

**PANEL DISCUSSION: INTEGRATING DATA FROM MULTIPLE SIMULATION MODELS OF  
DIFFERENT FIDELITY**

Derek Bingham

Department of Statistics and Actuarial Science  
Simon Fraser University  
Burnaby, BC, V5A 1S6, Canada

C. Shane Reese

Department of Statistics  
Brigham Young University  
Provo, UT 84602 USA

Brian Williams

Los Alamos National Laboratory  
Statistical Sciences, CCS-6  
PO Box 1663, MS-F600  
Los Alamos, NM, 87545 USA

**ABSTRACT**

Computer models are used to simulate physical processes in almost all areas of science and engineering. A single evaluation of these computation models (or computer codes) can take as little as a few seconds or as long as weeks or months. In either case, experimenters use the model outputs to learn something about the physical system. In some settings, outputs from several computational models, with varying levels of fidelity, are available to researchers. In addition, observations from the physical system may also be in hand. In this panel discussion we address issues relating to model formulation, estimation, prediction and extrapolation using multi-fidelity computer models are addressed. In the first presentation, Bayesian methods are used to build a predictive model using low and high fidelity computational models with different inputs and also field observations. The second presentation deals with the difficult computational issues facing computer model calibration and prediction using a Bayesian framework that are typically remedied through the use of Markov Chain Monte Carlo techniques. While the computational burden is substantial, we review faster alternatives to standard MCMC techniques that are particularly useful in the multi-fidelity simulator problem. In the final presentation, calibration of computational models is discussed in the context of validation and extrapolation, with introduction to developments in stochastic model calibration.

**1 PANELIST 1: DEREK BINGHAM, SIMON FRASER UNIVERSITY**

Computational models are used to simulate a wide variety of physical processes. A single run of such a model may take hours, days or even weeks. The potentially high computational cost of running the model makes it infeasible to continually exercise the simulator to carry out tasks such as solving inverse problems, parameter estimation and prediction. In cases where the computer code is slow to execute, experimenters are instead left to achieve their goals with only a limited number of calls to the computer model. The simulators are usually deterministic, and thus running the code twice with the same set of inputs yields the same outputs. To perform statistical inference in this deterministic setting (e.g., making predictions at

unsampled inputs with estimates of uncertainty), Sacks et al. (1989), proposed modelling the responses from a computer code as a realization of a Gaussian process (GP).

In some applications, experimenters have computer models with different levels of fidelity (e.g., Kennedy and O'Hagan 2000; Qian and Wu 2008) where the lower fidelity model can be run many more times than the higher fidelity model. While both models provide information about the physical process, the more computational, high fidelity model is assumed to yield more reliable information for the process of interest. Again, the model responses are typically modelled using a GP, but with the high fidelity model written as a function of the low fidelity model. However, even with multi-fidelity models in hand, there are limits as to what can be said about the process in the real world. To make predictions on the physical system, with associated estimates of uncertainty, field observations are typically needed. Several approaches have been proposed to combine field data and computer model output (e.g., Kennedy and O'Hagan 2001; Kennedy et al. 2008; Bayarri et al. 2007).

Another important goal of combining field observations and computer models is that of calibration or tuning (Kennedy and O'Hagan 2001). (We make the distinction that when we are estimating constants with meaning in the physical system, then we are doing model calibration, but if we are estimating constants to make the predictive model "good" in some sense, then we are tuning the model.) In this setting, the model runs and field observations are combined to form a predictive model and also estimate parameters that govern the computational model. The situation is even more challenging if there are multiple computer models, with varying levels of fidelity, in play.

Here methodology for combining, and tuning, multi-fidelity computer models with field observations is discussed. Briefly, suppose there are two simulators used to describe the physical system - one with lower fidelity than the other and denoted as the LF and HF simulators respectively. Assume that the LF simulator requires inputs  $(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_l)$ , where  $\mathbf{x} = (x_1, \dots, x_p)^T$  the observable inputs in the experiment,  $\mathbf{t}_f = (t_{f,1}, \dots, t_{f,m_f})^T$  are tuning parameters that are unique to the LF simulator and  $\mathbf{t}_l = (t_{l,1}, \dots, t_{l,m_l})^T$  are tuning parameters that appear in all simulators. Further assume that the HF simulator takes as inputs  $(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_h)$ , where  $\mathbf{t}_h = (t_{h,1}, \dots, t_{h,m_h})^T$  are tuning constants that appear in the HF simulator only. The model proposed is one where lower fidelity models are tuned to next higher fidelity model. In this setting, we tune the LF simulator to the HF simulator and the HF simulator to reality. The approach uses a Bayesian hierarchical model to integrate all sources of data to build a predictive model for the process.

We make a few key observations: (i) if the computer models do not exactly match the mean of the physical system, then model calibration is not likely possible; (ii) when the high and low fidelity models do not share the same input parameters, then one can only tune the computer models; and (iii) the multiple models can often be combined with observations to form a better predictive model than using any one source alone.

## 2 PANELIST 2: C. SHANE REESE, BRIGHAM YOUNG UNIVERSITY

In a common inverse problem, we wish to infer about an unknown spatial field  $x = (x_1, \dots, x_m)^T$  given indirect observations  $y = (y_1, \dots, y_n)^T$ . The observations, or data, are linked to the unknown field  $x$  through some physical system

$$y = \zeta(x) + \varepsilon$$

where  $\zeta(x)$  denotes the physical system and  $\varepsilon$  is an  $n$ -vector of observation errors. Examples of such problems include medical imaging (Kaipio and Somersalo 2004), geologic and hydrologic inversion (Stenerud et al. 2008), and cosmology (Jimenez et al. 2004). When a forward model, or simulator, of the physical process

$\eta(x)$  is available, one can model the data using the simulator

$$y = \eta(x) + e,$$

where  $e$  includes observation error as well as error due to the fact that the simulator  $\eta(x)$  may be systematically different from reality  $\zeta(x)$  for input condition  $x$ . Our goal is to use the observed data  $y$  to make inference about the spatial input parameters  $x$  – predict  $x$  and characterize the uncertainty in the prediction for  $x$ .

The likelihood  $L(y|x)$  is then specified to account for both mismatch and sampling error. We will assume zero-mean Gaussian errors so that

$$L(y|x) \propto \exp\{-\frac{1}{2}(y - \eta(x))^T \Sigma_e^{-1}(y - \eta(x))\}, \quad (1)$$

with  $\Sigma_e$  known. It is worth noting here that the data often come from only a single experiment. So there is no opportunity to obtain data from additional experiments for which some controllable inputs have been varied. Because of this, there is little hope in determining the sources of error in the error vector  $e$ , as is often done in the statistical analysis of complex computer code outputs (Kennedy and O’Hagan 2001). Therefore, the likelihood specification will often need to be done with some care, incorporating the modeler’s judgment about the appropriate size and nature of the mismatch term.

In many inverse problems we wish to reconstruct  $x$ , an unknown process over a regular 2-d lattice. We consider systems for which the model input parameters  $x$  denote a spatial field or image. The spatial prior is specified for  $x$ ,  $\pi(x)$  which typically takes into account modeling, and possibly computational considerations.

The resulting posterior is then given by

$$\pi(x|y) \propto L(y|\eta(x)) \times \pi(x).$$

This posterior can, in principle, be explored via Markov chain Monte Carlo (MCMC). However the combined effects of the high dimensionality of  $x$  and the computational demands of the simulator make implementation difficult, and often impossible, in practice. By itself, the high dimensionality of  $x$  isn’t necessarily a problem. MCMC has been carried out with relative ease in large image applications (Weir 1997). However, in these examples, the forward model was either trivial, or non-existent. Unfortunately, even a mildly demanding forward simulation model can greatly affect the feasibility of doing MCMC to solve the inverse problem.

We apply a standard single site updating scheme that dates back to (Metropolis et al. 1953) to sample from this posterior. While this approach has proven effective in a variety of applications, it has the drawback of requiring hundreds of thousands of calls to the simulation model. We also consider two MCMC schemes that use highly multivariate updates to sample from  $\pi(x|y)$ : the multivariate random walk Metropolis algorithm (Gelman et al. 1996) and the *distributed evolution*-MCMC (DE-MC) sampler of (ter Braak 2006). Such multivariate updating schemes are alluring for computationally demanding inverse problems since they have the potential to update many (or all) components of  $x$  at once, while requiring only a single evaluation of the simulator. Next, we consider augmenting the basic posterior formulation with additional formulations based on faster, approximate simulators. The faster, approximate simulators are created by altering the multigrid solver used to compute  $\eta(x)$ . These approximate simulators can be used in a delayed acceptance scheme (Fox and Nicholls 1997; Christen and Fox 2005, as well as in an augmented formulation (Higdon et al. 2002). Both of these recipes can be utilized with any of the above MCMC schemes, often leading to substantial improvements in efficiency.

### **3 PANELIST 3: BRIAN WILLIAMS, LOS ALAMOS NATIONAL LABORATORY**

#### **3.1 Introduction**

Computational models are increasingly utilized to provide predictive science capability for assessment or certification of complex systems. In many cases, some quantities of interest can be calculated by the model but cannot be directly validated due to lack of experimental data. For example, full system tests are often prohibitively expensive or perhaps even impossible to conduct. Traditional approaches to computational model validation are generally carried out in two steps:

1. Calibrate uncertain physics parameters so that the resulting model calculations provide an optimal match to reference experimental data (referred to as calibration data) with respect to pre-specified validation metrics. This calibration may be deterministic or probabilistic; producing optimal parameter values in the former case or resulting in probability density functions describing parameter uncertainty in the latter case.
2. Set physics variables at calibrated values and utilize the computational model to predict the outcomes of experiments held out of the calibration process (referred to as validation data). If these predictions compare favorably to their corresponding validation data as measured by the chosen validation metrics, the code is deemed valid for making predictions in the domain of applicability defined by the validation data.

The most common application of this validation process involves deterministic calibration, and the chosen validation metric judges the quality of the resulting point predictions from the computational model via comparing them to validation data with accommodation for irreducible uncertainty in these data (Oberkampf and Barone 2006). For example, chi-square statistics have been proposed for this purpose (Hills and Trucano 2002). Recently, systematic model bias was rigorously modeled and leveraged in a novel validation procedure (Wang et al. 2009). Probabilistic calibration involves solving an inverse problem to determine (or sample from) a probability law describing uncertainty in the physics parameters that is consistent with the observed irreducible uncertainty in the calibration data (Kennedy and O'Hagan 2001; Higdon et al. 2004; Higdon et al. 2008). The resulting uncertainty in the parameters is propagated through the code to produce predictions of the validation data with quantified uncertainty. Validation decisions are predicated on formal comparisons of these calibrated predictions with corresponding validation data that fully incorporate all quantified uncertainties. When the quantities of interest for computational model validation represent extrapolations beyond the domain of applicability in which experimental data is readily available, traditional validation approaches no longer apply, as direct comparisons to data are not available. We provide an overview of ongoing research efforts aimed at moving beyond the concepts of traditional validation to assess the predictive capability of computational models in extrapolation regimes.

#### **3.2 Validation and Extrapolation**

Novel enhancements of the traditional validation paradigm are being actively pursued by the computational engineering and sciences community (Oden et al. 2010a; Oden et al. 2010b) that accommodate validation of computational models for making predictions in scenarios with non-existent physical data. This approach is summarized as follows:

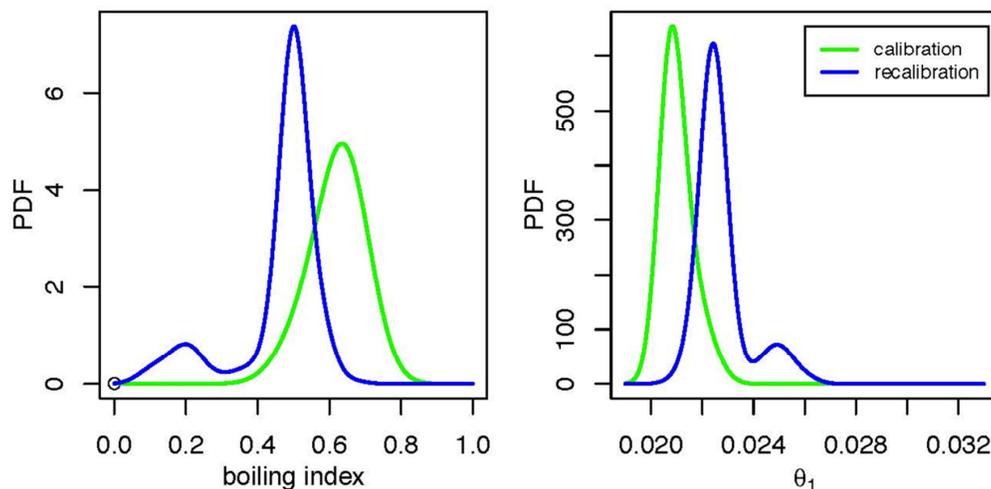
- Perform an initial probabilistic calibration of uncertain physics parameters using calibration data derived from observable phenomena known to be accurately represented by generally well-accepted physics models.
- Update the initial parameter calibration (referred to as recalibration) by introducing additional constraints imposed by validation data taken from complex phenomena anticipated to challenge the fidelity of computational model(s) simulating them.
- Calculate distributions for the extrapolative quantities of interest obtained from forward propagation of physics model parameter uncertainty derived from both the calibration and recalibration processes.
- Compare these predictive distributions according to pre-specified validation metrics and decide if the computational model should be invalidated for the extrapolations of interest.

The successful implementation of this approach requires a protocol for identifying and distinguishing calibration and validation data. As a general guideline, calibration data may be identified with separate effects tests that inform on single physics components of a computational model, while validation data may be identified with integral effects tests that challenge the fidelity of coupled multi-physics representations in the computational model. The essential idea is that breakdowns in a code's multi-physics capabilities are expected to manifest themselves in predictions of full system quantities of interest, particularly if they represent extrapolations beyond the domain of applicability for which calibration and validation data are available. However, validation data that allows these deficiencies to be exposed must be available and selected by knowledgeable experts for this purpose. The metrics upon which the model invalidation decision depends could involve comparison of the predictive distributions from calibration and recalibration in their entirety, or functionals of these distributions such as quantiles or thresholds of relevance to the application(s) for which extrapolations are required. Furthermore, even if a computational model is not invalidated based on this comparison of predictive distributions, additional concerns could arise that may need to be addressed prior to use of the model for extrapolation:

- The calibration data may insufficiently constrain the uncertain physics variables, allowing an unacceptable level of compensating errors among physics models.
- Uncertainties in quantities of interest may be too large even after recalibration.

Additional data may be required to mitigate one or both of these concerns prior to arriving at a final validation decision that provides sufficient confidence in the use of a computational model for extrapolation. We illustrate this process in a proof-of-concept study involving a thermal hydraulics code used to predict the proportion of surface area of a nuclear fuel assembly that experiences positive mass evaporation rate (sub-cooled boiling), referred to as the boiling index. Boiling index data from three assemblies running at different levels of average power are available: the two lowest power assemblies provide calibration data while the highest power assembly provides the validation data. Prediction of boiling index is desired for a fourth assembly. Although this assembly ran at the same average power as the validation assembly, we regard its boiling index prediction as an extrapolation due to other differences such as spatial location. Five uncertain parameters in physics models of the thermal hydraulics code were selected for calibration (and recalibration). The left panel of Figure 1 shows the predictive distributions of boiling index for the fourth assembly based on calibration and recalibration. These predictions would be deemed inconsistent by any reasonable validation metric, resulting in invalidation of the thermal hydraulics code for this extrapolation. The right panel of Figure 1 exposes the cause of this failure: calibration of the most sensitive physics variation depends strongly on average power of the assemblies providing experimental data. We conclude

that robustness of the corresponding physics model with respect to average power is in need of rigorous examination.



**Figure 1:** Boiling index predictive distributions (left) and calibrated marginal distributions of most sensitive physics variation  $\theta_1$  (right) resulting from calibration (green) and recalibration (blue).

### 3.3 Conclusions

As the scientific community increases its reliance on computational models for predicting real-world behavior in support of system certification, the execution of such models in extrapolation regimes will become more prevalent. Validating computational models in extrapolation regimes presents a unique challenge to the scientific community in that the validation must occur in the absence of experimental data. This document presented one method that sheds light on this clearly contradictory pursuit. This method employs a recently developed paradigm involving calibration and recalibration of computational models to validate (or invalidate) these models in extrapolation regimes. While this technique shows promise, there exists substantial subjectivity and thus its application to more practical systems of interest must be more fully explored. Recently, algorithms for collecting additional experiments to efficiently improve predictive inference for validation applications in the interpolation regime have appeared in the literature (Ranjan et al. 2008 and Williams et al. 2011). Given the complexity of probabilistic calibration algorithms, substantial future work will be required to examine potential relationships between sample size and efficiency gains, and extensions of such concepts to the extrapolation regime.

### REFERENCES

- Bayarri, M., J. Berger, R. Paulo, J. Sacks, J. Cafeo, J. Cavendish, C. Lin, and J. Tu. 2007. “A Framework for Validation of Computer Models”. *Technometrics* 49:138–154.
- Christen, J., and C. Fox. 2005. “Markov Chain Monte Carlo Using an Approximation”. *Journal of Computational & Graphical Statistics* 14 (4): 795–810.
- Fox, C., and G. Nicholls. 1997. “Sampling Conductivity Images via MCMC”. In *Proceedings of the Leeds Annual Statistical Research Workshop (LASR)*, 91–100. University of Leeds.

- Gelman, A., G. Roberts, and W. Gilks. 1996. "Efficient Metropolis jumping rules". In *Bayesian Statistics*, edited by J. M. Bernardo et al., Volume 5, 599. OUP.
- Higdon, D., J. Gattiker, B. Williams, and M. Rightley. 2008. "Computer model calibration using high-dimensional output". *Journal of the American Statistical Association* 103 (482): 570–583.
- Higdon, D., M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne. 2004. "Combining Field Data and Computer Simulations for Calibration and Prediction". *SIAM Journal of Scientific Computing* 26:448–466.
- Higdon, D., H. Lee, and Z. Bi. 2002. "A Bayesian Approach to Characterizing Uncertainty in Inverse Problems Using Coarse and Fine Scale Information". *IEEE Transactions in Signal Processing* 50:389–399.
- Hills, R., and T. Trucano. 2002. "Statistical validation of engineering and scientific models: a maximum likelihood based metric". Technical Report SAND2001-1783, Sandia National Laboratories, Albuquerque, NM.
- Jimenez, R., L. Verde, H. Peiris, and A. Kosowsky. 2004. "Fast cosmological parameter estimation from microwave background temperature and polarization power spectra". *Physical Review D* 70 (2): 23005.
- Kaipio, J. P., and E. Somersalo. 2004. *Statistical and Computational Inverse Problems*. New York: Springer.
- Kennedy, M., C. Anderson, A. O'Hagan, M. Lomas, I. Woodward, J. Gosling, and A. Heinemeyer. 2008. "Quantifying Uncertainty in the Biospheric Carbon Flux for England and Wales". *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171:109–135.
- Kennedy, M., and A. O'Hagan. 2000. "Predicting the Output from a Complex Computer Code when Fast Approximations are Available". *Biometrika* 87:11–13.
- Kennedy, M., and A. O'Hagan. 2001. "Bayesian Calibration of Computer Models". *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 63:425–464.
- Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. 1953. "Equations of state calculations by fast computing machines". *Journal of Chemical Physics* 21:1087–1091.
- Oberkampf, W., and M. Barone. 2006. "Measures of agreement between computation and experiment: validation metrics". *Journal of Computational Physics* 217 (1): 5–36.
- Oden, T., R. Moser, and O. Ghattas. 2010a. "Computer Predictions with Quantified Uncertainty, Part I". *SIAM News* 43 (9): 1–4.
- Oden, T., R. Moser, and O. Ghattas. 2010b. "Computer Predictions with Quantified Uncertainty, Part II". *SIAM News* 43 (10): 1–4.
- Qian, P., and C. Wu. 2008. "Bayesian Hierarchical Modeling for Integrating Low-accuracy and High-accuracy Experiments". *Technometrics* 50:192–204.
- Ranjan, P., D. Bingham, and G. Michailidis. 2008. "Sequential experiment design for contour estimation from complex computer codes". *Technometrics* 50 (4): 527–541.
- Sacks, J., W. J. Welch, T. J. Mitchell, and H. Wynn. 1989. "Design and analysis of computer experiments". *Statistical Science* 4:409–423.
- Stenerud, V., V. Kippe, K. Lie, and A. Datta-Gupta. 2008. "Adaptive multiscale streamline simulation and inversion for high-resolution geomodels". *SPE Journal* 13 (1): 99–111.
- ter Braak, C. F. J. 2006. "A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces". *Statistics and Computing* 16 (3): 239–249.
- Wang, S., W. Chen, and K. Tsui. 2009. "Bayesian validation of computer models". *Technometrics* 51 (4): 439–451.

- Weir, I. 1997. "Fully Bayesian Reconstructions from single photon emission computed tomography". *Journal of the American Statistical Association* 92:49–60.
- Williams, B. J., J. L. Loepky, L. M. Moore, and M. S. Macklem. 2011. "Batch sequential design to achieve predictive maturity with calibrated computer models". *Reliability Engineering & System Safety* 96 (9): 1208 – 1219.

#### **AUTHOR BIOGRAPHIES**

**DEREK BINGHAM** is an Associate Professor and Canada Research Chair in the Department of Statistics and Actuarial Science at Simon Fraser University. His research interests are computer experiments, experimental design and applications in industrial statistics and physics.

**C. SHANE REESE** is Professor and Associate Chair in the Department of Statistics at Brigham Young University. His research interests are Bayesian Hierarchical modeling applied to reliability, computer experiments, and sports. He is an Associate Editor for the *Journal of the American Statistical Association* and *Chance* magazine.

**BRIAN J. WILLIAMS** is a Scientist at Los Alamos National Laboratory. His research interests are computer experiments, experiment design and Bayesian analysis. He is an Associate Editor for the *American Statistician*.