

Design and analysis of variable fidelity experimentation applied to engine valve heat treatment process design

Deng Huang

Scientific Forming Technologies Corporation, Columbus, USA

and Theodore T. Allen

Ohio State University, Columbus, USA

[Received August 2003. Final revision July 2004]

Summary. When experimentation on a real system is expensive, data are often collected by using cheaper, lower fidelity surrogate systems. The paper concerns response surface methods in the context of variable fidelity experimentation. We propose the use of generalized least squares to generate the predictions. We also present perhaps the first optimal designs for variable fidelity experimentation, using an extension of the expected integrated mean-squared error criterion. Numerical tests are used to compare the performance of the method with alternatives and to investigate the robustness to incorporated assumptions. The method is applied to automotive engine valve heat treatment process design in which real world data were mixed with data from two types of computer simulation.

Keywords: Expected integrated mean-squared error; Optimal experimental design; Response surface methods; Surrogate systems

1. Introduction

Often, experimental data are analysed to predict outputs of a system as a function of system inputs. When experimentation on the real system of interest is expensive, the data sometimes are collected from cheaper surrogate experimental systems. For example, laboratory and pilot systems can be used to mimic production systems, and computer simulations can approximate physical experiments. Yet, the planning of experiments that includes an allocation of runs to different experimental systems has received relatively little attention. This paper describes perhaps the first attempt to plan experiments optimally by using more than a single experimental system for response surface generation. The goal is to derive an accurate prediction of the real system outputs, while achieving an acceptable total experimental cost.

Although experiments on a surrogate system are cheaper, the responses are usually associated with systematic errors. Thus, in planning the experiments, one has to balance between the cost and ‘faithfulness’ that are associated with each experimental run. We call this type of experimentation ‘variable fidelity experimentation’, where the term ‘fidelity’ relates to the magnitude of the systematic errors that each experimental system achieves in reproducing the input–output relationships of the real system. The real, often physical, system is called the ‘highest fidelity’ system.

Address for correspondence: Deng Huang, Scientific Forming Technologies Corporation, 5038 Reed Road, Columbus, OH 43220, USA.
E-mail: dhuang@deform.com

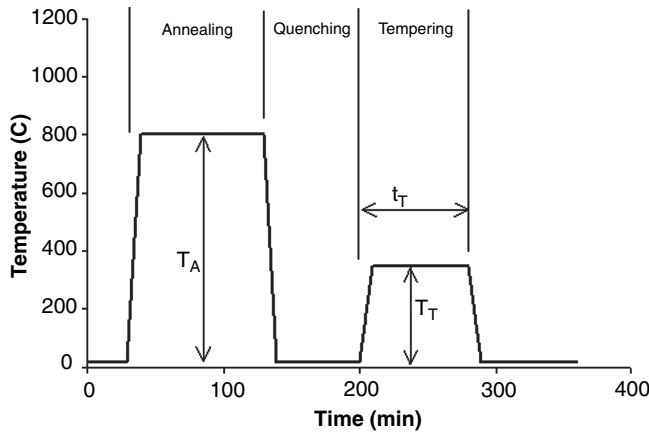


Fig. 1. Heat treatment schedule of an engine valve

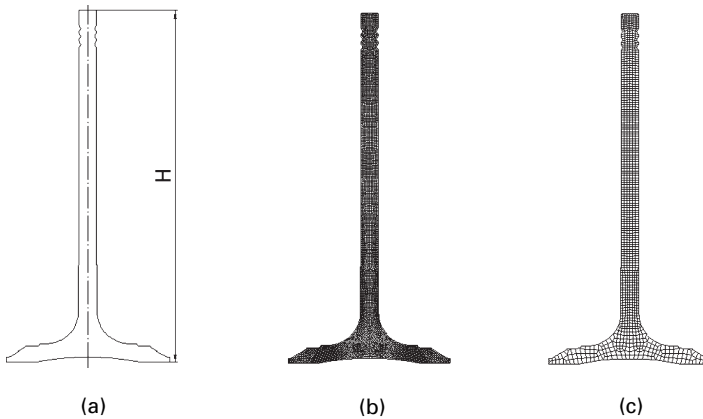


Fig. 2. Engine valve and its finite element analysis models: (a) geometry; (b) 6000-element model; (c) 1000-element model

The targeted application of this research relates to the design of the heat treatment process for an automotive engine valve. Heat treatment is applied to obtain good material hardness of the part, but it also induces distortion that increases the manufacturing cost in later steps of the process. Typically, process engineers trade off distortion for the hardness. To make informed trade-offs, it is desirable to predict accurately the distortion as a function of the heat treatment process parameters. A typical heat treatment schedule consists of the annealing, quenching and tempering procedures, as shown in Fig. 1. In the case that is presented here, the engineers were allowed to adjust the three parameters annealing temperature T_A , tempering temperature T_T and tempering time t_T to achieve the best process design. The distortion can be evaluated by the shrinkage in the distance between the top and the rim 'tip' ('H' in Fig. 2(a)) before and after heat treatment.

Because physical experiments on such a heat treatment process are expensive, engineers typically supplement the data by running finite element analysis (FEA) computer simulations, which can be done with different levels of sophistication and cost. Although FEA experiments generally cost less than physical experiments, their cost can be substantial. For example, a typical

consulting company might charge \$1200 for a physical experiment, \$400 for a relatively high resolution FEA run (Fig. 2(b)) and \$200 for a relatively low resolution FEA run (Fig. 2(c)). However, the total experimental budget may be fixed. In this case, this limit is set to \$5000, which is representative of factory floor optimization projects. We shall use this heat treatment distortion study to illustrate the application of the design and analysis methods proposed. The design proposed specifies the input combinations for the runs and the experimental system on which these runs will be performed.

In Section 2, we review the literature related to variable fidelity experimentation. In Section 3, following Kennedy and O'Hagan (2000), we assume that the systematic errors are approximated by Gaussian stochastic processes, and we propose the use of generalized least squares (GLS) for analysis. In Section 4, we propose an experimental design criterion, which is a generalization of the expected integrated mean-squared error (EIMSE) criterion that was proposed by Allen, Yu and Schmitz (2003). Section 5 includes a comparison of the methods proposed with alternatives by using numerical test examples. The application of the proposed methods to the engine valve heat treatment study is described in Section 6. Section 7 summarizes the contributions and suggestions for future research.

2. Literature review

Several methods have been proposed and used in the context of variable fidelity experiments. Probably the most common approach that is used by engineers is first to calibrate the surrogate system with data from the real system, and then to generate predictions by using data from the surrogate system only. The calibration is achieved often by tuning the surrogate system so that the outputs best match those of the real system at a single set of inputs. For example, this method was used in Koc *et al.* (2000). This approach has at least the following limitations.

- (a) A good 'alignment' of the surrogate system with the real system may be impossible over the entire region of interest.
- (b) This approach effectively discards the data from the real system in the analysis, which is intuitively problematic since these data contain the most valuable and expensive information.
- (c) This method does not take advantage of additional surrogate systems in cases in which more than one is available.

Knill *et al.* (1999), Vitali *et al.* (2002) and Kennedy and O'Hagan (2000, 2001) all explored the generation of prediction models from variable fidelity experiments in which the associated experimental designs had a particular structure. In their cases, the input combinations for the runs at the higher levels of fidelity were always a subset of the input combinations for lower fidelity levels. Therefore, if $\mathbf{D}_l$ was the experimental design for fidelity level l for $l = 1, \dots, m$ with 1 being the highest level of fidelity, their experimental plans satisfied $\mathbf{D}_1 \subseteq \mathbf{D}_2 \subseteq \dots \subseteq \mathbf{D}_m$. This 'hierarchical design restriction' permitted these researchers to model the differences between the outputs of various experimental systems directly. Specifically, Knill *et al.* (1999) and Vitali *et al.* (2002) investigated the use of linear regression models for this purpose. Kennedy and O'Hagan (2000) investigated the use of so-called 'Bayesian Gaussian stochastic process models', which are related to the 'kriging' models that have been proposed for computer experimentation by Sacks *et al.* (1989a) and others. Kennedy and O'Hagan (2001) also proposed the 'Bayesian calibration' method which attempts to correct any inadequacy of even the best-calibrated surrogate system. Although this 'Bayesian calibration' approach may overcome some shortcomings that are related to the traditional calibration approaches that were mentioned above, it also has the

hierarchical design restriction. The issue of design optimality has not been addressed in this context.

Additional Bayesian approaches have been explored to analyse variable fidelity data. Reese *et al.* (2004) simultaneously analysed the combined data (expert opinions, computer model and physical experiments) by using a hierarchical Bayesian integrated modelling approach. Posterior distributions generated by lower fidelity data are sequentially used as prior distributions in the next stage where higher fidelity data are used to construct posterior data. Experimental planning in the context of these hierarchical model fitting approaches has so far received little attention.

Other researchers have investigated the response surface generation from variable fidelity data by using regression. Etman (2000) recommended a central composite design (CCD) based structure, where a high fidelity analysis is carried out at the centre point of the design space, and low fidelity analyses are at the low and high levels and corners. Weighted least squares (WLS) was suggested for the analysis with higher weights subjectively assigned to higher fidelity runs. Rodriguez *et al.* (2001) explored two strategies for implementing designs based on the full factorial, CCD and orthogonal array structures. First, they suggested using random numbers to assign the runs from a design to different experimental systems, taking into account the availability of the total budget. Second, they investigated repeating runs of different systems at each point in the design structure.

In this paper, we focus on the design and analysis approaches for response surface generation. We explore the generation of new experimental plans without restricting ourselves either by imposing the hierarchical design restriction or by considering only designs that are similar to central composites. We propose the use of GLS for the analysis and the use of an extended EIMSE criterion for design.

3. Modelling and prediction

In this section, we define assumptions and models that are used for the analysis of experimental data and later for design generation. We begin by defining a general model of the responses from real and surrogate systems. An inspection of the properties of this model motivates the use of GLS. Next, we propose a covariance model of systematic errors partially extended from models in Kennedy and O'Hagan (2000). Finally, we discuss the physical interpretation of the hyperparameters in the model proposed, including relevant pre-estimations for engineering applications.

3.1. Response model

Assume that m experimental systems are available and the number of factors is d . The model for the response of the l th system (for $l = 1, \dots, m$) at the point $\mathbf{x} = (x_1, x_2, \dots, x_d)'$ is

$$\begin{aligned} Y(\mathbf{x}, l) &= \mathbf{f}_1(\mathbf{x})' \beta_1 + \mathbf{f}_2(\mathbf{x})' \beta_2 + Z_{\text{sys}}(\mathbf{x}, l) + \varepsilon_{\text{meas}}(l) \\ &= \mathbf{f}_1(\mathbf{x})' \beta_1 + \mathbf{f}_2(\mathbf{x})' \beta_2 + \varepsilon \end{aligned} \quad (1)$$

where $\mathbf{f}_1(\mathbf{x})' \beta_1 + \mathbf{f}_2(\mathbf{x})' \beta_2$ is the true function of the real systems, $Z_{\text{sys}}(\mathbf{x}, l)$ is the systematic error and $\varepsilon_{\text{meas}}(l)$ is the ordinary random error which might be 0 for computer experiments. For the true function, to allow for possible model misspecification, we assume that there are k_1 primary terms $\mathbf{f}_1(\mathbf{x})$ and k_2 potential terms $\mathbf{f}_2(\mathbf{x})$ and only the primary terms are included in the regression model to fit. Assuming that $l = 1$ corresponds to the real system of interest, $Z_{\text{sys}}(\mathbf{x}, 1) = 0$.

We follow Kennedy and O'Hagan (2001) in assuming that $Z_{\text{sys}}(\mathbf{x}, l)$ for $l > 1$ can be approximated by a realization of a Gaussian stochastic process. In this paper, we adopt the additional assumption that this Gaussian process has mean 0. Often, the experimenter will be comfortable with this assumption because the surrogate experimental systems are constructed to have minimal systematic errors and preliminary data have been used to validate them. Also, if it is known that systematic errors from a surrogate system are almost all positive (or negative), they could be approximated as a realization of a zero-mean Gaussian process with strong correlations.

More generally, the experimenters have limited information on the systematic errors and suspect that they may vary around a non-zero number, i.e. the responses are 'offset'. For these cases, we suggest a pre-step in which responses are collected from at least one point in the decision space by using the real and relevant surrogate systems. An offset value for each surrogate system is calculated from the observed differences. In subsequent experimentation, this offset value is added to all responses from that system such that the assumption of a zero mean becomes plausible.

An experimental design ξ is defined as a collection of n points in the input space, $\mathbf{x}_1, \dots, \mathbf{x}_n$, each of which is associated with a fidelity level labelled $l_1, \dots, l_n$ respectively. In equation (1), ε is defined to summarize all errors, both systematic and random, at the point $\mathbf{x}$ with fidelity level l . Let ε denote an n -dimensional vector of ε that is associated with the n runs in ξ having covariance $\mathbf{V} \equiv \text{var}(\varepsilon)$.

Assuming that $\mathbf{V}$ is known and $\mathbf{Y}$ is a vector that contains the data from the n experiments in ξ , the GLS estimator for β_1 is

$$\hat{\beta}_1 = (\mathbf{X}_1' \mathbf{V}^{-1} \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{V}^{-1} \mathbf{Y} \quad (2)$$

where $\mathbf{X}_1$ is the design matrix for the primary terms.

In this paper, we concentrate on resource-constrained situations such that the available data will be regarded as insufficient to estimate either the coefficients in β_2 or the covariance in $\mathbf{V}$. Thus, the model $\mathbf{f}_1(\mathbf{x})\hat{\beta}_1$ will be used for predicting real system outputs based on the estimator in equation (2), inserting a pre-estimated covariance matrix in place of the true covariance. A major concern is to develop methods yielding a performance that is robust to the true values of β_2 and $\mathbf{V}$.

3.2. Modelling covariance

We assume that the random errors $\varepsilon_{\text{meas}}(l)$ are independent of each other and independent of $Z_{\text{sys}}(\mathbf{x}, l)$. Therefore, the covariance of responses for run i at the point $\mathbf{x}_i$ with fidelity level l_i and run j at the point $\mathbf{x}_j$ with fidelity level l_j can be written

$$\text{cov}\{Y(\mathbf{x}_i, l_i), Y(\mathbf{x}_j, l_j)\} = \text{cov}\{Z_{\text{sys}}(\mathbf{x}_i, l_i), Z_{\text{sys}}(\mathbf{x}_j, l_j)\} + \delta_{i,j} \sigma(l_i)^2 \quad (3)$$

where $\sigma(l_i)$ is the standard deviation of the random error for the l_i th system and $\delta_{i,j} = 0$ for $i \neq j$ and $\delta_{i,j} = 1$ for $i = j$.

Kennedy and O'Hagan (2000) developed so-called 'autoregressive' assumptions about the systematic errors for different levels of fidelity, which satisfy

$$\text{cov}\{Z_{\text{sys}}(\mathbf{x}_i, l_i), Z_{\text{sys}}(\mathbf{x}_j, l_j)\} = \sigma_Z(l_i) \sigma_Z(l_j) \rho(l_i, l_j) R(l_i, l_j, \mathbf{x}_i, \mathbf{x}_j) \quad (4)$$

where $\sigma_Z(l_i)$ is the standard deviation of the systematic errors of system l_i . $\rho(l_i, l_j)$ can be understood as the correlation between systematic errors from system l_i and system l_j . $R(l_i, l_j, \mathbf{x}_i, \mathbf{x}_j)$ reflects the spatial correlation between point $\mathbf{x}_i$ and point $\mathbf{x}_j$, which also relates to system indices l_i and l_j . Note that Kennedy and O'Hagan (2000) focused on the case when responses at all levels

of fidelity come from computer experiments with different levels of sophistication. This caused them to make the additional assumption $\sigma(l) = 0$ for $l = 1, \dots, m$ which we do not necessarily make.

In this study, $R(l_i, l_j, \mathbf{x}_i, \mathbf{x}_j)$ is simplified to $R(\mathbf{x}_i, \mathbf{x}_j)$ for all $l_i, l_j = 2, \dots, m$ and is assumed to be given by

$$R(\mathbf{x}_i, \mathbf{x}_j) = \exp[-\theta\{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)\}] \quad (5)$$

where θ is a parameter that is associated with the ‘roughness’ of the systematic errors. In theory, both for different factors and for different systems, the roughness parameters could be different. However, we found that a fixed θ provided impressive results in our test problems, and system-dependent θ -parameters are difficult to obtain.

Taken together, the assumptions in equations (4) and (5) have properties that agree with our intuition about surrogate systems. $R(\mathbf{x}_i, \mathbf{x}_j)$, the correlation between systematic errors at point $\mathbf{x}_i$ and point $\mathbf{x}_j$, diminishes as the ‘distance’ between the two points becomes large. $\rho(l_i, l_j)$, the correlation between systematic errors from different systems, is also practically meaningful. For example, systematic errors from a 500-element and a 600-element FEA computer simulation (which have similar fidelity) are likely to be highly correlated. At the same time, systematic errors from a computer simulation and a real world laboratory experiment will probably have small or zero correlation.

3.3. Pre-estimation of hyperparameters

As mentioned earlier, we use a pre-estimated covariance matrix for the experimental design before data are collected as well as for analysis after data are collected. Pre-estimation of these parameters is feasible in engineering applications because engineers often have information relating to the repeatability and the faithfulness of their experimental system, e.g. through gauge repeatability studies and FEA theory.

The relevant hyperparameters include the $\sigma(l_i)$ for $l_i = 1, \dots, m$ in equation (3), the $\rho(l_i, l_j)$ and $\sigma_Z(l_i)$ in equation (4) and the θ in equation (5). By definition, $\sigma(l_i)$ is the standard deviation of the random errors. Therefore, if l_i corresponds to a computer experiment, often $\sigma(l_i) = 0$. The standard deviation of systematic error $\sigma_Z(l_i)$ reflects the general scale of systematic errors that are associated with surrogate system l_i . For example, an engineer may believe that outputs from a particular computer simulation contain less than 20% relative errors with respect to the physical experiments.

Similarly, engineers may be comfortable to guess the parameters $\rho(l_i, l_j)$ on the basis of their beliefs about how similar these experimental systems are. For example, if systems 1, 2 and 3 correspond to the physical and two computer experiments of similar type, we may assume that $\rho(1, 2) = \rho(1, 3) = 0$ and $\rho(2, 3) = 0.5$.

Sacks *et al.* (1989a) addressed the design of computer experiments with assumptions that the differences between a true model and a regression model are well approximated by Gaussian stochastic processes. (In this paper, the differences between the real and surrogate systems are approximated by Gaussian stochastic processes.) Using a robustness study, they motivated the choice $\theta = 1$ for planning start-up experiments, where all inputs were scaled to the $[-0.5, 0.5]^d$ hypercube. Here, if no better information is available, we also use $\theta = 1$ as the default pre-estimation.

Sometimes the readers might be able to pre-estimate θ on the basis of beliefs about the experimental systems. Fig. 3 shows one-dimensional realizations of stochastic processes for three different settings of θ . As mentioned previously, θ can be interpreted as measuring the degree of roughness of the systematic errors.

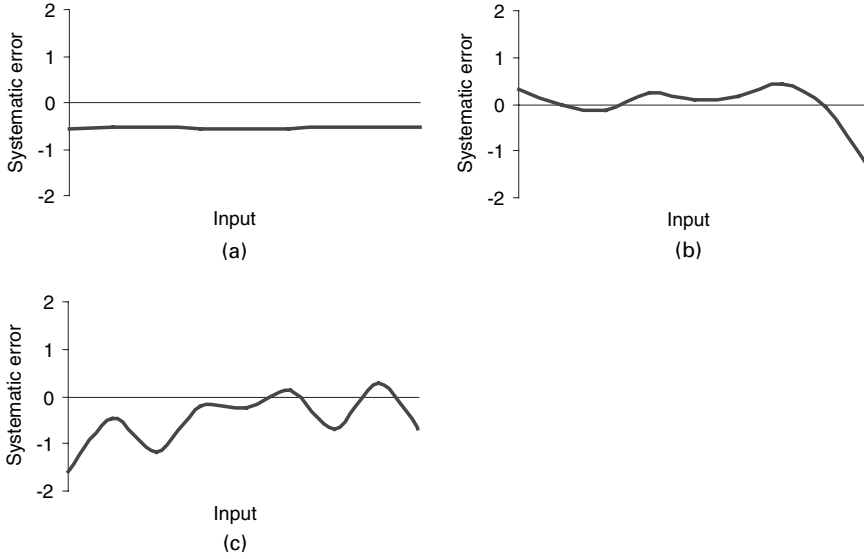


Fig. 3. Random realizations of the systematic errors for (a) $\theta = 0.1$, (b) $\theta = 1$ and (c) $\theta = 10$

We notice that, however, it may be possible to estimate or update these hyperparameters after the data have been collected. However, concerns include that the methods for estimating these parameters (for example, see Kennedy and O'Hagan (2000)) are associated with the hierarchical design restriction that was mentioned earlier. Also, the relevant number of parameters and hyperparameters can present a challenge for accurate estimation. For example, Kennedy and O'Hagan (2000) investigated the model involving terms for

- (a) the regression model,
- (b) the combined effects on the systematic errors of the factors and fidelity levels and
- (c) correlations between systematic errors for different levels of fidelity.

When the number of terms exceeds the number of runs, estimation can be difficult or impossible. We believe that this is an interesting topic for future study.

4. Experimental design

The experimental planning problem in the context of variable fidelity experimentation involves selecting the number of runs, n , the location of these runs in the input space, $\mathbf{x}_1, \dots, \mathbf{x}_n$, and the fidelity level that is associated with each point, $l_1, \dots, l_n$ respectively. In this section, we describe the criterion, additional assumptions and optimization methods that are used to generate experimental plans.

4.1. Design criterion

Of many possible formulations to generate experimental designs ξ , we use the EIMSE constrained by the total experimental cost TC. Therefore, the formulation can be written minimize

$$\text{EIMSE}(\xi) = E_{\eta, \mathbf{x}, \varepsilon} \{ \hat{y}(\mathbf{x}, \varepsilon, \xi, \eta) - \eta(\mathbf{x}) \}^2 \quad (6)$$

subject to

$$\sum_{i=1}^n C(l_i) \leq \text{TC}$$

where $\eta(\mathbf{x})$ and $\hat{y}(\mathbf{x}, \varepsilon, \xi, \eta)$ are the true function value and regression model prediction of the real system of interest at point $\mathbf{x}$ respectively. $C(l_i)$ is the cost per run of experimenting at fidelity level l_i .

The EIMSE criterion sums variance and bias errors in a single criterion and it has the simple interpretation of the expected squared errors. It was proposed in Allen and Yu (2002) and Allen, Yu and Schmitz (2003) for cases in which ordinary least squares (OLS) are applied for analysis. Allen, Bernshteyn and Kabiri (2003) showed that EIMSE optimal designs produced relatively low prediction errors in computer experiment case-studies from the literature.

Here, we extend the EIMSE formulation to consider GLS estimators of the form in equation (2) under the assumptions in equation (1). Denoting the design matrices for the primary terms and the potential terms as $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively, Appendix A shows the derivation of the generalized formula

$$\text{EIMSE}(\xi) = \text{tr}\{(\mathbf{A}'\boldsymbol{\mu}_{11}\mathbf{A} - 2\mathbf{A}'\boldsymbol{\mu}_{12} + \boldsymbol{\mu}_{22})E(\beta_2\beta_2')\} + \text{tr}\{\boldsymbol{\mu}_{11}(\mathbf{X}_1'\mathbf{V}^{-1}\mathbf{X}_1)^{-1}\} \quad (7)$$

where

$$\boldsymbol{\mu}_{ij} = \int_r \mathbf{f}_i(\mathbf{x}) \mathbf{f}_j(\mathbf{x})' d\mathbf{x} \quad \text{for } i, j = 1, 2 \quad (8)$$

and

$$\mathbf{A} = (\mathbf{X}_1'\mathbf{V}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}^{-1}\mathbf{X}_2$$

so that $\boldsymbol{\mu}_{ij}$ are the so-called ‘moment matrices’, r is the region of interest, $\mathbf{A}$ is the generalized ‘alias matrix’ and $\mathbf{V}$ is the covariance matrix.

4.2. Assumptions about the potential terms

The EIMSE is related to the coefficients of the potential terms only through the prior covariance matrix $E(\beta_2\beta_2')$. DuMouchel and Jones (1994) proposed assumptions about the prior covariance matrix that we also use. Their approach is based on adjustments of the magnitudes of each potential term so that they have comparable expected effects on the prediction errors. Chantararat (2003) showed that this assumption is equivalent to $E(\beta_2\beta_2') = \gamma^2\mathbf{B}^2$, where γ is a scalar that reflects the experimental planner’s estimation of the magnitude of the bias errors.

The matrix $\mathbf{B}$ is a diagonal scaling matrix with diagonal entries equal to the reciprocal of the difference between the minimum and the maximum values of the corresponding row of the matrix $(\mathbf{X}_2 - \mathbf{X}_1\mathbf{A})$ (which can be understood as the residual from $\mathbf{X}_2$ on $\mathbf{X}_1$). In the OLS context, a common choice is $\gamma = \sigma$, as suggested by DuMouchel and Jones (1994), where σ is the standard deviation of the random error. Yu (2000) indicated that an optimal EIMSE design will have a robust performance when the bias and variance components have comparable magnitudes.

4.3. Optimization methods

We use the genetic algorithm from Hadj-Alouane and Bean (1997) to generate the optimal designs, which are solutions to the program in equation (6). To permit flexibility about the shape of the search region, e.g. spherical or cuboidal, runs are selected from a finite set of

candidate points. Partly for simplicity in reporting, here we limit candidate points to nine evenly spaced levels in the $[-1, 1]^d$ hypercube.

5. Numerical test examples

5.1. Example 1

Consider a univariate example with a prespecified experimental design at two levels of fidelity, real (system 1) and surrogate (system 2). Assume that the true functions for the real and surrogate are both third-order polynomials as shown in Fig. 4(a). For simplicity, we also assume that the outputs for both systems have zero measurement errors. The prespecified eight-run design ξ and corresponding responses are shown in Fig. 4(b).

By definition, system 1 has no systematic error, i.e. $\sigma_Z(1)^2 = 0.0$. We assume that the pre-estimation on systematic errors of system 2 yields that $\sigma_Z(2)^2 = 0.01$. We also implement the default assumption that $\theta = 1$. Note that our assumptions imply that the total errors for the real system runs are 0, which causes $\mathbf{V}$ to be singular. To avoid this numerical problem, we implement a non-zero but negligible measurement error variance for the real system runs, e.g. $\sigma(1) = 10^{-6}$. It is easy to verify that the prediction results are insensitive to the particular choice of $\sigma(1)$.

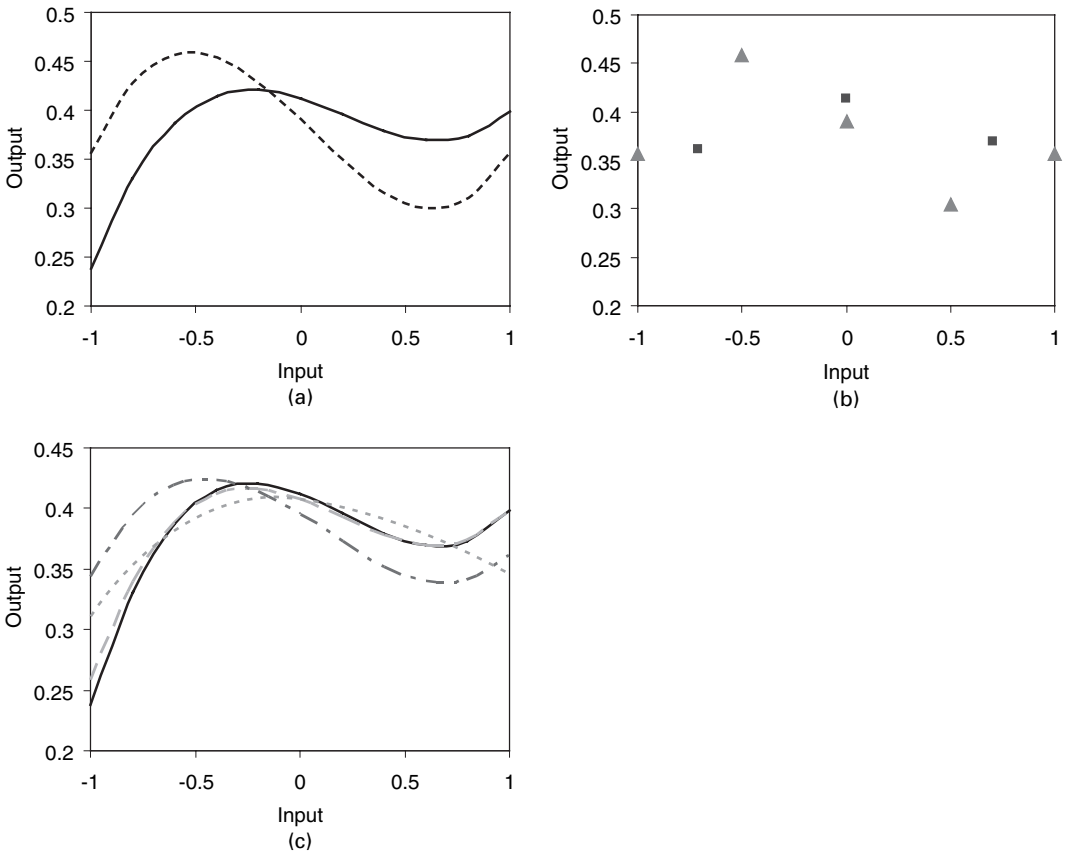


Fig. 4. (a) True functions (—, real system; ---, surrogate system), (b) experimental data (■, real system; ▲, surrogate system) and (c) predictions by using OLS (· · ·), WLS (-----) and GLS (— · —) (—, real system)

In this numerical study, to investigate the robustness of prediction with respect to the accuracy of the hyperparameter pre-estimates, we compute the maximum likelihood estimates (MLEs) of the hyperparameters by using the known true functions. Here the systematic errors are not a realization of a stochastic process, so we consider a stochastic process that would most probably generate the same systematic errors at nine evenly spaced points. From Sacks *et al.* (1989b), the likelihood is inversely proportional to $\det(\mathbf{V})^{1/n} \mathbf{Y}' \mathbf{V}^{-1} \mathbf{Y}$. Maximizing, we obtain estimates $\sigma_{Z, \text{MLE}}(2)^2 = 0.0044$ and $\theta_{\text{MLE}} = 1.396$. Therefore, the pre-estimates that are used in the analysis, $\sigma_Z(2)^2 = 0.01$ and $\theta = 1.0$, only approximately correspond to the likelihood optimal choices.

Fig. 4(c) shows predictions of the real system outputs based on the data in Fig. 4(b). In addition to the GLS prediction, we include the WLS and OLS predictions for comparison. (Note that WLS and OLS predictions can be derived from equation (2) by assuming that the off-diagonal terms in $\mathbf{V}$ are 0 and $\mathbf{V} = \mathbf{I}$ respectively.) Despite the fact that the pre-estimated hyperparameters are not ideal, GLS provides an apparent advantage in prediction accuracy. This robustness justifies the use of pre-estimated $\mathbf{V}$ to some degree.

5.2. Example 2

The second example is a hypothetical two-factor case with randomly generated true functions from McDaniel and Ankenman (2000). This example permits us to evaluate quantitatively the performance of the methods proposed and to compare them with alternative approaches. In this example, we assume a scenario where data can be collected from three experimental systems: the real, a higher fidelity surrogate and a lower fidelity surrogate. The systematic errors of the two surrogate systems are correlated to some degree. Each experimental run costs \$4000, \$2000 or \$1000 on systems 1, 2 or 3 respectively and the total budget is \$15000. The region of interest is a $[-1, 1]^2$ square.

Assume that the experimenter's knowledge yields the following 'best guesses': $\sigma(1)^2 = 0.2$, $\sigma(2)^2 = \sigma(3)^2 = 0.0$, $\sigma_Z(1)^2 = 0$, $\sigma_Z(2)^2 = 1$, $\sigma_Z(3)^2 = 4$, $\theta = 1$ and

$$\rho = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{pmatrix}. \quad (9)$$

In addition, we consider that the primary terms $\mathbf{f}_1(\mathbf{x})'/\beta_1$ correspond to full quadratic polynomials and the potential terms $\mathbf{f}_2(\mathbf{x})'/\beta_2$ contain all third-order effects. For experimental design generation, the parameter γ is adjusted to make the variance and bias roughly equal, which occurs when $\gamma = 2$. Using the EIMSE design criterion, the genetic algorithm of Hadj-Alouane and Bean (1997) generated the putatively optimal solution to the program in equation (6) in Fig. 5.

For comparison, we consider three alternative designs. The first design is a 'manually' partitioned central composite design (MPCCD) based on recommendations of Etman (2000). As shown in Fig. 6(a), the runs are allocated to the centre, star and factorial points in order of their fidelity level so that the highest fidelity runs are in the centre. The second approach is based on methodologies of Rodriguez *et al.* (2001), where CCD runs are randomly allocated to different experimental systems while satisfying the total budget constraint, as shown in Fig. 6(b). We call it the randomly partitioned central composite design (RPCCD). The third design is a 'shrunk' MPCCD, which we include to clarify the separate effects of model bias and systematic errors. By shrinking the design, bias errors are reduced. Note that the designs by Etman (2000) and Rodriguez *et al.* (2001) were proposed with the intended analysis methods of WLS and OLS, respectively. But in this paper, for generality, we evaluate all combinations of the design and analysis methods.

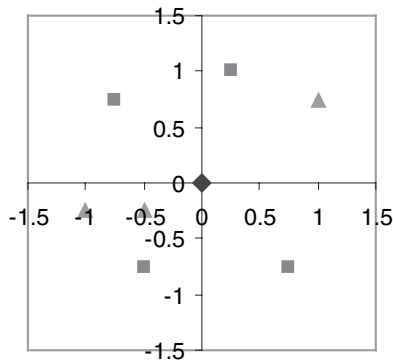


Fig. 5. Optimum design of the experiment (eight runs): $\blacklozenge$, system 1; $\blacksquare$, system 2; $\blacktriangle$, system 3

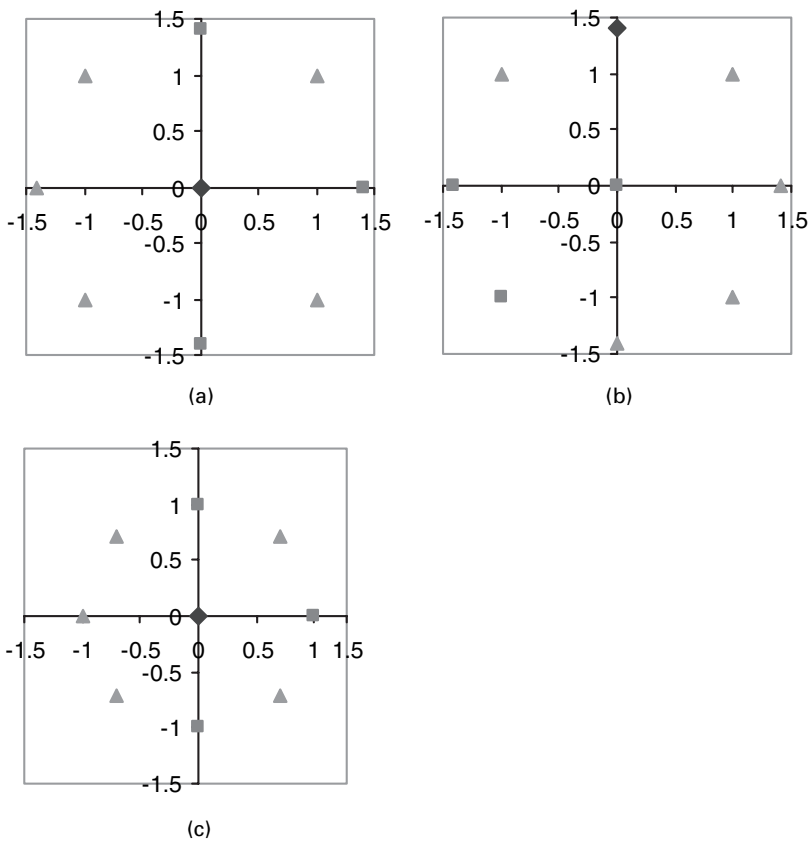


Fig. 6. (a) MPCCD, (b) RPCCD and (c) shrunk MPCCD: $\blacklozenge$, system 1; $\blacksquare$, system 2; $\blacktriangle$, system 3

To evaluate the performance of these designs, we assume that the true functions are polynomials randomly generated by using the ‘response surface test-bed’ of McDaniel and Ankenman (2000). We use the default flatness and effect heredity that were suggested by McDaniel and Ankenman (2000). The magnitude of the coefficients in their method is regulated by the ‘range-of-response’ parameters. In our study, we use the range of response $(-1.0, 1.0)$ to generate the true functions of the real system. The true functions of the surrogate systems are generated by

adding polynomials to the true function of the real system. These polynomials, which correspond to the systematic errors, are also generated by the ‘test-bed’, but with a relatively smaller response range.

We evaluate the method performances considering three classes of test functions. The first class derives from the scenarios in which the fitted model forms have non-negligible bias and the pre-estimates of the hyperparameters are approximately accurate. Thus, the responses of the real system (system 1) derive from third-order polynomials. The polynomials corresponding to the systematic errors of system 2 are also third order and are created by using a range of responses $(-0.1, 0.1)$. The polynomials for the systematic errors of system 3 are the sum of the error polynomial for system 2 and an additional randomly generated polynomial with response ranges $(-0.173, 0.173)$. Therefore, the systematic errors for system 3 were generated with 4.0 times the variance of system 2 and with correlation equal to 0.5. These correspond to the values assumed. The coefficients of all sets of true functions are in Table 1.

In the second class, we consider scenarios with zero bias errors and approximately accurate pre-estimates of the hyperparameters. Therefore, the responses of the real system derive from second-order polynomials. The responses of systems 2 and 3 are generated by using the same scheme as in the first class.

For the third class, we consider scenarios in which the bias errors are non-negligible and pre-estimates about the hyperparameters are very poor. Therefore, the responses of the real system derive from third-order polynomials, as for the first class. The systematic errors of system 2 are third-order polynomials generated with range of response $(-0.2, 0.2)$. The systematic errors of system 3 are generated independently with a range of response $(-0.1, 0.1)$. Therefore, counter to the belief that is used in experimental design generation and analysis, the systematic errors of system 3 are generally larger than those of system 2 and they are generated with zero correlation.

Following Sacks *et al.* (1989a) the prediction accuracies of the various method combinations are evaluated by using the empirical integrated squared error EISE at all points on a 101×101 grid of points in the region of interest:

$$\text{EISE} = 101^{-2} \sum_{i=1}^{101 \times 101} \{\hat{y}(\mathbf{x}_i) - \eta(\mathbf{x}_i)\}^2. \quad (10)$$

For every test set, we compute the EISE for all nine combinations of design and analysis methods. We define the ‘relative EISE’ as the ratio of the EISE of the method being evaluated to the EISE of the method proposed, i.e. the combination of an optimal EIMSE design and GLS analysis. Table 2 shows the EISE and averaged relative EISE over five test sets for each of the three classes.

Table 2 suggests the following findings about the performance of the various combinations of methods. For the first class of test functions, bias apparently causes the regular central-composite-based designs to deliver relatively poor prediction performance regardless of the method of analysis. For optimal EIMSE design and ‘shrunk MPCCD’, which address bias, GLS analysis incorporated with approximately correct information offers an improved performance. Yet, using regular central-composite-based designs, the benefits of GLS are not readily apparent. Therefore, there might be an interaction between the choice of design and the method of analysis on the average EISE performance. We suggest investigating the performance of GLS in the presence of model bias and alternative experimental designs as a topic for future study.

Test classes II and III represent two types of departure from the assumptions that are used for EIMSE optimal design generation and GLS analysis. Class II may be regarded as a ‘perfect model’ case in which the bias errors are negligible and the pre-estimates are approximately correct. For this case, it is not surprising that the performance of the design is predictable by using

Table 1. Coefficients of the true functions

Test set l		Coefficients									
		I	x_1	x_2	x_1^2	x_2^2	x_1x_2	x_1^3	x_2^3	$x_1^2x_2$	$x_1x_2^2$
<i>Class I</i>											
1	1	0.1029	-0.1489	-0.1432	-0.1648	0.0894	-0.1380	0.0116	-0.0376	0.0646	0.0713
	2	0.1061	-0.1532	-0.0810	-0.1611	0.0928	-0.1469	0.0115	-0.0464	0.0572	0.0744
	3	0.1145	-0.1578	-0.0788	-0.1605	0.0931	-0.1681	0.0202	-0.0435	0.0500	0.0709
2	1	-0.0084	0.0026	0.3568	-0.1248	0.0802	0.0014	-0.0068	-0.0147	-0.1536	0.0241
	2	-0.0243	0.0206	0.3477	-0.1030	0.0776	-0.0103	-0.0219	-0.0164	-0.1430	0.0329
	3	-0.0660	0.0574	0.3503	-0.0798	0.0748	-0.0010	-0.0375	-0.0124	-0.1454	0.0376
3	1	-0.1143	0.5447	0.0899	0.0342	-0.0012	-0.035	-0.1062	-0.0026	-0.0439	-0.1403
	2	-0.1157	0.5007	0.1211	0.0425	0.0052	-0.029	-0.0901	-0.0086	-0.0409	-0.1386
	3	-0.0853	0.4688	0.1270	0.0206	0.0183	-0.0361	-0.0785	-0.0061	-0.0555	-0.1329
4	1	0.0903	0.2130	0.1080	-0.0652	-0.0357	-0.0276	-0.0922	0.0060	-0.0036	-0.0207
	2	0.0629	0.2049	0.1213	-0.0547	-0.0438	-0.0285	-0.0906	0.0037	-0.0012	-0.0191
	3	0.1207	0.0976	0.0926	-0.0725	-0.0433	-0.0285	-0.0665	0.0089	0.0132	-0.0122
5	1	0.1396	-0.0016	-0.2508	-0.0501	0.0083	0.0776	-0.0156	-0.0063	0.02	0.0462
	2	0.1078	0.0097	-0.2746	-0.0479	0.0227	0.0803	-0.0161	0.0015	0.0166	0.0465
	3	0.1905	0.0945	-0.3207	-0.0488	-0.0112	0.1071	-0.0280	0.0196	0.0256	0.0242
<i>Class II</i>											
1	1	-0.2078	0.1548	-0.1888	0.042	0.0914	-0.0935	0	0	0	0
	2	-0.245	0.1392	-0.1865	0.0528	0.1009	-0.1016	0.0058	-0.0012	0.0047	0.0037
	3	-0.2032	0.2201	-0.1124	0.0279	0.0960	-0.1102	-0.0200	-0.0197	-0.0112	-0.0194
2	1	-0.0623	0.0113	-0.1094	0.0656	-0.0552	-0.1138	0	0	0	0
	2	-0.0490	0.0136	-0.1154	0.0566	-0.0537	-0.1119	-0.0016	0.0004	0.0049	0.0055
	3	-0.0580	-0.0423	-0.0750	0.0549	-0.0457	-0.1127	0.0131	-0.0044	-0.0044	0.0002
3	1	-0.1941	-0.0478	0.0042	0.1111	-0.0593	0.1547	0	0	0	0
	2	-0.1584	-0.0392	-0.0132	0.0852	-0.0710	0.1517	-0.0042	0.0042	0.0119	-0.0040
	3	-0.1406	-0.0649	-0.0380	0.0567	-0.0555	0.1278	-0.0022	-0.0023	0.0231	0.0084
4	1	-0.4564	-0.0549	-0.2178	0.1469	0.1615	0.0361	0	0	0	0
	2	-0.4401	-0.0694	-0.2043	0.1415	0.1505	0.0396	0.0066	-0.0034	-0.0078	-0.0014
	3	-0.4415	-0.0689	-0.1426	0.1199	0.1644	0.0398	0.0054	-0.0060	-0.0344	0.0028
5	1	-0.1011	-0.0437	0.0021	-0.0223	0.0468	-0.0111	0	0	0	0
	2	-0.1296	-0.0262	0.0143	-0.0189	0.0535	-0.0198	-0.0007	-0.0046	0.0035	0.0026
	3	-0.1366	-0.0355	0.0515	-0.0088	0.0487	-0.0125	-0.0038	-0.0110	0.0017	0.0111
<i>Class III</i>											
1	1	0.1635	-0.1443	0.1348	-0.0536	-0.1098	0.0344	0.0662	-0.0341	-0.0785	-0.0141
	2	0.1911	-0.0885	0.1038	-0.082	-0.1056	0.0175	0.0468	-0.0324	-0.0728	-0.0182
	3	0.1725	-0.1230	0.1456	-0.0601	-0.1134	0.0317	0.0570	-0.0335	-0.0789	-0.0162
2	1	-0.2853	0.1758	0.1225	0.0338	0.0673	-0.0876	-0.0069	-0.0462	0.0348	0.0261
	2	-0.3186	0.2018	0.065	0.0304	0.0943	-0.0925	-0.004	-0.0338	0.0423	0.0072
	3	-0.2714	0.1757	0.0974	0.0288	0.0682	-0.0799	-0.0085	-0.0468	0.0367	0.0308
3	1	-0.0929	-0.0545	0.183	0.0433	0.0449	-0.0552	-0.008	-0.068	-0.0753	0.0342
	2	-0.1033	-0.1192	0.2297	0.0414	0.0541	-0.0562	0.0089	-0.0735	-0.0862	0.028
	3	-0.0705	-0.0645	0.1930	0.0457	0.0410	-0.0632	-0.0062	-0.0750	-0.0698	0.0400
4	1	0.2745	0.2119	-0.2003	0.0076	-0.0793	0.1113	-0.0561	0.0039	0.1069	-0.1066
	2	0.2951	0.1821	-0.2289	-0.0253	-0.0614	0.0837	-0.0538	-0.0036	0.1199	-0.0924
	3	0.2776	0.2075	-0.1381	0.0113	-0.0758	0.1024	-0.0562	-0.0050	0.0995	-0.1036
5	1	0.1096	-0.2054	-0.1678	-0.0386	0.0337	-0.0576	0.0095	0.0131	0.0437	0.059
	2	0.108	-0.2049	-0.0964	-0.0635	0.0498	-0.0573	0.0081	0.0102	0.013	0.0638
	3	0.0937	-0.1875	-0.1768	-0.0168	0.0312	-0.0692	-0.0057	0.0114	0.0544	0.0678

Table 2. EISE for the two-factor test functions

Test function class	Test function	Analysis	EISEs for the following designs:			
			Optimal EIMSE design	MPCCD design of Etman (2000)	RPCCD design of Rodriguez et al. (1998)	Shrunk MPCCD
I	1	GLS	1.49×10^{-3}	5.49×10^{-4}	1.15×10^{-3}	7.21×10^{-4}
		WLS	1.57×10^{-3}	5.11×10^{-4}	1.20×10^{-3}	8.10×10^{-4}
		OLS	1.69×10^{-3}	7.61×10^{-4}	6.75×10^{-4}	1.18×10^{-3}
	2	GLS	1.26×10^{-3}	8.01×10^{-4}	1.00×10^{-3}	1.52×10^{-3}
		WLS	1.56×10^{-3}	8.09×10^{-4}	9.51×10^{-4}	1.69×10^{-3}
		OLS	1.40×10^{-3}	1.44×10^{-3}	1.28×10^{-3}	1.59×10^{-3}
	3	GLS	2.46×10^{-3}	7.43×10^{-3}	7.43×10^{-3}	2.30×10^{-3}
		WLS	2.59×10^{-3}	7.62×10^{-3}	7.55×10^{-3}	2.60×10^{-3}
		OLS	3.03×10^{-3}	8.03×10^{-3}	7.97×10^{-3}	2.56×10^{-3}
	4	GLS	8.45×10^{-4}	8.46×10^{-3}	6.98×10^{-3}	3.70×10^{-3}
		WLS	1.00×10^{-5}	8.33×10^{-3}	6.54×10^{-3}	3.88×10^{-3}
		OLS	2.17×10^{-5}	7.71×10^{-3}	6.18×10^{-3}	3.65×10^{-3}
	5	GLS	1.38×10^{-5}	5.14×10^{-4}	1.57×10^{-3}	9.77×10^{-4}
		WLS	1.50×10^{-5}	4.92×10^{-4}	1.69×10^{-3}	6.51×10^{-4}
		OLS	1.73×10^{-5}	7.46×10^{-4}	2.02×10^{-3}	1.19×10^{-3}
	Averaged relative†	GLS	100% (0%)	288% (370%)	280% (285%)	154% (144%)
		WLS	112% (8%)	286% (356%)	272% (265%)	160% (153%)
		OLS	146% (56%)	292% (326%)	270% (249%)	165% (134%)
II	1	GLS	4.10×10^{-4}	2.38×10^{-4}	2.53×10^{-4}	1.13×10^{-3}
		WLS	5.62×10^{-4}	2.35×10^{-4}	2.61×10^{-4}	7.55×10^{-4}
		OLS	1.01×10^{-3}	3.91×10^{-4}	4.16×10^{-4}	8.50×10^{-4}
	2	GLS	7.47×10^{-5}	1.57×10^{-4}	2.53×10^{-4}	4.17×10^{-4}
		WLS	9.88×10^{-5}	1.64×10^{-4}	2.61×10^{-4}	3.72×10^{-4}
		OLS	1.63×10^{-4}	2.60×10^{-4}	4.16×10^{-4}	4.45×10^{-4}
	3	GLS	3.94×10^{-4}	1.34×10^{-4}	1.21×10^{-3}	4.60×10^{-4}
		WLS	4.31×10^{-4}	1.40×10^{-4}	1.29×10^{-3}	5.58×10^{-4}
		OLS	5.96×10^{-4}	1.81×10^{-4}	1.34×10^{-3}	6.59×10^{-4}
	4	GLS	6.56×10^{-5}	6.56×10^{-5}	2.82×10^{-4}	7.03×10^{-5}
		WLS	5.95×10^{-5}	7.69×10^{-5}	3.15×10^{-4}	1.35×10^{-4}
		OLS	5.16×10^{-5}	1.70×10^{-4}	4.10×10^{-4}	3.72×10^{-4}
	5	GLS	4.61×10^{-4}	7.82×10^{-5}	1.10×10^{-3}	3.31×10^{-4}
		WLS	6.23×10^{-4}	9.91×10^{-5}	1.06×10^{-3}	4.40×10^{-4}
		OLS	7.15×10^{-4}	2.28×10^{-4}	1.07×10^{-3}	6.36×10^{-4}
	Averaged relative†	GLS	100% (0%)	84% (69%)	275% (123%)	226% (180%)
		WLS	121% (18%)	90% (73%)	290% (138%)	225% (141%)
		OLS	170% (58%)	160% (122%)	371% (196%)	335% (202%)
III	1	GLS	1.63×10^{-3}	3.21×10^{-3}	4.05×10^{-3}	1.45×10^{-3}
		WLS	1.60×10^{-3}	3.07×10^{-3}	3.47×10^{-3}	1.20×10^{-3}
		OLS	1.57×10^{-3}	2.56×10^{-3}	2.74×10^{-3}	7.78×10^{-4}
	2	GLS	1.12×10^{-3}	3.31×10^{-3}	2.68×10^{-3}	1.90×10^{-3}
		WLS	1.14×10^{-3}	2.94×10^{-3}	2.55×10^{-3}	1.52×10^{-3}
		OLS	1.16×10^{-3}	1.52×10^{-3}	1.81×10^{-3}	8.21×10^{-4}
	3	GLS	1.33×10^{-3}	1.25×10^{-3}	2.05×10^{-3}	1.69×10^{-3}
		WLS	1.36×10^{-3}	1.24×10^{-3}	2.12×10^{-3}	1.39×10^{-3}
		OLS	1.41×10^{-3}	1.35×10^{-3}	2.22×10^{-3}	8.68×10^{-4}

(continued)

Table 2 (continued)

Test function class	Test function	Analysis	EISEs for the following designs:			
			Optimal EIMSE design	MPCCD design of Etman (2000)	RPCCD design of Rodriguez et al. (1998)	Shrunk MPCCD
	4	GLS	1.92×10^{-3}	3.98×10^{-3}	2.13×10^{-3}	2.34×10^{-3}
		WLS	1.36×10^{-3}	3.66×10^{-3}	2.13×10^{-3}	1.67×10^{-3}
		OLS	9.71×10^{-4}	3.52×10^{-3}	3.54×10^{-3}	1.04×10^{-3}
	5	GLS	1.57×10^{-3}	1.66×10^{-3}	1.04×10^{-3}	1.38×10^{-3}
		WLS	1.46×10^{-3}	1.55×10^{-3}	1.02×10^{-3}	1.07×10^{-3}
		OLS	1.34×10^{-3}	1.12×10^{-3}	9.15×10^{-4}	5.36×10^{-4}
Averaged relative†		GLS	100% (0%)	180% (74%)	164% (71%)	119% (30%)
		WLS	93% (12%)	167% (64%)	155% (61%)	94% (24%)
		OLS	88% (20%)	130% (40%)	148% (45%)	55% (14%)

†Standard deviations are given in parentheses.

the integrated variance, which is the second term in equation (7) and constitutes the expected squared prediction errors assuming zero contributions from model bias. The integrated variance for the optimal EIMSE design, MPCCD, RPCCD and shrunk MPCCD were 0.80, 0.53, 0.94 and 0.85 respectively, assuming that GLS analysis is used. The relative performance of the methods of analysis also follows a predictable pattern as dictated by the degree to which the method capitalizes on the approximately correct information. GLS methods perform the best followed by WLS methods for all design and analysis combinations.

The third class of test functions provides evidence for some measure of the robustness of the proposed methods to poor pre-estimates of the covariance matrix $\mathbf{V}$. For all choices of design, the pattern that is observed for the second class is reversed such that GLS and WLS methods perform worse than OLS methods. This suggests that incorporating qualitatively incorrect information about the surrogate systems in analysis is generally worse than not utilizing any information. Yet, the losses might be regarded as small. For example, when EIMSE optimal designs are used, the losses are less than 12% from using GLS. Also, we observed some prediction advantages of optimal EIMSE designs over regular central-composite-based designs, despite the incorporation of the incorrect information. This occurs presumably because bias incorporated in the EIMSE criterion causes the design points to move closer to the centre of the design region. When bias errors dominate, this may offer advantages of prediction regardless of what method of analysis is applied. In addition, for the MPCCD, we also find that false prior information has negative effects on the analysis (i.e. OLS is better than GLS). Furthermore, an MPCCD plus OLS has the relatively lowest prediction error of all. This is probably because the MPCCD provides some protection against the bias, and, unlike the optimal EIMSE design, its generation is not affected by false prior information about the covariance.

To summarize, the results in Table 2 show that, when bias errors are not negligible and pre-estimates of hyperparameters are approximately accurate, the combination of optimal EIMSE designs and GLS analysis proposed can offer substantial advantages of prediction accuracy compared with the other combinations of methods that were studied. Further, the results suggest some degree of robustness of proposed methods to misspecification of the assumptions incorporated. We also found that, when the model bias is zero, the central-composite-based designs

that were proposed by Etman (2000) offer advantages of prediction compared with optimal EIMSE designs, presumably because of their lower integrated variance.

5.3. Example 3

The third example is a three-factor problem that was motivated by the engine valve heat treatment application that was described in Section 1. A pre-step is studied in which the surrogate system responses are adjusted to make more plausible the assumption that the systematic errors have zero mean.

The overall goal of the associated application was to generate an accurate response surface prediction of the distortion of the part after heat treatment. Data could be collected by using physical experiments and from the DEFORM[®] software, commercial FEA code. There were three levels of experimental fidelity:

- (a) physical,
- (b) high resolution simulation and
- (c) low resolution simulation

with associated experiment costs per run of \$1200, \$400 and \$200 respectively. The total budget was \$5000.

The response of interest was the change in dimension H before and after the heat treatment process (Fig. 2(a)). The region of interest was T_A (750–850 °C), T_T (350–450 °C) and t_T (30–200 min). Our experience indicated that the response was usually of the order of 0.1 (mm).

On the basis of previous experience on similar physical experiments, we estimated that $\sigma(1)^2 = 2.5 \times 10^{-5}$ (mm²), $\sigma(2)^2 = \sigma(3)^2 = 0.0$ and $\sigma_Z(1)^2 = 0$. The engineers also believed that the high resolution computer model should have less than 10% error and the low resolution model should have less than 15% error. Thus, we estimated $\sigma_Z(2)^2 = 1.0 \times 10^{-4}$ (mm²) and $\sigma_Z(3)^2 = 2.25 \times 10^{-4}$ (mm²). In addition, without sufficient prior information on the roughness parameter, we used the default assumption that $\theta = 1$. The engineers were confident that the systematic errors would be correlated between the computer codes and expressed comfort with the assumption

$$\rho = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{pmatrix}. \quad (11)$$

In addition, a full second-order polynomial regression model was used for the analysis, with the set of potential terms including all third-order interactions. To assure robustness of the design, we used $\gamma^2 = 2.25 \times 10^{-4}$ (mm²) so that the variance and bias were roughly equal. The putatively optimal EIMSE design has 11 runs as displayed in Table 3, where x_1 , x_2 and x_3 correspond to T_A , T_T and t_T respectively in coded units.

In this case-study, owing to our limited information on the errors of computer models, a pre-step calibration as described in Section 2 may be considered. The pre-step that we use involves a single test run from all systems at the centre of the region of interest. Then, all responses from surrogate systems are adjusted by the offsets that were determined from the pre-step responses.

Similar to the second numerical test example, we compare several alternative methods by using randomly generated true functions. Here, we shall compare approaches with and without the pre-step. And, for each approach, two design strategies are considered: the optimal EIMSE design and the shrunk MPCCD. The designs are listed in Table 3. For the with-pre-step optimal

Table 3. Three-factor, three-system experiment designs

<i>Run</i>	<i>System</i>	x_1	x_2	x_3	<i>Run</i>	<i>System</i>	x_1	x_2	x_3
<i>(a) No pre-step, optimal EIMSE design</i>					<i>(b) No pre-step, shrunk MPCCD</i>				
1	1	0	0	0	1	1	0	0	0
2	3	1	0.75	0	2	2	1	0	0
3	2	0	1	-1	3	2	-1	0	0
4	2	0.75	-0.25	-1	4	2	0	1	0
5	3	0.75	-1	0.5	5	2	0	-1	0
6	2	0.75	-0.75	1	6	2	0	0	1
7	3	0.75	0.75	1	7	2	0	0	-1
8	3	-1	1	0	8	3	0.58	0.58	0.58
9	3	1	-0.75	-0.5	9	3	-0.58	0.58	0.58
10	3	-1	-1	0.25	10	3	0.58	-0.58	0.58
11	2	-0.25	0.75	0.75	11	3	0.58	0.58	-0.58
12	3	-1	0.25	-1	12	3	0.58	-0.58	-0.58
13	3	-0.25	-1	-1	13	3	-0.58	0.58	-0.58
14	2	-0.75	-0.5	0.75	14	3	-0.58	-0.58	0.58
15	3	1	-0.5	0.75					
<i>(c) With pre-step, optimal EIMSE design</i>					<i>(d) With pre-step, shrunk MPCCD</i>				
1†	1	0	0	0	1†	1	0	0	0
2†	2	0	0	0	2†	2	0	0	0
3†	3	0	0	0	3†	3	0	0	0
4	3	0.75	-0.75	-1	4	2	1	0	0
5	3	1	1	-0.25	5	2	-1	0	0
6	2	-0.5	1	-0.25	6	2	0	1	0
7	2	1	0	0.5	7	2	0	-1	0
8	3	-1	0.5	-0.75	8	2	0	0	1
9	3	0	-1	1	9	2	0	0	-1
10	3	0.75	0.75	-1	10	3	0.58	0.58	0.58
11	2	0	-0.75	-1	11	3	-0.58	-0.58	0.58
12	3	1	-1	0.25	12	3	0.58	-0.58	-0.58
13	3	-1	0.25	0.75	13	3	-0.58	0.58	-0.58
14	2	0.25	0.75	1					
15	3	-1	-1	0					

†Pre-step design point.

EIMSE design (Table 3, part(c)), the experimental design is planned optimally, taking into account the runs at the centre from the pre-step.

For evaluation, we consider true functions that are full third-order polynomials, randomly generated by using the approach in McDaniel and Ankenman (2000). The true functions of the real system (system 1) are created by using the range of responses $(-1.0, 1.0)$. The polynomials corresponding to the systematic errors of system 2 are created by using the range of responses $(0, 0.2)$. The polynomials for the systematic errors of system 3 are generated as the sum of the error polynomial for system 2 and an additional randomly generated polynomial with response range $(0, 0.2)$. Note that both surrogate systems are positively biased. In addition, the variance of the systematic error for system 3 is about 2.0 times that of system 2, and the correlation between the systematic error for system 3 and for system 2 is about 0.707. Therefore, our prior information on sizes and correlations of the systematic errors are approximately accurate, but the zero-mean assumption on the systematic errors is poor.

The results in Table 4 suggest that the pre-step substantially improves the accuracy of prediction regardless of the approach of analysis. For all designs, GLS analysis provides better

Table 4. EISE for the three-factor test functions

Test function	Analysis	EISEs for the following designs:			
		No pre-step, optimal EIMSE design	No pre-step, shrunk MPCCD	With pre-step, optimal EIMSE design	With pre-step, shrunk MPCCD
1	GLS	7.25×10^{-3}	1.64×10^{-2}	1.86×10^{-3}	5.14×10^{-3}
	WLS	7.70×10^{-3}	1.88×10^{-2}	2.14×10^{-3}	5.16×10^{-3}
2	OLS	8.24×10^{-3}	2.20×10^{-2}	2.29×10^{-3}	5.20×10^{-3}
	GLS	9.31×10^{-3}	2.13×10^{-2}	9.65×10^{-4}	1.15×10^{-3}
3	WLS	9.83×10^{-3}	2.43×10^{-2}	9.98×10^{-4}	1.18×10^{-3}
	OLS	1.04×10^{-2}	2.75×10^{-2}	1.03×10^{-3}	1.27×10^{-3}
4	GLS	1.25×10^{-2}	1.66×10^{-2}	2.58×10^{-3}	3.90×10^{-3}
	WLS	1.31×10^{-2}	1.93×10^{-2}	2.72×10^{-3}	3.95×10^{-3}
5	OLS	1.27×10^{-2}	2.40×10^{-2}	3.03×10^{-3}	4.09×10^{-3}
	GLS	1.02×10^{-2}	1.92×10^{-2}	4.83×10^{-3}	9.95×10^{-3}
Averaged relative†	WLS	1.08×10^{-2}	2.15×10^{-2}	5.00×10^{-3}	9.95×10^{-3}
	OLS	1.18×10^{-2}	2.42×10^{-2}	5.19×10^{-3}	9.96×10^{-3}
	GLS	1.16×10^{-2}	2.28×10^{-2}	3.86×10^{-3}	7.34×10^{-3}
	WLS	1.14×10^{-2}	2.55×10^{-2}	3.73×10^{-3}	7.39×10^{-3}
	OLS	1.10×10^{-2}	2.83×10^{-2}	4.02×10^{-3}	7.56×10^{-3}
	GLS	470% (264%)	944% (650%)	100% (0%)	189% (53%)
	WLS	492% (281%)	1077% (742%)	105% (6%)	190% (53%)
	OLS	508% (298%)	1239% (835%)	112% (7%)	194% (50%)

†Standard deviations are given in parentheses.

accuracy, presumably because the underlying correlation structure can model the systematic errors to some extent. Still, in these cases, the benefit from using GLS is much less than the benefit from using the pre-step. In addition, optimal EIMSE designs produce better prediction accuracy than corresponding shrunk MPCCDs.

6. Application study: engine valve process design

The generation of the experimental designs for the engine valve heat treatment application is described in the numerical test example 3 of the previous section. The optimal EIMSE design with pre-step (Table 3, part(c)) combined with GLS analysis offered the best performance of all the methods considered. Therefore, physical and computer experiments were conducted according to this design.

The associated response data of the optimal EIMSE design, in the order of Table 3, part(c), are the following: 36.57, 38.86, 39.33, 57.05, 51.14, 29.69, 53.37, 33.66, 40.32, 47.30, 45.06, 61.48, 30.74, 33.84 and 40.96 ($\times 10^{-3}$ mm). From these data, we computed the following GLS prediction (in 10^{-3} mm) of the distortion:

$$\begin{aligned} \hat{y}(x_1, x_2, x_3) = & 36.54 + 10.46x_1 - 5.85x_2 - 2.82x_3 - 0.01x_1x_2 - 0.02x_1x_3 \\ & + 0.69x_2x_3 + 6.22x_1^2 + 0.21x_2^2 - 1.51x_3^2. \end{aligned} \quad (12)$$

where x_1 , x_2 and x_3 correspond to T_A , T_T and t_T respectively in coded units. A plot of this model in Fig. 7 shows the effects of the heat treatment process parameters on the distortion of the

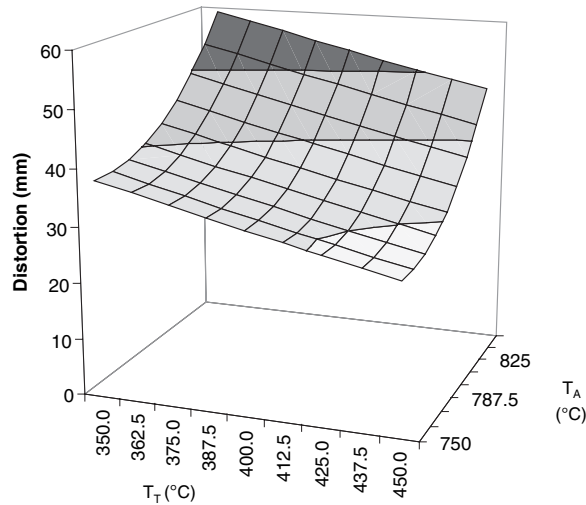


Fig. 7. Predicted distortion of the valve after heat treatment with $t_T=115$ min

part. This response surface prediction permits engineers to minimize distortion of the part with a constraint on hardness.

7. Conclusions and future work

In this paper, we proposed the use of GLS to generate prediction when data are from variable fidelity experiments. We also proposed perhaps the first criterion for design optimality of variable fidelity experimentation. The criterion is a GLS extension of the EIMSE criterion that was proposed in Yu (2000) and Allen, Yu and Schmitz (2003). We showed that the design methods proposed can achieve improved accuracy compared with alternatives from the literature by using numerical test examples. Also, we illustrated the application of the methods proposed in a heat treatment distortion study for fabricating engine valves.

A concern that is associated with the methods proposed relates to the required pre-estimation of hyperparameters relating to the systematic and measurement errors. We discussed in the context of engineering applications how these pre-estimations can be generated. We also investigated this concern by using numerical examples and concluded that the methods proposed are, to some degree, robust to quantitative inaccuracies in the pre-estimations. We suggest that future work should investigate the estimation or updating of the hyperparameters by using data collected from experiments.

Acknowledgements

We thank the reviewers for valuable suggestions that helped us to increase the generality of the methods proposed and our method comparison. This research was partially funded by Scientific Forming Technologies Corporation. We thank Wei-Tsu Wu and Allen Miller for many forms of support. We also thank William Notz and Thomas Santner for references, ideas and encouragement. Finally, we thank Mikhail Bernshteyn for help with developing the necessary computer code.

Appendix A

In this appendix, we derive the EIMSE formula assuming that GLS is used in the analysis. The derivation begins from the definition of the EIMSE in Allen, Yu and Schmitz (2003):

$$\text{EIMSE}(\xi) = E_{\eta, \mathbf{x}, \varepsilon} \{ \hat{y}(\mathbf{x}, \varepsilon, \xi, \eta) - \eta(\mathbf{x}) \}^2$$

where $\hat{y}(\mathbf{x}, \varepsilon, \xi, \eta)$ and $\eta(\mathbf{x})$ are prediction and the true function values at point $\mathbf{x}$ respectively. The experimental design is ξ and ε is a vector of the measurement errors.

We assume that the true function of the real system is

$$\eta(\mathbf{x}) = \mathbf{f}_1(\mathbf{x})' \beta_1 + \mathbf{f}_2(\mathbf{x})' \beta_2$$

where $\mathbf{f}_1(\mathbf{x})' \beta_1$ are the primary terms and $\mathbf{f}_2(\mathbf{x})' \beta_2$ are the potential terms. Only the primary terms are included in the fitted regression model. Let $\mathbf{X}_1$ and $\mathbf{X}_2$ denote the design matrix of primary and potential terms respectively; the data vector $\mathbf{Y}$ is

$$\mathbf{Y} = \mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2 + \varepsilon.$$

Note that $\varepsilon \sim N(0, \mathbf{V})$. So the GLS predictor of β_1 is

$$\begin{aligned} \hat{\beta}_1 &= (\mathbf{X}_1' \mathbf{V}^{-1} \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{V}^{-1} \mathbf{Y} \\ &= \beta_1 + (\mathbf{X}_1' \mathbf{V}^{-1} \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{V}^{-1} \mathbf{X}_2 \beta_2 + (\mathbf{X}_1' \mathbf{V}^{-1} \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{V}^{-1} \varepsilon. \end{aligned}$$

Let $\mathbf{A} = (\mathbf{X}_1' \mathbf{V}^{-1} \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{V}^{-1} \mathbf{X}_2$; then the IMSE is

$$\begin{aligned} \text{IMSE} &= E_{\mathbf{x}} [E_{\varepsilon} \{ \hat{y}(\mathbf{x}) - \eta(\mathbf{x}) \}^2] \\ &= \int_r \{ \mathbf{f}_1(\mathbf{x})' \mathbf{A} - \mathbf{f}_2(\mathbf{x}) \} \beta_2 \beta_2' \{ \mathbf{A}' \mathbf{f}_1(\mathbf{x}) - \mathbf{f}_2(\mathbf{x}) \} d\mathbf{x} \\ &\quad + \int_r \mathbf{f}_1(\mathbf{x})' (\mathbf{X}_1' \mathbf{V}^{-1} \mathbf{X}_1)^{-1} \mathbf{f}_1(\mathbf{x}) d\mathbf{x} \\ &= \text{tr} \{ (\mathbf{A}' \mu_{11} \mathbf{A} - 2\mathbf{A}' \mu_{12} + \mu_{22}) \beta_2 \beta_2' \} + \text{tr} \{ \mu_{11} (\mathbf{X}_1' \mathbf{V}^{-1} \mathbf{X}_1)^{-1} \} \end{aligned}$$

where $\mu_{ij} = \int_r \mathbf{f}_i(\mathbf{x}) \mathbf{f}_j(\mathbf{x})' d\mathbf{x}$ ($i, j = 1, 2$ and $i \leq j$) are the so-called 'moment matrices' and r is the region of interest (i.e. the input space).

Integrating the IMSE over the possible coefficients of the potential terms gives

$$\begin{aligned} \text{EIMSE} &= E_{\beta_2} (E_{\mathbf{x}} [E_{\varepsilon} \{ \hat{y}(\mathbf{x}) - \eta(\mathbf{x})^2 \}]) \\ &= \text{tr} \{ (\mathbf{A}' \mu_{11} \mathbf{A} - 2\mathbf{A}' \mu_{12} + \mu_{22}) E(\beta_2 \beta_2') \} + \text{tr} \{ \mu_{11} (\mathbf{X}_1' \mathbf{V}^{-1} \mathbf{X}_1)^{-1} \}. \end{aligned}$$

References

- Allen, T. T., Bernshteyn, M. A. and Kabiri, K. (2003) A comparison of alternative methods for constructing meta-models for computer experiments. *J. Qual. Technol.*, **35**, no. 2, 1–17.
- Allen, T. T. and Yu, L. (2002) Low cost response surface methods from simulation optimization. *Qual. Reliab. Engng Int.*, **18**, 5–17.
- Allen, T. T., Yu, L. and Schmitz, J. (2003) An experimental design criterion for minimizing meta-model prediction errors applied to die casting process design. *Appl. Statist.*, **52**, 103–117.
- Chantarat, N. (2003) Modern design of experiments methods for screening and experimentations with mixture and qualitative variables. *PhD Dissertation*. Ohio State University, Columbus.
- DuMouchel, W. and Jones, B. (1994) A simple Bayesian modification of D-optimal designs to reduce dependence on an assumed model. *Technometrics*, **36**, 37–47.
- Etman, L. F. P. (2000) Comments on 'Response surfaces for optimal weight of cracked composite panels: noise and accuracy' (by M. Paila and R. T. Haftka). *2nd International Society for Structural and Multi-disciplinary Optimization—American Institute of Aeronautics and Astronautics Internet Conf. Approximations and Fast Reanalysis in Engineering Optimization*. (Available from <http://socm.wbmt.tudelft.nl/~wwwboard/aor2>.)
- Hadj-Alouane, A. B. and Bean, J. C. (1997) A genetic algorithm for the multiple-choice integer program. *Ops Res.*, **45**, 92–101.

- Kennedy, M. C. and O'Hagan, A. (2000) Predicting the output of a complex computer code when fast approximations are available. *Biometrika*, **87**, 1–13.
- Kennedy, M. C. and O'Hagan, A. (2001) Bayesian calibration of computer models (with discussion). *J. R. Statist. Soc. B*, **63**, 425–464.
- Knill, D. L., Giunta, A. A., Baker, C. A., Grossman, B., Mason, W. H., Haftka, R. T. and Watson, L. T. (1999) Response surface model combining linear and Euler aerodynamics for supersonic transport design. *J. Aircraft*, **36**, 75–86.
- Koc, M., Allen, T. T., Jiratheranat, S. and Altan, T. T. (2000) The use of FEM and experimental design to investigate tube hydroforming of a simple geometry. *Int. J. Mach. Tools Manufact.*, **40**, 2249–2266.
- McDaniel, W. R. and Ankenman, B. E. (2000) A response surface test bed. *Qual. Reliab. Engng Int.*, **16**, 363–372.
- Reese, C. S., Wilson, A. G., Hamada, M., Martz, H. F. and Ryan, K. J. (2004) Integrated analysis of computer and physical experiments. *Technometrics*, **46**, 153–164.
- Rodriguez, J. F., Perez, V. M., Padmanabhan, D. and Renaud, J. E. (2001) Sequential approximate optimization using variable fidelity response surface approximation. *Struct. Multidisc. Optimizn*, **22**, 23–34.
- Sacks, J., Schiller, S. B. and Welch, W. (1989a) Design for computer experiments. *Technometrics*, **31**, 41–47.
- Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989b) Design and analysis of computer experiments (with discussion). *Statist. Sci.*, **4**, 409–435.
- Vitali, R., Haftka, R. T. and Sankar, B. V. (2002) Multi-fidelity design of stiffened composite panel with a crack. *Struct. Multidisc. Optimizn*, **23**, 347–356.
- Yu, L. (2000) Expected modeling errors and low cost response surface methods. *PhD Dissertation*. Ohio State University, Columbus.