Journal of Statistical Theory and Practice
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ujsp20

Follow-Up Experimental Designs for Computer Models and Physical Processes

Pritam Ranjan [a], Wilson Lu [a], Derek Bingham [b], Shane Reese [c], Brian J. Williams [d], Chuan-Chih Chou [e], Forrest Doss [e], Michael Grosskopf [f] & James Paul Holloway [g]

[a] Department of Mathematics and Statistics, Acadia University, Wolfville, NS, B4P2R6, Canada

[b] Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, V5A1S6, Canada

[c] Department of Statistics, Brigham Yong University, Provo, UT, 84602, USA

[d] Statistics Sciences, Los Alamos National Laboratory, Los Alamos, NM, 87545, USA

[e] Atmospheric Oceanic and Space Sciences, University of Michigan, Ann Arbor, Michigan, 48109, USA

[f] Atmospheric Oceanic and Space University of Michigan, Ann Arbor, Michigan, 48109, USA

[g] Department of Nuclear Engineering, University of Michigan, Ann Arbor, Michigan, 48109, USA

Available online: 01 Dec 2011

To cite this article: Pritam Ranjan, Wilson Lu, Derek Bingham, Shane Reese, Brian J. Williams, Chuan-Chih Chou, Forrest Doss, Michael Grosskopf & James Paul Holloway (2011): Follow-Up Experimental Designs for Computer Models and Physical Processes, Journal of Statistical Theory and Practice, 5:1, 119-136

To link to this article: http://dx.doi.org/10.1080/15598608.2011.10412055

Journal of
Statistical Theory and Practice
Volume 5, No. 1, March 2011

Follow-up Experimental Designs for Computer Models and Physical Processes

Pritam Ranjan, *Department of Mathematics and Statistics, Acadia University, Wolfville, NS, Canada B4P2R6. Email: pritam.ranjan@acadiau.ca*

Wilson Lu, *Department of Mathematics and Statistics, Acadia University, Wolfville, NS, Canada B4P2R6. Email: wilson.lu@acadiau.ca*

Derek Bingham, *Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A1S6. Email: dbingham@stat.sfu.ca*

Shane Reese, *Department of Statistics, Brigham Young University, Provo, UT 84602, USA. Email: reese@stat.byu.edu*

Brian J. Williams, *Statistics Sciences, Los Alamos National Laboratory, Los Alamos, NM 87545 USA. Email: brianw@lanl.gov*

Chuan-Chih Chou, *Atmospheric Oceanic and Space Sciences, University of Michigan, Ann Arbor, Michigan 48109 USA. Email: reese@stat.byu.edu*

Forrest Doss, *Atmospheric Oceanic and Space Sciences, University of Michigan, Ann Arbor, Michigan 48109 USA. Email: reese@stat.byu.edu*

Michael Grosskopf, *Atmospheric Oceanic and Space Sciences, University of Michigan, Ann Arbor, Michigan 48109 USA. Email: reese@stat.byu.edu*

James Paul Holloway, *Department of Nuclear Engineering, University of Michigan, Ann Arbor, Michigan 48109 USA. Email: hagar@umich.edu*

Received: April 15, 2010 *Revised:* November 4, 2010

Abstract

In many branches of physical science, when the complex physical phenomena are either too expensive or too time consuming to observe, deterministic computer codes are often used to simulate these processes. Nonetheless, true physical processes are also observed in some disciplines. It is preferred to integrate both the true physical process and the computer model data for better understanding of the underlying phenomena. In this paper, we develop a methodology for selecting optimal follow-up designs based on *integrated mean squared error* that help us capture and reduce prediction uncertainty as much as possible. We also compare the efficiency of the optimal designs with the intuitive choices for the follow-up computer and field trials.

AMS Subject Classification: 62K05; 62L05.

Key-words: Gaussian Process; Model calibration; Integrated Mean Squared Error.

* 1559-8608/11-1/$5 + $1pp - see inside front cover

1. Introduction

Deterministic computer simulators are often used to explain the behaviour observed in physical systems. Scientists are able to adjust inputs to computer codes in order to help understand the impact on the process. A sufficient mathematical description of the system is often computationally intensive, and the trials of a computer experiment must be selected with care. To this end, a variety of experimental designs, such as Latin hypercube and space filing designs (McKay, Conover and Beckman, 1979; Johnson, Moore and Ylvisaker, 1990) have been proposed.

In many cases, observations of the physical process are also available. Kennedy and O'Hagan (2001) proposed a Bayesian approach that directly incorporates both computer-model simulations and field observations to form a predictive model. This approach, called model calibration, explicitly models the computer simulator and the discrepancy between the field process and the computer code. In addition, their model also attempts to estimate unknown constants that govern the physical system. This model is of specific interest in this article.

To implement the approach of Kennedy and O'Hagan (2001), one needs both a suite of field observations and computer simulations. The field observations may have been the result of a designed experiment (e.g., the motivating example in the next section) or an observational study (e.g., climate studies). The initial computer experiment designs are typically Latin Hypercube designs with space filling properties like maximin and minimax criterion (Sacks *et al.,* 1989; Johnson, Moore and Ylvisaker, 1990).

In this article, our primary focus is to explore ways of obtaining good follow-up designs that help us capture and reduce prediction uncertainty as much as possible. Due to cost constraints and experimental settings, it is often preferred to choose follow-up trials in a batch of pre-specified number of trials. Such trials can be chosen in many different ways, for example, by optimizing a criterion that is based on prediction errors (Sacks, Schiller and Welch, 1992), entropy (Currin *et al.,* 1991) or expected improvement based (Jones, Schonlau and Welch, 1998; Schonlau, Welch and Jones, 1998). Santner, Williams and Notz (2003, Chapter 6) present more details on design selection criteria. In this article, we discuss two types of follow-up trial selection techniques: (a) optimal choice (in mean squared error sense) and (b) intuitive choice. The performance of different approaches is illustrated through simulation, and the proposed methodology is applied to an ongoing study.

This paper is organized as follows. In Section 2, we present a motivating example from a physics hydrodynamics application. A brief review of the integrated process model is presented in Section 3. The fundamental approach for selecting follow-up trials is developed and the implementation details are discussed in Section 4 . In Section 5, we present the results from a small scale simulation study that compares several different choices of follow-up trials, and in Section 6 we apply the new methodology for identifying new field trials to the hydrodynamics application. Finally, in Section 7 we present a brief discussion and recommendations for selecting follow-up trials for such integrated processes.

2. Motivating Example

At the Center for Radiative Shock Hydrodynamics (CRASH), model calibration is playing an important role in combining physics modeling (i.e., computer simulator) and physical

experiments. The application under study is one where a laser pulse irradiates a beryllium disk positioned at the mouth of a zenon-filled tube (see Figure 2.1). The laser pulse ignites the beryllium causing a shock wave to flow down the tube. This shock is such that a radiative cooling layer forms immediately behind the shock front. The radiation then travels in front of the shock wave and interacts with the tube, causing a second shock (the "wall shock"). This second shock then interacts with the primary shock. The physics in this application is dominated by complex interactions between the laser-driven radiative shock, the wall shock, and the xenon-beryllium interface behind the primary shock, and has implications in astrophysics and high-energy-density physics.

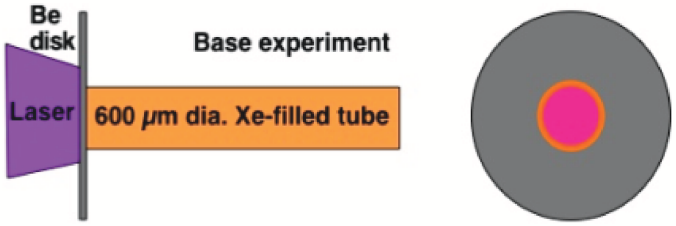


Figure 2.1. Illustration of the CRASH experiment.

An important endeavor for CRASH is to be able to accurately predict features of radiative shock with associated measures of uncertainty. The specific feature under study here is the distance the shock has traveled down the tube (i.e., *shock location*). A series of physical experiments were conducted, varying the settings of four experiment factors (see Table 2.1). There are nine such observations available for our modeling purposes.

Table 2.1. Experimentally controlled inputs.

Factor	Description	Average/Typical Values	Range for Study
x_1	Be drive disk thickness	21 μm	18 – 22 μm
x_2	Laser total energy	3870 J	3.23 – 4.37 kJ
x_3	Xe gas pressure	6.5 mg/cc (1.15 atm)	5.85 – 7.15 mg/cc
x_4	Observation time t_o	13, 14 and 16 ns	12.5 – 14.5 ns

In addition to physical observations, a deterministic computer simulator (the CRASH code) has been constructed to help understand radiative shock in this context. The CRASH code is computationally demanding, requiring a supercomputer to run, and often requires manual intervention to overcome numerical instabilities. A suite of 320 computer simulations were conducted, varying the levels of the same four factors in the physical experiments. Furthermore, four additional variables (see Table 2.2) that correspond to physical constants (e.g., *calibration parameters*) were also studied. The values of the physical constants are not

Table 2.2. Calibration inputs θ.

Description	Nominal Value	Range for Study	Symbolic Name
γ_{Be} Be gamma	5/3	1.4 – 5/3	θ_1
γ_{Xe} Xe gamma	1.2	1.1 – 1.4	θ_2
Be opacity scale factor	1	0.7 – 1.3	θ_3
Xe opacity scale factor	1	0.7 – 1.3	θ_4

known precisely, but some value must be given to the simulator for the code to run. So, an additional goal of the endeavor is to estimate the unknown physical constants. The choice of simulations was determined by using an orthogonal array based Latin hypercube with a space filling criterion (Owen, 1998; Tang, 1993; Johnson, Moore and Ylvisaker, 1990).

From a data analysis viewpoint, the aim is to combine the physical observations and computer simulations to build a predictive model of the physical system and also to estimate the calibration parameters. Data analysis will be discussed in the next section. The CRASH project has access to the laser facility where the experiments are conducted only once per year. A crucial question facing the experiments is "which physical experiments should be performed?" A secondary question that can be asked "if new computer simulations are to be run, which ones?" In the next several sections we introduce the methodology for analyzing such experiments and propose new methodology for identifying follow-up physical and computer trials to improve the predictive performance of the statistical model.

3. Modeling, Estimation and Prediction

In this section, we outline the standard approach to model calibration (Kennedy and O'Hagan, 2001). We follow the fully Bayesian implementation described in Higdon *et al.* (2004).

There are two types of variables that influence the system: the $p-$dimensional design variables, $\boldsymbol{x}$, that are the observable (often controllable) in the physical system and the $q-$dimensional calibration parameters. Without loss of generality, we will assume that the input space is a unit hypercube. The actual values of the calibration parameters, denoted $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$, are unknown. For each run of the computer simulator, some value, $\boldsymbol{t}$, is inserted in place of the unknown calibration parameters. Thus, the computer model has inputs $(\boldsymbol{x}, \boldsymbol{t})$, whereas the response in the physical system is governed by $(\boldsymbol{x}, \boldsymbol{\theta})$.

Two sources of data inform the predictive model for the physical system – the simulator and field observations. Denote y_{f_i} as the ith observation of the physical system and represented as a sum of three components

$$y_{f_i} = \eta(\boldsymbol{x}_i, \boldsymbol{\theta}) + \delta(\boldsymbol{x}_i) + \varepsilon(\boldsymbol{x}_i), \quad i = 1, \ldots, n_f,$$

where $\eta(\cdot,\cdot)$ is the computer model, $\delta(\cdot)$ is the discrepancy between the computer model and the physical observations (the aspects of the physical process not modeled by $\eta(\cdot,\cdot)$),

and $\varepsilon(\cdot)$ are iid observation error. Denote the jth simulator output, at inputs $(\boldsymbol{x}_j, \boldsymbol{t}_j)$, as

$$y_{c_j} = \eta(\boldsymbol{x}_j, \boldsymbol{t}_j), \quad j = 1, \cdots, n_c,$$

where $\eta(\cdot,\cdot)$ denotes the computer simulator.

Let $\boldsymbol{y}_c = (y_{c_1}, \ldots, y_{c_{n_c}})'$ be the vector of outputs for the n_c computer model runs and $\boldsymbol{y}_f = (y_{f_1}, \ldots, y_{f_{n_f}})'$ denote the vector of observations on the physical process. Denote the combined vector of responses as $\boldsymbol{d}^T = (\boldsymbol{y}_c^T, \boldsymbol{y}_f^T)$. Let $\boldsymbol{D}_1$ be the $n_c \times (p+q)$ experiment design matrix of the input settings for the simulator and $\boldsymbol{D}_2$ be the $n_f \times p$ matrix of the input vectors at which the physical process is observed.

Both $\eta(\cdot,\cdot)$ and $\delta(\cdot)$ are viewed as independent Gaussian spatial processes (Sacks *et al.,* 1989; Welch *et al.,* 1992). Specifically, the computer model, $\eta(\cdot,\cdot)$, is modeled as a Gaussian process (GP) with mean μ_η and power exponential covariance function,

$$c_1((\boldsymbol{x},\boldsymbol{t}),(\boldsymbol{x}',\boldsymbol{t}')) = \sigma_\eta^2 \cdot exp\left\{ -\sum_{k=1}^{p} \beta_k^\eta (x_k - x_k{}')^2 - \sum_{l=1}^{q} \beta_{p+l}^\eta (t_l - t_l')^2 \right\}. \tag{3.1}$$

Similarly, the discrepancy, $\delta(\cdot)$, is modeled as a GP with mean μ_δ and power exponential covariance,

$$c_2(\boldsymbol{x},\boldsymbol{x}') = \sigma_\delta^2 \cdot exp\left\{ -\sum_{k=1}^{p} \beta_k^\delta (x_k - x_k')^2 \right\}. \tag{3.2}$$

Thus the covariance of the combined vector of responses, $\boldsymbol{d}$, can be written as

$$\Sigma = \Sigma_\eta + \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_\delta \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_\varepsilon \end{pmatrix},$$

where Σ is an $(n_c + n_f) \times (n_c + n_f)$ matrix of covariances. The elements of Σ_η and Σ_δ follow the covariance models specified in (3.1) and (3.2), respectively, and $\Sigma_\varepsilon = \sigma^2 I_{n_f}$. The distribution of the combined vector of responses, $\boldsymbol{d}$, can be written as

$$[\boldsymbol{d}|\boldsymbol{\Omega}] \sim N(\boldsymbol{H}\boldsymbol{\mu}, \Sigma),$$

where $\Omega = (\boldsymbol{\theta}, \beta^\eta, \beta^\delta, \sigma_\eta^2, \sigma_\delta^2, \sigma_\varepsilon^2, \mu_\eta, \mu_\delta)$, $\boldsymbol{\mu} = (\mu_\eta, \mu_\delta)$, and

$$\boldsymbol{H} = \begin{pmatrix} \mathbf{1}_{n_c} & \mathbf{0} \\ \mathbf{1}_{n_f} & \mathbf{1}_{n_f} \end{pmatrix}.$$

The likelihood for $\boldsymbol{d}$ is then

$$L(\boldsymbol{d}|\boldsymbol{\Omega}) \propto |\Sigma^{1/2}| \exp\left(-\frac{1}{2} (\boldsymbol{d} - \boldsymbol{H}\boldsymbol{\mu}) \Sigma^{-1} (\boldsymbol{d} - \boldsymbol{H}\boldsymbol{\mu}) \right).$$

Completion of the Bayesian specification requires prior distributions for the unknown model parameters. We specify *a priori* that the unknown calibration parameter, θ, has a $N(0.5, (0.5)^2)$ distribution. This prior reflects reasonable uncertainty in the computer model $\eta(\cdot,\cdot)$. For purposes in this paper, without loss of generality we assume $\mu_\eta = \mu_\delta = 0$,

although we retain the notation for purposes of completeness. For the error variance, we assume a Jeffreys prior distribution, that is

$$\pi(\sigma_\varepsilon^2) \propto (\sigma_\varepsilon^2)^{-1},$$

a specification that allows the data to be the primary information source for the posterior distribution.

We specify the prior distribution for the parameters in the computer model $\eta(\cdot,\cdot)$ as independent distributions of the form

$$\pi(\beta^\eta) \propto \prod_{k=1}^{p+q}(1-e^{-\beta_k^\eta})^{-0.5}e^{-\beta_k^\eta}, \beta_k^\eta > 0,$$
$$\sigma_\eta^2 \propto (\sigma_\eta^2)^{-1}.$$

These GP governing priors allow for the computer model $\eta(\cdot,\cdot)$ to be relatively flat, allowing the data to dictate the overall response as a function of the inputs and calibration parameters. Specifically, the choice of the 0.5 exponent on the prior distribution of β^η follows the convention of Higdon *et al.* (2004) and has the effect of prior GP flattening. Furthermore, we adopt the Jeffreys prior for σ_η^2 analogous to the specification on σ_ε^2.

Similarly, we specify prior distributions for the discrepancy term $\delta(\cdot)$. That is,

$$\pi(\beta^\delta) \propto \prod_{k=1}^{p}(1-e^{-\beta_k^\delta})^{-0.6}e^{-\beta_k^\delta}, \beta_k^\delta > 0,$$
$$\sigma_\delta^2 \propto (\sigma_\delta^2)^{-1},$$

where the choice of the 0.6 exponent gives even stronger tendency for flat discrepancy functions, $\delta(\cdot)$, than the computer model, $\eta(\cdot,\cdot)$.

The main focus of this article is the prediction of the underlying physical process at unobserved trial locations. Let $y_f(\boldsymbol{x}_0)$ be the response of physical process value at $\boldsymbol{x}_0$. The expected response of the underlying physical process at $\boldsymbol{x}_0$ given the data $\boldsymbol{d}^T = (\boldsymbol{y}_c^T, \boldsymbol{y}_f^T)$ and the true values of parameters are given by

$$E\{y_f(\boldsymbol{x}_0)|\boldsymbol{d},\boldsymbol{\Omega}\} = \boldsymbol{h}_f^T\boldsymbol{\mu} + \boldsymbol{t}_f(\boldsymbol{x}_0,\boldsymbol{\theta})^T\Sigma^{-1}(\boldsymbol{d}-\boldsymbol{H}\boldsymbol{\mu}), \tag{3.3}$$

where $\boldsymbol{h}_f^T = (1,1)$ and $\boldsymbol{t}_f$ is the vector of correlations between the response at $\boldsymbol{x}_0$ and the data $\boldsymbol{d}$. Notice that the expected response in (3.3) is the same as the best linear unbiased predictor (BLUP) for $y_f(\boldsymbol{x}_0)$ in the likelihood setup if the parameters are replaced with their maximum likelihood estimates.

There are several ways one can imagine predicting $y_f(\boldsymbol{x}_0)$. One way is to average (3.3) over the posterior distribution of the model parameters. One can also imagine substituting a single value into (3.3) (e.g., the posterior mode or mean of the parameters). Indeed, because of the computational savings accompanying the plug-in approach, we use the posterior mean throughout.

4. Selecting Follow-up Runs

The primary interest of the application that motivates this work is building a predictive model for the physical system. Accordingly, the aim of our work is selecting new trials that improve the predictive ability of the model outlined in the previous section.

4.1. Integrated Mean Squared Error (IMSE) Criterion

A common approach to assessing the quality of a prediction is to consider the mean squared error. Given the existing responses $\boldsymbol{d}^T = (\boldsymbol{y}_c^T, \boldsymbol{y}_f^T)$ and designs $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ for the computer and field trials, the MSE for predicting the true physical process at an unobserved location, $\boldsymbol{x}_0$, under the model outlined in the previous section is

$$\begin{aligned} \text{MSE}[y_f(\boldsymbol{x}_0)|\boldsymbol{\Omega}] &= E\left(E\{y_f(\boldsymbol{x}_0)|\boldsymbol{d},\boldsymbol{\Omega}\} - y_f(\boldsymbol{x}_0)|\boldsymbol{\Omega}\right)^2 \\ &= \boldsymbol{h}_f^T\boldsymbol{W}\boldsymbol{h}_f - 2\boldsymbol{h}_f^T\boldsymbol{W}\boldsymbol{H}^T\Sigma^{-1}\boldsymbol{t}_f + \boldsymbol{t}_f^T(\Sigma^{-1}\boldsymbol{H}\boldsymbol{W}\boldsymbol{H}^T\Sigma^{-1} - \Sigma^{-1})\boldsymbol{t}_f + \sigma_\eta^2 + \sigma_\delta^2 + \sigma_\varepsilon^2, \end{aligned} \tag{4.1}$$

where $\boldsymbol{W} = (\boldsymbol{H}^T\Sigma^{-1}\boldsymbol{H})^{-1}$. We use a plug-in estimate (the posterior mean) of $\boldsymbol{\Omega}$ to evaluate (4.1). A full Bayesian approach where (4.1) is integrated over the posterior distribution of the model parameters could be implemented, but this approach would require considerably more computational effort.

It might be tempting to select new design points $\boldsymbol{x}_{new}$ where (4.1) is maximized. However, we are interested in improving predictions throughout the entire design space. Thus, we instead use the integrated mean squared error (hereafter, IMSE) criterion where (4.1) is integrated across the design space. That is,

$$\text{IMSE}(\boldsymbol{x}_{new}) = \int_{\mathscr{D}} \text{MSE}[y_f(\boldsymbol{x}_0)|\boldsymbol{x}_{new}]d\boldsymbol{x}_0, \tag{4.2}$$

where $\boldsymbol{x}_{new}$ is a new design point (either in the field or a computer model trial).

In principle, one could identify the computer model run, or field trial, that minimizes (4.2). However, in most settings it is not practical to run a single trial, update the model, identify a new trial, and continue in this manner. Instead a batch of new trials is identified. Let $\boldsymbol{X}_{new}$ be a batch of $m_f + m_c \geq 1$ new trials, where m_f and m_c are the number of new field trials and computer trials, respectively. Note that for runs corresponding to the field trials, the current estimate of $\boldsymbol{\theta}$ is inserted in the columns corresponding to the calibration parameters. The design criterion is then,

$$\text{IMSE}(\boldsymbol{X}_{new}) = \int_{\mathscr{D}} \text{MSE}[y_f(\boldsymbol{x}_0)|\boldsymbol{X}_{new}]d\boldsymbol{x}_0. \tag{4.3}$$

The design matrices, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$, are augmented with prospective design points when evaluating the integrand in (4.2) or (4.3).

Finally, the set of follow-up runs, $\boldsymbol{X}_{new}$, is chosen from the design region $\mathscr{X}$ (assumed to be a $[0,1]^p$ for the field trials and $[0,1]^{p+q}$ for the computer runs), such that the resulting

IMSE($\boldsymbol{X}_{new}$) is minimized. That is,

$$\boldsymbol{X}_{new}^{opt} = \underset{\boldsymbol{x}_{new} \in \mathscr{X}}{\arg\min}\ \text{IMSE}(\boldsymbol{X}_{new}). \tag{4.4}$$

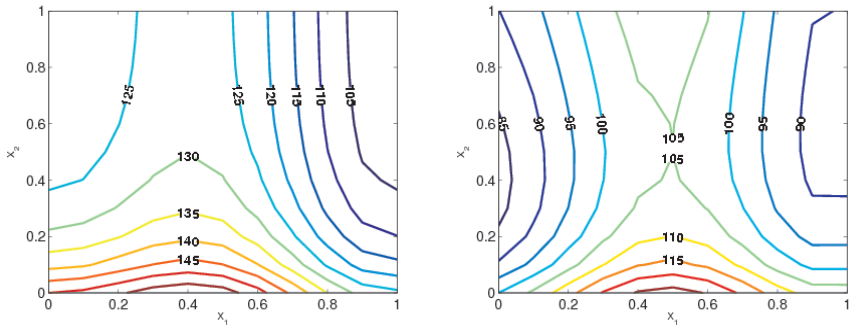
Although it looks simple, (4.4) represents a $(m_f p + m_c(p+q))$-dimensional optimization problem. It is computationally intensive to explore the entire region to find the global optimum. We will discuss the implementation of this approach in Section 4.3.

Example 4.1 below illustrates the selection of a batch follow-up trials (computer simulations only or field trials only) using the IMSE optimal strategy outlined above. For generating outputs from a computer simulator, we use a 2-dimensional test function used by Crary (2002) for demonstrating the prediction performance of GP based metamodels with the IMSE optimal designs. Note that Section 5 also presents simulation results for the scenario when the batch of desired follow-up runs is a mixture of both computer and field trials.

Example 4.1. Suppose the simulator output $\eta(x,t)$, is generated using

$$\eta((x_1,x_2),t) = (30 + 5x_1 \sin(5x_1)) \left(6t + 1 + e^{-5x_2}\right),$$

where $x_1, x_2, t \in [0,1]^3$ and the true value of the calibration parameter (replaced by t in the simulator) is $\theta = 0.5$. Suppose that the computer model matches the physical system up to a discrepancy $\delta(x_1,x_2)$ and a replication error $\varepsilon \sim N(0, 0.5^2)$. We arbitrarily chose the discrepancy function, $\delta(x_1,x_2) = -50e^{(-0.2x_1 - 0.1x_2)}$, that makes the simulator output, $\eta(\boldsymbol{x},t)$, easily distinguishable from the underlying physical process, $\eta(\boldsymbol{x},\theta) + \delta(\boldsymbol{x})$. Figure 4.1 displays the contour plots for both the processes.



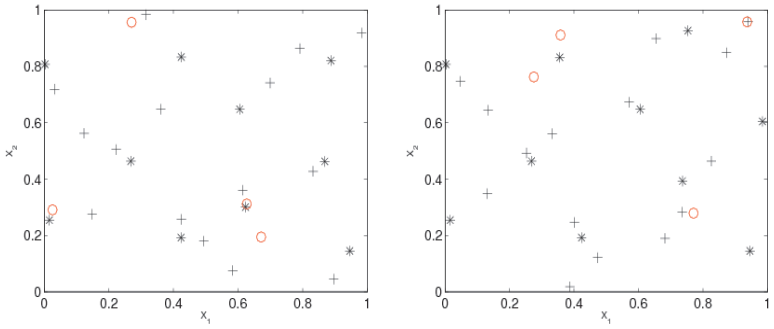
(a) Computer simulator (b) Computer simulator + discrepancy

Figure 4.1. Contour plots of the true computer simulator and the discrepancy function.

The implementation begins with finding two space-filling designs for the initial set of n_c computer trials and n_f field trials. We used a 15-point maximin Latin hypercube design (McKay, Conover and Beckman, 1979; Johnson, Moore and Ylvisaker, 1990) for computer

trials and a 10-point maximin coverage design (Johnson, Moore and Ylvisaker, 1990) for the initial set of field trial locations. Based on these 25 trials, the posterior distributions of all the parameters in $\mathbf{\Omega}$, and subsequently the plug-in estimates of $E\{y_f(\boldsymbol{x}_0)|\boldsymbol{d},\mathbf{\Omega}\}$ and $MSE[y_f(\boldsymbol{x}_0)|\mathbf{\Omega}]$, were obtained by fitting the model outlined in Section 3. Finally, a batch of four new follow-up trials are chosen by minimizing the IMSE criterion in (4.3). Next we present a few illustrations of the optimal choices for a batch of follow-up trials using the IMSE criterion based on different realizations of the physical process (differing only in observation error).

Figure 4.2 displays two realizations of the implementation when the desired batch of follow-up trials consists of *field trials only*. Although not always true, the optimal choice of new field trials tends to align a new trial with an existing field trial (see Figure 4.2(a)) and an existing computer trial (see Figure 4.2(b)).



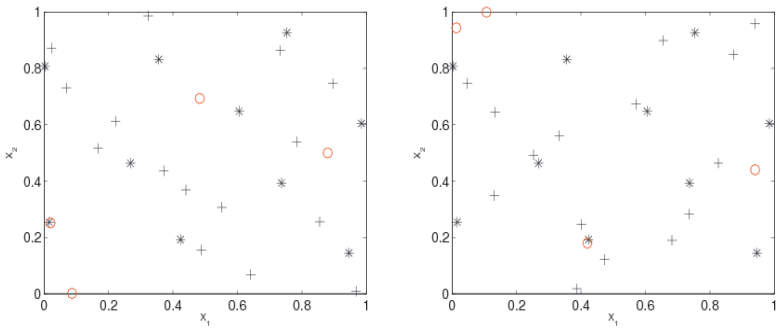
(a) Aligned with a field trial (b) Aligned with a computer trial

Figure 4.2. A batch of 4 new field trials (denoted by red circles) obtained by minimizing the IMSE criterion (4.3). Initial design: 10 field trials (denoted by $*$) and 15 computer trials (denoted by $+$).

Similarly, when adding a new batch of *computer trials only* found by minimizing the IMSE criterion (4.3), a few of the new trials in the optimal choice of follow-up trials can be aligned with the existing field trials. Since the computer simulator is assumed to be deterministic, the new computer trial locations will not be aligned with the initial computer trials. Figure 4.3 presents two realizations.

These illustrations indicate that the alignment of new field trials with the existing computer and/or field trials are sometimes recommended under the IMSE criterion. That is, the forced alignment and replication of the follow-up trials (at least for a fraction of the batch) may lead to good designs for minimizing the IMSE and hence for good prediction of the underlying physical process $y_f(\cdot)$. This feature can especially be very attractive to a practitioner because reducing the batch-size of the follow-up trials that are IMSE optimal reduces the dimensionality of the optimization problem (i.e., minimizing the IMSE criterion) which can be computationally burdensome.

Section 5 presents a simulation study to show that the forced alignment and/or replication of a fraction of the batch of new follow-up trials can be almost as efficient as the IMSE-



(a) Aligned with a field trial (b) Aligned with a computer trial

Figure 4.3. A batch of 4 new computer trials (denoted by red circles) obtained by minimizing the IMSE criterion (4.3). Initial design: 10 field trials (denoted by $*$) and 15 computer trials (denoted by $+$).

optimal choices for the follow-up runs. Although one can envision accommodating these features in the initial design itself, we choose to implement the notion of alignment and replication in the batch of follow-up trials as it fits well in our framework. It turns out that both the intuitive choices of the forced *replication* of the field trials and *alignment* of the computer trials with the field trials can be theoretically justified as advantageous for the prediction of the true underlying physical process.

4.2. Intuitive choices

Recall that the observations from the true physical process have observational error, and the replication of the field trials should be able to capture the associated uncertainty. Moreover, if the computer trials are aligned with the field trials, the discrepancy between the computer simulator and the physical process can be approximated reasonably well.

Let $x_{new,1}$ and $x_{new,2}$ be two field trials that are replicated, i.e., $x_{new,1} = x_{new,2}$. Then, the expected responses at $x_{new,1}$ and $x_{new,2}$ are given by

$$\begin{aligned} E(y_f(x_{new,1})|\boldsymbol{d},\boldsymbol{\Omega}) &= E(y_f(x_{new,2}|\boldsymbol{d},\boldsymbol{\Omega}) \\ &= \boldsymbol{h}_f^T\boldsymbol{\mu} + \boldsymbol{t}_f(\boldsymbol{x}_{new,1},\boldsymbol{\theta})^T\Sigma^{-1}(\boldsymbol{d}-\boldsymbol{H}\boldsymbol{\mu}). \end{aligned}$$

Note that $y_f(x_{new,1}) - y_f(x_{new,2})$ is conditionally normal with mean zero and variance $2\sigma_\varepsilon^2$, which is to say that the sample variance of the two field observations collected as replicates, $x_{new,1} = x_{new,2}$, can be used as a simple estimation procedure for σ_ε^2. Thus, forced replication of a fraction of the follow-up field trials can be very informative and may lead to good designs for prediction from the underlying physical process.

On the other hand, the alignment of a few field trials with the computer trials can be useful in directly accessing the information on the discrepancy between the computer simulator and the physical process. Let (x_{new}, t_{new}) be a follow-up computer trial that is aligned with a field

trial from the initial design. Then,

$$[\delta(x_{new})|\boldsymbol{d},\boldsymbol{\Omega}] \sim N\left(\mu_\delta + \boldsymbol{t}_\delta^T \Sigma^{-1}(\boldsymbol{d} - \boldsymbol{H}\boldsymbol{\mu}),\ \sigma_\delta^2 - \boldsymbol{t}_\delta^T \Sigma^{-1} \boldsymbol{t}_\delta\right),$$

where $t_\delta = t_f(x_{new},\theta) - t_c(x_{new},\theta)$. Specifically, one is gaining information on the discrepancy, conditional on $\boldsymbol{\theta}$.

If the primary objective is to estimate the uncertainty due to replication error or the discrepancy, one can use the outputs from the field and computer trials to fit the two models deduced here. However, we aim for good prediction from the underlying physical process; thus, we enforce the alignment and replication via the batch of follow-up trials and fit the integrated model proposed in Section 3. As we will briefly demonstrate in Section 5, these intuitive choices of follow-up trials often lead to efficient designs for prediction as compared to the IMSE optimal designs.

4.3. Implementation Details

In this section, we discuss the computational challenges and the steps taken to resolve these issues in implementing the methodology. The overall procedure for finding a batch of follow-up trials can be summarized as follows.

1. We assume that initial designs for the physical and computer experiments have been performed. In the previous examples, we used a maximin design (Johnson, Moore and Ylvisaker, 1990) for the field trials and a maximin Latin hypercube design (McKay, Conover and Beckman, 1979; Johnson, Moore and Ylvisaker, 1990) for the computer trials. The designs are: $\{x_1,...,x_{n_f}\}$ for the field trials and $\{(x_1,t_1),...,(x_{n_c},t_c)\}$ for the computer simulation runs. The simulator output at the initial computer trial locations is $\boldsymbol{y}_c = (y_1,...,y_{n_c})'$, and the physical process response for the initial field design is $\boldsymbol{y}_f = (y_1,...,y_{n_f})'$.
2. Fit the joint model outlined in Section 3 to the data $\{x_1,...,x_{n_f}\}$, $\{(x_1,t_1),...,(x_{n_c},t_c)\}$, and $\boldsymbol{d}^T = (\boldsymbol{y}_f^T, \boldsymbol{y}_c^T)$.
3. Choose a batch of $m_f + m_c$ follow-up trials, X_{new}, that minimizes the IMSE criterion (4.3). This search is a $(m_f p + m_c(p+q))$-dimensional optimization problem and can be very computationally expensive if one or more of three quantities (i) the batch size $(m_f + m_c)$, (ii) the input dimension p and (iii) the number of calibration parameters q, are large. We used an exchange algorithm (e.g., Miller, 1994) on a fine grid of the $p-$ and q-dimensional input space. As demonstrated in Section 5, one can obtain comparable efficiency in less time by forcing the alignment and/or replication of a fraction of the batch of follow-up trials with the initial design points.

In the implementation of the IMSE optimal design strategy, the computational cost of evaluating the IMSE criterion (4.3) for every candidate batch of follow-up trials takes a significant portion of the time. For large data sets, this can be of serious concern as the inversion of an $n \times n$ matrix takes $O(n^3)$ operations. The inversion time of the covariance matrix Σ for the augmented design can be substantially reduced by avoiding the direct inversion of Σ for every candidate batch, and instead using the inverse updating technique (Schur Complement method) proposed by Schur (1917).

5. Simulation Study

In this section, the proposed approach for identifying follow-up trials is investigated. We consider finding optimal IMSE designs and explore issues related to full/partial replication and full/partial alignment. The setting in Example 4.1 is used for illustration.

Consider the setup in Example 4.1 where the physical process is defined as $y_f(x) = \boldsymbol{\eta}(x, 0.5) + \boldsymbol{\delta}(x) + \boldsymbol{\varepsilon}(x)$, where $\boldsymbol{\delta}(x) = -50\, e^{-0.2x_1 - 0.1x_2}$, and $\boldsymbol{\varepsilon}(x) \sim N(0, 25)$. For the initial designs, a random maximin Latin hypercube design of size $n_c = 15$ is used for the computer model and a minimax design of size $n_f = 10$ is used for the physical process. Using these designs, computer model and field observations will be generated. We will do this 500 times. The goal is to find $m_f + m_c = 8$ follow-up runs for each simulated dataset.

To investigate the impact of adding field and computer follow-up trials suggested by the IMSE criterion and the intuitive choices, we find follow-up trials under the following schemes:

- **Field trials only**
  1. $m_f = 8$ randomly selected field trials.
  2. $m_f = 8$ IMSE optimal field trials.
  3. $m_f = 8$ replicated field trials (i.e., 8 of the $n_f = 10$ field trials are randomly selected to be repeated).
  4. $m_f = 8$ field trials where 4 are randomly selected to be replicated and 4 trials are selected as IMSE optimal.
- **Computer trials only**
  5. $m_c = 8$ randomly selected computer trials.
  6. $m_c = 8$ IMSE optimal computer trials.
  7. $m_c = 8$ computer trials that are aligned with 8 randomly selected field trials. The value of t for each simulation trial is also randomly selected.
  8. $m_c = 8$ computer trials where 4 are aligned with 4 randomly selected field trials and the other 4 trials are selected as IMSE optimal.. The value of t for each simulation trial is also randomly selected.
- **Field and Computer trials**
  9. $m_f = 4; m_c = 4$ randomly selected field and computer trials.
  10. $m_f = 4; m_c = 4$ IMSE optimal field and computer trials.
  11. $m_f = 4; m_c = 4$ where the 4 field trials are randomly selected to be replicated and the computer trials are IMSE optimal.
  12. $m_f = 4; m_c = 4$ where the 4 computer trials are aligned with 4 randomly selected field trials and the field trials are IMSE optimal.
  13. $m_f = 4; m_c = 4$ where the 2 computer trials are aligned with 2 randomly selected field trials, 2 field trials are randomly selected to be replicated and the remaining computer field trials are IMSE optimal.

Only field trials are added for the first four design schemes. The IMSE optimal design (scheme 2) will be compared to cases where we are replicating all or half of the follow-up runs. We include a random follow-up design (scheme 1) simply to investigate whether or not it is worth the effort to find good follow-up trials. Schemes $4-8$ consider adding only computer model runs. Notice that design schemes 7 and 8 explore the issue of alignment. The final design schemes find designs where the follow-up trials are equally divided between the computer model and the physical process. For instance, the final design has 2 field trials replicated, 2 computer simulations aligned with field trials, and the remaining trials selected according to the IMSE criterion.

The model in Section 3 was fit to each of the 500 simulated datasets, and the $m_f + m_c = 8$ follow-up trials were identified for each of the 13 above design schemes. The expected reduction in the IMSE was computed for each case. The results are summarized in Figure 5.1.

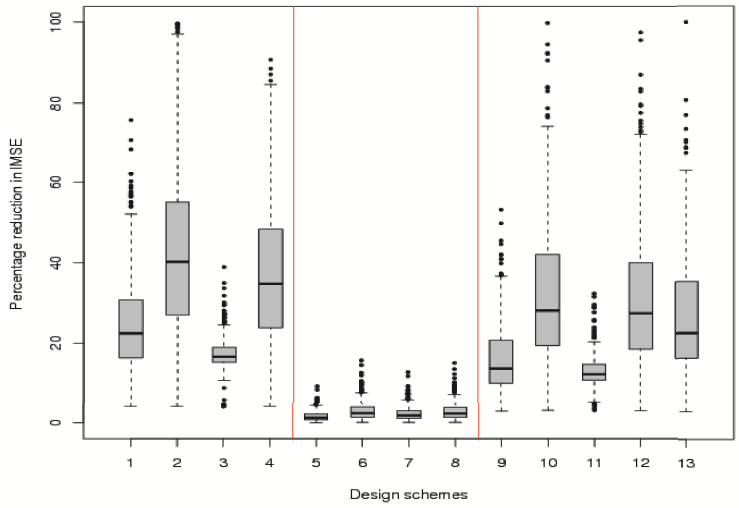


Figure 5.1. Percent reduction in the IMSE for 500 simulated datasets under 13 different follow-up design schemes.

There are several observations we can take from this plot. Consider first the cases where only field trials are performed (schemes $1-4$). The best performing design scheme is the IMSE optimal designs where the average reduction in the IMSE is 43%. The second best design, with an average reduction in the IMSE of 38%, is the scheme where half of the follow-up trials are replicated and the other half are IMSE optimal. Indeed, we would expect to see this since both schemes focus on reducing the IMSE. The random design (scheme 1) performs poorly, but better on average than the fully replicated design. This is not too surprising since the randomly selected trials will tend to fill in the gaps in the design region, whereas the fully replicated runs will only serve to reduce the impact, and improving the estimate, of observation error. Turning to the scenarios where only computer trials are

added (schemes $5-8$), the most important observation is that the gains in terms of IMSE are marginal. As before, the IMSE design performs best with an average reduction in the IMSE of 2.9%, and the design that is a combination of aligned and IMSE optimal design points performs comparably with an average percent reduction of 2.8%. Finally, turning to the cases where both field and computer trials are added, we see that the best performing cases are design schemes 10, 12, and 13, with average percentage reductions in the IMSE of 36%, 35%, and 29% respectively. Again, the random design and the design where the field trials are entirely replicated do comparatively worse.

Comparing all cases, we see that field trials are far more valuable than computer trials only. However, adding a combination of field trials and carefully chosen computer trials (either aligned or IMSE optimal) achieves substantial gains with only half the effort in the physical system. In general, in the simulations that we have performed, we have found that adding only computer model trials typically yields only modest gains in the IMSE (albeit when a good initial design is chosen). Furthermore, as one might expect, one should devote as much resources to experimentation on the physical process as possible. However, comparable gains are often seen in a combination of computer model and field trials, where some computer model runs are aligned with field experiments and some field experiments are replicated. In the next section, we note that there are advantages to considering replicated and aligned runs from an exploratory data analysis perspective.

6. Optimal Design for Radiative Shock

We now return to the motivating example from Section 2. An important goal for the CRASH researchers is to accurately predict features of radiative shock location (measured in meters). To this end, $n_f = 9$ field trials were conducted, varying the levels of the four design variables. In addition, a computer experiment with $n_c = 320$ trials was conducted. The experiment design for the first 256 simulation experiments was chosen using an orthogonal array based Latin hypercube with a space-filling criterion (Owen, 1998; Tang, 1993; Johnson, Moore and Ylvisaker, 1990). The next 64 runs were chosen so that they were aligned with the 8 field trials. That is, for each of the 8 field trial, 8 computer simulations were conducted at the same levels of the four design variables but at different levels of the four calibration parameters.

The joint model for the field and computer trials outlined in Section 3 was fit to the resulting data. At this point, interest lies in selecting new field trials aimed at improving the predictive capability of the model. The Omega facility where the experiments are conducted is available to the CRASH researchers once per year, and only a limited number of experiments can be performed. As a result, we are interested in finding the batch of $m_f = 10$ new field trials that are expected to give the most improvement in the predictive ability of the model as measured by the IMSE.

Before moving on to identify new trials to improve the predictive model, we take a moment to consider the impact of alignment and replication in these settings. As one would expect, the IMSE optimal designs give the most improvement in terms of prediction. From an exploratory data analysis viewpoint, there are sometimes advantages in aligning some computer model trials with field experiments and also replicating some of the field trials. Consider, for example, Figure 6.1 where the observed responses (the circles) are plotted

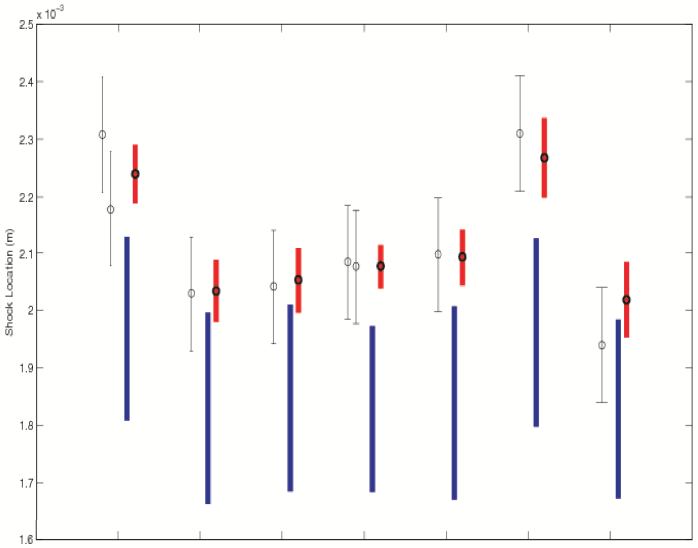


Figure 6.1. Observed field responses, with measurement uncertainty (black circles and bars), 95% posterior intervals for $\eta(x,\theta)$ (blue) and $\hat{y}_f(\boldsymbol{x}) = \eta(\boldsymbol{x},\theta) + \delta(\boldsymbol{x})$ (red).

with the estimated discrepancy and also the prediction at each of the $n_f = 9$ field locations. Notice that there are two field locations where experiments were replicated. The replicated field trials reveal that the experimental error is likely to be quite small – without replication in the design, it is not possible to visually assess this feature. Also, notice that 95% posterior prediction intervals are smallest where there is replication and where the computer model consistently underestimates the field responses.

Similarly, consider Figure 6.2 where the observed field responses (the circles) are plotted with the computer model responses (red x's) that are aligned at the same the four design variables. Notice that since two of the field trials are replicated, there are 16 computer trials at these locations. Also, note that the values of the four calibration parameters are varied for each simulation. Looking at the figure, we again observe that the computer simulator consistently underestimates the observations in the field. Indeed, notice that the range of the simulator responses fails to contain the aligned field observation in each case. This often implies that either the computer model is missing the necessary physics to adequately predict the physical process without discrepancy or the range of calibration parameters that is explored should be expanded. Since the sensitivity of the response to changes in the calibration parameters is fairly large – evident in the range of computer trial responses at each location – one might be tempted to expand the design range for the calibration inputs. In this case, however, the consensus among the scientists was that the simulator inadequacy was due to missing physics rather than misspecification of the design region for the calibration parameters.

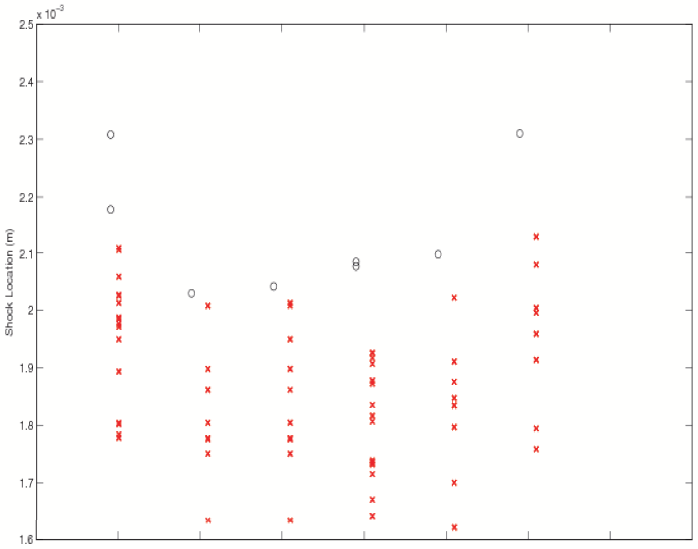


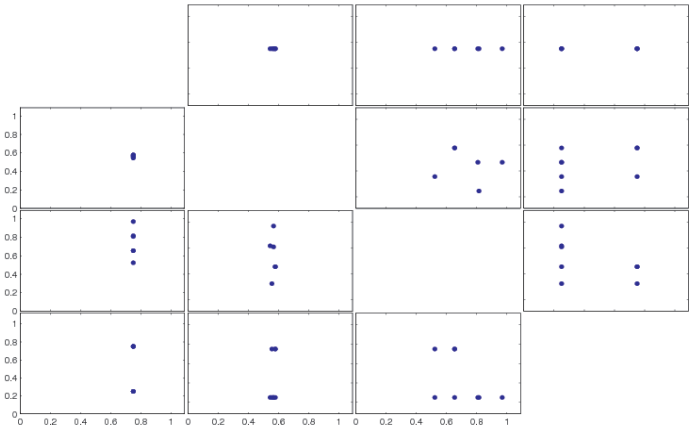
Figure 6.2. Observed field responses (black circles) and computer model responses (red x's).

The goal now is to identify new field trials to be run on the physical system. The IMSE optimal design with $m_f = 10$ was found. The expected percentage improvement in the IMSE is 21%, which is well worth the effort. Furthermore, Figure 6.3 displays the locations of the original field trials and the new IMSE optimal design. Notice that there are no replicated runs. The reason for this is likely that the follow-up trials are aimed at filling in the design space that was not adequately covered by the initial experiment design. Indeed, the initial field trials were not chosen for statistical purposes, but instead to validate specific conditions of particular interest to the scientists.

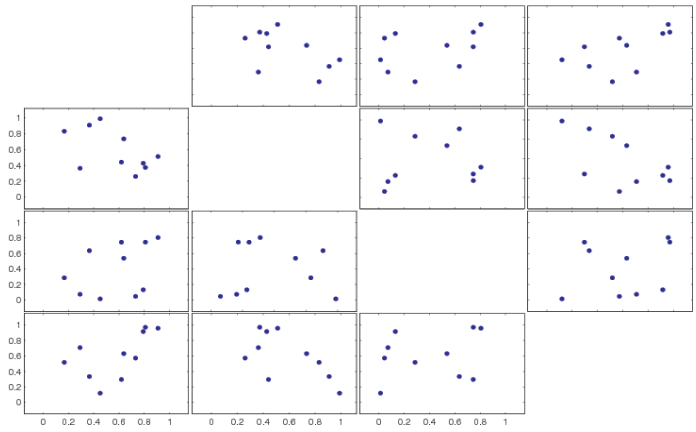
7. Final Comments

In this article we have introduced new methodology aimed at finding follow-up trials when the system under study is informed by both field and computer model data. The particular setting is one where calibration parameters are also present. We have found that a combination of field and computer model runs can substantially improve the predictive capability of the joint model outlined in Section 3.

By exploring a combination of potential design schemes, one is able to observe the impact of running new computer or field trials and also issues related to alignment and replication. Interestingly, we found that a random design is likely to improve the predictive model, in terms of IMSE, more substantially than replicating field trials alone. Of course, the IMSE



(a) Original field design



(b) IMSE optimal follow-up design

Figure 6.3. Two dimensional projections of (a) original field trials and (b) the IMSE optimal follow-up field trials for the CRASH experiments.

optimal design consisting of only field experiments gave the most improvement. However, a combination of replicated field trials, aligned computer trials and IMSE allocated field and computer trials still gave substantial gains in IMSE with fewer field experiments. We would recommend exploring such schemes routinely before conducting the IMSE optimal design with, say, field trials only because gains may be achievable with a fraction of the resources.

Acknowledgments

We would like to thank the reviewers for their comments and suggestions. This work was partially supported by Discovery grants from the Natural Sciences and Engineering Research Council of Canada and a National Science Foundation Collaborations in Mathematical Geosciences Grant (NSF 0934490). This research was also supported by the DOE NNSA/ASC under the Predictive Science Academic Alliance Program by grant number DEFC52-08NA28616.

References

Crary, S.B., 2002. Design of computer experiments for metamodel generation. *Analog Integrated Circuits and Signal Processing*, 32, 7–16.

Currin, C., Mitchell, T., Morris, M., Ylvisaker, D., 1991. Bayesian prediction of deterministic functions with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86, 953–963.

Harris C.M., Hoffman K.L., Yarrow L-A., 1995. Obtaining minimum-correlation Latin hypercube sampling plans using an IP-based heuristic. *OR Spektrum*, 17, 139–148.

Higdon, D., Kennedy, M., Cavendish, J.C., Cafeo, J.A. Ryne, R.D., 2004. Combining field data and computer simulation for calibration and prediction. *SIAM Journal on Scientific Computing*, 26, 448–466.

Johnson, M.E., Moore, L.M., Ylvisaker, D., 1990. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26, 131–148.

Jones, D.R., Schonlau, M., Welch, W.J., 1998. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13, 455–492.

Kennedy, M.C., O'Hagan, A., 2001. Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society, Series B*, 63, 424–462.

McKay, M.D., Conover, W.J., Beckman, R.J., 1979. A comparison of three methods for selecting the input variables in the analysis of the output from a computer code. *Technometrics*, 21, 239–245.

Miller, A.J., 1994. A Fedorov exchange algorithm for D-optimal design. *Journal of the Royal Statistical Society, Series B*, 43, 669–677.

Owen, A.B., 1998. Latin supercube sampling for very high-dimensional simulations. *ACM Trans. Model. Comput. Simul.*, 8, 71–102.

Sacks, J., Schiller, S.B., Welch, W.J., 1992. Design for computer experiments. *Technometrics*, 31, 41–47.

Sacks, J., Welch, W.J., Mitchell, T., Wynn, H.P., 1989. Designs and analysis of computer experiments (with discussion). *Statistical Science*, 4, 409–435.

Santner, T.J., Williams, B.J., Notz, W.I., 2003. *The Design and Analysis of Computer Experiments*. Springer, New York.

Schonlau, M., Welch, W.J., Jones D.R., 1998. Global versus local search in constrained optimization of computer models. In *New Developments and Applications in Experimental Design, 34*, Flournoy, N., Rosenberg, W.F., and Wong W.K. (Editors), 11–25, Institute of Mathematical Statistics.

Schur, I., 1917. Potenzreihn im innern des einheitskreises. *J. Reine. Angew. Math*., 147, 202–232.

Tang, B., 1993. Orthogonal array-based Latin hypercubes. *Journal of the American Statistical Association*, 88, 1392–1397.

Welch, W., Buck, R., Sacks, J., Wynn, H., Mitchell, T., Morris, M., 1992. Screening, predicting, and computer experiments. *Technometrics*, 34, 15–25.