SAND2018-11643C

The Need for Credibility Guidance for Analyses Quantifying Margin and Uncertainty

Benjamin B. Schroeder[+], Lauren Hund, and Robert S. Kittinger

Sandia National Laboratories
P.O. Box 5800, MS 0828
Albuquerque, NM 87185-0828
[+]bbschro@sandia.gov
(505) 284-3796

ABSTRACT

Current quantification of margin and uncertainty (QMU) guidance lacks a consistent framework for communicating the credibility of analysis results. Recent efforts at providing QMU guidance have pushed for broadening the analyses supporting QMU results beyond extrapolative statistical models to include a more holistic picture of risk, including information garnered from both experimental campaigns and computational simulations. Credibility guidance would assist in the consideration of belief-based aspects of an analysis. Such guidance exists for presenting computational simulation-based analyses and is under development for the integration of experimental data into computational simulations (calibration or validation), but is absent for the ultimate QMU product resulting from experimental or computational analyses. A QMU credibility assessment framework comprised of five elements is proposed: requirement definitions and quantity of interest selection, data quality, model uncertainty, calibration/parameter estimation, and validation. Through considering and reporting on these elements during a QMU analysis, the decision-maker will receive a more complete description of the analysis and be better positioned to understand the risks involved with using the analysis to support a decision. A molten salt battery application is used to demonstrate the proposed QMU credibility framework.

Keywords: Credibility, Margin, Uncertainty, QMU

1 INTRODUCTION

The purpose of this paper is to describe the need for credibility guidance in quantification of margins and uncertainty (QMU) analyses and provide a potential structure for such guidance. Credibility is defined as "the quality or power of inspiring belief"[1], so credibility guidance should assist in the consideration of belief-based aspects of an analysis. A QMU credibility assessment framework comprised of five elements is proposed: requirement definitions and quantity of interest (QoI) selection, data quality, model uncertainty, calibration/parameter estimation, and validation. Through considering and reporting relevant aspects of these elements during a QMU analysis, the decision-maker will receive a more complete description of the analysis and be better positioned to understand the risks involved with using the analysis to support a decision.

This paper will be structured as follows. The remained of this section (Section 1) will provide a history of QMU, motivation for why a credibility assessment framework is needed, and highlight similar efforts in the CompSim domain. Section 2 will outline the proposed framework for gathering and organizing QMU credibility evidence. How to use the evidence to evaluate analysis credibility is then discussed in Section 3. A demonstration of the process applied to a molten salt battery example problem is provided in Section 4. Lastly, Section 5 provides a summary of the paper.

1.1 What is QMU

QMU originated at the national laboratories in the early 2000s to address risk in nuclear weapon stockpile stewardship in the absence of full system testing [2]. QMU was originally posed as a risk assessment framework for nuclear weapons, addressing the three elements of the risk triplet (what can occur? how likely is it? and what are the consequences?)[3]; this QMU formulation also included a fourth element, credibility, defined as the answer to the question 'how much confidence do we have in our risk assessment?' [4]. Historically at Sandia National Laboratories (Sandia), QMU was largely applied to experimental data-based problems, but it appears likely that an integration of computational simulation (Comp-

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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.

Sim) results and experimental data will be the paradigm of the future. While processes for conducting QMU have developed over time (e.g., [5, 6]), there are still no formal processes for evaluating the credibility of a QMU analysis.

QMU entails comparing a performance measure to a performance requirement to determine the likelihood of functioning as intended, considering all relevant uncertainties. Implementing a QMU analysis requires building a team with the relevant expertise; identifying performance measures and requirements; assimilating relevant data; running an analysis; and communicating the results. Considering these steps of a QMU analysis, a corresponding QMU credibility assessment should address many of the inherent aspects of the analysis such as relevance of the performance measure and requirement, data quality, and analysis limitations.

1.2 Why Measure Credibility?

There is currently a gap in guidance within Sandia National Laboratories (Sandia) for assessing the credibility of QMU analyses. New guidance for QMU was recently released as internal documents within Sandia in two sections: 1) an overview of high-level QMU concepts and processes and 2) descriptions of statistical tools that can be used to derive QMU results, with a focus on QMU for experimental data. This new guidance pushed for broadening the analyses supporting QMU results beyond extrapolative statistical models and advocated for a more holistic picture of risk, including information garnered from both experimental and CompSim campaigns. Although this new guidance improves the informational basis of QMU analyses, it does not provide a consistent framework for communicating the credibility of analysis results.

Credibility assessment guidance for QMU is needed because:

- Decision-makers are increasingly asking for credibility assessments when being provided analysis results. Decision-makers are learning that they must understand the level of confidence they should invest in the results to better utilize the analysis that they commissioned.
- Failing to provide guidance for communicating credibility may lead to overconfidence in results. A question that should be posed to QMU analysts is, "What is the credibility of your results?" Without asking this question, the decision-maker may believe results are more reliable than is warranted and make ill-informed decisions.
- A unified QMU credibility framework would result in greater consistency in information presentation. When credibility results are analyst-specific and/or analysis-specific, decision-makers will interpret results differently depending on who conducted the analysis.
- Streamlined documentation of important auxiliary information (e.g., metadata, methods) is integral to understanding and reproducing QMU results. Summary QMU results (for example, margin over uncertainty ratios) always rely on auxiliary supporting information about the QMU process and supporting experimental data.

Without a consistent credibility assessment framework, decision-makers must rely on source credibility, or their belief in the source of the information. Although not specific to the reception of QMU results, psychological research has explored the role of source credibility in other information distribution areas. Across the psychology literature, source credibility is typically attributed to a person providing a message. Key aspects of source credibility include the source's trustworthiness and expertise [7]; to a lesser degree, composure, dynamism, sociability [8] and even accents of voices [9]. Chaiken and Maheswaran (1994) found source credibility can affect decisions in two ways: 1) by serving as a peripheral cue for simple acceptance or rejection of an argument, and 2) by biasing the strength of the decision-maker's argument processing[10].

While biasing the belief in results based on the source is potentially problematic in itself, Heesacker et al., (1983) found that as source credibility increases, persuasion also increases[11]. They attribute this phenomenon to more credible sources eliciting greater thinking about the message (improved information presentation, not informational content).

Across psychological research a theme persists: human judgment is persuaded and biased by a variety of minute factors. As humans participate in high stakes decision making, it is important to understand how small changes in presentation of the message (or data) can unintentionally bias the decision-maker. To mitigate such bias, credibility frameworks may help through providing consistency, transparency, and structure.

1.3 History of Credibility in CompSim

The concepts of credibility continue to be developed for presenting CompSim results as evidence to support a decision as well as for the incorporation of experimental data into CompSim analyses. Reviewing the progress of credibility guidance for these fields provides a starting point for the analogous guidance for QMU analyses.

For institutions that utilize CompSim to support decisions regarding complex engineering questions such as national laboratories, the aerospace and defense industries, and space agencies, the credibility associated with CompSim predictions must be understood. Methods for assessing and communicating the credibility of CompSim based evidence are being developed by many organizations [12]. As an example, the Predictive Capability Model Maturity (PCMM)[13] has been developed at Sandia over the last decade to provide a consistent framework for evaluating CompSim credibility. PCMM was developed as a method of directing discussion about

and communication of the many assumptions, errors, biases, and uncertainties ever present in CompSim predictions. A broad spectrum of CompSim activities are covered by elements of PCMM including code verification, physics and material model fidelity, representation and geometric fidelity, solution verification, validation, and uncertainty quantification. Those elements are perceived to encompass the majority of error/uncertainty sources that may impact a CompSim analysis. An approach for grading a simulation's performance in the different elements is also provided, which includes guidance describing the expected attributes needed to achieve a specific maturity level for each element. This grading is meant to foster gap identification and resource allocation.

PCMM can be to be used as a results credibility communication tool as well as an initial analysis planning aid. Applications using PCMM as a prediction credibility assessment tool have been demonstrated [14, 15].

In the CompSim community, experimental credibility is currently being developed from the perspective of using experimental data for model validation and calibration[16, 17]. Through providing structure for the assessment of experiments used for CompSim and experimental integration activities, consistency between modeling activities can be increased. A common difficulty when comparing experimental and CompSim results is ensuring that the scenarios captured by each are similar enough to not be the cause of significant discrepancy. When such discrepancies occur, it may be difficult to determine the source. Through capturing information about the experimental setup from the perspective of how that information will be used in CompSim analyses, more information can be gained from the comparisons. This same framework can be used to increase an experimental campaign's value through incorporating knowledge about how the data will be utilized into the test planning process. Outcomes of these experimental credibility processes include characterization of experimental uncertainties, assessment of model validation or calibration quality, and assessment of experimental alignment with modeling goals.

2 IMPORTANT ELEMENTS FOR QMU CREDIBILITY

Following the strategy for developing a credibility framework laid out by the CompSim community, potential sources of error, uncertainty, bias, or assumptions that could impact a QMU analysis are categorized into elements. It is proposed that QMU credibility can be assessed using the following five elements.

QMU Credibility Elements

1. **Requirement Definition and QoI Selection**
   Defining the requirement against which performance is compared and selecting the appropriate quantity of interest that can be used to represent performance
2. **Data Quality**
   Evaluating the available data and its attributes
3. **Model Uncertainty**
   Describing any models used to analyze the data and associated assumptions
4. **Calibration / Parameter Estimation**
   Considering how the model is fit or calibrated
5. **Validation**
   Determining if the model is a sufficient representation of the data with respect to making the prediction of interest

The five elements are described in more detail in the subsections below. At the end of each element-specific section, suggested documentation is provided that would support credibility statements for each element.

2.1 Requirement Definition and QoI Selection

Requirements may sometimes be clearly specified and the mapping from available data to that requirement may be simple, but this is not strictly true. Requirements may need interpretation that comes from consultation with a subject matter expert or simply from the QMU analyst. Available data often requires additional assumptions and/or processing to be comparable with the requirement. The quantity compared against the requirement is referred to as the QoI. QoIs are typically physical quantities, while requirements may be functions of these physical quantities. Determining how the requirement and QoI definitions will be compared is a necessary step of a QMU analysis.

Suggested documentation. What is the requirement? Are there any perceived ambiguities in the requirement definitions? What is the QoI? What is the relevance of the QoI relative to the requirement?

2.2 Data Quality

A great deal of qualitative information lives with the dataset that may impact the value of the dataset. Specifically, metadata about a dataset should be documented and preserved, so that important information about the data-generating mechanism can be evaluated when the data are analyzed. Metadata may include:

- When was the data gathered?
- What method was used to capture the measurements?
- How well developed was the measurement/experimental method?

- Where was the test conducted?
- Who conducted the test?
- What tester(s) was used?
- How well characterized are the experimental conditions?

Transparently evaluating metadata reduces the risk of omitting information that may impact the conclusion of the analysis. The following four categories are common categories of such auxiliary data (but should not be considered all-encompassing).

1. ***Sparsity*** The amount of data available impacts how much sampling uncertainty will exist in an estimate. Further, some estimands require more data than others to avoid extrapolative inferences; for instance, estimating a mean typically requires much less data than estimating an extreme percentile or rare exceedance probability to avoid extrapolation. Issues with presenting distributional tail extrapolation have been highlighted in [18].

   Suggested documentation. How much data is available? Is the data sufficient to empirically validate any estimates being made?

2. ***Representativeness*** The QoI often cannot be directly measured given the available data. Therefore, the analyst must consider how the available data map onto the QoI. For example, are we interested in environment A, but only have data tested in a similar, but less stressing environment B?

   Suggested documentation. What is the representativeness of the data relative to the application space (including tested environments, age, etc.), as defined by the QoI?

3. ***Noisiness / Measurement uncertainty*** Most measurements contain some error. This error can arise from many different sources. A common source of error is the tester or instrument's measurement error. In addition, errors can be introduced during data processing steps to convert a signal captured by a measurement device to a physical quantity. Uncertainty in the measurement can also be injected into the data through uncertainty about what is truly being measured. For example, measurement devices may be placed in orientations and exposed to boundary conditions that deviate from those specified for the experiment.

   Suggested documentation. What are the magnitudes and hypothesized sources of the measurement errors?

4. ***Bad data / Outliers*** Rejection of bad (inaccurate) data or non-physical outliers is an aspect of data analysis. Omitting outliers is often acceptable, but only when the root cause of the outlier is known. Understanding the root cause of impactful outliers often requires investigation into manufacturing and/or measurement process.

   Suggested documentation. How much data was rejected (not included in the final analysis)? Why it was rejected?

2.3 Model Uncertainty

Models, whether physics-based or statistical, are an important aspect of QMU analyses, particularly when data are sparse or are not representative. Information about the types of models, underlying assumptions, and additional uncertainties associated with modeling activities must be considered and communicated. If the model is purely physics-based, then existing predictive maturity methods like PCCM[13] can be used to assess the model credibility. If the model is empirical or statistical, then the credibility for these types of models should be evaluated, though we are not aware of any formal frameworks for evaluating model credibility. Goodness-of-fit methods are not sufficient metrics for evaluating model credibility[19], due to only testing if the distribution form hypothesis can be rejected. A typical means of assessing a statistical model's prediction capabilities is to demonstrate the model's ability to predict data not used to train the model. While such activities may be used to support model validation (as will be address later in Section 2.5), this does not probe the underlying model uncertainties we deem to be essential to model credibility. We recommend assessing two components of model credibility: the causal structural and functional assumptions of the model.

1. ***Causal structural assumptions*** The inability to accurately represent the collected data in the empirical model will introduce bias in QoI estimates. Causal structural assumptions concern whether causal or physics-based relationships can be learned from the available data by comparing how the data were generated to an underlying causal model for the data. Specifically, causal analysis concerns establishing underlying causal relations between variables and then determining if the collected data are sufficient to infer the QoI under these causal relations [20]. Common sources of bias include [21]:

   - Omitted variable bias: important variables were not measured in the dataset that should be included in the model to accurately capture the physics in the empirical model.
   - Selection bias: the data are not a random sample from the population, but the model assumes a random sample.

   Suggested documentation. Was the causal structure of the model considered? Is the fitted model consistent with an underlying causal model for the data? Is selection or omitted variable bias present? To what fidelity is the causal structure understood?

2. ***Functional assumptions*** Given a set of causal structural assumptions, statistical models are then specified to represent the empirical relationship between the inputs

and outputs. Functional assumptions specify this relationship, conditioning on the set of causal assumptions. Stated differently, causal assumptions pertain to whether all of the necessary inputs are included in the modeling to address biases in the data; functional assumptions pertain to whether the empirical model is correct, conditioning on having the correct inputs in the model. Examples of functional assumptions include normality or other distributional assumptions, linearity between inputs and outputs, and no interaction between inputs on the output, i.e., independence of effects. The complexity of the selected model is often limited by the available data. Further, the importance of the functional assumptions often varies. For instance, normality assumptions will often have minimal impact when estimating means, but can have a significant impact on tail extrapolations, which are common in reliability and QMU analyses. If multiple model forms provide similar fits to the data, this model form uncertainty should be considered.

Suggested documentation. What functional assumptions were made? Were the assumptions tested empirically, based on subject matter data, or required due to lack of data? To what fidelity are the functional relationships understood? How sensitive is the QoI estimate to the functional assumptions?

2.4 Calibration / Parameter Estimation

The act of updating model parameters using data (including both estimating and quantifying uncertainty in the parameters) is called calibration when models are physics-based and parameter estimation when models are empirical/statistical. These definitions are not universally accepted, but will be used within this framework. When data are sparse, calibration and statistical estimation procedures can perform poorly (e.g., maximum likelihood, bootstrapping), and limitations to calibration procedures should be addressed. Bayesian calibration processes incorporate additional knowledge into parameter estimates in the form of prior distributions. When using Bayesian techniques, the sensitivity of the calibration result to the prior should be considered and acknowledged if significant.

Additionally, not all calibration parameters are equally important to consider; specifically, the degree of consideration paid to an updated parameter should scale with the model's sensitivity to that parameter. Model sensitivity analysis typically refers to evaluating the magnitude of change in a prediction caused by changes to an input parameter's value.

Suggested documentation. What is the accuracy of the selected calibration/estimation procedure for important model parameters in the application? Was additional information incorporated into the parameter estimates? What is the sensitivity to updated and non-updated parameters?

2.5 Validation

Once the dataset is understood, the model selected, and the model fit to best represent the data, the model's validity should be judged with regards to the prediction of interest (quantity deemed comparable to requirement). Comparing model predictions with experimental data allows for the model's predictive capability, in regards to the prediction quantities of interest, to be quantitatively assessed. Model validation should occur when using physical-based models or statistical models[6]. Model validation has become a major area of emphasis in the CompSim domain[22, 23], and is well developed for statistical cross-validation of models[24, 25, 26]. It should be noted that validation cannot prove a model's predictive capability, only provide supporting evidence. If data sufficiently relevant to the requirement was available, then this data would be used to make the QMU assessment, and models would not be needed.

Suggested documentation. How well does the model predict the available data? Can the model be compared to an 'external test set,' i.e., data that were not used to build or calibrate the model? If so, what is the fidelity of the validation data? Are the model predictions consistent with subject matter knowledge? How relevant to the requirement is the validation?

3 EVALUATING CREDIBILITY

Once these elements of credibility have been evaluated, then these elements can be assimilated to provide an overall assessment of credibility. Each of the five elements should undergo a **peer review** of the analysis decisions and **documentation** of those decisions so that the analysis can be fully understood at a later date. Both peer review and documentation are also included in the aforementioned experimental[17] and CompSim credibility approaches[13]. When presenting QMU results to a decision-maker, overviews of these five elements should be included in order to allow the decision-maker to better understand the value of information contained.

Whether to develop a quantitative scale for scoring analysis credibility remains an open question. Many 'predictive maturity' frameworks assign numeric scores to sub-elements and combine the sub-scores to create an overall score. For instance, in PCMM, sub-elements are assigned an ordinal score from 0 to 3, and the PCMM authors suggest methods for combining sub-element scores into a single overall score, though advise against this collapsing of information due to interpretability issues [13]. Zeng et al., (2017) score the 'trustworthiness' of methods using a decision model based on the analytic hierarchy process[27]. Hemez et al., (2010) developed a predictive maturity index, emphasizing the need to go beyond goodness of fit and consider a more wholistic picture of credibility when evaluating the predictive maturity of modeling and simulation based results[19].

We do not score credibility herein, instead favoring a more qualitative synopsis of the credibility supporting evidence. Following [13], we argue that there is not a natural 'weighting' of the subelements that can produce a meaningful overall score. Further, our experience suggests that quantitative scoring can become highly politicized and arbitrary. When presenting credibility results, we recommend that, instead of collapsing information into a quantitative score, information should be collapsed into a set of key points. Specifically, the sensitivity of the QMU predictions to the model assumptions should be qualitatively or quantitatively assessed. Elements with particularly low credibility and potentially high impact should be highlighted. In areas of concern, sensitivity studies [28] can be conducted to determine the potential quantitative impact of an assumption. If the QoI results hinge on assumptions that are highly uncertain, then the analysis naturally lacks credibility.

4 EXAMPLE APPLICATION

To demonstrate our concept of QMU credibility, the proposed framework is applied to a simulated molten-salt battery dataset. The example is meant to resemble a real-world equivalent that could be generated from a production facility. Conditions varied within the dataset are the environmental temperature (-35°C to 65°C), intensity of the current loading profile (characterized as varying intensities between 0 and 1), and production lot number (1 through 7). Typical means of visualizing this dataset against a requirement are shown in Figure 1. The performance requirement specified for this dataset is for the baseline voltage to remain above 30 volts for a specified time. The dataset is comprised of observed baseline voltages at the required time. Examination of the plots would seem to show that the requirement would be met, but it is also difficult to know how useful the information provide is in answering the question "what is the margin to the requirement and what are the associated uncertainties in that estimate?"

In order to better answer this margin and uncertainty question, suggested documentation from the proposed QMU credibility assessment framework when applied to the molten salt battery example is now provided.

4.1 Requirement Definition and QoI Selection

Requirement: Once activated, 99.5% of batteries must maintain a baseline voltage above 30 volts for YY seconds.
QoI: Predict the baseline voltage at the requirement time, in the battery population at their harshest temperature and loading conditions at end of life.
Rationale: The requirement must be met in the current battery stockpile for all application environments. Therefore, environments that impact battery voltage, such as temperature, loading profile, and age, must be considered.
Uncertainties: The QoI is defined at the worst-case load and

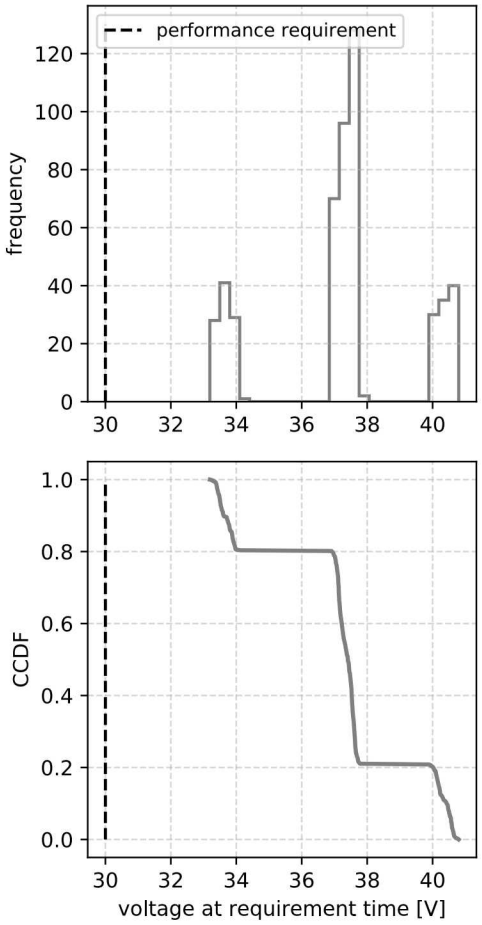

FIGURE 1: Raw battery performance data (500 samples) available for comparison to requirement. The top figure shows data's battery performance measure as a frequency density, where the three subpopulations come from the three testing temperatures. A cumulative complementary density function of the same performance measure is shown in the second plot as an example of another common method of visualizing QMU datasets.

temperature, but the requirement is ambiguous about those effects. It is improbable that the battery will ever experience these environmental extremes in use-environments. Therefore, predictions to the worst-case setting may be too conservative and sensitivity studies should explore the impact of this conservatism. The requirement specifies a 0.995 reliability, so variability in the battery population must be considered to reach comparable terms.

4.2 Data Quality

Metadata / Source: The data was gathered during production 15 years ago and captured using 2 high quality testers by 3 operators (equally distributed). The measurement method was

developed during battery design process.
Uncertainty: Uncertainty in the data primarily stems from the lack of representativeness. Individual sources of uncertainty are detailed below.
Sparsity: 500 units were tested at different environmental conditions. The quantity of data was deemed sufficient, but may require extrapolation from a statistical model to characterize tail behavior and estimate the percentile of interest and corresponding uncertainty.
Representativeness: Tested units were randomly sampled from all produced units and are therefore representative of all production lots. Tested units span the full temperature and loading conditions of interest, but were tested immediately after production and therefore do not provide any information about battery aging.
Noisiness / Measurement uncertainty: Loading and voltage measurements are sufficiently precise. Experimental temperature conditions are within $\pm 0.5°$C. Unit to unit variability is expected due to manufacturing tolerances of components, but will need to be characterized.
Bad data / Outliers: No outliers found.

4.3 Model Uncertainty

Type of model: For this dataset a statistical model is used, so causal and functional assumptions can be assessed.
Causal structural assumptions: An underlying causal model for the molten-salt battery system was elicited from experts and shown below in a causal network format.

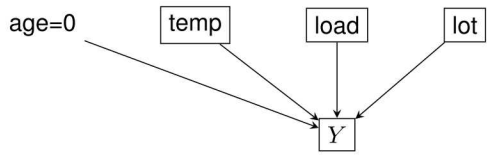

FIGURE 2: Causal network for molten salt battery. Boxes indicate observed variables. In causal language the boxes would be known as interventions due to those variable values being manipulable for a single battery test. Age is believed to potentially have a causal relationship with the QoI (Y, voltage), but is always 0 in the available data.

Age, temperature, loading conditions, and manufacturing lot number are all covariates that have a causal relationship with voltage. This causal structure assumes that the covariates are not confounders (no association between inputs).

To estimate the QoI (the voltage at the worst-case temperature and load) using the data, we define the QoI of interest as:

$$\begin{aligned} Y_Q(lot) &= (Y|age, load, temp, lot) \\ Y_Q &= (Y|age, load, temp) \\ &= \int Y_Q(lot)P(lot)\mathrm{d}lot \end{aligned}$$

An omitted variable bias exists, because all production data was collected on un-aged batteries. Expert judgement can be leveraged to determine the potential impact of this bias. To estimate the QoI, we assume:

$$Y_Q(lot) = (Y|age = 0, load, temp, lot) + \delta_A$$

where δ_A is an additive shift due to age that is elicited from experts or an auxiliary source of information.

Selection bias may also be present in the dataset due to a great number (≈60%) of the samples coming from room temperature tests versus the tests at temperature extremes (≈20% each). Because we condition on temperature in the QoI, this selection bias should not impact the ability to make inference about the QoI, though it does increase the variance of the estimated effect of temperature on voltage.

Functional assumptions: To model $Y_Q(lot)$, a linear model is assumed to be an appropriate method of representing the input-output relationships:

$$\begin{aligned} Y_Q(lot) &= \alpha_0 \times temp. + \alpha_1 \times load + \sum_{i=1}^{lots} \alpha_{1+i} I(i = lot) + \epsilon \\ \epsilon &\sim N(0, \sigma^2) \end{aligned}$$

This model assumes linear relationships between the input parameters and output; and assumes that interactions between parameters are insignificant. Further, residual variability due to manufacturing tolerances is modeled using an normal distribution. Because age $= 0$ in the dataset, age-effects cannot be estimated in the fitted statistical model.

4.4 Calibration / Parameter Estimation

Parameter estimation procedure: Ordinary least-squares (OLS) minimization is used to fit the statistical model. Because the sample size is large ($n = 500$) and model is simple, there are no meaningful uncertainties associated with the parameter estimation procedure. Fit results are shown below with standard errors and manufacturing variability captured as the model residual.

TABLE 1: OLS model parameter fits and standard errors.

parameter	OLS fit	std. err.
α_0	0.067	9.9E-5
α_1	-0.784	0.013
α_2	36.48	0.012
α_3	36.56	0.014
α_4	36.60	0.013
α_5	36.63	0.014
α_6	36.57	0.013
α_7	36.63	0.013
α_8	36.71	0.012
σ	0.0734	

Parameter sensitivities: Variability due to lot number is found through the calibration, but the major model sensitivities are due to the load and temperature. Sensitivity to aging cannot be inferred from the data, leaving an unknown in the analysis.

4.5 Validation

Prediction performance: With ample data the functional assumptions of linearity and no interaction can be evaluated from the data, as shown in Figure 3. Comparing the data fit

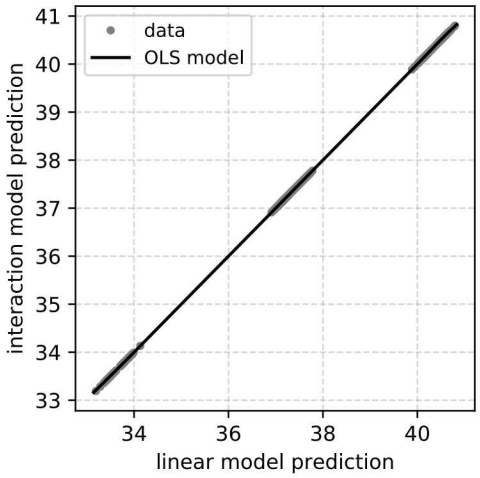

FIGURE 3: Validating linear assumption by comparing an OLS fit containing no-interaction terms to one with first order interactions.

of the model without any interactions to one with all possible interactions demonstrates that no improvement in fit occurred.

Prediction assessment: The model's predictions are consistent with behaviors anticipated by subject matter experts. The normal-residuals assumption can be empirically checked for inaccuracy. However, because we are using the model to predict a 99.5^{th} percentile from data collected at multiple loads and temperatures, we cannot directly confirm this assumption for the temperature/load condition where we are predicting (50 out of 500 samples were at the worst-case conditions).

4.6 Summary of credibility assessment

Key assumptions that were identified include: relevance of the QoI, normality of the residuals, and no battery aging. Sensitivity studies can be conducted to evaluate the impact of these assumptions. For instance, worst case temperature (-35°C) and loading conditions (1) have been assumed for the QoI. Figure 4 compares the calibrated model's prediction battery performance uniformly sampled across all conditions of interest with the predictions only for the worst case scenario, demonstrating how this QoI assumption significant impacts on our QMU conclusions. Model predictions for the worst-case

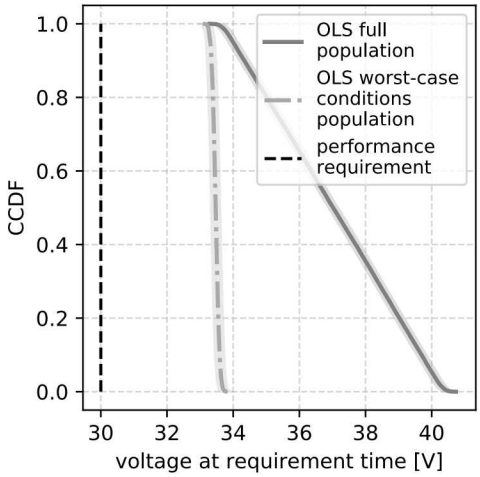

FIGURE 4: Illustration of potential QoI bias from using most strenuous conditions to define performance, as compared to general population. Shaded regions show 95% confidence intervals.

scenario are still a distribution, due to including lot and unit-to-unit variability. Comparing the worst-case distribution with the full potential distribution shows the degree of conservatism being added to the analysis. Experts may also have some knowledge about the appropriateness of a normal approximation to represent unit to unit variability in battery performance; to elicit such information, analysts can inquire about subpopulations or non-linearities in manufacturing tolerances that would result in a multi-modal, skewed, or heavy tailed distribution. Subject matter experts can be consulted to determine the impact of age on voltage over time, resulting in sensitivity information such as: aging will reduce the performance in a linear manner by at worst 3 volts by the end of lifetime of the battery population.

In order to estimate the 99.5^{th} percentile of the battery population at the worst case conditions, extrapolation using the model is needed due to only limited data available for those conditions. Figure 5 illustrates both the sensitivity to aging and the extent of extrapolation through plotting the experimental data and model predictions in terms of return level[29]. Return level is $\frac{1}{1-percentile}$; for instance, a return-level of 200 can be interpreted as the average number of units necessary to detect 1 failure (or, similarly, to inform a 0.995 reliability requirement)[18]. While the raw data trend and model predictions for the worst-case temperature and loading conditions show significant margin (~3.25 V) with minimal uncertainty (~0.14 V), when the worst case aging impact is considering, the margin becomes small (~0.25 V). Here margin is defined as the distance from the model's mean estimate

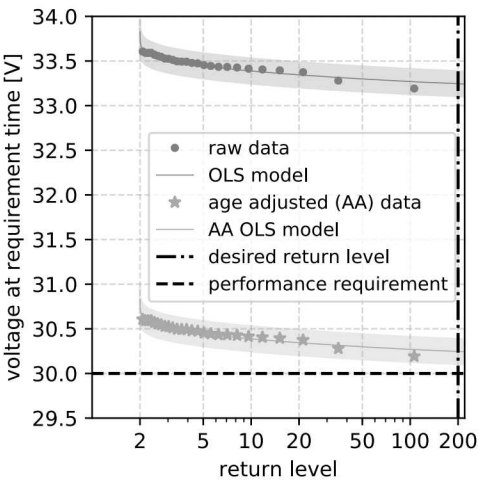

FIGURE 5: Comparison of return level trends for raw experimental data at most strenuous temperature and loading conditions with ($\star$) and without ($\bullet$) inclusion of worst case aging impact. Estimates of percentiles based on the raw data with 95% confidence intervals (shaded regions) are compared with those based on the fitted model (line). Variability in predictions is due to lot differences and unit-to-unit variability.

of the 99.5^{th} percentile and uncertainty is distance from the mean 99.5^{th} percentile estimate to the lower 95% confidence interval bound. With such a small margin, the results have an increased sensitivity to the assumptions used to extrapolate with the model. Where the uncertainty in the margin prediction was insignificant when age effects were neglected, it becomes potentially significant once that effect is considered. In a standard QMU analysis, the potential impact of an unquantified variable like age would likely not be presented.

5 SUMMARY

Following the recent revamping of the QMU process at Sandia and current emphasis on prediction credibility, guidance for assessing the credibility of QMU analyses is needed. Direction for how to communicate credibility of CompSim and experimental gathering campaigns (designed to support CompSim analysis) is already being developed. The future QMU paradigm will likely look more like experimentally supported CompSim than the historic model that was largely experimental based. With this change in QMU paradigm comes the need to provide credibility evidence with any QMU result. Five elements have been proposed as the basis for QMU credibility assessment framework: requirement definition and QoI selection, data quality, model uncertainty, calibration/parameter estimation, and validation. Through considering those elements and proposed subelements, documentation and communication of such information should be included in the communication of any QMU results. With this information the decision-maker receives a greater appreciation of the assumptions that went into generating the results as well understanding of the utility of the information provided. The application of this QMU credibility framework has been demonstrated on a molten-salt battery dataset.

ACKNOWLEDGMENTS

Review: We would like to thank John R. Lewis and Aubrey C. Eckert-Gallup for their helpful comments that allowed us to refine and improve this paper.

Funding: This work was supported by a Sandia National Laboratories Laboratory Directed Research and Development (LDRD) grant. Sandia National Laboratories is a multimission 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.

REFERENCES

[1] **https://www.merriam-webster.com/dictionary/credibility**, *Merrian-Webster Dictionary*.

[2] **National Research Council**, Evaluation of Quantification of Margins and Uncertainties Methodology for Assessing and Certifying the Reliability of the Nuclear Stockpile Committee on the Evaluation of Quantification of Margins and Uncertainties Methodology for Assessing and Certifying the Reliability of the Nuclear Stockpile, National Academies Press, Washington, 2008.

[3] **Kaplan, S.** and **Garrick, B. J.**, *On the quantitative definition of risk*, Risk analysis, Vol. 1, No. 1, pp. 11–27, 1981.

[4] **Pilch, M., Trucano, T.** and **Helton, J.**, *Ideas underlying quantification of margins and uncertainties (QMU): A white paper.*, Technical Report SAND2006-5001, Sandia National Laboratories, 2006.

[5] **Newcomer, J.**, *A New Approach to Quantification of Margins and Uncertainties for Physical Simulation Data*, Technical Report SAND2012-7912, Sandia National Laboratories, 2012.

[6] **Hund, L., Schroeder, B., Rumsey, K.** and **Murchison, N.**, *Robust Approaches to Quantification of Margin and Uncertainty for Sparse Data*, Technical Report SAND2017-9960, Sandia National Laboratories, 2017.

[7] **Roy Dholakia, R.** and **Sternthal, B.**, *Highly credible sources: Persuasive facilitators or persuasive liabilities?*, Journal of Consumer Research, Vol. 3, No. 4, pp. 223–232, 1977.

[8] **Gass, R. H.** and **Seiter, J. S.**, Persuasion: Social influence and compliance gaining, Routledge, 2015.

[9] **Lev-Ari, S.** and **Keysar, B.**, *Why don't we believe non-native speakers? The influence of accent on credibility*, Journal of experimental social psychology, Vol. 46, No. 6, pp. 1093–1096, 2010.

[10] **Chaiken, S.** and **Maheswaran, D.**, *Heuristic processing can bias systematic processing: effects of source credibility, argument ambiguity, and task importance on attitude judgment.*, Journal of personality and social psychology, Vol. 66, No. 3, pp. 460, 1994.

[11] **Heesacker, M., Petty, R. E.** and **Cacioppo, J. T.**, *Field dependence and attitude change: Source credibility can alter persuasion by affecting message-relevant thinking*, Journal of personality, Vol. 51, No. 4, pp. 653–666, 1983.

[12] **Mehta, U., Romero, V., Eklund, D., Pearce, J.** and **Keim, N.**, *The JANNAF Simulation Credibility Guide on Verification, Uncertainty Propagation and Quantification, and Validation*, *53rd AIAA Aerospace Sciences Meeting*, AIAA SciTech Forum, 2015.

[13] **Oberkampf, W. L., Pilch, M.** and **Trucano, T. G.**, *Predictive Capability Maturity Model for Computational Modeling and Simulation*, Technical Report SAND2007-5948, Sandia National Laboratories, 2007.

[14] **Beghini, L.** and **Hough, P.**, *Sandia Verification and Validation Challenge Problem: A PCMM-Based Approach to Assessing Prediction Credibility*, Journal of Verification, Validation and Uncertainty Quantification, Vol. 1, pp. 011002, 2016.

[15] **Schroeder, B., Silva, H.** and **Smith, K.**, *Separability of Mesh Bias and Parametric Uncertainty for a Full System Thermal Analysis*, *In ASME 2018 Verification and Validation Symposium*, pp. V001T04A003–V001T04A003, American Society of Mechanical Engineers, May 16 2018.

[16] **Oberkampf, W. L.** and **Smith, B.**, *Assessment Criteria for Computational Fluid Dynamics Model Validation Experiments*, Journal of Verification, Validation and Uncertainty Quantification, Vol. 2, pp. 031002–1, 2017.

[17] **Kieweg, S. L.** and **Witkowski, W. R.**, *Experimental Credibility and Its Role in Model Validation and Decision Making*, *Model Validation and Uncertainty Quantification, Volume 3*, pp. 31–36, Springer, 2019.

[18] **Hund, L., Schroeder, B., Rumsey, K.** and **Huerta, G.**, *Distinguishing between model- and data-driven inferences for high reliability statistical predictions*, Reliability Engineering & System Safety, Vol. 180, pp. 201–210, 2018.

[19] **Hemez, F., Atamturktur, H. S.** and **Unal, C.**, *Defining predictive maturity for validated numerical simulations*, Computers & Structures, Vol. 88, No. 7-8, pp. 497–505, 2010.

[20] **Pearl, J., Glymour, M.** and **Jewell, N. P.**, Causal inference in statistics: A primer, John Wiley & Sons, 2016.

[21] **Bareinboim, E.** and **Pearl, J.**, *Causal inference and the data-fusion problem*, Proceedings of the National Academy of Sciences, Vol. 113, No. 27, pp. 7345–7352, 2016.

[22] **Oberkampf, W.** and **Barone, M.**, *Measures of agreement between computation and experiment: Validation metrics*, Journal of Computational Physics, Vol. 217, No. 1, pp. 5–36, 2006.

[23] **Liu, Y., Arendt, P.** and **Huang, H.**, *Toward a Better Understanding of Model Validation Metrics*, Journal of Mechanical Design, Vol. 133, No. 7, pp. 071005, 2011.

[24] **Stone, M.**, *Cross-validatory choice and assessment of statistical predictions*, Journal of the royal statistical society. Series B (Methodological), pp. 111–147, 1974.

[25] **Efron, B.**, *Estimating the error rate of a prediction rule: improvement on cross-validation*, Journal of the American statistical association, Vol. 78, No. 382, pp. 316–331, 1983.

[26] **Hastie, T., Tibshirani, R.** and **Friedman, J.**, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Series in Statistics, Springer, NY, 2nd edn., 2009.

[27] **Zeng, Z., Di Maio, F., Zio, E.** and **Kang, R.**, *A hierarchical decision-making framework for the assessment of the prediction capability of prognostic methods*, Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, Vol. 231, No. 1, pp. 36–52, 2017.

[28] **EricksonKirk, M. et al.**, *Sensitivity Studies of the Probabilistic Fracture Mechanics Model Used in FAVOR Version 03.1*, NUREG-1808, US Nuclear Regulatory Commission, ADAMS ML, Vol. 61580349, 2004.

[29] **Coles, S., Bawa, J., Trenner, L.** and **Dorazio, P.**, An introduction to statistical modeling of extreme values, Vol. 208, Springer, 2001.