realization across all random and systematic uncertainty sources particular to a given experiment, model/simulation, and optimization procedure are used to perform a DD calibration. Ultimately, K groups of N calibration parameter sets for N tests are produced (illustrated as K rows of N=4 calibration parameter sets at the bottom of Figure 16). Each group of calibration parameter sets is propagated to next-level model predictions. Each group produces a corresponding set of N response QOI values. These comprise a generic version of a vertical set of QOI response values as in Figure 13b where N=3 in that example. K such vertical sets of QOI values are produced, as illustrated in Figure 13b for K=5. The calibration-UQ methodology then proceeds from generic versions of Figure 13b to Figure 13c.

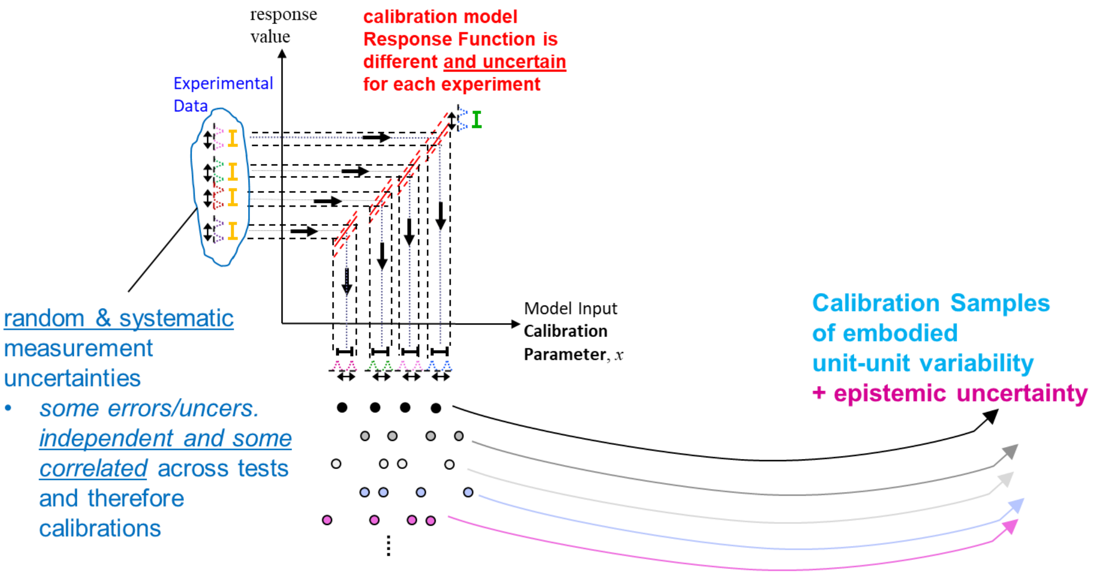

response value
calibration model Response Function is different and uncertain for each experiment
Experimental Data
random & systematic measurement uncertainties
• some errors/uncers. independent and some correlated across tests and therefore calibrations
Model Input Calibration Parameter, x
Calibration Samples of embodied unit-unit variability
+ epistemic uncertainty

Figure 16. Random and systematic measurement uncertainties on the test outputs included in the DD calibration-UQ methodology.

5.3 Extension of UQ to the Simultaneous Discrete-Direct method

As already established, each submodel to be calibrated in the SDD method will incur a cost of N replicate tests and nominally 5N calibrations (for K=5 uncertainty sampling rounds to account for the various experimental, model/simulation, and optimization procedure uncertainties involved). For typically low numbers (N = 2 to 5) of replicate experiments affordable, this would typically involve order- 10 to 25 calibrations per submodel.

System-level analysis model propagations will involve N system model simulations to obtain N values of the system-level QOIs (see e.g. Figure 7). From the N values per QOI, reasonable bounds are estimated on the desired QOI statistics using sparse-sample 1-D UQ techniques. The N system simulations use each of N calibration parameter sets per submodel once-and-only-once for reasons explained in Section 3. However, when submodel uncertainties are accounted for, there are multiple (say K=5) realizations of N parameter sets per submodel. For example, Figure 16 could represent a submodel with K=5 rows of N=4 calibration parameter sets. Any row 1,…,K would be combined at random with any row 1,…,K of the other submodels to perform N system-model simulations and get N response values per QOI to form epistemically conditional bounding estimates on desired statistics. The conditional bounds depend on the particular row of N calibration parameters selected from the K options for each submodel. This exercise can be performed say L=5 times (5 marks the knee in the TI cost-effectiveness curve in Figure 9), each time with a different random pick of row 1,…,K for each submodel (where the different row realizations reflect that submodel's uncertainties). Each performance of this exercise will cost N system-model simulations and will yield a conditional bounding estimate for each desired QOI statistic. The L=5 such estimates for a given QOI statistic are random realizations from a population of conditional bounding estimates that reflect the experimental, model/simulation, and optimization procedure related uncertainties in the submodels of the system model. One could form a 95/90 TI from the L=5 conditional bounding estimates to reasonably well encompass the full population of estimates (per Footnote 1 in Section 2.1).

This methodology is yet to be implemented and empirically evaluated in a SDD application, but is similar to handling of epistemically conditional bounding estimates via sparse-sampling UQ methodology confirmed in DD calibration-UQ efforts [6] and [9], and used in model validation efforts [37], [40], [45]. The proposed methodology requires L*N system-model simulations. Note that the uncertainty propagation cost does not scale with the number of submodels or number of calibration or other parameters in the system model. For the recommended L=5, the SDD method cost is 5N system-model runs. For the usual small number of replicate tests affordable to inform each submodel, say N = 2 to 5, the SDD method cost would be 10 to 25 runs of the system model. This is very reasonable and would not normally involve the complication and error/uncertainty sources with surrogate models and/or other propagation cost reduction methods like dimension reduction, reduced-order physics models, multi-fidelity modeling, etc.

6. Conclusions

The DD and Simultaneous DD approaches are well equipped to straightforwardly handle the many heterogeneous aleatory and epistemic uncertainty types and sources in calibration-prediction-UQ application problems. The methodologies better preserve the fundamental information from calibration experiments in a way that enables calibration parameter sets and model predictions to be more directly tied to the supporting experimental data than other calibration-UQ approaches enable. In particular, the approaches have several advantages over Bayesian and other calibration-UQ approaches in terms of capturing and utilizing the information obtained from the typically small number of replicate experiments in model calibration situations, especially when sparse realizations of random function data are involved. When stochastic phenomena and sparse replicate tests are involved, the DD and SDD mappings to 1-D UQ problems and efficient and effective methods for estimating conservative bounds on response QOI statistics enable objectively better control and characterization of expected accuracies than other calibration-UQ approaches appear to offer.

The DD and SDD approaches are conceptually and implementationally straightforward. They also typically incur less computational cost than other calibration-propagation-prediction-UQ methodologies.

Surrogate models are not foreseen to be needed either in the calibration procedures or in uncertainty propagation to next-level analysis models and predictions. This portends less cost and complexity and greater accuracy with the DD and SDD approaches relative to other calibration-UQ approaches.

In addition, it is argued in Section 2 that less analyst expertise and prior knowledge is needed to competently implement the DD and SDD approaches. The approaches do not need good prior information on calibration parameter values as Bayesian methods do to give reliable results under the realities of sparse calibration data. For this reason and for reasons of lower complexity because of no likelihood functions or surrogates to construct, no joint probability density functions to construct, no Markov Chain Monte Carlo sampling procedures, etc., the DD and SDD approaches are less subject to analyst-to-analyst variability of implementations and results. Some evidence of the latter was presented in Section 2.2 and more is presented in the Appendix.

The DD and SDD model calibration and uncertainty propagation methodologies are part of a pragmatic systems approach to end-to-end UQ for stochastic phenomena (with non-stochastic phenomena as a simplification). The methods directly feed a complementary model validation and bias correction method for extrapolative prediction beyond the validation conditions ([40]).

Appendix: Evidence of analyst variability in Bayesian sparse-sample estimates and their lower accuracy than sparse-sample UQ method estimates

This appendix summarizes some results from an End-to-End UQ Test problem [46] worked within and outside Sandia in 2017 and 2018. The problem statement was sent out to around 50 academic and applied VVUQ research groups in the AIAA, ASME, and SAE communities. Teams inside and outside Sandia made varying progress on the problem, with only references [9] and [40] demonstrating methodology and results through the model validation, extrapolative prediction, and tail probability risk-estimation portions of the problem.

A few teams reported on the model calibration portion of the E2E UQ problem. In [47] a maximum-likelihood based approach was used to optimize hyperparameters of several types of PD model forms for the calibration parameters. The problem description was changed somewhat so the accuracy of the methodology could be studied as a function of the number of experimental data samples. Therefore, results were not directly comparable to other efforts like the DD one [9] that used the calibration data specified in [46]. A Bayesian calibration method applied by an internal Sandia team [48] that included two engineers and two statisticians each formally trained in Bayesian methods produced an unexpected trend of decreased prediction uncertainty with added experimental measurement uncertainty. The project ended before this could be investigated.

Nonetheless, several teams' results are compared for the first part of the E2E UQ problem. This involved directly using sparse experimental deflection data supplied in the problem to estimate the cantilever-beam population central 95% of deflections and tail-probability above a specified threshold. Even though not involving model calibration directly, the results support the themes in the title of this appendix and in this document more broadly as follows.

The rectangular beams have nominally the same geometry (length, width, and height) within small perturbations about design values L, W, and H. The material strength property (modulus, E) also varies randomly over the population of beams. An affordably small number (N = 4) of nominally identical beams and replicate tests are used to infer response variability in the large full population of beams (asymptotically ∞) in a particular loading configuration and load magnitude Po. The loading exhibits small control variations from test to test about the set point Po. Information is given to estimate uncertainty to account for this.

Figure A.1 shows three teams' estimated bounding ranges on the central 95% of deflection responses in the asymptotically large population of beams. Estimates are shown for the cleanest case in [46] of no random or systematic measurement errors on the deflections or experimental loads P_i in the four tests.

The Sandia Bayesian estimate and the sparse-sample 1-D UQ 95/95 TI estimate ([49]) are at advertised 95% confidence levels of the methods. A non-Sandia Bayesian estimate [50] is at an un-specified confidence level. Both Sandia and non-Sandia Bayesian methods involved a Normal probability model for population variability with uncertain hyperparameters mean μ and standard deviation σ whose uncertainty distributions were first proposed as priors and then updated to posteriors based on the data from the four beam deflection tests and Bayesian methodology. The prior distributions chosen by the two Bayesian teams were different. Details of the likelihood formulations, surrogates, sampling procedures, etc. were not reported in [50], so associated differences between the two teams' Bayesian methodologies could not be determined. In any case, very different results emerged from the two Bayesian variants, as shown in Figure A.1. Relative to the true central 95% range in the UQ test problem, one Bayesian variant produced very conservative results, while the other variant did not bound the true 95% range in two of the three random trials.

In contrast, the TI results are conservative in the three trials, but not excessively conservative as the Sandia Bayesian results are. Moreover, the TI results would be repeatable across analysis teams, having very little room for analyst–analyst variability of results (given the same confidence level sought).

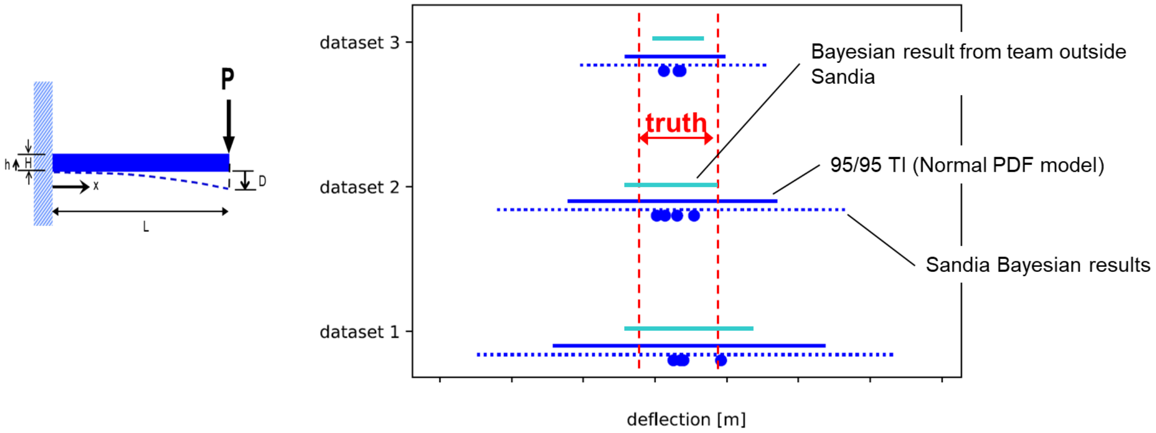


Figure A.1. Estimated bounds on the central 95% of beam deflection responses estimated by three methods indicated.

Figure A.2 shows results of tail probability estimation for the Sandia Bayesian and some 1-D UQ sparse-sample methods. The non-Sandia Bayesian estimate [50] of the right-tail exceedance probabilities in the problem are usually non-conservative, as one might expect from the results in Figure A.1, so are not included in the figure.

Bayesian results are shown for the Sandia variant previously discussed, as well as a Bayesian Multiple Model Averaging variant that enlists Normal, Log-Normal, t, and Weibull forms of probability distributions for deflection variability and infers their hyperparameter posteriors and then weights everything in the final estimate. It was found that N = 4 data samples were insufficient to differentiate between the various model forms with any significant degree of preference. The Bayesian tail probability estimates in the figure correspond to the 95th percentile largest (conservative-leaning) estimates given the proposed priors and the treatment of sparse-sample epistemic uncertainty in the estimation methodology. In the three trials, the estimates from the two Bayesian variants trade 2nd and 3rd places as the most conservative estimates of the methods presented.

The uniformly most-conservative estimate in the three trials was an Inverse Binomial method. This determined the binomial distribution parameters that give the 95% highest propensity for threshold exceedance that a Normal population could have and still be consistent with no exceedances observed (none of the data sets contained an actual defection case that exceeded the prescribed threshold).

Results from the Super Distribution (SD) and TI Equivalent-Normal (TIEN) 1-D UQ sparse-sample tail-probability estimation methods developed and assessed in [49] are the least overly-conservative. In particular, the SD method is consistently the most accurate of all the methods. It is still very conservative though, by one to three orders of magnitude depending on the trial.

Large conservatism is the tradeoff for high reassurance that the tail probability magnitude will not be under-estimated based on sparse sample data from a large variety of potential distributions types and shapes (see Footnote 3 in subsection 2.1 of the present paper). The SD method distribution, derived from Frequentist concepts and methodology for sampling from a Normal distribution, was later found to be approximately equivalent to a scaled t-distribution posterior probability density derived from Bayesian concepts and methodology that is empirically shown in [26] to perform similarly to SD for sparse-sample estimation of tail probabilities of small magnitude. Depending on the problem conditions of number of samples, tail probability magnitude, etc., the SD, TIEN, new modifications of the scaled-t distribution method, and hybrids with statistical Jackknifing alternately become the most efficient estimators with high reliability of not underestimating. Ref. [26] studies and maps out these dynamics and identifies which of the said methods perform best under different problem conditions.

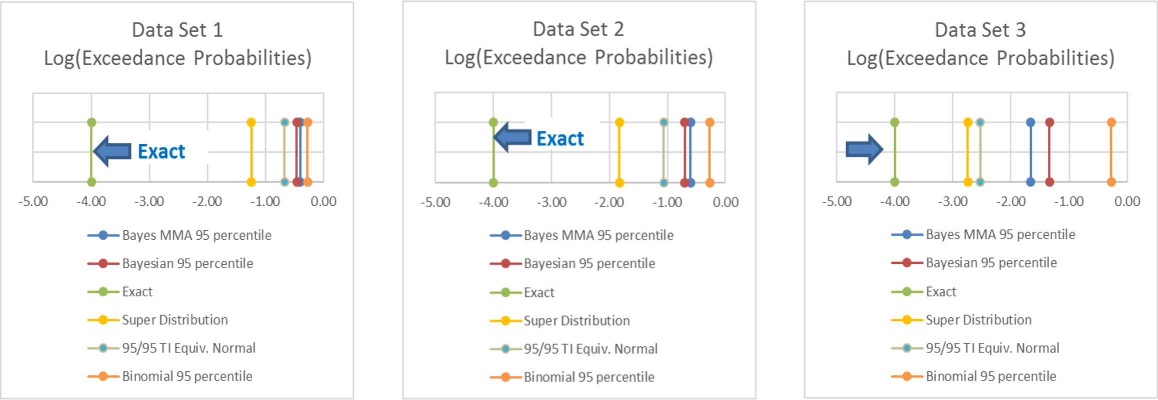


Figure A.2. Beam deflection tail probabilities estimated by the various methods indicated.

References

[1] Jamison, R., V. Romero, M. Stavig, T. Buchheit, C. Newton, “Experimental data uncertainty, calibration, and validation of a viscoelastic potential energy clock model for inorganic seal glasses,” Sandia National Laboratories document SAND2016-4635C, Albuquerque, NM, presented at ASME Verification & Validation Symposium, Las Vegas, NV, May 18-20, 2016.

[2] Romero, V., J. Winokur, G. Orient, J.F. Dempsey, “Discrete-Direct Model Calibration and Uncertainty Propagation Method confirmed on Multi-Parameter Plasticity Model calibrated to Sparse Random Field Data,” *ASCE-ASME J. Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, doi:10.1115/1.4050371, vol. 7 no. 2, June 2021.

[3] Scherzinger, W., V. Romero, K. Karlson, Sandia National Laboratories ASC V&V Research Portfolio project (2018-19): study of metal plasticity model calibration under sparse tension test data and material model-form uncertainty applied to predictions of pressure vessel maximum load.

[4] Romero, V., K. Karlson, G. Bergel, “Simultaneous Discrete-Direct sparse-data model calibration, propagation, and tail-probability estimation confirmed on weld modeling problem,” Sandia National Laboratories report in preparation.

[5] Russ, J., V. Romero, R. Hopkins, component structural dynamics model calibration and uncertainty quantification project, Sandia National Laboratories, 2016-17.

[6] Romero, V., C. Sanders, T. Walsh, “Methods and Performance Characterization for Uncertainty Quantification and Propagation in Inverse Estimation of Structural Dynamics Parameters given Material Property Uncertainties and Limited Sensor Data,” Sandia National Laboratories paper to be submitted for 2022 ASME V&V Symposium, May 23-24, location TBD..

[7] Romero, V., B. Paskaleva, A. Mar, I. Wilcox, “Sparse-Data Model Calibration, Uncertainty Propagation, and QMU demonstrated on a Multi-Component Circuit,” Official Use Only/Export Controlled Sandia National Laboratories presentation, Oct. 27, 2020.

[8] Romero, V., L. Swiler, A. Mar, B. Paskaleva, “Sparse-Data Model Calibration, Propagation, and Interpolative and Extrapolative Prediction/UQ with Surrogates and Multiple Fidelities of Rad-Electrical Models,” Official Use Only/Export Controlled Sandia National Laboratories report in preparation.

[9] Romero, V., “Discrete Direct Model Calibration and Propagation Approach addressing Sparse Replicate Tests and Material, Geometric, and Measurement Uncertainties,” Soc. Auto. Engrs. 2018 World Congress (WCX18) paper 2018-01-1101 (doi:10.4271/2018-01-1101), April 10-12, Detroit, MI.

[10] Schutte, J.F., R.T. Haftka, B.J. Fregly, “Improved global convergence probability using multiple independent optimizations,” *Intn’l. J. Num. Mthds. in Engrng*., 2007, 71:678-702.

[11] Xiong, Y., W. Chen, K. Tsui, D.W. Apley, “A better understanding of model updating strategies in validating engineering models,” *Computer Methods in Applied Mechanics and Engineering*. 198 (15–16) (2009) 1327–1337.

[12] Youn, B.D., B.C. Jung, Z. Xi, S.B. Kim, W. Lee, “A Hierarchical Framework for Statistical Model Calibration in Engineering Product Development,” *Computer Methods in Applied Mechanics and Engineering*, 200:1421-1431, 2011.

[13] Jung, B.C., H. Youn, H. Oh, G. Lee, M. Yoo, B.D. Youn, Y.C. Huh, “Hierarchical Model Calibration for Designing Piezoelectric Energy Harvester in the Presence of Variability in Material Properties and Geometry,” *Structural and Multidisciplinary Optimization*, 53:161-173, 2016.

[14] Lee, G., G. Yi, B.D. Youn, “A Comprehensive Study on Enhanced Optimization-Based Model Calibration Using Gradient Information,” *Structural and Multidisciplinary Optimization*, 57:2005-2025, 2018.

[15] Emery, J.M., R.V. Field, J.W. Foulk, K.N. Karlson, M.D. Grigoriu, “Predicting laser weld reliability with stochastic reduced-order models,” *Int. J. Numer. Meth. Engng*. 2015; 103:914–936.

[16] Swiler, L.P., B.M. Adams, M.S. Eldred, “Model Calibration under Uncertainty: Matching Distribution Information.” paper AIAA-2008-5944 in Proceedings of the 12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Victoria, BC, Canada, Sept 2008.

[17] Kennedy, M.C., and A. O’Hagan, “Bayesian calibration of computer models,” *J. Royal Statistical Society Series B, Statistical Methodology*, 63(3) pp. 425-464, 2002.

[18] Rizzi, F., R.E. Jones, J.A. Templeton, J.T. Ostien, B.L. Boyce, “Plasticity models of material variability based on uncertainty quantification techniques,” Sandia National Laboratories document SAND2017-12030 J (Nov. 2017).

[19] Mullins, J., and S. Mahadevan, “Bayesian uncertainty integration for model calibration, validation, and prediction,” *ASME J. Verification, Validation and Uncertainty Quantification*, 1.1 (2016): 011006.

[20] Xi, X., and R.J. Yang, “Reliability Analysis with Model Uncertainty Coupling with Parameter and Experimental Uncertainties: A Case Study of 2014 Verification and Validation Challenge Problem,” *ASME J. Verification, Validation and Uncertainty Quantification*, 1.1 (2016), 011005.

[21] Li, W., S. Chen, Z. Jiang, D. Apley, Z. Lu, W. Chen, “Integrating Bayesian Calibration, Bias Correction, and Machine Learning for the 2014 Sandia Verification and Validation Challenge Problem,” *ASME J. Verification, Validation and Uncertainty Quantification*, 1.1 (2016), 011004.

[22] Romero, V., M. Bonney, B. Schroeder, V.G. Weirs, “Evaluation of a Class of Simple and Effective Uncertainty Methods for Sparse Samples of Random Variables and Functions,” Sandia National Laboratories report SAND2017-12349, Nov. 2017.

[23] Romero, V., B. Schroeder, J.F. Dempsey, N. Breivik, G. Orient, B. Antoun, J.R. Lewis, J. Winokur, “Simple Effective Conservative Treatment of Uncertainty from Sparse Samples of Random Variables and Functions,” *ASCE-ASME Journal of Uncertainty and Risk in Engineering Systems: Part B. Mechanical Engineering*, DOI 10.1115/1.4039558, Dec. 2018, vol. 4, pp. 041006-1 – 041006-17.

[24] Jekel, C., and Romero, V., “Bootstrapping and Jackknife Resampling to Improve Sparse-Data UQ Methods for Tail Probability Estimates with Limited Samples,” paper https://doi.org/10.1115/VVS2019-5127, ASME 2019 Verification and Validation Symposium VVS2019, May 15-17, 2019, Las Vegas, NV.

[25] Jekel, C., and Romero, V., “Conservative Estimation of Tail Probabilities from Limited Sample Data,” Sandia National Laboratories report SAND2020-2828, March 2020.

[26] Jekel, C., and Romero, V., “Conservative and efficient tail probability estimation from sparse sample data,” Sandia National Laboratories document SAND2020-7572 J, July 2020.

[27] Wang, L., D. Beeson, G. Wiggs, M. Rayasam, “A Comparison of Meta-modeling Methods Using Practical Industry Requirements,” paper AIAA 2006-1811, 47th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, 1 - 4 May 2006, Newport, RI.

[28] Viana, F.A.C., R.T. Haftka, V. Steffan Jr., “Multiple surrogates: How cross-validation errors can help us obtain the best predictor. *Structural and Multidisciplinary Optimization*, 39(4):439–457, 2009.

[29] Eldred, M.S., and J. Burkardt, “Comparison of non-intrusive polynomial chaos and stochastic collocation methods for uncertainty quantification,” 47th AIAA Aerospace Sciences Meeting and Exhibit, paper AIAA-2009-0976, January 5-8, 2009, Orlando, FL.

[30] Swiler, L.P., V.J. Romero, “A Survey of Advanced Probabilistic Uncertainty Propagation and Sensitivity Analysis Methods,” Chapter 6 of Joint Army/Navy/NASA/Air Force (JANNAF) e-book: *Simulation Credibility—Advances in Verification, Validation, and Uncertainty Quantification*, U. Mehta (Ed.), D. Eklund, V. Romero, J. Pearce, N. Keim, document NASA/TP-2016-219422 and JANNAF/GL-2016-0001, Nov. 2016.

[31] Field, Jr. R.V., M.D. Grigoriu, J.M. Emery, “On the efficacy of stochastic collocation, stochastic Galerkin, and stochastic reduced order models for solving stochastic problems,” *Probabilistic Engineering Mechanics*, 41 (2015) pp. 60-72.

[32] Grigoriu, M.D., “Reduced order models for random functions, Application to stochastic problems.” *Applied Mathematical Modelling* (2009); 33:161–175.

[33] Romero, V., L. Swiler, A. Urbina, J. Mullins, “A Comparison of Methods for Representing Sparsely Sampled Random Quantities,” Sandia National Laboratories report SAND2013-4561, Sept. 2013.

[34] Winokur, J., and V. Romero, “Optimal Design of Computer Experiments for Uncertainty Quantification with Sparse Discrete Sampling,” Sandia National Laboratories document SAND2016-12608, 2016.

[35] Conover, W. M, “On a better method for selecting input variables,” (1975) unpublished Los Alamos National Laboratory manuscript reproduced as Appendix A of “Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems” by J.C. Helton and F.J. Davis, Sandia National Laboratories report SAND2001-0417 printed November 2002.

[36] Lee, G., W. Kim, H. Oh, B.D. Youn, N.H. Kim, “Review of Statistical Model Calibration and Validation – From the Perspective of Uncertainty Structures,” *Structural and Multidisciplinary Optimization*, 2019, 60:1355–1372.

[37] Romero, V., and Black, A., “Processing Aleatory and Epistemic Uncertainties in Experimental Data from Sparse Replicate Tests of Stochastic Systems for Real-Space Model Validation,” *ASME J. Verification, Validation and Uncertainty Quantification*, paper VVUQ-20-1037 doi:10.1115/1.4051069 May, 2021.

[38] DIRECT Version 2.0 User Guide, J.M. Gablonski, North Carolina State U. Dept. of Mathematics, Raleigh, NC, April18, 2001.

[39] Adams, B.M., Bauman, L.E., Bohnhoff, W.J., Dalbey, K.R., Ebeida, M.S., Eddy, J.P., Eldred, M.S., Hough, P.D., Hu, K.T., Jakeman, J.D., Stephens, J.A., Swiler, L.P., Vigil, D.M., and Wildey, T.M.,

"Dakota, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 6.0 User's Manual," Sandia Technical Report SAND2014-4633, July 2014.

[40] Romero, V., "Real-Space Model Validation and Predictor-Corrector Extrapolation applied to the Sandia Cantilever Beam End-to-End UQ Problem," paper AIAA-2019-1488, 21st AIAA Non-Deterministic Approaches Conference, AIAA SciTech 2019, Jan. 7-11, San Diego, CA.

[41] Romero, V.J., "Model-Discretization Sizing and Calculation Verification for Multipoint Simulations over Large Parameter Spaces," paper AIAA2007-1953, 9th AIAA Non-Deterministic Methods Conference, April 23 - 26, 2007, Honolulu, HI.

[42] Schroeder, B., H. Silva III, K.D. Smith, "Separability of Mesh Bias and Parametric Uncertainty for a Full System Thermal Analysis," Sandia National Laboratories report SAND2018-1007, January 2018.

[43] Roache, P. J., *Verification and Validation in Computational Science and Engineering*, Hermosa Publishing, 1998.

[44] Oberkampf, W. L., and Roy, C. J., *Verification and Validation in Scientific Computing*, Cambridge University Press, 2010.

[45] Romero, V., F. Dempsey, B. Antoun, "Application of UQ and V&V to Experiments and Simulations of Heated Pipes Pressurized to Failure," Chapter 11 of Joint Army/Navy/NASA/Air Force (JANNAF) e-book: *Simulation Credibility—Advances in Verification, Validation, and Uncertainty Quantification*, U. Mehta (Ed.), D. Eklund, V. Romero, J. Pearce, N. Keim, document NASA/TP-2016-219422 and JANNAF/GL-2016-0001, Nov. 2016.

[46] Romero, V., B. Schroeder, M. Glickman, "Cantilever Beam End-to-End UQ Test Problem: Handling Experimental and Simulation Uncertainties in Model Calibration, Model Validation, Extrapolative Prediction, and Risk Assessment," Sandia National Laboratories document SAND2017-4689 O, version BeamTestProblem-34.docx, 2017.

[47] Gaymann, A., M. Pietropaoli, L.G. Crespo, S.P. Kenny, F. Montomoli, "Random Variable Estimation and Model Calibration in the Presence of Epistemic and Aleatory Uncertainties," Soc. Auto. Engrs. 2018 World Congress (WCX18) paper 2018-01-1105 (doi:10.4271/2018-01-1105), April 10-12, Detroit, MI.

[48] Sandia Cantilever Beam End-to-End UQ Methods Investigation Team (in alphabetical order):
P. Hough, L. Hund, J.R. Lewis, J. Mullins, V. Romero, B. Schroeder, V.G. Weirs.

[49] Romero, V., and V.G. Weirs, "A Class of Simple and Effective UQ Methods for Sparse Replicate Data applied to Cantilever Beam End-to-End UQ Problem," Sandia National Laboratories document SAND2017-12365 C, 20th AIAA Non-Deterministic Approaches Conference, AIAA SciTech 2018, Jan. 8-12, Kissimmee, FL.

[50] Kim, T., G. Lee, L. Kim, B.D. Youn, "Expectation-Maximization Method for Data-Based Estimation of the Cantilever Beam End-to-End Problem," doi:10.2514/6.2018-1666, 20th AIAA Non-Deterministic Approaches Conference, AIAA SciTech 2018, Jan. 8-12, Kissimmee, FL.

[51] Hu, K.T., and Orient, G.E., "The 2014 Verification and Validation Challenge Problem Statement," *ASME J. Verification, Validation and Uncertainty Quantification*, 1.1 (2016), 011001.

[52] Dowding, K.J, M. Pilch, R.G. Hills, “Formulation of the thermal problem,” *Computer Methods in Applied Mechanics and Engineering*, 197 (29–32) (2008) 2385–2389.

[53] Kass, R.E., and Wasserman, L., “The Selection of Prior Distributions by Formal Rules,” *J. American Statistical Association*, Vol. 91, No. 435, Sept. 1996.