**Proceedings of IDETC/CIE 2006**
**ASME 2006 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference**
**September 10-13, 2006, Philadelphia, Pennsylvania, USA**

# IDETC2006-99599

# SOME METRICS AND A BAYESIAN PROCEDURE FOR VALIDATING PREDICTIVE MODELS IN ENGINEERING DESIGN

**Wei Chen**[1], **Ying Xiong**
Department of Mechanical Engineering
Northwestern University

**Kwok-Leung Tsui, Shuchun Wang**
School of Industrial & Systems Engineering
Geogia Institute of Technology

## ABSTRACT

Even though model-based simulations are widely used in engineering design, it remains a challenge to validate models and assess the risks and uncertainties associated with the use of predictive models for design decision making. In most of the existing work, model validation is viewed as verifying the model accuracy, measured by the agreement between computational and experimental results. However, from the design perspective, a good model is considered as the one that can provide the discrimination (good resolution) between design candidates. In this work, a Bayesian approach is presented to assess the uncertainty in model prediction by combining data from both physical experiments and the computer model. Based on the uncertainty quantification of model prediction, some design-oriented model validation metrics are further developed to guide designers for achieving high confidence of using predictive models in making a specific design decision. We demonstrate that the Bayesian approach provides a flexible framework for drawing inferences for predictions in the intended but may be untested design domain, where design settings of physical experiments and the computer model may or may not overlap. The implications of the proposed validation metrics are studied, and their potential roles in a model validation procedure are highlighted.

## KEYWORDS

Model validation, Bayesian approach, Predictive modeling, Uncertainty quantification, Validation metrics, Design

## NOMENCLATURE

| | |
|---|---|
| $Y^e(\mathbf{x})$ | physical experimental observation |
| $\varepsilon(\mathbf{x})$ | experimental error, assumed as Normal |
| $Y^r(\mathbf{x})$ | true response outcome |
| $Y^m(\mathbf{x})$ | outcome of computer model |
| $\delta(\mathbf{x})$ | the bias (or error) of computer model |
| $f(\mathbf{x})$ | design objective function |
| $\mathbf{x}$ | $\mathbf{x}=(x_1,\cdots,x_p)^T$, design in a $p$-dimensional space |
| $D_e$ | $D_e=\{\mathbf{x}_1,\cdots,\mathbf{x}_{n_e}\}$, for physical experiments |
| $n_e$ | size of $D_e$, the number of physical experiments |
| $D_m$ | $D_m=\{\mathbf{x}'_1,\cdots,\mathbf{x}'_{n_m}\}$, for computer experiments |
| $n_m$ | size of $D_m$, the number of computer experiments |
| $\mathbf{y}^m$ | $\mathbf{y}^m=(y^m(\mathbf{x}'_1),\cdots,y^m(\mathbf{x}'_{n_m}))^T$, the deterministic model outputs |
| $\sigma_\varepsilon^2$ | variance parameter of $\varepsilon(\mathbf{x})$ |
| $\sigma_m^2$, $\sigma_\delta^2$ | variance parameter of the prior Gaussian process $Y^m(\mathbf{x})$ and $\delta(\mathbf{x})$ |
| $\phi_m, \phi_\delta$ | correlation parameter of the prior Gaussian process $Y^m(\mathbf{x})$ and $\delta(\mathbf{x})$ |
| $\tau$ | ratio of $\sigma_\varepsilon^2$ to $\sigma_\delta^2$ |
| $n_{\delta\|e,m}, n_{m\|m}$ | degree of freedom of $t$ distribution |
| $\mu_{\delta\|e,m}(\mathbf{x}), \mu_{m\|m}(\mathbf{x})$ | noncentrality parameter of $t$ distribution |
| $\sigma^2_{\delta\|e,m}(\mathbf{x}), \sigma^2_{m\|m}(\mathbf{x})$ | scale parameter of $t$ distribution |
| $P_{ij}$ | probability for pair-wise comparison |
| $M_D^{Multip}(\mathbf{x}_i)$ | 'multiplicative' design validation metric |
| $M_D^{Average}(\mathbf{x}_i)$ | 'average (additive)' design validation metric |
| $M_D^{Worstcase}(\mathbf{x}_i)$ | 'worst-case' design validation metric |
| $M_D$ | general term of design validation metric |
| $k$ | number of design candidates space |

## 1 INTRODUCTION

With rapid increase of computational capability, modeling and simulation based design has been increasingly used for designing new engineering systems. However, it remains a challenge on assessing the risks and uncertainties associated with the use of predictive models in engineering design. Even

[1] Corresponding author. Department of Mechanical Engineering, Northwestern University, 2145 Sheridan Road, Tech B224, Evanston, IL 60208-3111, Email: weichen@northwestern.edu, Phone: (847) 491-7019, Fax: (847) 491-3915.

though there is growing interest from both government and industries in developing fundamental concepts and terminology for model validation (DoD; Ang et al. 1996, Doebling, et al. 2002; Oberkampf et al, 2003; Cafeo and Thacker 2004; Gu and Yang, 2003), model validity and model validation are poorly understood in engineering design. In most of the existing work, validation is viewed as verifying the model accuracy, i.e., a measure of the agreement between computational results and experimental results. Model validation has been primarily carried out from the perspective of model builders (or analysts) but not from that of designers (model users).

Model validation in practice mirrors the status of its limited development in research. In industry, product design has become a systems engineering activity that involves the integration of various analysis models, often owned by different disciplines or even different vendors. In current practice, validation is restricted to providing maturity scores by individual model builders through physical tests. Often these scores are obtained based on a very limited number of tests without considering the potential design space from the system perspective and the various sources of uncertainties. In summary, the existing approaches for validating analysis models cannot be directly used for validating design models in engineering decision making.

In the engineering design research community, special attentions have been given to how models and information are used in design decision making (McAdams and Dym, 2004). Preliminary efforts have been made on characterizing and assessing the validity of behavior models and their predictions in design (Malak and Paredis, 2004). Hazelrigg (2003) is the first one to have brought up the notion that the validation of a predictive model can be accomplished only in the context of a specific decision, and only in the context of subjective input from the decision maker, including preferences. As noted by Hazelrigg (2003), what really matters to designers is whether a model generates design choices whose real outcomes are better than other design choices. The concept is illustrated in Fig. 1.1. Both design alternatives A and B have prediction uncertainty associated with their outcomes. For making the right design choice (right means that the real outcome of the selected choice is better than those of the others), a good model is the one that can provide the discrimination (good resolution) between the two alternatives, e.g., $f(\mathbf{x}_A)$ and $f(\mathbf{x}_B)$, where $f(\mathbf{x}_A)$ and $f(\mathbf{x}_B)$ stand for the design objective function of alternatives $\mathbf{x}_A$ and $\mathbf{x}_B$, respectively. From the probabilistic point of view, to identify the model validity, it is important to have the capability of assessing the probability $P_{AB}$ of design alternative i to produce an outcome that is preferred to or indifferent to another alternative j, i.e., $P_{AB} = \mathrm{P}\left( f(\mathbf{x}_A) < (\mathbf{x}_B) \right)$, assuming smaller-the-better scenario.

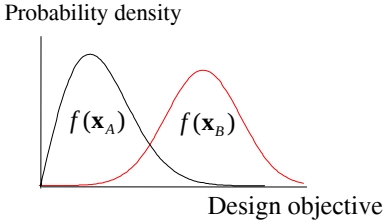


Figure 1.1 Design resolution

Here we differentiate a design objective $f(\mathbf{x})$ from $y(\mathbf{x})$, which stands for a single or multiple responses from computer model(s). To quantify the uncertainty of $f(\mathbf{x})$, tatistical inference techniques must be developed to quantify the uncertainty associated with the prediction of $y^r(\mathbf{x})$ based on the results from both models and physical experiments. As experiments are seldom available for a new design, this requires merging model and test data from a variety of single and multiple phenomena into an inference about prediction at the intended design. This is the "inference bridge" in Fig. 1.2. The greater the distance, the larger the prediction uncertainty normally is.

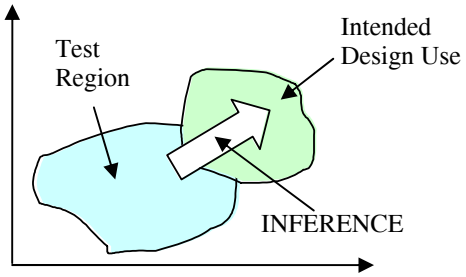


Figure 1.2 Inferring predictive capability

Although the need for validating models from the perspective of engineering design has been brought up in the existing model validation work, few have developed quantitative means to define and to assess model validity for specific decisions. In author's earlier work, an approach was developed to provide stochastic assessment of the validity of a model (Chen et al. 2004; Buranathiti et al. 2004). However, the approach is more useful for rejecting (invalidating) a model rather than accepting (validating) a model. In recent work of Mahadevan and Rebba 2005, a Bayes network approach is proposed for validating the reliability assessment made by computational models in design. Validation is treated as a hypothesis testing problem, with which prediction uncertainty cannot be quantified. Again, the emphasis in on validating the modeling accuracy at tested design points, but not in the context of a new design. In order to accept a design solution with good confidence, *a design validation metric needs to be developed to provide a confidence measure of a candidate design being better than other design choices.*

In this paper, we present a model validation approach that provides quantitative assessments of uncertainty in using predictive models in engineering design and further develop some validation metrics that guide designers for achieving high confidence of using predictive models in design decision making. A Bayesian procedure is presented to combine the data from physical experiments and computer models for predictive modeling. The Bayesian approach provides a framework for drawing inferences for predictions in the intended but untested design domain. The approach is generic enough to handle cases where design settings of physical experiments and the computer model may or may not overlap. When limited amount of physical data is available, the approach is capable of taking into account scientific knowledge and past information

in the form of prior distributions of model parameters. With the obtained uncertainty quantification of prediction of $y^r(\mathbf{x})$ and thus the uncertainty quantification of design objective $f(\mathbf{x})$, we further develop validation metrics to provide confidence measures of accepting a candidate design solution. The implications of using such validation metrics are examined.

## 2. BACKGROUND AND GENERAL APPROACHES

### 2.1 Uncertainties in Model Prediction and the Mathematical Framework

Predicting the amount by which a model output may differ from the true value is often complicated by the presence of uncertainties and errors from various sources, such as model (lack of knowledge), parametric, algorithmic, computational, and system variability, as well as testing data that are used to compare with the model prediction. Different ways of classifying uncertainties in model prediction are seen in the literature (Apostolakis 1994; Trucano, 1998; Hazelrigg, 1999; Oberkampf et al., 1999). Using *x* to represent design variables and *y* stand for model response, the relationship between the experimental observations $Y^e(\mathbf{x})$, the true outcome $Y^r(\mathbf{x})$, and the prediction generated by a computer model $Y^m(x)$ can often be generalized as follows:

$$Y^e(\mathbf{x}) = Y^r(\mathbf{x}) + \varepsilon(\mathbf{x}) = Y^m(\mathbf{x}) + \delta(\mathbf{x}) + \varepsilon(\mathbf{x}), \quad (2.1)$$

where $\varepsilon(\mathbf{x})$ is the random variable representing the experimental error (relating to both experimental setup and measurement) that may depend on $\mathbf{x}$, and $\delta(\mathbf{x})$ is the error of the model , or called the prediction bias, i.e.,

$$\delta(\mathbf{x}) = Y^r(\mathbf{x}) - Y^m(\mathbf{x}), \quad (2.2)$$

which captures the model inadequacy.

For the purpose of verifying model accuracy, it is essential to estimate the prediction bias $\delta(\mathbf{x})$ and characterize its uncertainty. If the emphasis is on comparing the outcomes of different design candidates, it is then important to estimate the true model output $Y^r(\mathbf{x})$ and characterize its uncertainty. Based on Eqn. (2.2), it is noted that estimating the prediction bias $\delta(\mathbf{x})$ is an intermediate step for estimating the true model output $Y^r(\mathbf{x})$. Statistical approaches for characterizing the probability distributions of these quantities are generally divided into two categories, *classical statistical* (Easterling and Berger, 2002) and *Bayesian* (Bayarri et al., 2002) approaches. The fundamental difference between the two is that the former draws confidence intervals of prediction based on statistical data analysis, while the latter assumes that the model parameters themselves are random and follow a prior distribution, specified based on model builder/designers' prior knowledge. The prior distribution will be updated once data is available and becomes posterior distribution. The Bayesian approach is preferred to the classical statistical approach when it is too expensive to obtain statistically sufficient amount of data.

### 2.2 General Model Validation Approaches

The need for relating model validation to the intended design use was brought up in the AIAA Guide for the Verification and Validation of Computational Fluid Dynamics Simulations (1998), where model validation is defined as "a 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*". However, existing validation metrics are mostly associated with the measures of model accuracy based on limited tested points. Many of existing approaches cannot provide stochastic measurements with regard to the confidence in using a model. For instance, graphical comparisons through visual inspection of x-y plots, scatter plots and contour plots are often subjective and not sufficient (Oberkampf and Trucano, 2000). Quantitative comparisons (Marczyk et al. 1997) that rely on the measures of correlation coefficient and other weighted and non-weighted norms to quantify the distance between the two "clouds" cannot provide statistical judgment of model validity. Various statistical inference techniques, such as $\chi^2$ (Chi-square) test on residuals between model and experimental results (Freese, 1960; Reynolds, 1984; Gregoire and Reynolds, 1988) require multiple evaluations of the model and experiments, and many statistical assumptions that are difficult to satisfy. In the area of Department of Energy applications, examples of statistical analysis of physics models and experiments are given in Hills and Trucano (1999) and Easterling and Berger (2002).

An extensive discussion of validation literature is given by Oberkampf and Trucano (2000). Recent approaches for quantitatively comparing computations and experiments can be divided into two categories, namely classical frequentist approach (Oberkampf and Barone, 2004) and Bayesian approach (Kennedy and O'Hagan, 2001; Bayarri et al., 2002; Buslik, 1994; Hanson, 1999; Wang, et al., 2006). Easterling and Berger 2002 provide an extensive review on classical statistical approaches for model validation and a simple case study. A review of Bayesian approaches can be found in Bayarri et al. (2002).

Oberkampf and Barone (2004) proposed a frequentist approach to the comparison of computer outputs and physical observations. They first fitted a nonlinear regression model to the physical data and then evaluated a validation metric based the differences between computer outputs and the fitted curve to measure the agreement between computations and experiments. Their approach has several limitations. First, the function form chosen for the nonlinear regression model has a large impact on the results obtained. A complicated nonlinear model may require a large amount of data to have a good fit. In reality, only few physical observations are often available. Second, the calculation of confidence intervals is rather complicated with a nonlinear model and often requires approximations. Third, their approach treats computer outputs and physical observations separately in the sense that the computer outputs play no roles in fitting the regression model based on the physical data. Last, with their approach, it is not clear how to improve or remedy a predictive model when the validation metric suggests a large disagreement between computations and experiments. Even though the idea of extending model validation to untested design sites/regions was presented, Oberkampf and Barone's work focuses on validating the pure accuracy of models, but not on the validity of using a model for making a specific design decision.

On the contrary, the Bayesian approach (e.g., Kennedy and O'Hagan, 2001; Wang et al., 2006) integrates computer outputs and physical observations together to improve the predictions of computer models using physical observations. Wang et al.

(2006) focus on characterizing the behavior of the prediction bias $\delta(\mathbf{x})$ while the emphasis of Kennedy and O'Hagan's work is on the *calibration* of computer models based on physical observations, but not on model validation. Their assumption on the relationship between computer outputs and physical observations is similar to the mathematical framework considered in this work, with the term $Y^m(x)$ in Eqn. (2.1) replaced by $\rho Y^m(\mathbf{x},\Theta)$, where $\rho$ is an unknown regression parameter, and $\Theta$ is the vector of calibration parameters. Their method for model calibration is aimed at finding the value of $\Theta$ that brings computer outputs as closely as possible to the physical observations rather than characterizing the difference between the two. Our focus in this work is on model validation with an emphasis on studying the validity of using a model for making a specific design decision.

## 3. THE BAYESIAN VALIDATION PROCEDURE

Most research in validating computer models had focused on estimating *prediction bias* and improving accuracy of the computer model. Much less work had been done on characterizing *prediction uncertainty* and *prediction bias* under general situations. From the engineering design perspective, both the predictive capability (accuracy) of a model as well as the confidence of using the model in choosing the best design candidate are of interest to the designer. The prediction bias $\delta(\mathbf{x})$ is more closely related to the assessment of model accuracy, while the prediction of the true model output $Y^r(\mathbf{x})$ is essential to assess the probability that a design alternative will produce an outcome that is preferred to or indifferent to other alternatives.

Referring to Eqns. (2.1) and (2.2), the relationship between $Y^e(\mathbf{x})$ and $Y^m(x)$ is given by $Y^e(\mathbf{x}) = Y^m(\mathbf{x}) + \delta(\mathbf{x}) + \varepsilon(\mathbf{x})$. Based on the experimental data, outputs of the computer model, and the specified experimental error $\varepsilon(\mathbf{x})$, the estimated prediction error, $\hat{\delta}(\mathbf{x})$, and its probability distribution can be obtained and used for validating the accuracy as well as other predictive capabilities of the model. Let $\hat{Y}^r(\mathbf{x})$ be the estimator of $Y^r(\mathbf{x})$, which can be obtained by $\hat{Y}^r(\mathbf{x}) = \hat{Y}^m(\mathbf{x}) + \hat{\delta}(\mathbf{x})$. The estimated prediction, $\hat{Y}^r(\mathbf{x})$, and its associated uncertainty quantification will be used to predict $f(\mathbf{x})$ and quantify its uncertainty.

In this work, a Bayesian approach is used to provide uncertainty quantification of both $\hat{\delta}(\mathbf{x})$ and $\hat{Y}^r(\mathbf{x})$. For complex validation metrics and design decision making, Bayesian inferences may be preferred as they require fewer assumptions and are more flexible for applications. In engineering applications where it may be too expensive to obtain experimental data, Bayesian methods may be preferable as additional information can be incorporated through prior distributions. Below, we describe the steps of the Bayesian procedure. Mathematical details of steps (1)~(4) for prediction and uncertainty quantification of $\hat{Y}^m(\mathbf{x})$ and $\hat{\delta}(\mathbf{x})$ can be found in Wang et al. (2006).

**(1) Collect both physical and computer model data.**

Both physical observations and model outputs are essential to model validation. Physical observations are desired to be as many as possible and close to the intended design region. Compared to physical observations, model outputs are less costly and should be simulated at design settings where the physical observations are available and close if not within the intended design regions. Let $\mathbf{x} = (x_1, \cdots, x_p)^T$ be a point in a $p$-dimensional design variable space. Let $D_e = \{\mathbf{x}_1, \cdots, \mathbf{x}_{n_e}\}$ be the design settings for physical experiments, and $\mathbf{y}^e = (y^e(\mathbf{x}_1), \cdots, y^e(\mathbf{x}_{n_e}))^T$ be the corresponding experimental observations. Let $D_m = \{\mathbf{x}'_1, \cdots, \mathbf{x}'_{n_m}\}$ be the design settings for computer experiments, and $\mathbf{y}^m = (y^m(\mathbf{x}'_1), \cdots, y^m(\mathbf{x}'_{n_m}))^T$ be the corresponding deterministic model outputs. $D_e$ and $D_m$ may or may not overlap.

**(2) Determine priors of Gaussian process parameters.**

Priors should be chosen to reflect existing scientific knowledge and past information. For example, Wang et al. (2006) assume the following priors for the location and variance parameters of Gaussian processes $Y^m(x)$ and $\delta(\cdot)$:

$$\sigma_m^2 \sim IG(\alpha_m, \gamma_m), \quad \sigma_\delta^2 \sim IG(\alpha_\delta, \gamma_\delta), \quad \sigma_\varepsilon^2 \sim IG(\alpha_\varepsilon, \gamma_\varepsilon),$$
$$\beta_m \mid \sigma_m^2 \sim N(\mathbf{b}_m, \sigma_m^2 \mathbf{V}_m), \quad \beta_\delta \mid \sigma_\delta^2 \sim N(\mathbf{b}_\delta, \sigma_\delta^2 \mathbf{V}_\delta),$$

where $IG(\alpha, \gamma)$ denotes the inverse gamma distribution.

**(3) Compute the posterior of computer model.**

As indicated in Eqn. (2.2), the posterior of computer model $Y^m(\cdot)$ is needed in the model validation procedure as an intermediate step to obtain the prediction of the true behavior $Y^r(\mathbf{x})$. Such information is also needed for calculating the posterior of prediction bias $\delta(\mathbf{x})$ when the design settings of the computer experiments do not completely overlap with those of physical experiments. In other words, model outputs at some points in $D_e$ are not available. Although the original computer model can be used directly to obtain $Y^m(\cdot)$, computer simulations may still be expensive and time-consuming and may not be available wherever we need them. In those cases, the posterior means of $Y^m(\cdot)$ at those points are used instead. The posterior of $Y^m(\mathbf{x})$ is given by Wang et al. (2006) as a noncentral $t$ distribution, with degree of freedom $n_{m|m}$, noncentrality parameter $\mu_{m|m}(\mathbf{x})$, and scale parameter $\sigma^2_{m|m}(\mathbf{x})$, i.e.,

$$Y^m(\mathbf{x}) \mid \mathbf{y}^m, \phi_m \sim T(n_{m|m}, \mu_{m|m}(\mathbf{x}), \sigma^2_{m|m}(\mathbf{x})), \tag{3.1}$$

where

$$n_{m|m} = n_m + 2\alpha_m, \tag{3.2}$$

$$\mu_{m|m}(\mathbf{x}) = \mathbf{f}_m^T(\mathbf{x})\mathbf{A}_m\mathbf{v}_m + \mathbf{r}_m^T(\mathbf{x})\mathbf{R}_m^{-1}(\mathbf{y}^m - \mathbf{F}_m\mathbf{A}_m\mathbf{v}_m), \tag{3.3}$$

$$\sigma^2_{m|m}(\mathbf{x}) = \frac{Q_m^2}{n_{m|m}} \cdot \left(1 - \begin{bmatrix} \mathbf{f}_m(\mathbf{x}) \\ \mathbf{r}_m(\mathbf{x}) \end{bmatrix}^T \begin{bmatrix} -\mathbf{V}_m^{-1} & \mathbf{F}_m^T \\ \mathbf{F}_m & \mathbf{R}_m \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{f}_m(\mathbf{x}) \\ \mathbf{r}_m(\mathbf{x}) \end{bmatrix}\right) \tag{3.4}$$

$$Q_m^2 = 2\gamma_m + (\mathbf{y}^m)^T \mathbf{R}_m^{-1}\mathbf{y}^m + \mathbf{b}_m\mathbf{V}_m^{-1}\mathbf{b}_m - \mathbf{v}_m^T\mathbf{A}_m\mathbf{v}_m, \tag{3.5}$$

$$\mathbf{A}_m^{-1} = \mathbf{F}_m^T\mathbf{R}_m^{-1}\mathbf{F}_m + \mathbf{V}_m^{-1}, \tag{3.6}$$

$$\mathbf{v}_m = \mathbf{F}_m^T\mathbf{R}_m^{-1}\mathbf{y}^m + \mathbf{V}_m^{-1}\mathbf{b}_m. \tag{3.7}$$

In the above equations, $\mathbf{F}_m = (\mathbf{f}_m(\mathbf{x}'_1), \cdots, \mathbf{f}_m(\mathbf{x}'_{n_m}))^T$ is the $n_m \times q_m$ design matrix, $\mathbf{R}_m$ is the $n_m \times n_m$ correlation

(parameterized by $\phi_m$) matrix of $\mathbf{y}^m$, and $\mathbf{r}_m(\mathbf{x})$ is the correlation (parameterized by $\phi_m$) between $Y^m(\mathbf{x})$ and $\mathbf{y}^m$. To get the marginal posterior of $Y^m(\mathbf{x})$, we need to integrate out $\phi_m$, which is computationally prohibitive. Alternatively, $\phi_m$ is estimated and treated as its true value. Methods such as the Maximum Likelihood Estimates (MLE) (Hastie et al., 2000), Markov Chain Monte Carlo (MCMC) (Geyer, 1992), and Minimum Mean Squared Error Estimates (MMSE) (Hastie et al., 2000), can be used to estimate $\phi_m$.

**(4) Compute the posterior of prediction bias.**

Similar to $Y^m(\mathbf{x})$ , the posterior of the prediction bias $\delta(\mathbf{x})$ is given as (Wang et al., 2006)

$$\delta(\mathbf{x}) \mid \mathbf{y}^e, \mathbf{y}^m \phi_\delta, \tau \sim T(n_{\delta|e,m}, \mu_{\delta|e,m}(\mathbf{x}), \sigma^2_{\delta|e,m}(\mathbf{x})), \tag{3.8}$$

$$\text{where } n_{\delta|e,m} = n_e + 2\alpha_\delta, \tag{3.9}$$

$$\mu_{\delta|e,m}(\mathbf{x}) = \mathbf{f}_\delta^T(\mathbf{x})\mathbf{A}_\delta\mathbf{v}_\delta + \mathbf{r}_\delta^T(\mathbf{x})(\mathbf{R}_\delta + \tau\mathbf{I}_{n_e})^{-1}(\mathbf{y}^e - \mathbf{y}^m_{n_e} - \mathbf{F}_\delta\mathbf{A}_\delta\mathbf{v}_\delta), \tag{3.10}$$

$$\sigma^2_{\delta|e,m}(\mathbf{x}) = \frac{Q^2_\delta}{n_{\delta|e,m}} \cdot (1 - \begin{bmatrix} \mathbf{f}_\delta(\mathbf{x}) \\ \mathbf{r}_\delta(\mathbf{x}) \end{bmatrix}^T \begin{bmatrix} -\mathbf{V}_\delta^{-1} & \mathbf{F}_\delta^T \\ \mathbf{F}_\delta & \mathbf{R}_\delta + \tau\mathbf{I}_{n_e} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{f}_\delta(\mathbf{x}) \\ \mathbf{r}_\delta(\mathbf{x}) \end{bmatrix}), \tag{3.11}$$

$$Q^2_\delta = 2\gamma_\delta + (\mathbf{y}^e - \mathbf{y}^m_{n_e})^T(\mathbf{R}_\delta + \tau\mathbf{I}_{n_e})^{-1}(\mathbf{y}^e - \mathbf{y}^m_{n_e}) + \mathbf{b}_\delta\mathbf{V}_\delta^{-1}\mathbf{b}_\delta - \mathbf{v}_\delta^T\mathbf{A}_\delta\mathbf{v}_\delta, \tag{3.12}$$

$$\mathbf{A}_\delta^{-1} = \mathbf{F}_\delta^T(\mathbf{R}_\delta + \tau\mathbf{I}_{n_e})^{-1}\mathbf{F}_\delta + \mathbf{V}_\delta^{-1}, \tag{3.13}$$

$$\mathbf{v}_\delta = \mathbf{F}_\delta^T(\mathbf{R}_\delta + \tau\mathbf{I}_{n_e})^{-1}(\mathbf{y}^e - \mathbf{y}^m_{n_e}) + \mathbf{V}_\delta^{-1}\mathbf{b}_\delta. \tag{3.14}$$

The denotations used above are analogues to those used in Eqns (3.1)~(3.7). We note that, $\phi_\delta$ is the correlation parameter underlying $\mathbf{R}_\delta$ and $\mathbf{r}_\delta^T$; $\tau$ is the ratio of $\sigma^2_\varepsilon$ to $\sigma^2_\delta$, i.e., $\tau = \sigma^2_\varepsilon / \sigma^2_\delta$, where $\sigma^2_\delta$ denotes the process variance of $\delta(\mathbf{x})$ while $\sigma^2_\varepsilon$ denotes the variance of $\varepsilon(\mathbf{x})$. Unlike $\delta(\mathbf{x})$ and $Y^m(\mathbf{x})$ which are assumed to be the Gaussian process with spatial correlation structure, $\varepsilon(\mathbf{x})$ follows identical independent normal distribution at different design site $\mathbf{x}$. The methods used to estimating $\phi_\delta$ and $\tau$ are similar to that of $\phi_m$.

**(5) Compute the prediction of the true behavior.**

Combining the results from Steps (3) and (4), the true behavior $Y^r(\mathbf{x})$ is predicted using the following equations on the estimations of the mean and variance,

$$\hat{Y}^r(\mathbf{x}) = \hat{Y}^m(\mathbf{x}) + \hat{\delta}(\mathbf{x}), \tag{3.15}$$

$$Var[\hat{Y}^r(\mathbf{x})] = Var[\hat{Y}^m(\mathbf{x})] + Var[\hat{\delta}(\mathbf{x})] = \sigma^2_{m|m}(\mathbf{x}) + \sigma^2_{\delta|e,m}(\mathbf{x}). \tag{3.16}$$

Under certain assumptions, $\hat{Y}^m(\mathbf{x})$ and $\hat{\delta}(\mathbf{x})$ are independent. The covariance between $\hat{Y}^r(\mathbf{x}_i)$ and $\hat{Y}^r(\mathbf{x}_j)$ is given by:

$$\begin{aligned} &Cov\left[\hat{Y}^r(\mathbf{x}_i), \hat{Y}^r(\mathbf{x}_j)\right] \\ &= Cov\left[\hat{Y}^m(\mathbf{x}_i) + \hat{\delta}(\mathbf{x}_i), \hat{Y}^m(\mathbf{x}_j) + \hat{\delta}(\mathbf{x}_j)\right] \\ &= Cov[\hat{Y}^m(\mathbf{x}_i), \hat{Y}^m(B_j)] + Cov[\hat{\delta}(B_i), \hat{\delta}(\mathbf{x}_j)] \\ &= \sigma^2_{m|m}(\mathbf{x}_i, \mathbf{x}_j) + \sigma^2_{\delta|e,m}(\mathbf{x}_i, \mathbf{x}_j) \end{aligned} \tag{3.17}$$

where

$$\sigma^2_{m|m}(\mathbf{x}_i, \mathbf{x}_j) = \frac{Q^2_m}{n_{m|m}} \cdot (R_m(\mathbf{x}_i, \mathbf{x}_j) - \begin{bmatrix} \mathbf{f}_m(\mathbf{x}_i) \\ \mathbf{r}_m(\mathbf{x}_i) \end{bmatrix}^T \begin{bmatrix} -\mathbf{V}_m^{-1} & \mathbf{F}_m^T \\ \mathbf{F}_m & \mathbf{R}_m \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{f}_m(\mathbf{x}_j) \\ \mathbf{r}_m(\mathbf{x}_j) \end{bmatrix}), \tag{3.18}$$

$$\sigma^2_{\delta|e,m}(\mathbf{x}_i, \mathbf{x}_j) = \frac{Q^2_\delta}{n_{\delta|e,m}} \cdot (R_\delta(\mathbf{x}_i, \mathbf{x}_j) - \begin{bmatrix} \mathbf{f}_\delta(\mathbf{x}_i) \\ \mathbf{r}_\delta(\mathbf{x}_i) \end{bmatrix}^T \begin{bmatrix} -\mathbf{V}_\delta^{-1} & \mathbf{F}_\delta^T \\ \mathbf{F}_\delta & \mathbf{R}_\delta + \tau\mathbf{I}_{n_e} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{f}_\delta(\mathbf{x}_j) \\ \mathbf{r}_\delta(\mathbf{x}_j) \end{bmatrix}). \tag{3.19}$$

When $\mathbf{x}_i = \mathbf{x}_j = \mathbf{x}$, Eqns (3.18) and (3.19) reduce to Eqns (3.4) and (3.11); Eqn. (3.17) reduces to Eqn. (3.16).

In the following section, we present some design validation metrics that utilize the information the predicted objective function $\hat{f}(\mathbf{x})$ at multiple design sites to select the best design candidate under model uncertainty and determine the confidence associated with the design decision.

## 4. SOME DESIGN VALIDATION METRICS

Different from the existing validation metrics that assess the predictive capability (accuracy) of a model, the design validation metrics $M_D$ are proposed and examined in this work to provide a probabilistic measure of whether a candidate design is better than other design choices with respect to a particular design objective. A few metrics that share the similar concept are developed to provide a direct measure of how reliable is the decision of choosing one design candidate versus the other design alternative, therefore to provide the confidence associated with a design decision with consideration of model uncertainty. Such metrics are desired to be useful in guiding validation activities. If large uncertainty exists in model response *y*, as well as the design objective *f*, the achieved $M_D$ may be too low to meet the design validity requirements, forcing designers to add new experiments to reduce model uncertainty or to lower the validity requirement.

Distinguishing neighboring designs in a continuous design space with the consideration of model uncertainty is mathematically more challenging than separating discrete and distinctive design choices. In this work, we start our investigation by defining design validity for a finite number (*k*) of design alternatives. Assuming a smaller design objective value is preferred, the following three forms of design validation metrics are considered and compared in this work:

(1) The Multiplicative Metric:

$$M_D^{Multip}(\mathbf{x}_i) = \left\{ \prod_{j=1, j\neq i}^{k} P\{\hat{f}(\mathbf{x}_i) < \hat{f}(\mathbf{x}_j)\} \right\}^{1/(k-1)} \tag{4.1}$$

(2) The Average (Additive) Metric:

$$M_D^{Average}(\mathbf{x}_i) = \frac{1}{k-1} \sum_{j=1, j\neq i}^{k} P\{\hat{f}(\mathbf{x}_i) < \hat{f}(\mathbf{x}_j)\} \tag{4.2}$$

(3) The Worst-Case Metric:

$$M_D^{Worstcase}(\mathbf{x}_i) = \min_{j=1,\ldots,k,\ j\neq i} P\{\hat{f}(\mathbf{x}_i) < \hat{f}(\mathbf{x}_j)\} \tag{4.3}$$

The proposed $M_D(\mathbf{x}_i)$ metrics in Eqns. (4.1) and (4.2) provide an averaged measure of the probability that the real outcome of $\mathbf{x}_i$ is better than or indifferent from other design choices, representing the confidence of using a predictive model to select $\mathbf{x}_i$ as the optimal design choice. If $M_D(\mathbf{x}_i)$=1, it indicates that a designer should have full confidence of taking $x_i$ as the optimal design. The $M_D(\mathbf{x}_i)$ metric in Eqn. (4.3) stands for the worst case of *P* be used instead of the average. It is our interest in this work to compare these several different metrics and determine to what extent these validity assessments

are useful to provide design differentiation and to guide model validation and design decision making.

## 5. EXAMPLE: ENGINE PISTON DESIGN

We consider the vehicle engine piston design case study previously analyzed in Jin et. al (2005). The Noise, Vibration and Harshness (NVH) characteristic of the vehicle engine is one of the critical elements of customer dissatisfaction. The goal of the design is to optimize the geometry of the engine piston to obtain the minimal piston slap noise. To graphically illustrate the results and better explain the concepts of the proposed method, only one design variable is considered. The same approach can be applied to high-dimensional problems. Previous results shows that the skirt profile (SP) strongly affects the response (slap noise), therefore SP is considered the design variable. Skirt profile is represented by characteristic ratios of the shape of an engine piston, ranging continuously from 1 to 3. Piston slap noise is the engine noise resulting from piston secondary motion, which can be simulated using ADAMS/Flex, a finite element based multi-body dynamics code. Thirty-four (34) hypothetical physical experiments are considered. Ten (10) computer experiments are conducted using the finite element model. It should be pointed out that ten computer experiments are sufficient for this one-dimensional case, although normally computer outputs are expected to be more than physical observations. All these data are provided in Tables 5.1 and 5.2, respectively. Note the design variable $\mathbf{x}$ = SP has been normalized to the unit interval [0,1].

Table 5.1 Thirty-four (34) physical experiments

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $\mathbf{x}_i \in D_e$ | 0.000 | 0.100 | 0.200 | 0.300 | 0.400 | 0.500 | 0.600 |
| $y^e(\mathbf{x}_i)$ | 56.332 | 56.077 | 55.875 | 55.542 | 55.159 | 54.840 | 54.682 |
| i | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| $\mathbf{x}_i \in D_e$ | 0.700 | 0.800 | 0.900 | 1.000 | 0.500 | 0.540 | 0.580 |
| $y^e(\mathbf{x}_i)$ | 55.039 | 55.183 | 55.774 | 56.749 | 54.867 | 54.646 | 54.748 |
| i | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| $\mathbf{x}_i \in D_e$ | 0.620 | 0.660 | 0.700 | 0.740 | 0.780 | 0.000 | 0.070 |
| $y^e(\mathbf{x}_i)$ | 54.576 | 54.614 | 54.623 | 54.978 | 54.923 | 56.224 | 56.228 |
| i | 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| $\mathbf{x}_i \in D_e$ | 0.140 | 0.210 | 0.280 | 0.350 | 0.420 | 0.490 | 0.560 |
| $y^e(\mathbf{x}_i)$ | 55.767 | 55.676 | 55.583 | 55.214 | 55.185 | 54.902 | 54.894 |
| i | 29 | 30 | 31 | 32 | 33 | 34 | |
| $\mathbf{x}_i \in D_e$ | 0.630 | 0.700 | 0.770 | 0.840 | 0.910 | 0.980 | |
| $y^e(\mathbf{x}_i)$ | 54.611 | 54.831 | 54.947 | 55.352 | 55.765 | 56.560 | |

Table 5.2 Ten (10) computer experiments

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $\mathbf{x}_i \in D_m$ | 0.050 | 0.150 | 0.250 | 0.350 | 0.450 | 0.550 | 0.650 |
| $y^m(\mathbf{x}_i)$ | 56.033 | 55.584 | 55.417 | 55.402 | 55.278 | 54.957 | 54.641 |
| i | 8 | 9 | 10 | | | | |
| $\mathbf{x}_i \in D_m$ | 0.750 | 0.850 | 0.950 | | | | |
| $y^m(\mathbf{x}_i)$ | 54.656 | 55.191 | 56.193 | | | | |

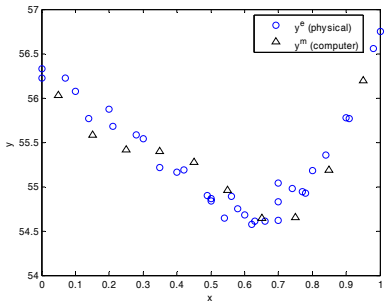


Figure 5.1 Physical and computer experiment data (circles: physical experiments; triangles: computer experiments)

### 5.1 Prediction and uncertainty quantification

Based on the available data, the Bayesian approach described in Section 3 is implemented. For the purpose of comparison, the predictive models are established in two stages. *In the first stage, we only use the first 19 points out of 34 physical experiment points in Table 5.1***.** The remaining 14 points are added in the second stage.

**Prediction and uncertainty quantification of $\hat{Y}^m(\mathbf{x})$**

From the data shown in Table 5.1, it is found that there is no overlap between $D_e$ and $D_m$, indicating that the settings of the design variable (x) for computer outputs are different from those for physical experiments. We first calculate the posterior of computer model $Y^m(x)$, $p(Y^m(\mathbf{x})|\mathbf{y}^m,\phi_m)$, through Eqn. (3.1). To do this, we need to estimate the correlation parameter $\phi_m$ using the eleven available computer experiments. Because of the small amount (10) of computer outputs available, leave-one-out cross validation strategy is used. Fig. 5.2 shows the plot of Rooted Mean Squared Error (RMSE) from the cross-validation vs. $\phi_m$ ranging from 0.5 to 50. The minimum RMSE is identified at $\phi_m$ =2.2.

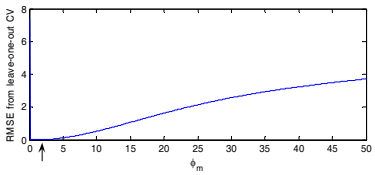


Figure 5.2 RMSE from leave-one-out cross validation vs. $\phi_m$ (optimal $\phi_m$ =2.2)

Given $\phi_m$, the prediction of $\hat{Y}^m(\mathbf{x})$ and the associated 95% confidence interval are calculated through the posterior of $Y^m(\mathbf{x})$. From Fig. 5.3, it is noted that $\hat{Y}^m(\mathbf{x})$ passes all ten computer experiment points and there is no prediction uncertainty at each sampling site. Furthermore, owing to the smooth behavior of the computer model, ten sampling points are sufficient; hence the uncertainty due to the use of Gaussian process model replacing the computer model is small across the design range.

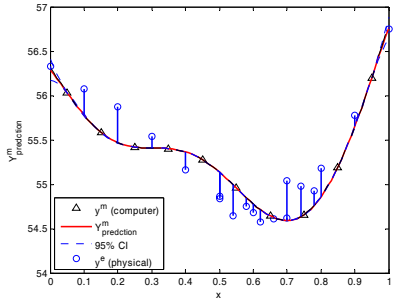


Figure 5.3 Prediction of $\hat{Y}^m(\mathbf{x})$ and 95% confidence interval

**Prediction and uncertainty quantification of $\hat{\delta}(\mathbf{x})$**

From Eqn. (3.8), the prediction of $\hat{\delta}(\mathbf{x})$ and the associated uncertainty is characterized by the posterior of $\delta(\mathbf{x})$, given $\phi_\delta$ and $\tau$. Ten-fold cross validation is used to determine the optimal values of $\phi_\delta$ and $\tau$ in the similar way as in estimating $\phi_m$. The results show the optimal setting at $\tau$=2, $\phi_\delta$=22. Fig. 5.4 displays the prediction of $\hat{\delta}(\mathbf{x})$ and the 95% confidence interval. Note the sampling points illustrated in Figure 5.4 represent the difference between the physical experiment $Y^e(\mathbf{x})$ and the model prediction $\hat{Y}^m(\mathbf{x})$ (the magnitude of the vertical line segments shown in Figure 5.3). $\hat{\delta}(\mathbf{x})$ has a relatively small variance in the range of $\mathbf{x} \in [0.6, 0.8]$ compared with the region of $\mathbf{x} \notin [0.6, 0.8]$. This can be explained by the fact that more physical observations are available for $\mathbf{x} \in [0.6, 0.8]$.

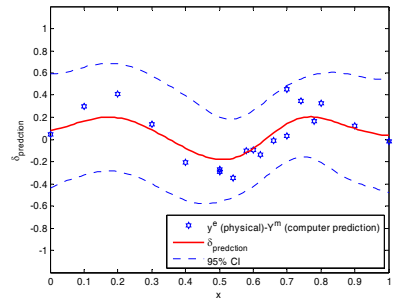


Figure 5.4 Prediction of $\hat{\delta}(\mathbf{x})$ and 95% confidence interval

**Prediction and uncertainty quantification of $\hat{Y}^r(\mathbf{x})$**

Having obtained the posteriors of $Y^m(\mathbf{x})$ and $\delta(\mathbf{x})$, the prediction of $\hat{Y}^r(\mathbf{x})$ is simply the addition of $\hat{Y}^m(\mathbf{x})$ and $\hat{\delta}(\mathbf{x})$. The variance of $\hat{Y}^r(\mathbf{x})$ is the addition of the two. The prediction and 95% confidence interval is illustrated in Fig. 5.5. In the range of $\mathbf{x} \notin [0.6, 0.8]$, where less sampling points are available for both physical and computer experiments, the uncertainty of $\hat{Y}^r(\mathbf{x})$ is higher accordingly. Comparing Figs. 5.3 and 5.4, it is found that the uncertainty of $\hat{\delta}(\mathbf{x})$ dominates the uncertainty of $\hat{Y}^r(\mathbf{x})$.

**Prediction and uncertainty quantification of $f(\mathbf{x})$**

Design objective function $f(\mathbf{x})$ is defined based on the design scenario and designers' preference. In this work we consider a typical robust design objective, where the design variable $\mathbf{x}$ is assumed random (e.g. $\mathbf{x}$ has a normal distribution $\mathbf{x} \sim N(\mu_x, 0.05)$). Therefore $f(\mathbf{x}) = \mu_y + k \cdot \sigma_y$ where $\mu_y$ and $\sigma_y$ are the mean and standard deviation of $y$ (engine slap noise), and the weighting factor $k$ is set at $k$=3. The robust design objective is utilized to reduce the impact of the uncertainty associated with the randomness of $\mathbf{x}$. On the other hand, since the uncertainty of $\hat{Y}^r(\mathbf{x})$ is reducible with more experiment data are added, essentially, it is the inherent uncertainty in design objective function $f(\mathbf{x})$, due to the model uncertainty, that influences the confidence in making any design decision.

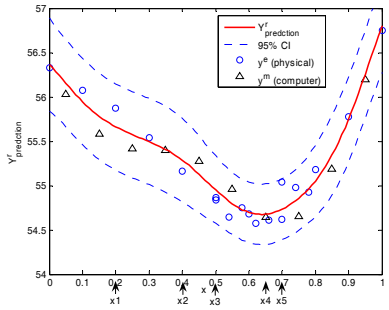


Figure 5.5 Prediction of $\hat{Y}^r(\mathbf{x})$ and 95% confidence interval

Prediction of $\hat{f}(\mathbf{x})$ and quantification of its uncertainty is computationally challenging. Approximation of the mean and variance of $\hat{f}(\mathbf{x})$ in analytical way is discussed by Apley et al (2005). Monte Carlo simulation approach is used in this work. Based on the mean, variance and covariance of $\hat{Y}^r(\mathbf{x})$ given in Eqns. (3.15)~(3.17), one can simulate a large amount (e.g. 100) of realizations of the random process $\hat{Y}^r(\mathbf{x})$. For simplicity, only three of such realizations are selected and shown in Fig. 5.6. Each single realization of $\hat{Y}^r(\mathbf{x})$ determines the corresponding realization of $f(\mathbf{x})$ subject to the randomness of $\mathbf{x}$. As a result, the prediction of $\hat{f}(\mathbf{x})$ and its uncertainty is quantified, as shown in the bold lines in Figure 5.6.

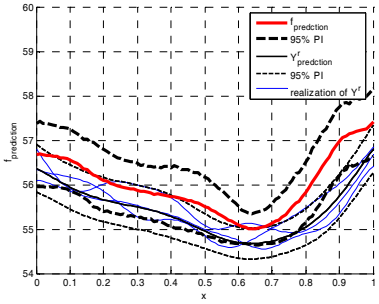


Figure 5.6 Prediction of $f(\mathbf{x})$ and 95% confidence interval (19 physical experiments)

## 5.2 Application of Design Validation Metrics

In this section, we apply the design validation metrics $M_D$ proposed in Section 4 to the engine piston design. Suppose $k$=5 design candidates have been identified as $\mathbf{x}_i$={0.2, 04, 0.5, 0.65, 0.7} (see Fig. 5.5). To establish the basic

notion of probability based comparison, we first explore the pair-wise comparison involving two design candidates.

### 5.2.1 Probability Based Pair-wise Comparison: $P_{ij}$

With the consideration of model uncertainty, differentiating the predicted performance at two sites $x_i$ and $x_j$ is to examine the probability of one performance being smaller or larger than the other. Under the 'smaller the better' scenario, if $\hat{f}(\mathbf{x}_i) < \hat{f}(\mathbf{x}_j)$, we could measure the probability based comparison $P_{ij}$ as

$$P_{ij} = P\{\hat{f}(\mathbf{x}_i) < \hat{f}(\mathbf{x}_j)\} \quad (5.1)$$

The larger the $P_{ij}$, the larger capability the predicative model $\hat{Y}^r(\mathbf{x})$ has in differentiating two designs. Based on the calculated mean, variance of $\hat{f}(\mathbf{x})$, and assuming $\hat{f}(\mathbf{x}_i)$ and $\hat{f}(\mathbf{x}_j)$ jointly follow the multivariate Gaussian distribution, Monte Carlo simulation is conducted to sample a relatively large number (e.g., $N_s$=1000) of two-dimensional points. $P_{ij}$ is calculated by $N\{Y^r{}_n(\mathbf{x}_i) < Y^r{}_n(\mathbf{x}_j)\}/N_s$, where $N\{Y^r{}_n(\mathbf{x}_i) < Y^r{}_n(\mathbf{x}_j)\}$ represents the number of two-dimensional sampling points among which $Y^r{}_n(\mathbf{x}_i)$ is smaller than $Y^r{}_n(\mathbf{x}_j)$.

### 5.2.2 Design validation metrics $M_D$

From Eqns. (4.1)~(4.3), the calculation in each of the three types of $M_D(\mathbf{x}_i)$ depends on the probability level in the pair-wise comparison of design site $\mathbf{x}_i$ against other designs $\mathbf{x}_j$ $(j \neq i)$. The points generated by Monte Carlo simulation of $\hat{Y}^r(\mathbf{x}_i)$ are illustrated in Fig. 5.7. Table 5.3 provides the calculated values of three types of $M_D$ described in Eqns. (4.1)~(4.3) for each design candidate $\mathbf{x}_i$.

Table 5.3 $M_D$ (19 physical experiments)

| Design $i$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $M_D^{Multip}(\mathbf{x}_i)$ | 0 | 0.1057 | 0.3379 | **0. 8870** | 0. 6938 |
| $M_D^{Average}(\mathbf{x}_i)$ | 0. 0842 | 0. 2715 | 0. 4830 | **0. 8957** | 0. 7655 |
| $M_D^{Worstcase}(\mathbf{x}_i)$ | 0 | 0. 0150 | 0. 1010 | **0. 6990** | 0. 3010 |
| $J_{worst}(\mathbf{x}_i)$ | 4 | 4 | 4 | 5 | 4 |

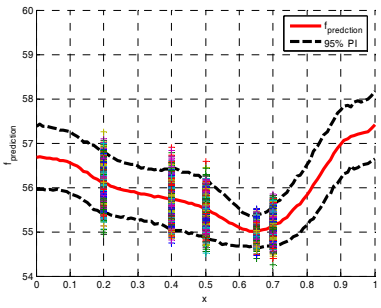


Figure 5.7 Comparison between five design sites (19 physical experiments)

From Table 5.3, it is found that the three different types of $M_D$ at design site $\mathbf{x}_4$ consistently achieve the largest $M_D$ value among the five design alternatives. Note that $\mathbf{x}_4$ is also the optimal design from the predicted $\hat{f}(\mathbf{x}_i)$ (the mean value). In fact, the ranking order of $M_D(\mathbf{x}_i)$ among the five candidate design matches (inversely) the ranking order of $\hat{f}(\mathbf{x}_i)$. $M_D(\mathbf{x}_i)$ provides the confidence of choosing the optimal design $\mathbf{x}_4$ against the other alternatives.

Recall the relation $\hat{Y}^r(\mathbf{x}) = \hat{Y}^m(\mathbf{x}) + \hat{\delta}(\mathbf{x})$. To enhance the accuracy of the predictive model, both $\hat{Y}^m(\mathbf{x})$ and $\hat{\delta}(\mathbf{x})$ can be refined. Figures 5.3 shows that $\hat{Y}^m(\mathbf{x})$ has already reached a high accuracy. In contrast, $\hat{\delta}(\mathbf{x})$ contributes much larger uncertainty to the predictive model than $\hat{Y}^m(\mathbf{x})$. Therefore, to refine the predictive model $\hat{Y}^r(\mathbf{x})$, additional physical experiments need to be conducted to reduce the uncertainty of $\hat{\delta}(\mathbf{x})$ (at the same time to enhance the accuracy of $\hat{\delta}(\mathbf{x})$). In the 2nd stage of testing, the remaining fifteen (15) physical experiments in Table 5.1 are used. The procedure described in Section 5.1 is repeated with 19+15=34 in total physical experiment points.

The updated objective function $\hat{f}(\mathbf{x})$ and representative realizations of $\hat{Y}^r(\mathbf{x})$ are shown in Figure 5.8. Compared with Figure 5.6, they are more accurate with reduced uncertainty. The reduced uncertainty has an impact on the values of the $M_D$ metrics. The updated $M_D$ values for the five selected design sites are summarized in Table 5.4. Because $\hat{f}(\mathbf{x}_4)$ achieves the smallest predicted performance again, $M_D(\mathbf{x}_4)$ continues to be the largest one among the five alternatives as in Table 5.3. It is noted that the values of all three types of $M_D(\mathbf{x}_4)$ have increased, indicating larger confidence in differentiating design alternatives.

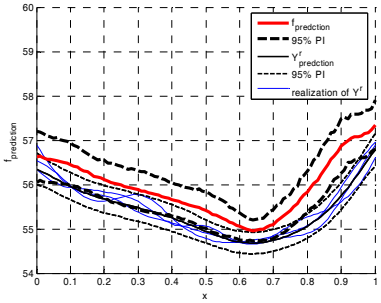


Figure 5.8 Prediction of $f(\mathbf{x})$ and 95% confidence interval (19+15=34 physical experiments)

Table 5.4 $M_D$ (19+15=34 physical experiments)

| Design $i$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $M_D^{Multip}(\mathbf{x}_i)$ | 0 | 0 | 0. 2137 | **0. 9101** | 0. 7170 |
| $M_D^{Average}(\mathbf{x}_i)$ | 0. 0160 | 0. 2850 | 0. 4807 | **0. 9192** | 0. 7990 |
| $M_D^{Worstcase}(\mathbf{x}_i)$ | 0 | 0 | 0. 0310 | **0. 7080** | 0. 2920 |
| $J_{worst}(\mathbf{x}_i)$ | 4 | 4 | 4 | 5 | 4 |

### 5.2.3 Implications of three types of $M_D$

Although the pair-wise probability comparison is used in all three forms (see Eqns. 4.1~4.3) of design validation metrics $M_D(\mathbf{x}_i)$, they have different implications, which are explained as follows. Apparently, $M_D(\mathbf{x}_i)$ in all three forms ranges from 0 to 1.

**(1) Multiplicative Metric**

By Eqn. (4.1), $M_D^{Multip}(\mathbf{x}_4) = \{ \prod_{j=1,2,3,5} P[\hat{f}(\mathbf{x}_4) < \hat{f}(\mathbf{x}_j)] \}^{1/4}$. Due to the multiplication, $M_D^{Multip}(\mathbf{x}_4)$ is sensitive to each probability value $P\{\hat{y}(\mathbf{x}_4) < \hat{y}(\mathbf{x}_j)\}$, implying that $M_D^{Multip}(\mathbf{x}_4)$ can reflect the local refinement of predictive model.

**(2) Average (Additive) Metric**

By Eqn. (4.2), $M_D^{Average}(\mathbf{x}_4) = \sum_{j=1,2,3,5} P\{\hat{f}(\mathbf{x}_4) < \hat{f}(\mathbf{x}_j)\} / 4$. Unlike the multiplicative metric, $M_D^{Average}(\mathbf{x}_4)$ is less sensitive to each constituent value of $P\{\hat{f}(\mathbf{x}_4) < \hat{f}(\mathbf{x}_j)\}$. If the number of alternative designs is large, due to averaging, the local refinement of the model might not be reflected in the small change of $M_D^{Average}(\mathbf{x}_i)$.

**(3) The Worst-Case Metric**

$M_D^{Worstcase}(\mathbf{x}_4)$ takes the worst case (minimum) of $P\{\hat{f}(\mathbf{x}_4) < \hat{f}(\mathbf{x}_j)\}$. Unlike the other metrics that provide an overall confidence involving all the other design alternatives, $M_D^{Worstcase}(\mathbf{x}_4)$ only concerns the most competitive design ($2^{nd}$ best design). In Tables 5.3 and 5.4, the last row ($J_{worst}(\mathbf{x}_i)$) displays the index of the most competitive design site. For instance, $J_{worst}(\mathbf{x}_4)$=5 indicates that design $\mathbf{x}_5$ is the $2^{nd}$ best design to design $\mathbf{x}_4$, or $\mathbf{x}_5$ is the most difficult to be differentiated from design $\mathbf{x}_4$. $M_D^{Worstcase}(\mathbf{x}_4)$ is equal to $P\{\hat{f}(\mathbf{x}_4) < \hat{f}(\mathbf{x}_5)\}$, which is the lowest probability compared to the other three.

## 6. CLOSURE

In this work, a Bayesian approach to model validation is presented to provide quantitative assessments of uncertainty in using predictive models in engineering design. Design-oriented validation metrics are further developed to guide designers for achieving high confidence of using predictive models. In engineering applications where it is too expensive to obtain experimental data, the Bayesian inference approach offers much flexibility as it requires fewer assumptions and additional design knowledge and information can be easily incorporated through prior distributions.

Compared to the existing work, our work results in a full Bayesian analysis model for predicting computer model bias and true model output, that are both accurate and economically sound. Our approach provides quantitative means to define and to assess model validity from the perspective of design decision making with the consideration of various sources of uncertainties. It offers rigorous methods for quantifying the model uncertainty in an intended design domain that may interpolate as well as extrapolate from a tested domain. In addition, our work offers a new and improved way of viewing model validation by relating its definition to a specific design choice. The proposed measure for assessing design validity provides some probabilistic measurements with regard to the confidence of using a model for making a specific design choice; they can be used to overcome the limitations of many existing model validation approaches while providing direct estimate of the global impact of uncertainty sources on the confidence in a design decision. Even though our approach is demonstrated for a simplified one dimensional engineering design problem for ease of visualization, the same approach can be applied to problems with multidimensional design inputs and the interest is always to provide the probabilistic assessment on whether the *performance* (measured by the design objective) of one particular design is better than the others.

In this work, the proposed model validation metrics are only applied to design cases with a finite number of candidate design alternatives. Our study lays the ground for distinguishing neighboring designs in a continuous design space, a much challenging topic that is being investigated. A more general model validation framework is currently under development to determine how an optimal should be picked along with the activities in model validation. Future research is also planned for particularizing the proposed Bayesian validation procedure and statistical inferences for specific engineering applications where the natures of available experimental and computational data vary. The estimation of prediction bias will be extended to develop validation metrics that measure the predictive capability of a model considering both tested and untested regions. The role of design validation metrics in engineering design will be further extended by introducing not only product design decisions but also decisions in allocating the resources for physical and computer experiments. This will require the incorporation of decision analysis techniques to study the tradeoffs involved in model refinement and uncertainty reduction by considering designers' preference.

## ACKNOWLEDGMENTS

The grant support from National Science Foundation (NSF) to this collaborative research between Northwestern University (DMI – 0522662) and Georgia Tech (DMI-0522366) is greatly appreciated. The views expressed are those of the authors and do not necessarily reflect the views of the sponsors.

## REFERENCES

Ang, J.A., Trucano, T.G., and Luginbuhl, D.R., "Confidence in ASCI Scientific Simulations", Ninth Nuclear Explosives Code Developers' Conference, San Diego, CA, Oct. 22-25, 1996.

Apley, D. W., Liu, J., Chen, W., "Understanding the Effects of Model Uncertainty in Robust Design with Computer Experiments," accepted by ASME Journal of Mechanical Design, 2005.

Apostolakis, G., "A Commentary on Model Uncertainty", Model Uncertainty: Its Characterization and Quantification, editors: Mosleh, A, Siu, N, Smidts, C. and Lui, C,

 

NUREG/CP-0138, U.S. Nuclear Regulatory Commission, 1994.

Bayarri, M.J.., Berger, J.O., Higdon, D., Kennedy, M.C., Kottas, A., Paulo, R., Sacks, J., Cafeo, J.A., Cavendish, J., Lin, C.H., and Tu, J., "A Framework for Validation of Computer Models", Foundations for Verification and Validation in 21st Century Workshop, Johns Hopkins University, October 22-23, 2002.

Bayarri, M.J.., Berger, J.O., Higdon, D., Kennedy, M.C., Kottas, A., Paulo, R., Sacks, J., Cafeo, J.A., Cavendish, J., Lin, C.H., and Tu, J., "A Framework for Validation of Computer Models", Foundations for Verification and Validation in 21st Century Workshop, Johns Hopkins University, October 22-23, 2002.

Buranathiti, T., Cao, J., Chen, W., Baghdasaryan, L., and Xia, Z.C., "Approaches for Model Validation: Methodology and Illustration on a Sheet Metal Flanging Process", ASME Journal of Manufacturing Science and Engineering, in press, 126, November, 2004.

Buslik, A., "A Bayesian Approach to Model Uncertainty", Model Uncertainty: Its Characterization and Quantification, editors: Mosleh, A, Siu, N, Smidts, C. and Lui, C, U.S. Nuclear Regulatory Commission, NUREG/CP-0138, 1994.

Cafeo, J.A., and Thacker, B.H., "Concepts and Terminology of Validation for Computational Solid Mechanics Models", SAE 2004 World Congress & Exhibition, Detroit, MI, March, 2004.

Chen, W., Baghdasaryan, L., Buranathiti, T., and Cao, J., "Model Validation via Uncertainty Propagation and Data Transformations", AIAA Journal, 42(7), 1406-1415, 2004.

DoD, DoD Directive No. 5000.61, "Modeling and Simulation (M&S) Verification, Validation, and Accreditation (VV&A)", Defense Modeling and Simulation Office, www.dmso.mil/docslib.

Doebling, S.W., Hemex, F.M., Schultz, J.F., Girrens, S.P., "Overview of Structural Dynamics Model Validation Activities at Los Alamos National Laboratory", Proc AIAA/ASME/ASCE/AHS/ASC 43rd Structures, Structural Dynamics, and Materials Conf., AIAA 2002-1643, Denver, CO, April 22-25, 2002.

Easterling, R. G. & Berger, J. O., Statistical Foundations for The Validation of Computer Models, presented at Computer Model Verification and Validation in the 21st Century Workshop, Johns Hopkins University, 2002.

Freese, F., Testing Accuracy, Forest Science, 6(2), 139-145, 1960.

Geyer, C. J., "Practical Markov Chain Monte Carlo," Statistical Science, 7(4):473-511, 1992.

Gregoire, T., G., and Reynolds, M., R., Accuracy Testing and Estimation Alternatives, Forest Science, 34(2), 302-320, 1988.

Gu, L., Yang, R.J., "Recent Applications on Reliability-based Optimization of Automotive Structures," SAE Technical Paper Series: 2003-01-0152, 2003.

Hanson, K.M., "A Framework for Assessing Uncertainties in Simulation Predictions", Physica D. 133, 179-188, 1999.

Hastie, T., Tibshirani, R., Friedman, J., The Elements of Statistical Learning, Springer-Verleg, 2000.

Hazelrigg, G.A., "Thoughts on Model Validation for Engineering Design", ASME Design Technical Conference, DETC2003/DTM-48632, Chicago, IL, Sept. 2-6, 2003.

Hazelrigg, G.A., "On the Role and Use of Mathematical Models in Engineering Design", Transactions of ASME, Journal of Mechanical Design, 121(3), 336-341, 1999.

Hills, G. R., and Trucano, T. G., "Statistical Validation of Engineering and Scientific Models: Background", SAND99-1256, 1999.

Jin, R., Chen, W., Sudjianto, A., "Analytical metamodel-based global sensitivity analysis and uncertainty propagation for robust design," to be published in Journal of Quality Technology, 2005

Kennedy, M. C. and O'Hagan, A., "Bayesian Calibration of Computer Experiments", Journal of the Royal Statistical Society, B. 63, 425-464, 2001.

Mahadevan, S., and Rebba, R., "Validation of Reliability Computational Models using Bayes Networks," Journal of Reliability Engineering and System Safety, 87, pp. 223-232, 2005.

Malak, R.J. and Paredis, "On Characterizing and Assessing the Validity of Behavioral Models and Their Predictions", 2004 ASME Design Technical Conferences, DETC2004-57452, Salt Lake City, Utah, Sept. 28-Oct. 2, 2004.

Marczyk, J., Holzner, M., et.al., "Stochastic Automotive Crash Simulation; A new Frontier in Virtual prototyping," Proceedings of the Pam 97 user Conference, Pragu, Czech Republic, October 16-17, 1997.

McAdams, D.A. and Dym, C.L., "Modeling and Information in the Design Process", 2004 ASM Design Technical Conferences, DETC2004-57101, Salt Lake City, Utah, Sept. 28-Oct. 2, 2004.

Oberkampf, W. and Barone, M., "Measures of Agreement Between Computation and Experiment: Validation Metrics", 34th AIAA Fluid Dynamics Conference and Exhibit, AIAA-2004-2626, Portland, Oregon, June 28-1, 2004.

Oberkampf, W.L., Deland S.M., Rutherford, B.M, Diegert, K.V. and Alvin, K.F., "A New Methodology for the Estimation of Total Uncertainty in Computational Simulation", AIAA System Dynamics, Material, Conference, AIAA-99-1612, 3061 –3083, 1999.

Oberkampf, W.L. and Trucano, T.G., "Validation Methodology in Computational Fluid Dynamics", AIAA 2000-2549, Fluids 2000, Denver, CO, 2000.

Oberkampf, W.L., Trucano, T.G., and Hirsch, C., "Verification, Validation, and Predictive Capability in Computational Engineering and Physics", SAND2003-3769, February, 2003.

Reynolds, M.R., Jr., "Estimating the Error in Model Predictions", Forest Science, 30(2), 454-469, 1984.

Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P., "Design and analysis of computer experiments," Statistical Science, 4(4), pp. 409-435, 1989

Santner, T.J., B.J. Williams, and W.I. Notz (2003), "The Design and Analysis of Computer Experiments," New York: Springer-Verlag.

Trucano, T. G., "Prediction and Uncertainty in Computational Modeling of Complex Phenomena, A Whitepaper", Sandia report, SAND98-2776, 1998.

Wang, N., Ge, P., "Study of Metamodeling Techniques and Their Applications in Engineering Design," ASME-MED Manufacturing Science and Engineering, Vol. 10, pp. 89-95, 1999.

Wang, S., Chen, W., and Tsui, K, "Bayesian Validation of Computer Models," working paper, 2006.

 