A Bayesian Analysis of the Thermal Challenge Problem

F. Liu, M. J. Bayarri, J. Berger, R. Paulo, J. Sacks

Duke University, Università de Valencia, Duke University,
ISEG-Technical University of Lisbon, National Institute of Statistical Sciences

Abstract

A major question for the application of computer models is *Does the computer model adequately represent reality?* Viewing the computer models as a potentially biased representation of reality, Bayarri *et al.* (2007) develop the Simulator Assessment and Validation Engine (`SAVE`) method as a general framework for answering this question. In this paper, we apply the `SAVE` method to the challenge problem which involves a thermal computer model designed for certain devices. We develop a statement of confidence that the devices modeled can be applied in intended situations.

Keywords: Bayesian analysis; Computer model validation; Gaussian stochastic process; Thermal computer model.

1 Introduction

We view the most important question for the evaluation of a computer model to be

Does the computer model adequately represent reality?

Because a computer model can never be said to be a completely accurate representation of the real process being modeled, we do not focus on answering the yes/no question "Is the model correct?", although this question can be addressed within our framework. In the vast majority of cases, the relevant question is, instead, "Does the model provide predictions that are accurate enough for the intended use of the model?" While there are several concepts within this question deserving careful definition, the central issue is simply that of assessing the accuracy of model predictions. This will be

done by presenting *tolerance bounds,* such as 803 ± 76, for a model prediction 803, with the interpretation that there is a specified chance (e.g., 80%) that the corresponding true process value would lie within the specified range. Such tolerance bounds should be given *whenever predictions are made*, i.e., they should routinely be included along with any predictions making use of the model.

This focus on giving tolerance bounds arises for three reasons:

1. Models rarely give highly accurate predictions over the entire range of inputs of possible interest, and it is important to characterize regions of accuracy and inaccuracy.

2. The degree of accuracy needed can vary from one application of the computer model to another.

3. Tolerance bounds incorporate *model bias*, the principal symptom of model inadequacy; accuracy of the model cannot simply be represented by a variance or standard error.

These concerns are obviated by routinely presenting tolerance bounds along with model predictions. Thus, at a different input value, the model prediction and tolerance bound might be 650 ± 155, and it is immediately apparent that the model is considerably less accurate at this input value. Either of the bounds, 76 or 155, might be acceptable or unacceptable predictive accuracies, depending on the intended use of the model.

Bayesian analysis: Producing tolerance bounds is not easy. A list of hurdles includes:

- Uncertainties in model inputs or parameters of different varieties: based on data, expert opinion, or simply an "uncertainty range."

- Model runs are expensive and only limited model-run data may be available.

- Field data of the actual process being modeled may be limited and noisy.

- Data may be of a variety of types, including functional data.

- Model runs and field data may be at different input values.

- We may need to 'tune' and calibrate parameters and inputs of the computer model based on field data, and at the same time (because of sparse data), apply the validation methodology.
- The computer model is typically highly non-linear.
- Accounting for possible model bias is challenging.
- Validation should be viewed as an accumulation of evidence to support confidence in the model outputs and their use, and the methodology needs to be able to update its current conclusions as additional information arrives.

Overcoming these hurdles requires a powerful and flexible methodology; the only one we know that can accommodate all of these different factors is a Bayesian approach, following the work in Kennedy and O'Hagan (2001), to assessment and analysis of uncertainty, together with its modern computational implementation via Markov Chain Monte Carlo analysis (see, e.g., Robert and Casella (1999)).

When a bias in the model is detected, the methodology allows one to adjust the model prediction by the estimated bias, creating a "reality prediction", and provides tolerance bounds for this prediction. In specific applications this can result in considerably more accurate predictions than use of the model alone (or use of the field data alone) and, importantly, responds to questions where prediction of reality is required.

Strictly speaking, the presence of bias would call into question the suitability of the model. However, the amount of bias may be small compared to the uncertainty in model output generated by measurement error or uncertainty in inputs. In such instances it is plausible that the model may retain substantial utility. The tolerance bounds for model and reality predictions provide such indications.

Prediction in specific application and assessment of the model respond to seemingly different questions. But they are two manifestations of the same principle: predictions must be accompanied by

measures of accuracy, the tolerance bounds, which can then be used for answers.

1.1 The thermal challenge problem

In this paper, we apply the Simulator Assessment and Validation Engine (`SAVE`) approach (Bayarri *et al.*, 2007) to the thermal challenge problem (Dowding *et al.*, 2007), produce predictions with uncertainty estimates, and interpret the implications of the results.

The output of the thermal computer model is

$$y^M(\kappa, \rho, T_0, L, q; x, t) = T_0 + \frac{qL}{\kappa}\left[\frac{\kappa t/\rho}{L^2} + \frac{1}{3} - \frac{x}{L} + \frac{x^2}{2L^2} - \sum_{N=1}^{6} \frac{2}{\pi^2 n^2} \exp\left(-\frac{n^2\pi^2\kappa t}{L^2\rho}\right) \cos\left(n\pi\frac{x}{L}\right)\right], \quad (1)$$

where κ is the thermal conductivity of the device, ρ is the volumetric heat capacity, q is applied heat flux, L = thickness, x = distance from the surface, T_0 = initial temperature and t is time. The inputs (κ, ρ) are physical properties varying from specimen to specimen; they are unknown for a particular device. The input T_0 is fixed at 25°C for all data and analyses, and is therefore ignored. The controllable inputs (L, q) are assumed to be known exactly and their specification is called a configuration.

Let $y^R(\kappa, \rho, L, q; x, t)$ be the real temperature at time t for a specimen with properties κ, ρ under the associated experimental configuration. The principal application is to predict the (real) temperature at $x = 0$, $t = 1000$ under the regulatory configuration $(L = 0.019, q = 3500)$, and determine whether

$$\mathrm{P}\{y^R(\kappa, \rho, L = 0.019, q = 3500; x = 0, t = 1000) > 900\} < .01, \quad (2)$$

the stated regulatory requirement. Because κ, ρ are unknown, interpretation of this probability must be dealt with. In fact, the Bayesian analysis we use treats these unknowns as random and their

distribution is incorporated into the calculation of the probability.

There are three sets of field (experimental) data. The material characterization data (MC) are used to provide prior distributions for the κ, ρ's that are associated with each specimen. The ensemble data (EN) are used to produce assessments of the bias as well as tolerance bounds on model and reality predictions and are then further used to compare the predictive distribution

$$\pi\left(y^R\left(\kappa, \rho, L = .019, q = 3000; x = 0, t\right) \mid \text{EN}\right)$$

with the accreditation configuration data (AC). The EN and AC data are then taken together and lead to a follow-on analysis providing new assessments of bias and tolerance bounds for predictions. This second analysis is used to predict temperature at the regulatory configuration.

Each of the EN and AC data has its own (unknown) κ, ρ and so there are as many parameters κ_i, ρ_i as there are EN and AC measurements. These many unknowns are assumed to have a common prior distribution.

The analyses are carried out for two situations: the so-termed medium-level data and the high-level data; the medium-level data is a subset of the high-level data. There are some limited data with $x \neq 0$ in the accreditation data set but we ignore them because only surface temperature ($x = 0$) is involved in the intended application (regulatory condition) and little benefit is expected by including them. In all that follows, x is fixed at 0. We thus remove x from the input list.

1.2 The Simulator Assessment and Validation Engine (SAVE)

SAVE (Bayarri *et al.*, 2007) is a Bayesian-based analysis that combines computer simulation results with output from field experiments to produce assessment of a simulator (computer model). The method follows these six steps:

1. Specify the Input/Uncertainty (I/U) map, which consists of prior knowledge on uncertainties or

ranges of the computer model inputs and parameters. The I/U map for the thermal challenge problem is in Table 1.

2. Set the evaluation criteria for intended applications.

3. Collect data – both field and computer runs;

4. Approximate, if necessary, computer model output;

5. Compare computer model output with field data using Bayesian statistical analysis;

6. Feed back the analysis to improve the current validation scheme and computer model, and feed forward to future validation activities.

The central technical issues for `SAVE` lie in implementing Steps (4) and (5). We bypass (4) because the thermal computer model in Equation (1), is fast and can be evaluated as many times as we wish. The *statistical structure* for implementing (5) is built as follows: View the computer model $y^M(\cdot)$ as a possibly biased representation of the underlying real physical phenomenon $y^R(\cdot)$ by defining a bias process, b, to satisfy $y^R(\cdot) = y^M(\cdot) + b(\cdot)$. Field data $y^F(\cdot)$ are realizations of the real process,

$$y^F(\cdot) = y^M(\cdot) + b(\cdot) + e(\cdot), \tag{3}$$

where $e(\cdot)$ is (field) measurement error. Arguments (inputs) of $y^R(\cdot), y^M(\cdot), b(\cdot), e(\cdot)$ will differ in kind depending on the specific model. In many problems (including the thermal challenge problem), the vector of inputs to the computer model $\boldsymbol{z}$ can be written as $\boldsymbol{z} = (\boldsymbol{u}, \boldsymbol{v})$, where $\boldsymbol{u}$ consists of unknown (tuning/calibration) parameters and $\boldsymbol{v}$ is a vector of controllable inputs. When the model output is a function of time, as it is in the thermal problem, Bayarri *et al.* (2005) treat time, t, as a controllable input, here kept separate from $\boldsymbol{v}$ in the notation.

When there are replicate field data $y_i^F(\cdot)$, we have corresponding $e_i(\cdot)$ but no replicates in $y^M(\cdot)$, unless the replicates have variations in one or more inputs, e.g., different samples of material being

tested so material properties that are inputs to the computer model will vary. These must be taken into account. In the thermal problem a replicate i has an associated $\boldsymbol{u}_i$, and the statistical model in (3) becomes

$$y_i^F(\boldsymbol{v}, t) = y^M(\boldsymbol{u}_i, \boldsymbol{v}, t) + b_i(\boldsymbol{v}, t) + e_i(t) \ . \tag{4}$$

In Equation (4), we have a different bias function b_i for each replication in the field. Also, there is confounding between each b_i and the corresponding $\boldsymbol{u}_i$, and, in general, no amount of data can sort this out. Moreover, there is very little data to infer about all of these different bias functions b_i, so we make the simplifying assumption that the bias depends only on the controllable inputs. But we then add in a different "nugget" for each replication to accommodate possible errors:

$$b_{\boldsymbol{u}_i}(\boldsymbol{v}, t) = b(\boldsymbol{v}, t) + e_i^b(t), \tag{5}$$

and incorporate this nugget into $e_i(t)$. This leads to the model,

$$y_i^F(\boldsymbol{v}, t) = y^M(\boldsymbol{u}_i, \boldsymbol{v}, t) + b(\boldsymbol{v}, t) + e_i(t) \ . \tag{6}$$

1.3 Technical challenges

Applying the `SAVE` methodology to the thermal challenge problem faces:

1. Many unknowns. The κ's and ρ's are unknown and vary from specimen to specimen leading to a large number of unknown parameters. The temptation to merely use estimates of κ and ρ based on the MC data ignores information inherent in the EN and AC data about the individual specimen-specific values. This can have a non-trivial effect on predictions and must be dealt with.

2. Sparse design. There are only four configurations in the ensemble experimental design and a

fifth from the accreditation data. This limits the number of parameters that can be treated by `SAVE`, and requires using reasonable simplifying assumptions.

3. Multiple resolutions. In Bayarri *et al.* (2005), all outputs (computer runs and field data) are assumed to be on the same time grid, enabling computational efficiency by utilizing a Kronecker product specification for the correlation matrices of the involved Gaussian processes. In the thermal problem, the AC data are observed on a finer grid than the EN data. Instead of investing in extra computational effort, we choose an innocuous simplification by only using the AC data on the common, albeit coarser, time grid.

1.4 Organization; Conclusions

The paper is organized as follows. In Section 2 we discuss the material characterization data and use it to generate prior distributions for κ, ρ. In Section 3 we sketch the analysis and the assumptions required; details are placed in the Appendix. In Section 4, we carry out the analysis for the EN (medium- and high- level) data set, the EN + AC (medium- and high- level) data set, and assess the regulatory probabilities under each level.

Our conclusions: In neither the medium-level nor high-level case is the regulatory requirement met by the appropriate (reality) predictor: the estimates of .02 and .04 (for medium- and high-level respectively) in Section 4.3 do not meet the regulatory requirement of .01. Though the histograms (Figure 6) of the predicted temperature at the regulatory configuration are centered near 700 (well below the critical 900) there is high variability in the key material characteristics κ, ρ.

2 Material Characterization

The MC data are used to obtain posterior distributions for κ and ρ that are then used as priors in later analyses. Liu *et al.* (2006) shows the quantile-quantile plots of the normalized data for all MC data.

Though the plots suggest that κ and ρ might be assumed to be normally distributed, $\kappa \sim \mathrm{N}(\mu_\kappa, \sigma_\kappa^2)$ and $\rho \sim \mathrm{N}(\mu_\rho, \sigma_\rho^2)$, closer examination of the data for κ indicates that the assumption of constancy is not tenable; κ is, more plausibly, a linear function of temperature. But replacing the constant κ by a linear function in Equation (1) does not conform to the physics. Therefore, we only use the data at temperatures 500^oC or higher, to estimate $(\mu_\kappa, \sigma_\kappa^2, \mu_\rho, \sigma_\rho^2)$, in the hope that doing so will make $\pi(\kappa \mid MC)$ and $\pi(\rho \mid MC)$ close to the distributions under the higher temperatures of the regulatory condition.

The assumption that κ and ρ are independent is borne out by the data. The parameters of the normal distribution of κ are estimated in the traditional way as $\bar{\kappa}^{\mathrm{MC}}, \sum_i (\kappa_i^{\mathrm{MC}} - \bar{\kappa}^{\mathrm{MC}})^2/(n-1)$ and similarly for ρ. Equation (7) gives the priors $\pi(\kappa)$ and $\pi(\rho)$ for medium and high level experimental data,

$$\begin{aligned} &\pi(\kappa \mid \text{Medium MC data}) = \mathrm{N}(0.0671, 0.0070^2) \quad \pi(\kappa \mid \text{High MC data}) = \mathrm{N}(0.0687, 0.0072^2) \\ &\pi(\rho \mid \text{Medium MC data}) = \mathrm{N}(405420, 38432^2) \quad \pi(\rho \mid \text{High MC data}) = \mathrm{N}(398220, 33690^2)\,. \end{aligned} \tag{7}$$

A naive (and generally wrong) answer to the question "Does the device meet the regulatory requirement?" is to sample κ and ρ from these priors, plug them, along with the regulatory configuration, into Equation (1), and count the proportion of times the regulatory criterion is violated. In Figure 1, we show the histograms (for medium- and high-level data) of the temperatures so obtained. The proportions above 900 are 0.08 and 0.06 for the two levels respectively. An initial conclusion is that the device might not be safe for the intended application. But we are not predicting the correct quantity: the correct quantity is *reality* at the regulatory configuration, and bias, if present, must be accounted for before making conclusions.

3 Assumptions and Sketch of the Analysis

The controllable inputs defining a configuration are $\boldsymbol{v} = (v_1, v_2) \equiv (L, q)$. The field data are assumed to be without measurement error, but that is irrelevant for us since the $e(\cdot)$ term takes into account both that error, if present, and the nugget arising from our assumption about b in Equation (5).

The unknowns in the structure of Equation (6) are $(\boldsymbol{u}_i, b(\cdot,\cdot), \epsilon(\cdot))$. The Bayesian analysis proceeds by placing prior distributions on the unknowns and then produces a posterior distribution of the unknowns given the data. The posterior distribution provides all the necessary information for prediction, tolerance bounds, etc.

3.1 Prior distributions

Prior for $\boldsymbol{u}$: We use the priors given in Equation (7).

Prior for $b(\boldsymbol{v}, t)$: We use a Gaussian process (GP) as the prior distribution for the bias function $b(\boldsymbol{v}, t)$. The GP is characterized by its mean and covariance function. We take the mean of the GP as an unknown value μ_b. The covariance function of the GP is from a family whose parameters (such parameters of a prior distribution are called hyper-parameters) become part of the unknowns and are incorporated into the Bayesian analysis. Specifically,

$$b(\cdot,\cdot) \sim \mathrm{GP}\left(\mu_b, \tau^2 C\left((\cdot,\cdot),(\cdot,\cdot)\right)\right),$$

where τ^2 is the variance of the GP, and the correlation function is assumed to be $C\left(b(\boldsymbol{v}; t), b(\boldsymbol{v}'; t')\right) = c_v\left(\boldsymbol{v}, \boldsymbol{v}'\right) c_t\left(t, t'\right)$, with $c_v\left(\boldsymbol{v}, \boldsymbol{v}'\right) = \exp\left(-\sum_1^2 \beta_k |v_k - v_k'|^{\alpha_k}\right)$ and $c_t\left(t, t'\right) = \exp\left(-\beta^{(t)} |t - t'|^{\alpha^{(t)}}\right)$. Here $\boldsymbol{\alpha} = \left(\alpha_1, \alpha_2, \alpha^{(t)}\right)$ are roughness parameters associated with the smoothness of the process realizations, and $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \beta^{(t)}\right)$ are range parameters controlling the decay of the spatial correlations.

These GP priors have been consistently effective for treating computer model output since their

introduction in Sacks *et al.* (1989) and Currin *et al.* (1991). The correlation structure implies that close-by values of inputs lead to high correlations, i.e., close relation of the outputs, while far apart inputs lead to near zero correlations or lack of relation between the outputs. These are typical features of smooth functions, the kind expected to come from computer models like Equation (1).

Prior for $\epsilon_i(t)$: For the prior distribution of $\epsilon_i(\cdot)$, we use GP $\left(0, \sigma^2 c_t(\cdot,\cdot)\right)$ independent of $b(\cdot,\cdot)$ and $\epsilon_j(\cdot)(j \neq i)$. (Note that, following Bayarri *et al.* (2005), we impose the simplifying assumption that the correlation structure in the time component is the same for ϵ_i and for $b(\cdot)$.)

Priors for hyper-parameters: We will elicit the priors of the hyper-parameters μ_b, $\alpha^{(t)}$, $\boldsymbol{\beta}$, σ^2, and τ^2 in Appendix B.

3.2 Posterior distribution

The likelihood function, when combined with the prior distribution of the unknowns, leads to the posterior distribution, following Bayes Theorem. Let $(t_1, \ldots, t_n)$ be the time grid for the observations, $\boldsymbol{u}_i$ the calibration parameters for the i^{th} specimen, and $\boldsymbol{v}_i$ be the configuration associated with this specimen. Note that $\boldsymbol{u}_i$ is unique for each specimen while specimens in the same configuration may have the same $\boldsymbol{v}$ values. The likelihood is obtained from

$$\boldsymbol{y}^F \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{b}, \mu_b, \tau^2, \sigma^2 \sim \mathrm{N}\left(\boldsymbol{y}^M + \mu_b \mathbf{1}, \left(\tau^2 \boldsymbol{\Sigma}_b + \sigma^2 \mathbf{I}\right) \otimes \boldsymbol{\Sigma}_t\right) \tag{8}$$

where

$$\boldsymbol{y}^F = \left(y_i^F(\boldsymbol{v}_i, t_j), i \in \{1, \ldots, m\}, j \in \{1, \ldots, n\}\right)^t$$

$$\boldsymbol{y}^M = \left(y^M(\boldsymbol{u}_i, \boldsymbol{v}_i, t_j), i \in \{1, \ldots, m\}, j \in \{1, \ldots, n\}\right)^t \tag{9}$$

$$\boldsymbol{b} = \left(b(\boldsymbol{v}_i, t_j), i \in \{1, \ldots, m\}, j \in \{1, \ldots, n\}\right)^t \tag{10}$$

$$(\boldsymbol{\Sigma}_b)_{i,i'} = c_v(\boldsymbol{v}_i, \boldsymbol{v}_{i'})$$

$$(\boldsymbol{\Sigma}_t)_{j,j'} = c_t(t_j, t_{j'})$$

The Kronecker product operation $\otimes$ is defined in Appendix A along with some of its properties. These properties are crucial for easy evaluation of the likelihood in Equation (8).

To obtain the posterior distribution we use a modular MCMC approach in order to reduce the confounding between $\boldsymbol{b}$ and $\{\boldsymbol{u}_i\}$. First, we fix the $\boldsymbol{u}$'s at their prior means in Equation (7), then run an MCMC for the other parameters $\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{b}, \mu_b, \tau^2, \sigma^2\right)$, extract a sample from $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and, finally, run an MCMC on the other parameters, including $\boldsymbol{u}$, with $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ fixed at their posterior medians. The details are in Appendix C. This modular MCMC approach results in a sequence of 10000 draws from the posterior distribution of all unknowns given the data. Statistical inference is based on these posterior draws.

3.3 Inference

The MCMC produces a sequence of draws $\left(\{\kappa_i^h, \rho_i^h\}, \mu_b^h, \tau^{2h}, \sigma^{2h}\right)$. The vector $\boldsymbol{y}^{Mh}$ is obtained by plugging the draws $\left(\kappa_i^h, \rho_i^h\right)$ into Equation (1) and evaluating, forming a vector as in (9). Obtain a draw from the posterior distribution of the vector (10), $\boldsymbol{b}^h$, by sampling from the multivariate normal distribution with mean vector

$$(\boldsymbol{y}^F - \boldsymbol{y}^M) - \tau^2 \boldsymbol{\Sigma}_b \left(\tau^2 \boldsymbol{\Sigma}_b + \sigma^2 \mathbf{I}\right)^{-1} \otimes \mathbf{I}(\boldsymbol{y}^F - \boldsymbol{y}^M - \mu_b \mathbf{1}) \tag{11}$$

and covariance matrix

$$\left[\sigma^2(\sigma^2 \mathbf{I} + \tau^2 \boldsymbol{\Sigma}_b)^{-1} \tau^2 \boldsymbol{\Sigma}_b\right] \otimes \boldsymbol{\Sigma}_t \ , \tag{12}$$

where all the parameters involved in the calculations of the formula above are $(\{\kappa_i^h, \rho_i^h\}, \mu_b^h, \tau^{2h}, \sigma^{2h})$ and $\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}$.

We obtain 10000 such draws. With these MCMC draws, we obtain the model prediction (sometimes called pure-model prediction) by averaging $\boldsymbol{y}^{Mh}$ over h. Call the result $\widehat{\boldsymbol{y}}^M$. Because the (κ_i, ρ_i)'s are different for each replicate, the predictions will differ from replicate to replicate. Reality at a specimen in the experiment is the same as the field value, since there is no measurement error. It follows that a 95% point-wise tolerance bound at time t for the model prediction at such a specimen is a quantity $\delta^M(t)$ such that 95% of the $y^{Mh}(t)$ for this specimen satisfy $|y^{Mh}(t) - \widehat{y}^M(t)| < \delta^M(t)$.

For predicting at a new configuration, $\boldsymbol{v}_{new}$ (and, therefore, a new specimen with parameters κ_{new}, ρ_{new}), we first generate $\boldsymbol{b}^h(\boldsymbol{v}_{new}) = \left(b^h(\boldsymbol{v}_{new}, t_j), j = 1, \ldots, n\right)^T$ by drawing from the multivariate normal generated by the GP assumption on b while conditioning on the data and the draws on all parameters including $\boldsymbol{b}^h$. Then, we add $\boldsymbol{\epsilon}^h$ to $\boldsymbol{b}^h(\boldsymbol{v}_{new})$, where $\boldsymbol{\epsilon}^h \sim \mathrm{N}(\mu_b^h \mathbf{1}, \sigma^{2h}\widehat{\boldsymbol{\Sigma}}_t)$, and generate $\boldsymbol{y}^{Mh}(\kappa_{new}^h, \rho_{new}^h, \boldsymbol{v}_{new})$ by drawing $\kappa_{new}^h, \rho_{new}^h$ from their prior distributions and plugging them into Equation (1). Draws of reality, $\boldsymbol{y}^R = \boldsymbol{y}^M(\kappa_{new}, \rho_{new}, \boldsymbol{v}_{new}) + \boldsymbol{b}(\boldsymbol{v}_{new}) + \boldsymbol{\epsilon}$, can also be obtained. Letting $\boldsymbol{y}^{Rh}$ be the MCMC draws of $\boldsymbol{y}^R$, we have reality prediction $\widehat{\boldsymbol{y}}^R$, as the average of the $\boldsymbol{y}^{Rh}$ and a 95% point-wise tolerance bound can be obtained as the quantity $\delta^R(t)$ such that 95% of the h satisfy $|y^{Rh}(t) - \widehat{y}^R(t)| < \delta^R(t)$.

4 Results

4.1 Ensemble Analysis

For the EN data, summaries of the posterior distributions of the parameters $\boldsymbol{\theta} = \left(\mu_b, \boldsymbol{\alpha}, \boldsymbol{\beta}, \tau^2, \sigma^2\right)$ based on the analysis sketched above and detailed in the Appendix are given in Table 2. We fix (α_1, α_2) at 2 to reflect belief in the smoothness of the outputs as functions of L and q; this also helps computationally by reducing the number of unknowns.

Figure 2 shows the marginal posterior distributions of $\boldsymbol{\kappa}$ and those of $\boldsymbol{\rho}$ under the high level EN data on the left and right panel, respectively. The priors correspond to the solid lines. Clearly, each tested specimen should be associated with its own posterior. The posteriors of $\boldsymbol{\kappa}$, in particular, differ considerably as configuration settings vary. For instance, the posteriors at $q = 1000$ (top two panels) center at larger values than those at $q = 2000$ (bottom two panels). With the same q, the posteriors at $L = 0.0127$ have smaller values of variance than those at $L = 0.0254$. We observe less variability for the posteriors of $\boldsymbol{\rho}$ across configuration settings.

The pure-model prediction and the bias function for each of the replicates and each configuration can be calculated as in Section 3.3. To illustrate, the upper left panel of Figure 3 shows the bias prediction (solid black) with 95% uncertainty bounds for the first replicate of the high level EN data at configuration $L = 0.0127, q = 1000$. The lower left panel has the model prediction for the same setting with 95% tolerance bounds; the red curve plots the experimental data. For the same configuration and a new specimen, with parameters $(\kappa_{new}, \rho_{new})$ the upper right panel gives the results for the bias and the lower right panel gives the reality prediction (not the model prediction) as the solid black line with dashed lines as tolerance bounds; the red curves are the experimental data for the four replicates.

For predicting (extrapolating) at the AC configuration $L = 0.019, q = 3000$, a new pair $(\kappa_{new}, \rho_{new})$ is also involved. We can use the same $\boldsymbol{\theta}^h$'s found above but draw κ^h, ρ^h from their prior distribution. In addition, we must draw $\boldsymbol{\epsilon}^h$ from its distribution (see Section 4.1) and draw from the distributions of $\boldsymbol{b}(0, 0.019, 3000)$ given the four values of $\boldsymbol{b}^h$ at the EN configuration. Everything else is done as above and produces the bias function in Figure 4, the model prediction in the left panel of Figure 5, and reality prediction for the AC configuration in the right panel.

4.2 Accreditation Analysis

With the addition of the accreditation data, a reprise of the analysis of Section 4.1 is summarized as follows:

- The posterior distribution of unknown parameters is shown in Table 2.

- Figure 7, the counterpart to Figure 3, displays model prediction and bias prediction for the first AC replicate as well as the reality prediction and bias prediction for a new specimen with parameters κ_{new}, ρ_{new}.

Each column in Table 2 corresponds to the estimates of the hyper-parameters given the indicated data set. These estimates differ considerably, except for those associated with time, $\alpha^{(t)}$ and $\beta^{(t)}$. In large part, this is due to the additional design point of the accreditation configuration at some distance from the sparse design of the EN data, producing substantially more variability. The higher variability in the high-level data compared to the medium-level data is reflected in the differing parameter values corresponding to the two levels.

Bias appears negligible in the EN analysis but emerges as a non-trivial matter in the EN+AC analysis, an indication that the model may be less accurate at higher temperatures, which is a fact of considerable relevance for predicting at the regulatory condition. Moreover, the presence of bias in the EN+AC analysis has a strong effect on κ, ρ, and this is reflected in the considerable difference between the pure model predictions of Figure 5 and Figure 7.

4.3 The regulatory assessment

Let y_R^M, b_R, y_R^R be the model prediction, bias function and reality prediction, respectively, under the regulatory configuration $L_R = 0.019$, $q_R = 3500$ at time $t_R = 1000$. We get posterior draws for y_R^M, b_R, y_R^R following the prescription in Section 3.3.

Figure 6 gives the posterior histograms of the reality prediction of the device surface temperature under the regulatory configuration at time 1000 given the EN+AC data at both medium and high levels. The distributions are summarized in Table 3.

The proportion of values that exceed 900 is the estimate of the probability that the regulatory

requirement is unmet. For the medium level this number is 0.02; for high-level data the number is 0.04. Compared with the pure model predictions discussed in Section 2, the chance of failure is decreased but the requirement of 0.01 is still not met.

5 Discussion and Summary

The formulation of the problem and the process described above provides complete answer to the question of how to assess a computer model. In particular, for the thermal problem it is clear from Figure 8 that bias is present so that the model is not fully reliable. By producing "legitimate" estimates of reality (the reality predictions described above) the ability to use the model is enhanced. In the case at hand, there may be too much variability in the material characterizations to overcome and assure the necessary certainty for regulatory compliance. The assumption of constancy for the material properties is, likely, the key flaw.

Acknowledgement

This research was supported in part by the National Science Foundation Grant DMS-0103265 and Spanish Ministry of Science and Technology Grant MTM2004-03290. Any opinions, findings and conclusions or recommendations expressed in this publication are entirely those of the authors and do not necessarily reflect the views of the National Science Foundation nor the Spanish Ministry of Science and Technology.

A Kronecker product

The Kronecker product of two matrices $A = (a_{ij})_{i,j}$ and $B = (b_{ij})_{i,j}$ is defined as,

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

It has the following properties.

1. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ if A and B are both invertible.

2. $|A \otimes B| = |A|^{d_2}|B|^{d_1}$, where d_1 and d_2 are the dimensions of A, B.

3. $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ if the dimensions are matched.

   If we assume $A = U_1 U_1^T$, $B = U_2 U_2^T$, then $(A \otimes B) = (U_1 \otimes U_2)(U_1 \otimes U_2)'$.

B Prior distributions for the hyper-parameters

Data limitations and the belief that the responses are smooth functions of input lead us to fix the α_i's at 2 (but not $\alpha^{(t)}$). We use the standard noninformative prior for μ_b, $\pi(\mu_b) \propto 1$, and specify prior distributions for the other parameters as:

$$\pi(\sigma^2) \propto 1/\sigma^2, \qquad \pi(\tau^2) \propto \exp(-1000/\tau^2),$$

$$\pi(\beta_1) \propto \exp(-0.001\beta_1), \qquad \pi(\beta_2) \propto \exp(-10^5\beta_2),$$

$$\pi(\alpha^{(t)}) \propto I_{(1,2)}(\alpha^{(t)}), \qquad \pi(\beta^{(t)}) \propto \exp(-100\beta^{(t)}) I_{[10^{-4},\infty)}(\beta^{(t)}).$$

where $I_A(x)$ equals one if $x \in A$ and zero otherwise. The priors $\pi(\sigma^2)$ and $\pi(\alpha^{(t)})$ are the standard objective priors. The prior for τ^2 has to be informative due to the limitations of the data. The choice of

1000 reflects belief about the scale of the bias. Because the data are scarce the other hyper-parameters are given informative priors with scales matching the scales of the variables. For instance, $\pi(\beta_1)$ is chosen to reflect the fact that the scale of L^2 is about 0.001. The prior on $\beta^{(t)}$ is truncated to guarantee the non-singularity of Σ_t and avoid numerical issues.

C Modular MCMC approach

We first describe how to estimate $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$. Approximate the computer model output $y^M(\boldsymbol{u}_i, \boldsymbol{v}, t)$ by $\widehat{y}^M = y^M(\widehat{\boldsymbol{u}}, \boldsymbol{v}, t)$, where $\widehat{\boldsymbol{u}}$ is the vector of the prior means (nominal values) for $\boldsymbol{u}$. Then approximate the `SAVE` formula in Equation (6) as,

$$y_i^F(\boldsymbol{v}, t) \approx \widehat{y}^M(\widehat{\boldsymbol{u}}, \boldsymbol{v}, t) + b(\boldsymbol{v}, t) + e_i(t). \tag{13}$$

Use a MCMC algorithm to draw from $\pi\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mu_b, \sigma^2, \tau^2 \mid \boldsymbol{y}^F, \widehat{\boldsymbol{y}}^M\right)$. At the end of the MCMC, $\{\boldsymbol{\alpha}^h, \boldsymbol{\beta}^h, h = 1, \ldots, 10000\}$ are produced. We take the posterior medians of the MCMC draws as our estimates $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$. Details of the MCMC can be found in Liu *et al.* (2006).

The modular approach depends on samples from the posterior distribution

$$\pi\left(\boldsymbol{b}, \{\boldsymbol{u}_i\}, \mu_b, \tau^2, \sigma^2, \mid \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \boldsymbol{y}^F\right).$$

These are obtained by iteratively sampling (Gibbs sampler) from $\pi\left(\boldsymbol{b}, \{\boldsymbol{u}_i\}, \mu_b \mid \tau^2, \sigma^2, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \boldsymbol{y}^F\right)$ and $\pi\left(\sigma^2, \tau^2 \mid \boldsymbol{b}, \{\boldsymbol{u}_i\}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \boldsymbol{y}^F\right)$. The first term is dealt with as follows:

Write

$$\begin{aligned}\pi(\{\boldsymbol{b}, \{\boldsymbol{u}_i\}, \mu_b \mid \tau^2, \sigma^2, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \boldsymbol{y}^F) = \pi(\{\boldsymbol{b}\} \mid \{\boldsymbol{u}_i\}, \mu_b, \tau^2, \sigma^2, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \boldsymbol{y}^F) \\ \times \pi(\mu_b, \{\boldsymbol{u}_i\} \mid \tau^2, \sigma^2, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \boldsymbol{y}^F).\end{aligned} \tag{14}$$

The first factor in Equation (14) can be sampled from the multivariate normal distribution, with mean vector and covariance matrix given, respectively, by

$$(\boldsymbol{y}^F - \boldsymbol{y}^M) - \tau^2\boldsymbol{\Sigma}_b\left(\tau^2\boldsymbol{\Sigma}_b + \sigma^2\mathbf{I}\right)^{-1} \otimes \mathbf{I}(\boldsymbol{y}^F - \boldsymbol{y}^M - \mu_b\mathbf{1}), \quad \left(\sigma^2(\sigma^2\mathbf{I} + \tau^2\boldsymbol{\Sigma}_b)^{-1}\tau^2\boldsymbol{\Sigma}_b\right) \otimes \boldsymbol{\Sigma}_t\,.$$

Samples from the second factor in Equation (14) are obtained by another Gibbs sampler by iteratively sampling from $\pi(\mu_b \mid \{\boldsymbol{u}_i\}, \tau^2, \sigma^2, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \boldsymbol{y}^F)$, which is

$$\mathrm{N}\left(\frac{\mathbf{1}^t\left(\boldsymbol{y}^F - \boldsymbol{y}^M\right)}{\mathbf{1}^t\left(\tau^2\boldsymbol{\Sigma}_b + \sigma^2\mathbf{I}\right)^{-1} \otimes \left(\boldsymbol{\Sigma}_t\right)^{-1}\mathbf{1}}, \frac{1}{\mathbf{1}^t\left(\tau^2\boldsymbol{\Sigma}_b + \sigma^2\mathbf{I}\right)^{-1} \otimes \left(\boldsymbol{\Sigma}_t\right)^{-1}\mathbf{1}}\right), \tag{15}$$

and using a Metropolis-Hastings algorithm to sample from $\pi(\{\boldsymbol{u}_i\} \mid \mu_b, \tau^2, \sigma^2, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \text{Data})$. The Metropolis-Hastings algorithm usually results in highly correlated samples. Therefore, within each iteration of the Gibbs loop, we run the Metropolis-Hastings algorithm 200 times, and keep the last one as our sample.The details of the Metropolis-Hastings algorithm are in Liu *et al.* (2006).

To deal with the second term, $\pi\left(\sigma^2, \tau^2 \mid \boldsymbol{b}, \{\boldsymbol{u}_i\}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \boldsymbol{y}^F\right)$, we note that, conditional on $\{\boldsymbol{b}\}, \{\boldsymbol{u}_i\}, \mu_b$, $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$, the posteriors of τ^2 and σ^2 are independent, with

$$\begin{aligned}
\left(\tau^2 \mid \boldsymbol{b}, \{\boldsymbol{u}_i\}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \boldsymbol{y}^F\right) &\sim \mathrm{IG}\left(\frac{n}{2}\mathrm{rank}(\boldsymbol{\Sigma}_b) - 1, \frac{1}{2}(\boldsymbol{b} - \mu_b\mathbf{1})^t\left(\boldsymbol{\Sigma}_b \otimes \boldsymbol{\Sigma}_t\right)^{-1}(\boldsymbol{b} - \mu_b\mathbf{1}) + 1000\right), \\
\left(\sigma^2 \mid \boldsymbol{b}, \{\boldsymbol{u}_i\}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \boldsymbol{y}^F\right) &\sim \mathrm{IG}\left(\frac{mn}{2}, \frac{1}{2}\left(\boldsymbol{y}^F - \boldsymbol{y}^M - \boldsymbol{b}\right)^t\left(\mathbf{I} \otimes \boldsymbol{\Sigma}_t\right)^{-1}\left(\boldsymbol{y}^F - \boldsymbol{y}^M - \boldsymbol{b}\right)\right).
\end{aligned}$$

and drawing from the Inverse Gamma distributions is straightforward.

The above Gibbs sampler was run for 10000 iterations (within each iteration, we run 200 iterations for the Metropolis-Hastings algorithm), and standard diagnostic measures were used to check for convergence. These MCMC draws from the Modular MCMC algorithm, together with the draws for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ are summarized in Table 2 for the EN data and in Table 2 for the AC data.

D Extrapolation of the bias

We can extrapolate the bias to a new configuration $\boldsymbol{v}_{new}$. Denote the bias at new configuration by $\boldsymbol{b}(\boldsymbol{v}_{new}) = (b(\boldsymbol{v}_{new}, t_j), j \in (1, \ldots, n))^T$, and the biases for all the field data configurations by $\boldsymbol{b} = \{\boldsymbol{b}_i\}$. The posterior distribution for $\boldsymbol{b}(\boldsymbol{v}_{new})$ is,

$$\pi(\boldsymbol{b}(\boldsymbol{v}_{new}) \mid \boldsymbol{y}^F) = \int \pi(\boldsymbol{b}(\boldsymbol{v}_{new}) \mid \boldsymbol{b}, \boldsymbol{\theta}) \pi(\boldsymbol{b}, \boldsymbol{\theta} \mid \boldsymbol{y}^F) \mathrm{d}\boldsymbol{b} \mathrm{d}\boldsymbol{\theta}.$$

We can make draws from this distribution as follows. At iteration h, in the MCMC described in Appendix C, we have draws from $\boldsymbol{b}, \boldsymbol{\theta} \mid \boldsymbol{y}^F$. It then suffices to draw $\boldsymbol{b}^h(\boldsymbol{v}_{new})$ from $\pi\left(\boldsymbol{b}(\boldsymbol{v}_{new}) \mid \boldsymbol{b}^h, \boldsymbol{\theta}^{(h)}\right)$, which is a normal distribution with mean and covariance given , respectively, by

$$\mu_b \mathbf{1} + \mathbf{c}^t \left(\boldsymbol{\Sigma}_b \otimes \boldsymbol{\Sigma}_t\right)^- (\boldsymbol{b} - \mu_b \mathbf{1}) \quad \text{and} \quad \tau^2 \left(1 - \mathbf{c}^t \boldsymbol{\Sigma}_b^{-1} \mathbf{c}\right) \boldsymbol{\Sigma}_t,$$

where $\boldsymbol{c} = (c_v(\boldsymbol{v}_1, \boldsymbol{v}_{new}), \ldots, c_v(\boldsymbol{v}_m, \boldsymbol{v}_{new}))^T$ is the correlation vector between the new configuration and the experimented configurations.

References

Bayarri, M., Berger, J., Kennedy, M., Kottas, A., Paulo, R., Sacks, J., Cafeo, J., Lin, C., and Tu, J. (2005). Bayesian validation of a computer model for vehicle crashworthiness. Tech. rep., National Institute of Statistical Sciences, http://www.niss.org/technicalreports/tr163.pdf.

Bayarri, M., Berger, J., Paulo, R., Sacks, J., Cafeo, J., Cavendish, J., Lin, C., and Tu, J. (2007). A framework for validation of computer models. *Technometrics* **49**, 2, 138–154.

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991). Bayesian prediction of deterministic

functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association* **86**, 953–963.

Dowding, K., Pilch, M., and Hills, R. (2007). Formulation of the thermal problem. *CMAME special issue* .

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society B* **63**, 425–464.

Liu, F., Bayarri, M. J., Berger, J. O., Paulo, R., and Sacks, J. (2006). A bayesian analysis of the thermal challenge problem. Tech. rep., National Institute of Statistical Sciences, http://www.niss.org/technicalreports/tr166.pdf.

Robert, C. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag, New York.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments (C/R: p423-435). *Statistical Science* **4**, 409–423.

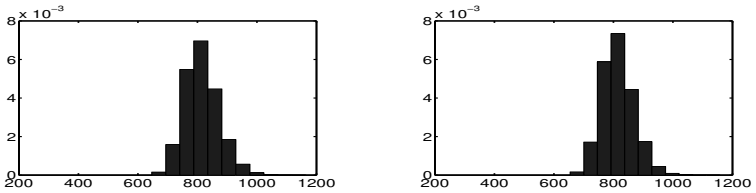

Figure 1: Computer model predictions for surface temperature at the regulatory configuration based on medium (left)- and high (right)- level MC data

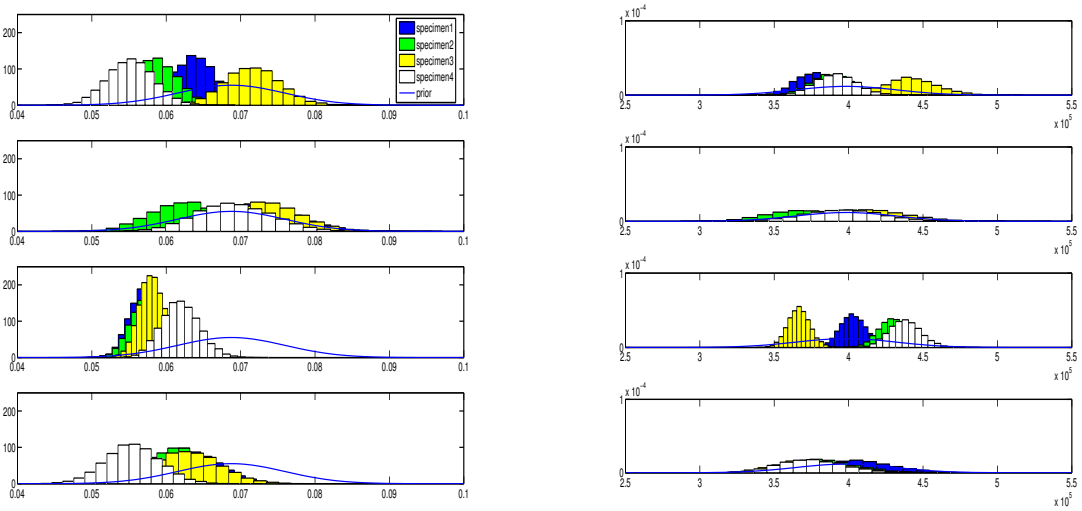

Figure 2: Histograms of the κ's (Left) and ρ's (Right) given EN data (high level) with the configurations (from up to bottom): $L = 0.0127, q = 1000$, $L = 0.0254, q = 1000$, $L = 0.0127, q = 2000$, $L = 0.0254, q = 2000$. The color represents the specimen as indicated. Priors are plotted in blue lines in the panels.

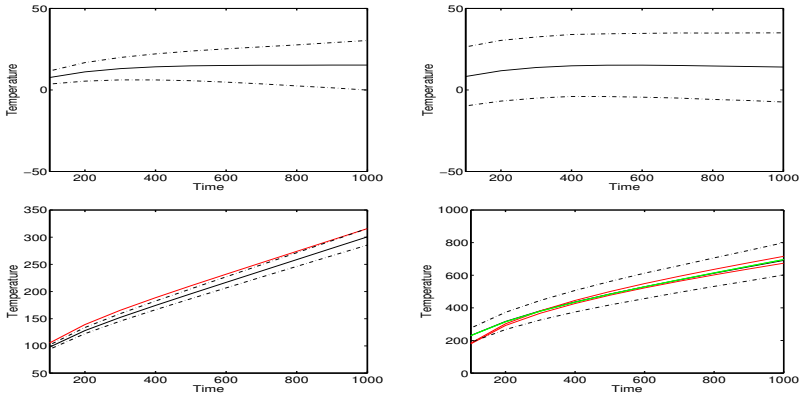

Figure 3: Upper-left: bias function for first run in first ensemble configuration ($L = 0.0127, q = 1000$). Upper-right: bias function for a new specimen at this configuration. Lower-left: model prediction for first run of this configuration. Lower-right: reality prediction for a new specimen at this configuration. Observations are plotted in red, posterior medians as solid black lines, and the 2.5% and 97.5% pointwise uncertainty bounds as dashed black lines. The results are obtained conditional on the high-level EN data.

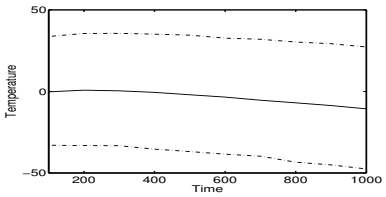

Figure 4: Bias function at accreditation configuration ($L = 0.019, q = 3000$) given high-level EN data.

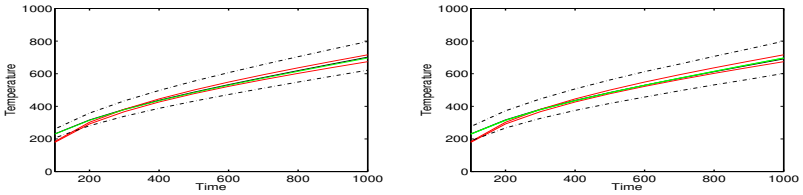

Figure 5: Pure model prediction (left) and Reality prediction (right) at accreditation configuration given high-level EN data. Tolerance bounds are dashed lines, experimental data are in red and the green line is the prediction by plugging the prior means of κ and ρ into Equation (1).

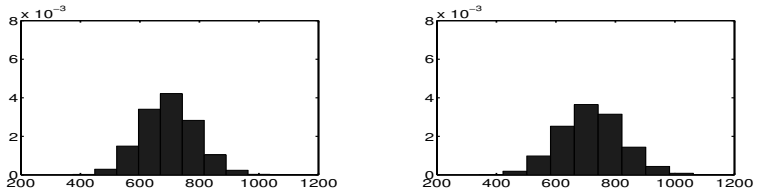

Figure 6: Histograms of the device surface temperature under regulatory configuration with medium (left)- and high (right)- level AC + EN data.

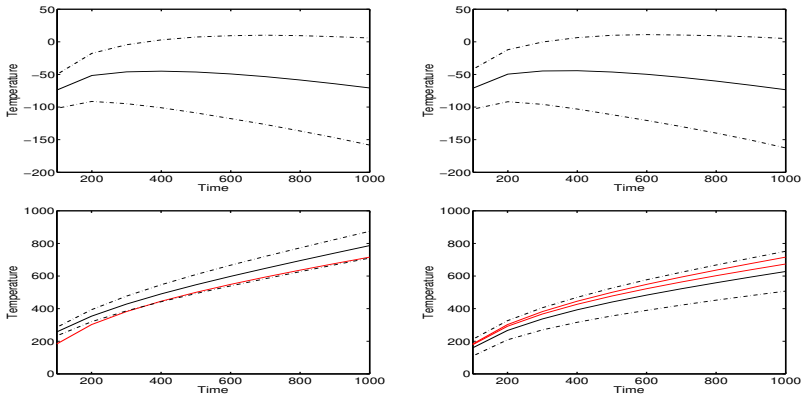

Figure 7: Bias function (upper-left) for the first run in the accreditation configuration ($L = 0.019, q = 3000$); bias function (upper-right) for a new run at this configuration; pure model prediction (lower-left) for the first run in this configuration; and reality prediction (lower-right) for a new run at this configuration. Red lines are the experimental data. The results are obtained conditional on the high-level EN+AC data.

Input	Impact	Uncertainty	Current status
κ	5	$\pi(\kappa)$	unknown
ρ	5	$\pi(\rho)$	unknown
q	5	None	1000, 2000, 3000
L	5	None	0.0127, 0.019, 0.0254
x	1	None	0
T_0	1	None	25
t	5	None	0, 50, 100, ..., 1000

Table 1: the Input/Uncertainty Map for thermal computer model. $\pi(\kappa), \pi(\rho)$ are given in section 2.

Parameter	Medium EN	High EN	Medium EN+AC	High EN+AC
β_1	549.96 $(47.67, 2314.27)$	851.41 $(227.30, 2660.50)$	11.93 $(0.92, 45.52)$	17.42 $(4.22, 53.20)$
β_2	1.80×10^{-7} $(0.37, 6.21) \times 10^{-7}$	3.25×10^{-7} $(1.23, 8.91) \times 10^{-7}$	1.30×10^{-6} $(0.62, 3.68) \times 10^{-6}$	1.19×10^{-6} $(0.59, 3.43) \times 10^{-6}$
$\alpha^{(t)}$	1.9987 $(1.9969, 1.9995)$	1.9989 $(1.9980, 1.9994)$	1.9967 $(1.9918, 1.9988)$	1.9983 $(1.9970, 1.9989)$
$\beta^{(t)}$	1.33×10^{-3} $(1.09, 1.63) \times 10^{-3}$	1.02×10^{-3} $(0.89, 1.15) \times 10^{-3}$	1.33×10^{-3} $(1.09, 1.71) \times 10^{-3}$	1.05×10^{-3} $(0.94, 1.17) \times 10^{-3}$
μ_b	-0.57 $(-23.10, 23.38)$	-11.56 $(-41.52, 18.15)$	-62.92 $(-140.65, 18.56)$	-95.25 $(-245.31, 52.60)$
τ^2	269.72 $(148.86, 512.02)$	473.89 $(251.25, 906.29)$	6500.37 $(4121.96, 10786.89)$	22423.07 $(14239.77, 36243.67)$
σ^2	22.59 $(15.10, 35.53)$	66.02 $(50.97, 87.30)$	9.84 $(6.50, 15.96)$	43.97 $(34.12, 57.82)$

Table 2: Posterior medians with 95% credible intervals for the indicated parameters given the indicated data set.

Value	Medium Level	High Level
Mean	697.13	719.20
Median	695.93	717.90
Standard deviation	91.52	105.32

Table 3: Summary of the device surface temperature under regulatory configuration given the indicated data set.