Reliability Engineering and System Safety 148 (2016) 32–43

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Role of calibration, validation, and relevance in multi-level uncertainty integration

Chenzhao Li, Sankaran Mahadevan *

Department of Civil and Environmental Engineering, Vanderbilt University, Nashville, TN, USA

ARTICLE INFO

Article history:
Received 10 February 2015
Received in revised form
26 October 2015
Accepted 7 November 2015
Available online 3 December 2015

Keywords:
Calibration
Validation
Uncertainty
Bayesian
Model reliability metric
Sobol indices

ABSTRACT

Calibration of model parameters is an essential step in predicting the response of a complicated system, but the lack of data at the system level makes it impossible to conduct this quantification directly. In such a situation, system model parameters are estimated using tests at lower levels of complexity which share the same model parameters with the system. For such a multi-level problem, this paper proposes a methodology to quantify the uncertainty in the system level prediction by integrating calibration, validation and sensitivity analysis at different levels. The proposed approach considers the validity of the models used for parameter estimation at lower levels, as well as the relevance at the lower level to the prediction at the system level. The model validity is evaluated using a model reliability metric, and models with multivariate output are considered. The relevance is quantified by comparing Sobol indices at the lower level and system level, thus measuring the extent to which a lower level test represents the characteristics of the system so that the calibration results can be reliably used in the system level. Finally the results of calibration, validation and relevance analysis are integrated in a roll-up method to predict the system output.

1. Introduction

Parameters of computational models are often calibrated using experimental data. For a complicated system it may be difficult to conduct full-scale experiments, but it may be possible to obtain data at lower levels of complexity (e.g., isolated physics or simpler configurations). Fig. 1 shows such a multi-level problem with two lower levels (G_1, G_2) and a system level (H). The lower levels and the system level constitute a hierarchy, and different levels have the same set of model parameters ($\boldsymbol{\theta}_m$) that need to be calibrated.

In order to predict the system level output when data are only available at lower levels, a reasonable route is to quantify the model parameters using lower level data, and propagate the results through the computational model at the system level. Several issues need to be addressed in realizing such a multi-level parameter estimation problem. First, even if model input and output are measured in the lower level tests, thereby forming pairwise input–output data, the calibration result can still be uncertain due to several sources, including: 1) model errors in the lower level computational models; 2) measurement errors in the experiments; and 3) sparse experimental data. Second, the existence of *multiple* lower levels provides multiple possibilities to conduct model calibration and leads to multiple calibration results. In a multi-level problem, model calibration can be conducted using the data from a single level or multiple levels. For the problem in Fig. 1 with two lower levels, 3 calibration options are possible: 1) calibration using the data and model from Level 1 alone; 2) calibration using the data and model from Level 2 alone; and 3) calibration using the data and models from both Level 1 and Level 2. Generally, if data are available at n different levels, $2^n - 1$ model calibration options are possible to quantify the uncertainty of model parameters [1].

This paper uses Bayesian inference for model calibration, thus the result of model calibration is a joint posterior distribution of model parameters. As Kennedy and O'Hagan [2] pointed out, the posterior distribution is the "best-fitting" results in the sense of representing the calibration data faithfully, not necessarily representing the true physical values. *The main objective of this paper is to determine the appropriate distribution for model parameters $\boldsymbol{\theta}_m$ to be used in system level prediction.* One possibility is to use all the lower level data in model calibration and propagate the resultant posterior distribution to predict the system level output. However, this result is conditioned on the event that both the models at Level 1 and Level 2 are valid, which may or may not be true [3]. This paper answers this question by assigning a "confidence" measure to each posterior distribution. Note that this paper is not using the term "confidence" in the same sense as is used in statistics (as in confidence interval). This "confidence" measure

* Corresponding author. Tel.: +1 615 322 3040.
E-mail address: sankaran.mahadevan@vanderbilt.edu (S. Mahadevan).

http://dx.doi.org/10.1016/j.ress.2015.11.013
0951-8320/

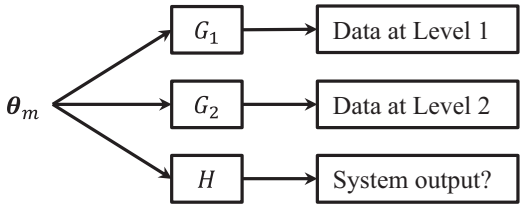

Fig. 1. Multi-level parameter estimation problem.

constitutes of two components: 1) the model validity at the corresponding lower level (one can think of this as local confidence regarding each lower level); 2) the relationship between the lower level and the system level, i.e., the relevance of the posterior distribution obtained at the lower level to the system level prediction problem (one can think of this as inter-level confidence). The relationship between two lower levels can be also important. However, this relationship is not considered here since in this paper the obtained information in a lower level is extrapolated to the system level, but not to another lower level.

Before quantifying the local confidence, the relationship between model calibration and model validation should be clarified. The purpose of model calibration is to adjust a set of parameters associated with a computational model so that the agreement between model prediction and experimental observation is maximized [4]. The term "model validation" has had different interpretations in different studies, and this paper follows the AIAA definition [5], i.e., model validation is the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model. Generally model validation is realized by comparing the model prediction against experimental data. Both model calibration and model validation are conducted in this paper, but they use different sets of experimental data (no calibration data is used in model validation). Comprehensive reviews on model validation can be found in [5–8]. A methodology for integrating model validation results from multiple experiments, each of which tests one part of the physics in the target application, can be found in [9].

Model calibration and model validation are distinct activities. Theoretically, for a computation model $F(\boldsymbol{\theta}_m;\boldsymbol{x})$ where $\boldsymbol{x}$ is a set of model inputs and $\boldsymbol{\theta}_m$ is a set of model parameters, model validation can be conducted exclusive of any model calibration [5] if the model parameters are assumed to be known. However, the model parameters $\boldsymbol{\theta}_m$ are often unknown. Therefore, prior to model validation, model calibration can be conducted to quantify the values of $\boldsymbol{\theta}_m$ or reduce the uncertainty about their values. The KOH framework [2] of model calibration used in this paper not only reduces the analyst's uncertainty about $\boldsymbol{\theta}_m$ by Bayesian inference, but also quantifies the model error $\delta(\boldsymbol{x})$ which is defined as the difference between model prediction and reality. The corrected prediction model under the KOH framework is $F(\boldsymbol{\theta}_m;\boldsymbol{x})+\delta(\boldsymbol{x})$. Compared to the original computational model, the new model is different in two aspects: 1) reduced uncertainty in $\boldsymbol{\theta}_m$; and 2) introduction of model error $\delta(\boldsymbol{x})$. In this paper, the model to be assessed in model validation is this "corrected" model. Thus validation is a subsequent and distinct activity after calibration in this paper. In other words, we consider model calibration and model validation as two distinct activities, and use two different sets of experimental data for these two activities, as suggested in [10,11]. Thus the calibration results of $\delta(\boldsymbol{x})$ and $\boldsymbol{\theta}_m$ within a single level do not change as a result of model validation in our approach.

With the calibration and validation perspectives to be used in this paper defined as above, the reason to use model validation to quantify the local confidence is explained next. In model validation, the assessed model validity of the corrected prediction model $F(\boldsymbol{\theta}_m;\boldsymbol{x})+\delta(\boldsymbol{x})$ at a lower level is a combined effect of three components: 1) $F(\boldsymbol{\theta}_m;\boldsymbol{x})$; 2) $\delta(\boldsymbol{x})$; and 3) the posterior distribution of $\boldsymbol{\theta}_m$. The third aspect corresponds to the "local confidence" (not to be confused with confidence intervals used in statistics), thus this paper takes the model validity as one factor affecting our confidence in extrapolating the posterior distribution of the model parameter from the lower level to the system level. This is reasonable since the model parameter has been calibrated with a model corresponding to the lower level experiment, and it is important to know whether the model was calibrated accurately; the calibration result is obviously affected by how accurately the lower level model represents the physics in the lower level experiment.

Model validation is about comparing the model prediction against experimental data, and a model validation metric is needed to quantify this comparison. Among the validation metrics in the literature, classical hypothesis testing gives an acceptance/rejection decision. Confidence intervals have also been calculated for the difference between model prediction and observed data [5]. Although the confidence intervals may provide a quantitative measure of the model validity at a single level, it is not possible to apply them in uncertainty propagation and integration across multiple levels, since the concept of propagation of confidence interval does not exist in classical statistics. Validation metrics resulting in a single quantitative value indicating the degree of model validity have also been developed. In Bayesian hypothesis testing [10,12], the posterior distribution obtained by model calibration is used as the null hypothesis and an alternative distribution is selected for the alternative hypothesis. The result of Bayesian hypothesis testing is a Bayes factor (the likelihood ratio between the null and alternate hypotheses), measuring the support from validation data to the null and alternate hypotheses. This is a relative measure significantly depending on the choice of distribution of the alternate hypothesis. In contrast, Ferson et al. [13,14] proposed an area metric, which is the difference between CDFs and has the same unit as the prediction/data. For the case that the model output is stochastic at fixed model input, this metric measures the area between the CDF of model output and the EDF (empirical distribution function) of experimental data at a fixed model input. (Note that in this paper model inputs $\boldsymbol{x}$ and model parameters $\boldsymbol{\theta}_m$ are different quantities, thus the model output can be stochastic at fixed model inputs $\boldsymbol{x}=\boldsymbol{x}^*$ if the model parameters $\boldsymbol{\theta}_m$ are still uncertainty. In addition, uncertain model errors, surrogate model uncertainty are other reasons that the model output can be stochastic at fixed model inputs) If data are from experiments with different inputs, this metric is still applicable by building a single EDF for all the data with u-pooling method [13].

The model validation metric used in this paper is the model reliability metric proposed by Rebba and Mahadevan [15] and further developed by Sankararaman and Mahadevan [16]. This metric measures the model validity by "model reliability", which is defined as the probability that the difference between model prediction and observed data is less than a pre-defined tolerance. Here the model prediction is stochastic, whose uncertainty is caused by the uncertainty in the posterior distribution of model parameters as well as the uncertainty regarding the model error. In other words, the model reliability metric considers the combined effect of these two sources of uncertainty. The value of model reliability is between 0 and 1, thus it can be conveniently used as a weighting term in subsequent uncertainty integration across multiple levels.

For a given validation data point, the model reliability is a deterministic value. However, its value is different for different data points. To capture this variability in model reliability, this paper proposes a stochastic model reliability metric where the model reliability is treated as a random variable instead of a