Sequential Design for Functional Calibration of Computer Models

Ahmed Aziz Ezzat[a], Arash Pourhabib[b], and Yu Ding[a]

[a]Department of Industrial & Systems Engineering, Texas A & M University, TX, U.S.A

[b]Walmart Global eCommerce, San Bruno, CA, U.S.A

Abstract

The calibration of computer models using physical experimental data has received a compelling interest in the last decade. Recently, multiple works have addressed the functional calibration of computer models, where the calibration parameters are functions of the observable inputs rather than taking a set of fixed values as traditionally treated in the literature. While much of the recent works on functional calibration was focused on estimation, the issue of sequential design for functional calibration still presents itself as an open question. Addressing the sequential design issue is thus the focus of this paper. We investigate different sequential design approaches and show that the simple separate design approach has its merit in practical use when designing for functional calibration. Analysis is carried out on multiple simulated and real world examples.

KEY WORDS: Calibration, computer experiments, functional calibration, physical experiments, sequential design.

1. Introduction

Computer models are often used to mimic, understand and predict the behavior of complex physical processes. Oftentimes, the computer models contain a set of physically unobservable variables referred to as the "*calibration parameters*." The true values of these parameters are unknown, but they can be estimated using physical and computer experimental data such that the computer model aligns itself with its respective physical system. Such procedure is known in the literature as "*calibration*." Traditionally treated in the literature (Kennedy and O'Hagan 2001; Tuo and Wu 2015), the calibration parameters take a set of unknown fixed values. In more recent works, multiple researchers consider the more complicated circumstances where the calibration parameters are functions of the observable inputs in physical reality (Pourhabib et al. 2015; Atamturktur et al. 2015; Plumlee et al. 2016; Pourhabib et al. 2016). In a physical experiment, the settings of the input variables are to be chosen. In the computer experiments, however, the input variables and the calibration parameters can be controlled independently. The focus of our paper is on the design issue in the context of functional calibration, i.e, to investigate on how to sequentially generate designs for the physical and computer experiments, such that the estimation of the functional calibration parameters leads to a good alignment between the computer model outputs and their physical counterparts.

Let us illustrate the problem setting using the buckypaper fabrication experiments (Wang et al. 2004). Buckypaper is made of carbon nanotubes and has desired properties such as high tensile strength, relative to the thinness and light weight of the resulting buckypaper. To enhance the material's tensile strength, the polyvinyl acid (PVA) is added to the buckypaper and functions like glue. An important task in the fabrication process is to understand and test the effect of PVA on the strength of the buckypapers. For that purpose, a finite element analysis (FEA) model was developed to simulate the response of the buckypaper's tensile strength under different PVA levels (Wang et al. 2017). This computer model is to be calibrated using the outputs from a set of physical experiments.

The nominal amount of PVA is one of the dominating input factors, affecting the resulting tensile strength. Materials engineers running the experiments realize that in addition to

the nominal PVA amount, its absorption rate by the host material also affects the resulting tensile strength significantly (Pourhabib et al. 2015). The absorption rate, however, is not physically observable, and instead of being a fixed value over the whole input spectrum, it depends on the nominal PVA amount mixed with the host material. As such, the absorption rate is in fact a function of the observable input, i.e., the PVA amount. This dependency is understandable, because as more PVA is mixed with the hosting material, the absorption rate tends to decrease and the PVA effect appears saturated. The exact functional relationship between the PVA amount and the absorption rate is unknown but is likely of a nonlinear form. In the physical experiment, only the PVA amount can be chosen, while in the computer model, the absorption rate can also be freely and independently set, just like another input, in addition to the PVA amount.

In general terms, the design problem under consideration is as follows: **a)** The physical experiment has a set of input variables, $\mathbf{x} = (x_1, x_2, ..., x_p)^T$, that can be observed and adjusted during experimentation. Selecting the settings of $\mathbf{x}$ of dimension p constitutes a design, denoted by $D^P(\mathbf{x})$, where the superscript P refers to the design peculiar to the physical experiment. **b)** The computer experiment on the other hand, in addition to the same input vector $\mathbf{x}$, includes a set of extra inputs, $\theta = (\theta_1, \theta_2, ..., \theta_c)^T$, referred to as the calibration parameters. The design of the computer experiment is to select the settings for both $\mathbf{x}$ and θ of a combined dimension $p + c$ and the design is denoted by $D^S(\mathbf{x}, \theta)$, where the superscript S refers to the design associated with the computer experiment. **c)** In physical reality, the calibration parameters θ are functions of the observable inputs $\mathbf{x}$, i.e. they are functional parameters, rather than just taking on a set of fixed scalar values. In the buckypaper example, $\mathbf{x}$ = (PVA amount), θ = (absorption rate), $p = 1$, and $c = 1$.

Closely related to our problem in the sequential design literature is the sequential design of multi-fidelity (or multi-accuracy) computer-vs-computer experiments, in which sequential nested design-based strategies (Xiong et al. 2013; Le Gratiet and Cannamela 2015) are proposed. The core idea of nested designs is to generate small designs that are nested within larger designs. Small design(s) suggest sampling locations for the more expensive, high-fidelity computer experiments, whereas larger design(s) correspond to cheaper low-fidelity computer experiments. Specifically, Xiong et al. (2013) propose a sequential design

method for two-fidelity computer experiments based on the properties of augmentation and nesting of Latin hypercube designs (LHD). Le Gratiet and Cannamela (2015) propose a co-kriging-based nested design method in which not only the sampling locations for each fidelity level of the computer experiments are sequentially suggested, but also the decision concerning which fidelity level to execute in the next step is considered.

The nested designs are well suited for the circumstance of the computer-vs-computer experiments, where the input design regions for computer experiments at different fidelity levels have the same number of input variables. The physical-vs-computer experiments circumstance, as we are considering here, have different-sized input spaces for the two types of experiments, due to the presence of calibration parameters. Furthermore, as in the buckypaper example explained above (Pourhabib et al. 2015), as well as in several engineering applications reported in the literature (Bayarri et al. 2007; Atamturktur et al. 2015; Brown and Atamturktur 2018; Plumlee et al. 2016; Pourhabib and Balasundaram 2015; Pourhabib et al. 2016), the calibration parameters can be functions of the observable inputs. This setting creates a sequential design problem that the nested designs could not adequately handle. In fact, our analysis shows that the traditional sequential designs applied separately to each type of experiment could be more effective in many practical circumstances. Our study investigates the conditions under which a separate design may be preferred in practice. The insight garnered in this study, namely the merit of the separate designs for functional calibration, is generalizable to the case where the outputs of a physical experiment are integrated with multiple levels of computer codes, as in Goh et al. (2013).

The remainder of this paper is organized as follows. Section 2 provides a review of the calibration literature as well as the sequential design of experiments. Section 3 discusses the limitations of the nested designs and the merits of the separate design approach for functional calibration. In Section 4, we introduce a simple extension of Xiong et al. (2013) to account for the existence of calibration parameters. Two case studies are presented in Section 5 to demonstrate the merits of sequential design in real-world problems. Section 6 concludes the paper.

2. Literature Review

The scope of this work lies at the intersection of sequential experimental design and calibration of computer models. This section provides a review of the related background.

2.1 Calibration of computer models

In many real world applications, data could come from more than one source. For instance, data could come from multi-accuracy computer codes (Kennedy and O'Hagan 2000; Qian et al. 2006; Le Gratiet and Garnier 2014), or computer codes with tunable precision (Picheny et al. 2013; Tuo et al. 2013, 2014), or multi-resolution physical processes (Xia et al. 2011), or the situation where physical and computer simulation data are integrated (Kennedy and O'Hagan 2001; Reese et al. 2004; Qian and Wu 2008; Joseph and Melkote 2009; Li et al. 2016). Goh et al. (2013) further consider the integration of the output of a physical process with the outputs of multiple computer codes. In this paper, we are specifically concerned with the design issues that arise during the integration of computer experiment data with those coming from a physical system for the purpose of calibrating the computer model.

Denote by $\mathbf{x} \in \Omega$ the vector of explanatory and physically-observable variables, where Ω is a compact and convex subset of $\mathbb{R}^p$. Denote by y^P and y^S, respectively, the responses of the physical and computer experiments. Kennedy and O'Hagan (2001) propose a linkage model, as in Equation (1) below, to connect the two responses:

$$y^P(\mathbf{x}_i) = \rho y^S(\mathbf{x}_i, \theta) + \gamma(\mathbf{x}_i) + e_i \qquad i = 1, \ldots, n, \tag{1}$$

where ρ is a scale coefficient, $\gamma(\cdot)$ is a bias correction term, e_i is the i.i.d zero-mean normally distributed random variable representing the observational noise, and θ is the calibration parameter. Recall that the calibration parameters are model attributes that cannot be physically measured or observed, but can be included in the computer model and easily manipulated in computer experiments. Hence, y^S becomes a function of $\mathbf{x}$ and θ. Technically, y^P is also a function of $\mathbf{x}$ and θ, but because θ is not physically observable, the convention is to express y^P only in terms of $\mathbf{x}$.

In most applications, the computer model's response cannot be evaluated an infinite number of times. A common approach is to interpolate the missing values by constructing the unknown function $y^S(\cdot)$ using a surrogate model. Oftentimes, $y^S(\cdot)$ and $\gamma(\cdot)$ are both modeled as independent Gaussian processes. Gaussian process regression is a common tool in the computer experiment literature to reconstruct an unknown function based on a set of observations (Santner et al. 2003). A key issue in fitting a Gaussian process is to model the covariance structure across the domain of $\mathbf{x} \in \Omega$. Assuming isotropy, the squared exponential covariance function is a popular choice (Rasumussen and Williams 2006).

In Kennedy and O'Hagan (2001), θ is estimated through a Bayesian framework. Assuming $\rho = 1$, Tuo and Wu (2015) proposed to estimate θ by minimizing the distance between the two responses evaluated at a set of commonly sampled locations $\mathbf{x}^P$, that is,

$$\theta^* = \arg\min_{\theta \in \Theta} ||y^P(\mathbf{x}^P) - y^S(\mathbf{x}^P, \theta)||_{\mathcal{L}_2(\Omega)}, \tag{2}$$

where $||.||$ is the $\mathcal{L}_2$ norm. The prediction at any untried location $\mathbf{x}'$ is computed by simply plugging the values of $\mathbf{x}'$ and θ^* into the model of Equation (1).

Equation (2) assumes that the calibration parameter takes on a fixed value, regardless of the values of the observable inputs $\mathbf{x}$. We refer to this case, where $\theta = \theta^*, \forall\, \mathbf{x} \in \Omega$, as *fixed-value calibration*. However, several engineering applications have been recently reported in the literature that explicitly mention the existence of a functional relationship between calibration parameters and observable variables (Bayarri et al. 2007; Pourhabib et al. 2015; Pourhabib and Balasundaram 2015; Atamturktur et al. 2015; Brown and Atamturktur 2018; Plumlee et al. 2016). Setting it apart from the *fixed-value calibration*, this case is referred to as *functional calibration*, since θ becomes a function of $\mathbf{x}$ but the functional form is unknown prior to the execution of experiments, data collection, and parameter estimation. Obviously, fixed-value calibration, which assumes θ is a constant function of $\mathbf{x}$, is a special case of functional calibration.

For functional calibration, Pourhabib et al. (2015) and Atamturktur et al. (2015) propose parametric approaches in which $\theta = \theta(\mathbf{x})$ assumes a parametric functional form and the parameters therein are estimated by minimizing a distance function. Recently, several

lines of research have been conducted for developing nonparametric functional estimations of $\theta(\mathbf{x})$ (Plumlee et al. 2016; Brown and Atamturktur 2018; Pourhabib and Balasundaram 2015; Pourhabib et al. 2016). Specifically, Pourhabib et al. (2016) extends Equation (2) to the functional calibration case by replacing the scalar parameter θ by the functional parameter $\theta(\mathbf{x})$, employing the sample average instead of the $\mathcal{L}_2$ integral and adding a regularization term. As such, the functional calibration formulation proposed by Pourhabib et al. (2016) is to minimize a penalized distance function given $\{(\mathbf{x}_i^P, y_i^P, y_i^S); i = 1, 2, .., n^P\}$, as in Equation (3):

$$\hat{\theta}(\cdot) = \underset{\theta(\cdot)}{\arg\min} \ \frac{1}{n^P} \ \sum_{i=1}^{n^P} \left\{ y^P(\mathbf{x}_i^P) - y^S(\mathbf{x}_i^P, \theta(\mathbf{x}_i^P)) \right\}^2 + \lambda \sum_{j=1}^{q} ||\theta_j||^2_{\mathcal{N}_{K_j}}, \tag{3}$$

where each θ_j lies in a reproducing kernel Hilbert space such that $\mathcal{N}_{K_j(\Omega)}$ is the native space generated by the kernel function $K_j(\cdot, \cdot)$ with the corresponding norm $||\theta_j||_{\mathcal{N}_{K_j}}$ as a measure of roughness of the j^{th} component of $\theta(\cdot)$ and λ is a smoothing parameter. While the functional estimation in Equation (3) may not be universally accepted, it is still a useful and sensible estimator. Therefore, we adopt this specific nonparametric functional calibration method to be used for our later design tasks.

If y^P and y^S are not evaluated at the same set of locations, a reasonable approximation is to use the surrogate model predictions as a substitute for the true value of the computer model response in Equations (2) and (3). Such a situation occurs when the computer model's response is not sufficiently cheap or when the design generation for the physical and computer experiments is conducted separately.

2.2 Sequential Design of Experiments

Sequential sampling strategies can be classified, depending on the experimental objective, into designs for optimization, which attempt to sample at locations that optimize a target response, and designs for exploration, which are concerned with optimally exploring a response surface. For instance, the expected improvement criterion is a design strategy for optimization that sequentially selects points to maximize/minimize a response (Jones et al. 1998; Williams et al. 2000). On the other hand, Maximum Entropy Sampling (MES)

(Shewry and Wynn 1987), Active Learning Mackay (Mackay 1992) and Active Learning Cohn (Cohn et al. 1996) are information-driven criteria that optimally explore a response surface. Latin hypercube designs (Mckay et al. 1979), distance-based designs (Johnson 1990) and uniform designs (Fang 1979) spatially fill the input space for exploration purposes.

In particular, we review the Maximum Entropy Sampling (MES) criterion for its relevance to the learning framework of this paper, its suitability for designing both physical and computer experiments, and its easy adjustability to accommodate sequentiality. MES suggests sampling at locations which maximize the information gain about the model parameters. The expected information gain achieved by conducting an experiment is quantified as the expected difference between the amount of information prior to and post sampling. Shewry and Wynn (1987) prove that the maximal information gain is achieved by sampling at locations which maximize the entropy of the observed responses, hence absorbing more in-sample variability. When using a Gaussian process to model the response, maximizing entropy simplifies to maximizing the determinant of the correlation matrix (Shewry and Wynn 1987; Gramacy and Lee 2009; Tuo et al. 2014). A *sequential* MES strategy is to fix the already-sampled locations and optimize over the remaining sample space.

An emerging field that is related to our problem setting, is the sequential design of multi-fidelity computer-vs-computer experiments. Xiong et al. (2013) propose a sequential design framework for a pair of computer experiments with varying accuracies based on nested Latin hypercube design. Nested LHDs were originally proposed by Qian (2009) allowing smaller LHD's of n runs to be augmented to larger LHDs with $m = sn$ runs, where $s \in \mathbb{Z}_+$ is the data ratio of the small-to-large LHDs. In their proposed method, Xiong et al. (2013) suggest that small LHDs correspond to the locations at which the high-fidelity computer experiment response is evaluated, whereas the low-fidelity computer experiment response, being cheaper, is obtained at the locations suggested by larger LHDs. The high and low-fidelity outputs are then integrated using a linkage model, as the one proposed in (Kennedy and O'Hagan 2000), and the nesting step is repeated in a sequential experimental fashion, until the integrated model reaches a pre-set accuracy or a pre-determined design size.

3. Separate and Nested Designs for Functional Calibration

Section 3.1 discusses the limitations of nested designs in the presence of functional calibration parameters. Section 3.2 takes a closer look at the separate design approach, a simple design strategy that is effective for functional calibration. In Section 3.3 and 3.4, numerical analyses are conducted to demonstrate the points thus discussed.

3.1 Nested design and its limitations

With the presence of calibration parameters, $D^S(\mathbf{x}, \theta)$ has a higher dimension than $D^P(\mathbf{x})$. In order to apply the nested design, a straightforward treatment is to leave out the calibration parameters in the design stage and restrict the nested design to the common set of variables in $\mathbf{x}$. Such a design enables the evaluation of $y^P(\mathbf{x})$ but not of $y^S(\mathbf{x}, \theta)$, as θ is not specified in the design stage. A simple adjustment is to "fill out" the missing samples of θ by using randomly selected values.

The nested design approach is not desirable when designing physical-vs-computer experiments, for three reasons. First of all, a common feature of computer experiments is that the input factors can be set to any arbitrary level at no cost. In a nested design, the high-fidelity computer experiment design is a less dense version of its low-fidelity counterpart and both responses would be evaluated at the exact sampling locations suggested by the high-fidelity design. In a physical experiment, however, it is not easy to set the factors as accurate specific values, which makes continuous designs like the LHDs, and naturally the nested LHDs as well, less practical in the design of physical experiments. The continuous designs could suggest sampling locations for which the physical response cannot be evaluated.

Second, calibration parameters are present in the physical-vs-computer experiments. The existence of calibration parameters creates a dimensionality mismatch between the two sets of designs, deeming the nested design in its current form impractical, and a random sampling treatment inevitable, as the one described above. Moreover, the calibration parameters can take on a functional form, as the buckypaper example shows, and as such, their functional estimation could heavily rely on the quality of the input designs.

Third, the data ratio parameter s in a nested design can only take positive integer values, namely $s = \{2, 3, \ldots\}$. Given the high cost that is often associated with the physical experiments as in the case of the buckypaper fabrication, it is impractical to set the parameter s to a value higher than 2. A large value for s makes the size of the design for the physical experiments increase drastically and the design would quickly become infeasible, because the current design for the computer experiment becomes the design for its physical counterpart in the next iteration of the sequential procedure. Practically, the data ratio parameter s is fixed at the value of 2 throughout the sequential design process, but doing so takes away the flexibility of varying the data amount ratio, when needed, to achieve a better physical-to-simulation data balance.

3.2 Separate designs for calibration

As its name implies, *separate design* handles the physical and computer experiment designs independently through two separate sequential design processes. The integration between the two designs takes place in the linkage modeling step as opposed to in the design phase. The steps of generating a separate design are explained in Algorithm 1. Here, one could use two different design methods, $\mathcal{M}^P$ and $\mathcal{M}^S$, to generate the physical and computer simulation designs, respectively, depending on the application of interest. For instance, an experimenter can choose a classical factorial design for the physical experiment, while using a space-filling design for the computer experiment.

Admittedly, the separate design is a naive approach. But surprisingly, when designing for functional calibration, separate designs could have desirable properties in practice. Separate designs generate the physical and computer experiment designs without concerning the dimensionality constraint, enabling reliable calibration, whether constant or functional, using Equations (2) and (3), respectively. Moreover, as shown in Algorithm 1, separate designs give the experimenter the choice of different design strategies for the physical and computer experiments. Given the different nature of the physical and computer experiments, such advantage can be meaningful in practice. Additionally, separate designs entertain a full flexibility in varying the physical-to-simulation data amount ratio at each iteration of the sequential procedure, which is a flexibility lacking in the nested designs.

Algorithm 1 Separate sequential design for physical-vs-computer experiments with functional calibration parameters.

0. Set the design threshold N^P.
1. Let $d = p + c$, where p and c are the number of observable variables and calibration parameters, respectively.
2. Specify the sampling methods, $\mathcal{M}^P$ and $\mathcal{M}^S$, used to generate the physical and computer experiment design points, respectively.
3. At $i = 1$, randomly initialize a p-dimensional design for D_i^P and a d-dimensional design for D_i^S.

repeat

4. Evaluate y^P and y^S associated with D_i^P and D_i^S, respectively.
5. Estimate $\theta(\mathbf{x})$ by solving Equation (3).
6. Integrate the physical and simulation data using Equation (1).
7. Use $\mathcal{M}^P$ to find $\mathbf{x}_{new}^P$ and $\mathcal{M}^S$ to find $(\mathbf{x}^S, \theta^S)_{new}$. Concatenate D_i^P and $\mathbf{x}_{new}^P$ to form D_{i+1}^P, and D_i^S and $(\mathbf{x}^S, \theta)_{new}$ to form D_{i+1}^S. Now, n^P is set to the current number of design points in D_{i+1}^P.

until $n^P \geq N^P$

One of the drawbacks of separate designs, however, is that such designs cannot ensure that y^P and y^S are evaluated at the same locations $\mathbf{x}^P$, which is required in the linkage phase. When there is no observed y^S at a location, the prediction based on its surrogate model is then used as a substitute in the linkage model. Nested designs, on the other hand, were specifically devised to ensure the co-location of y^P and y^S at $\mathbf{x}^P$'s through the nesting property, so that the substitution of y^S by $\hat{y}^S$ is not needed. On the other hand, nested designs are not entirely free of involving a surrogate model, as the inclusion of calibration parameters into nested designs requires the use of surrogate model predictions to enable the estimation of calibration parameters. Separate designs also do not consider the relative location of $D^S(\mathbf{x}) \setminus D^P(\mathbf{x})$ with respect to $D^P(\mathbf{x})$, while nested designs account for such relative design configuration which provides a higher explorative capability for the input space of the observable variables.

It is apparent that in the presence of calibration parameters, separate and nested designs could have their own pros and cons. A question of interest is to investigate the merits of each strategy with respect to the calibration framework. Our research suggests that three performance influencing factors should guide the choice of the appropriate design approach. The three factors are: the computer experiment data amount n^S, the physical data amount

n^P, and the form of the calibration parameter θ.

As more computer experiment data become available, that is, n^S gets larger, the confidence in a trained surrogate model to accurately reconstruct the unknown function $y^S(\cdot)$ increases. When $\hat{y}^S \approx y^S$, the adverse impact due to lacking co-location of inputs that separate designs suffer from, is reduced. This observation suggests that the performance of separate designs improves when n^S is large. This condition can be easily satisfied when the computer experiment is cheap, because it can be executed in a large number of runs. The total flexibility of the physical-to-simulation data ratio in separate designs makes it possible to take full advantage of a cheap computer experiment at each iteration in a sequential procedure, as opposed to nested designs that restrict the data ratio to 1-to-2.

The performance of separate designs also tends to improve when the amount of physical data, n^P, is large. Considering the cost of physical experiments relative to their computer experiment counterparts, the amount of computer experiment data will be even greater when there are plenty of physical data. Following the logic explained above, the disadvantage of separate designs in terms of lacking the co-location is then greatly alleviated. In the meanwhile, larger sample sizes of both physical and computer experiment data improve the quality of estimation of the calibration parameters, especially if they are of a functional form, e.g., the $\theta(\mathbf{x})$ in the buckypaper example. In general, functional estimations are known to be data-dependent/intensive. Nested designs, which do not account for the presence of calibration parameters in the design stage and do not make use of a cheaper computer experiment by acquiring more data, can be prone to poor functional estimation. It is therefore not surprising that nested designs are often followed by estimating a constant θ^* through the fixed-value calibration. In the separate designs, on the other hand, the calibration parameters are considered as part of the design inputs and the physical-to-simulation data balance could be adjusted accordingly to ensure a good estimation capability even for functional calibration.

The discussion on the above two factors leads us to expect that when θ is of a functional form, the separate designs allow for a better performance than the nested designs as the sequential procedure progresses and the data amounts get greater, which we do observe in our numerical analyses. On the other hand, in the simple and traditional case, where θ

has a fixed and unknown value, we should not expect a notable difference between the two approaches.

The third performance-influencing factor is the form of the calibration parameter in the application of interest. Naturally, a complex functional relationship $\theta(\mathbf{x})$ needs more care in selecting the sampling locations. In the meanwhile, it becomes less reasonable to assume it a fixed constant as in the fixed-value calibration formulation. We therefore expect separate designs which are suitable for the more general functional calibration setting to show a greater advantage over nested designs when $\theta(\mathbf{x})$ is of a nonlinear form with increasing complexity.

3.3 Numerical analysis

To demonstrate the above-mentioned points, consider the following simulated example where the physical response is given by $y^P = \exp(\frac{x}{10})\sin(x) + \epsilon$, and $\epsilon \sim \mathcal{N}(0, 0.05)$. The associated computer model is imperfect and its response is given by $y^S = \exp(\frac{x}{10})\sin(x)(\frac{x}{2\theta}) - (\frac{x-1}{9})^2$. The true function relating the observable variable $x \in [\pi, 3\pi]$ to the calibration parameter $\theta \in [0.5\pi, 1.5\pi]$ is $\theta(x) = \frac{x}{2}$. The physical, computer simulation and calibration response surfaces are illustrated in Figure 1.

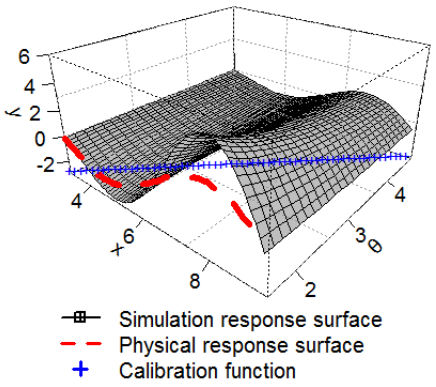

Figure 1: The simulation response as a function of both x and θ in 3D, the physical response as a function of x only (dashed line), and the calibration function $\theta(x)$ (crosses).

We compare three approaches, which are: a separate design followed by functional calibration, a separate design that uses cheap computer experiments followed by functional

calibration, and a nested design followed by fixed-value calibration. The three approaches are simply referred to as the separate design, the cheap separate design, and the nested design, respectively.

For the separate design approaches, the MES criterion is used to generate two individual designs of different dimensions, namely a 1-D design for x and a 2-D design for (x, θ). The two designs are then fed into the functional calibration by solving Equation (3). Prediction at a new location x' is computed by plugging $(x', \theta(x'))$ into a Gaussian process model, trained on the computer experiment data $\{(x_i^S, \theta_i, y_i^S); i = 1, .., n^S\}$ and adjusted by the physical responses. For the nested design approach, we generate a 1-D nested LHD for x. A vector of random values is chosen for the calibration parameter for the purpose of evaluating $y^S(x, \theta)$. We then use the physical and computer experiment responses to find a fixed constant θ^* that minimizes Equation (2). Prediction at a new location x' is then made by plugging (x', θ^*) into the Gaussian process model, trained on $\{(x_i^S, \theta_i, y_i^S); i = 1, .., n^S\}$ and adjusted by the physical responses.

In the approach besides the cheap separate design, the prediction performance is evaluated at sample sizes of proportional physical-to-simulation ratios, namely $(n^P, n^S) = \{(1, 2); (2, 4); (4, 8); (8, 16); (16, 32); (32, 64); (64, 128)\}$. Those sample sizes are chosen to maintain the amount ratio constrained by the nested LHDs. In the cheap separate design, because obtaining computer experiment data is assumed to be relatively cheaper, the physical-to-simulation data ratio is no longer restricted to 1-to-2. Specifically, we fix $n^S = 128$ at each iteration regardless of the physical data amount used.

For the purpose of comparison, 3 random test sets, each with 10 points uniformly spread over the input region and 3 random design sets, are used. Given a set of design points, the physical and computer experiment data are generated and the trained linkage model of Equation (1) is used to make predictions at each of the 10 points in one of the test sets. Then, a root mean squared error (RMSE) is computed for this design-test combination. Repeat this process for all nine design/test combinations for each design scenario. Eventually, the average root mean squared error from all nine combinations along with its standard deviation, are reported as the performance measure of a design approach.

Figure 2 presents the performance of each design approach as the data amount increases,

in which the horizontal axis is the amount of the physical experiment data, n^P. The separate design and nested design approaches have a specific corresponding n^S that increases proportionally with n^P. We note that the nested design outperforms the separate design at the small sample sizes. As the data amount increases, however, this order is reversed. The superiority of a functionally calibrated separate design over a fixed-value calibrated nested design is much as anticipated. The cheap separate design further demonstrates the advantage of separate designs: in relaxing the data ratio constraint, RMSEs can be further lowered.

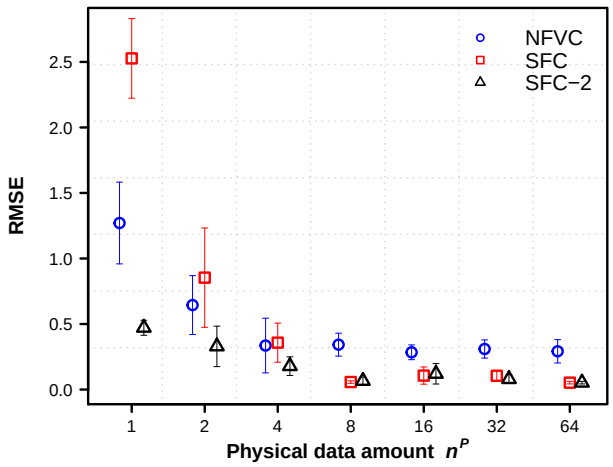

Figure 2: The average RMSE and its standard deviation of different design approaches as the data amount increases. SFC: Separate design followed by functional calibration; SFC-2: Separate design that uses cheaper computer experiments followed by functional calibration; NFVC: Nested design followed by fixed-value calibration.

We would like to further demonstrate the message using examples of higher dimensional inputs. We consider two additional simulated examples of higher dimensions. For the second simulated example, the physical design consists of two observable variables ($p = 2$), denoted by x_1 and x_2, and the physical response is given by $y^P = \exp(\frac{x_1}{10})\sin(x_2) + \epsilon$, $\epsilon \sim \mathcal{N}(0, 0.05)$. The simulation design consists of two observable variables and one calibration parameter, where the simulation response is given by $y^S = \exp(\frac{x_1}{10})\sin(x_2)(\frac{x_2}{2\theta}) - \sqrt{\frac{x_2}{5}}$, such that $\theta(x_2) = \frac{x_2}{2}$ is a linear univariate function, $(x_1, x_2) \in [\pi, 3\pi]$, and $\theta \in [0.5\pi, 1.5\pi]$. In the third simulated example, the physical design consists of four observable variables ($p = 4$), where the physical response is given by $y^P = 0.25x_1^2 + 0.75x_2 + 0.5x_3 + 0.25x_4 + \epsilon$, $\epsilon \sim \mathcal{N}(0, 0.05)$.

The simulation design consists of four observable variables and one calibration parameter, where the associated response is given by $y^S = \frac{(x_1^2+3x_2)^2}{8\theta} + 0.5x_3 + 0.25x_4 - \frac{x_1 x_2}{5}$, such that $\theta(x_1, x_2) = \frac{x_1^2+3x_2}{2}$ is a bivariate quadratic function, $(x_1, x_2, x_3, x_4) \in [1, 2]$, and $\theta \in [2, 3.5]$.

Figure 3 presents the comparison between the separate design and the nested design for these high-dimensional examples. It is apparent that the findings suggested in the single dimension simulated examples are extendable to their high-dimensional peers. Moreover, it appears that, as the calibration function $\theta(\mathbf{x})$ takes on a more complicated functional form, as the one in the third simulated example, the advantage gained by adopting the separate design appears to be more notable.

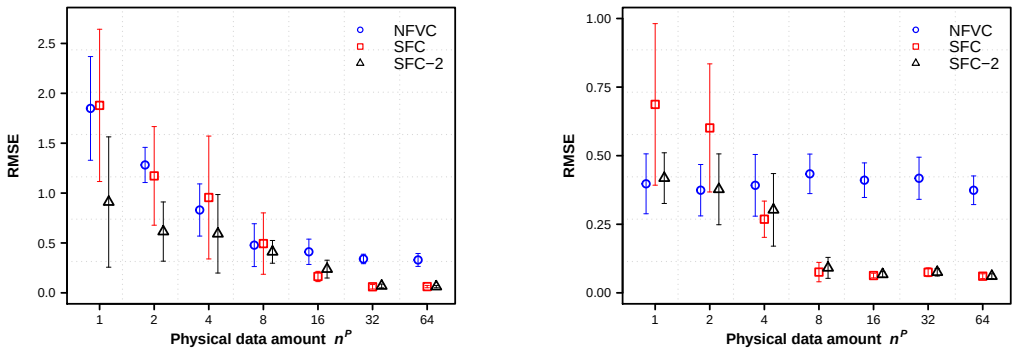

Figure 3: The left panel is for $p = 2$ and the right panel is for $p = 4$. SFC: Separate design followed by functional calibration; SFC-2: Separate design that uses cheaper computer experiments followed by functional calibration; NFVC: Nested design followed by fixed-value calibration.

To summarize the analysis, a separate design is expected to be more advantageous than a nested design in the intermediate and later stages of a sequential design procedure. The merits of a nested design are continuously taxed by its inability to capture the nature of the calibration parameter and its constrained physical-to-simulation data ratio. Generally speaking, only when data-scarcity of both sets of experiments is the dominant trait would a nested design provide a relatively competitive performance. On the other hand, if the computer experiment data is significantly cheaper than their physical counterparts, a separate design is preferred for its ability to exploit the information in a denser computer experiment design grid. Moreover, if $\theta(x)$ is known *a priori* to be sufficiently complex, the advantage brought by separate designs in capturing this relationship ultimately favors its

Scenario	Formula
1.	$y^S = \exp(\frac{x}{10})\sin(x)(\frac{x}{2\theta})$
2.	$y^S = \exp(\frac{x}{10})\sin(x)(\frac{x}{2\theta}) - 0.25$
3.	$y^S = \exp(\frac{x}{10})\sin(x)(\frac{x}{2\theta}) - (\frac{x-1}{9})^2$

Table 1: Response functions for the computer models.

use over nested designs. In practice, a calibration parameter $\theta(\cdot)$ is oftentimes known to be functional but its specific form is unknown. Then, it is safer to employ the separate design strategy under such circumstance.

3.4 Further Discussion

In this subsection, we discuss the issue of identifiability, which has been repeatedly raised in the calibration literature (Loeppky et al. 2006; Arendt et al. 2012; Tuo and Wu 2015). Simply speaking, identifiability arises in the context of calibration, when the computer model's deviation from the physical system is sufficiently sizeable, deeming the effect of the calibration parameter and that of the model bias indistinguishable. In order to explore that issue, we compare three variants of the first simulated example presented in Section 3.3. The three variants have the same physical response given by $y^P = \exp(\frac{x}{10})\sin(x) + \epsilon$, and $\epsilon \sim \mathcal{N}(0, 0.05)$, but the associated computer models are different and their response functions are given in Table 1.

In the first scenario, the computer model has no model bias and the only source of discrepancy between the simulation and physical system comes from the unknown calibration parameter. Should the true $\theta(x)$ be known, the outputs from the computer model would exactly match those of the physical system. The second and third scenarios represent the response functions of an "imperfect" computer model, in which there is an inherent bias between the simulation and physical responses. The third scenario includes a model bias term of a complex functional form; this third scenario is what was used in the study in Section 3.3. Given these three scenarios, we take a deeper look into the functional calibration estimation procedure, as well as the resulting predictive performance.

Specifically, we compare the three scenarios in Figure 4 using two metrics: the calibration error and the model prediction error. The calibration error, denoted by Err_c,

represents the discrepancy between the estimated and the true calibration parameter, as defined in Equation (4).

$$Err_c = \sqrt{\frac{\sum_{i=1}^{n^P} \left[\theta(x_i^P) - \hat{\theta}(x_i^P)\right]^2}{n^P}}. \tag{4}$$

The prediction error, in RMSE, measures how well the calibrated response surface predicts the true response at the test locations. This is the same metric presented in Figures 2 and 3 of Section 3.3.

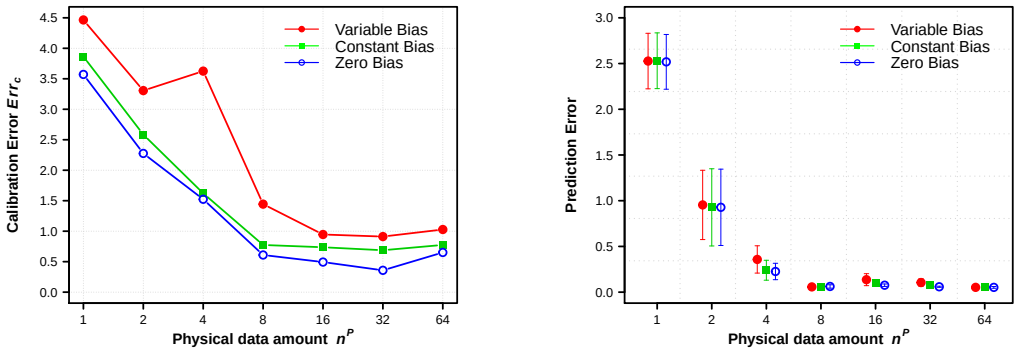

Figure 4: The left panel presents the calibration error, whereas the right panel presents the prediction error for the three scenarios. Zero bias scenario: blue; constant bias scenario: green; and variable bias scenario: red.

By looking at the left panel of Figure 4, we note that the calibration error increases as the model's bias takes on a more severe and complex functional form, with the lowest error associated with the unbiased computer model across all sample sizes. As the sample size gets larger, i.e., more information about both the physical and simulation models is acquired through sequentially adding more design points, it appears that this effect does not vanish, but it is mitigated. This is understandable, since identifiability issues are not directly resolved by simply adding more design data, as concluded by Arendt et al. (2012). In the right panel, the three scenarios perform comparably when it comes to prediction accuracy.

The observation leads us to conclude that for imperfect computer models with sizeable model bias, a perfect estimation of the calibration parameters is not necessarily achievable;

that is to say, making reliable inference about the calibration parameters cannot be guaranteed. We understand that making inference about calibration parameters, especially when they are physically meaningful, is sometimes desirable in engineering applications. On the other hand, the goal of calibration, in many other applications such as originally introduced in Kennedy and O'Hagan (2001), is to align the computer model with the physical system. Such objective appears achievable, even in the presence of sizeable computer model bias. For the buckypaper fabrication example, an enhanced prediction as achieved by the calibrated computer model can help decide the optimal amount of PVA to be added, while a biased, imperfect estimation of the absorption parameter, in and by itself, does not harm the decision making process.

4. Nested Designs with Functional Calibration Parameters

One of the motivations behind using separate designs is its ability to account for the dimensionality difference between the physical and computer experiment data in the presence of calibration parameters by generating two respective optimal designs. Here we propose a simple extension of the nested design by Xiong et al. (2013) so that the extended version can accommodate calibration parameters in the design phase, thus exploiting other advantages offered by the nested designs.

Recall that we have p input variables and c calibration parameters. We are interested in designing a sequential selection strategy that generates a p-dimensional physical experiment design and a d-dimensional computer experiment design at each iteration, such that $d = p + c$. We call the resulting approach a functional calibration nested design.

Specifically, at the first iteration of the sequential procedure, a pair of d-dimensional nested LHDs are generated using the method of Xiong et al. (2013), such that $D_1^P \subset D_1^S$. Then, D_1^P is projected onto the p-dimensional design subspace of input variables to form the physical design, $\hat{D}_1^P$. The physical and computer experiment responses are evaluated at the projected $\hat{D}_1^P(\mathbf{x})$ and $D_1^S(\mathbf{x}, \theta)$, respectively. The two sets of responses are then fed to Equation (3) to estimate the functional calibration. Subsequent experiments in the sequential procedure can be conducted using the augmentation property, namely

augmenting D_1^S to a larger LHD to form D_2^S, and then, letting D_1^S be D_2^P, whose projection into the p-dimensional subspace forms again the corresponding physical design $\hat{D}_2^P$. The steps of generating a functional calibrated nested design are presented in Algorithm 2.

Algorithm 2 Functional calibrated nested design

1: Set design threshold N^P. Let $d = p + c$, where p and c are the number of observable variables and calibration parameters, respectively.
2: At $i = 1$, initialize a pair of d-dim nested LHDs $\{D_i^P, D_i^S\}$ such that $D_i^P \subset D_i^S$.
3: Project the d-dimensional design D_i^P onto a p-dimensional space to obtain $\hat{D}_i^P(\mathbf{x})$.
4: **repeat**
5: Evaluate y^P and y^S associated with $\hat{D}_i^P$ and D_i^S, respectively.
6: Feed $\hat{D}_i^P$ and D_i^S into a functional calibration estimation to estimate $\theta(\mathbf{x})$ by solving Equation (3).
7: Integrate the physical and simulation data using Equation (1).
8: Let $D_{i+1}^P = D_i^S$, project D_{i+1}^P onto the p-dimensional space to form $\hat{D}_{i+1}^P$ and enlarge D_i^S to form a larger LHD D_{i+1}^S. Now, n^P is set to the current number of design points in $\hat{D}_{i+1}^P$.
9: **until** $n^P \geq N^P$

We illustrate the merit of the functional calibration nested design approach in terms of its predictive performance, likewise defined in Section 3.3. At sample sizes of (n^P, n^S) = {(1, 2); (2, 4); (4, 8); (8, 16); (16, 32); (32, 64); (64, 128)}, Algorithm 2 is employed to generate the respective physical and computer experiment designs, which are fed to a functional calibration, as in Equation (3).

Figure 8 presents the comparison between the functional calibration nested design with aforementioned three other approaches, using the three sets of simulated data from Section 3.3. Overall, the functional calibration nested design outperforms the earlier version of the nested design in nearly all data sizes. The most noticeable benefit of the functional calibration nested design is that it can improve upon the separate design at the regions where physical data size is small, so that the functional calibration nested design performs more robust than the separate design. The functional calibration nested design also competes with the cheap separate design in most of data sizes, except where physical data size is small. The functional calibration nested design keeps the data ratio restriction and does not take the full advantage of the abundance of cheap computer experiment data even if they are available; this restriction apparently hurts all forms of nested design, albeit at an alleviated degree for the functional calibration nested design.

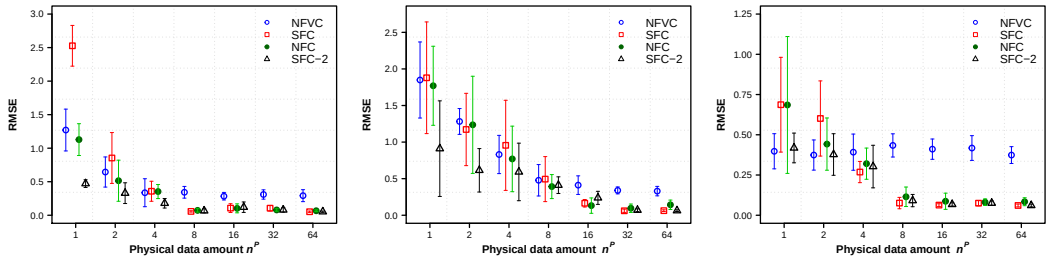

Figure 5: Comparison of the functional calibration nested design versus the other three design approaches using the three simulated datasets, where $p = 1$ (left panel), $p = 2$ (middle panel) and $p = 4$ (right panel). SFC: Separate design followed by functional calibration; SFC-2: Separate design that uses cheaper computer experiments followed by functional calibration; NFVC: Nested design followed by fixed-value calibration; NFC: functional calibration nested design.

5. Case Studies

In this section, the merit of the sequential separate designs in the presence of functional calibration parameters is demonstrated using two real-world examples. The first example is the fabrication process of PVA-treated buckypaper, which involves a 1-D physical design and a 2-D computer simulation design. The second example is a resistance spot-welding experiment, which involves a 2-D physical design and a 3-D computer simulation design.

5.1 Buckypaper fabrication

We briefly explained the buckypaper fabrication process in Section 1, the full details can be found in Wang (2013). Its computer simulation model (Wang et al. 2017) is based on a finite element analysis, in which each carbon nanotube is treated as a bar with certain length and diameter. Tens of thousands of the nanotubes are randomly simulated with varying sizes, locations and orientations in the hosting epoxy resin matrix. The effect of PVA is simulated as the key binding mechanism that glues the nanotubes together, so that the nanotubes form a network enhancing the tensile strength of the resulting buckypaper. To estimate the functional calibration parameter (PVA absorption rate), we use the functional calibration approach in Equation (3), which was proposed in Pourhabib et al. (2016).

The data for the physical and computer experiments have already been collected in pre-

vious studies. The physical experiment was run at 17 sampling locations uniformly chosen over the range of $x \in [0.4, 1.2]$ by an increment of 0.5. As for the computer simulation data, 149 data points were obtained from the computer experiments through a non-rectangle, 2-D space-filling design; for details, please see Pourhabib et al. (2015). In this case study, the computer experiments are sufficiently cheap such that the number of simulation data points is one order of magnitude greater than that of the physical data.

Since the physical and simulation data have already been obtained from the previous experiments and they do not have the nested structure, it is not possible for us to generate the nested designs for the case study in hand. For this reason, we here test the predictive performance of the separate design strategy, specifically, the cheap separate design, by applying the MES criterion on a discrete search space. We compare the performance of the separate sequential design over a single-stage design strategy that does not account for the sequentiality but rather generates the physical design all at once by randomly sampling from the pool of available data with a designated sample size. We reserve four design points for testing and ensure that the test points are a mix of interior and peripheral points of the design region, so that we test both the interpolation and extrapolation ability of the model.

The left panel of Figure 6 compares the predictive RMSE of the cheap separate design versus multiple realizations of the single-stage design. In the cheap separate design, the physical data is increased through an increment of one data point at a step, whereas the simulation data is fixed at $n^S = 149$.

In the left panel of Figure 6, there is a significant improvement in the performance of the cheap separate design when the number of physical data points increases from 1 to 5 but the improvement beyond that is marginal. In a sequential manner, the cheap separate design quickly learns the physical response surface, as well as the functional calibration relationship $\theta(x)$, so that the experimenter could possibly stop after exhausting six or seven physical data points. In doing so, it saves the time and expense in the experimental stage. By contrast, the single-stage designs suffer from a large variability in performance; in some extreme cases, say realization #3, it can exhaust nearly all the available physical resources to reach the desired level of accuracy.

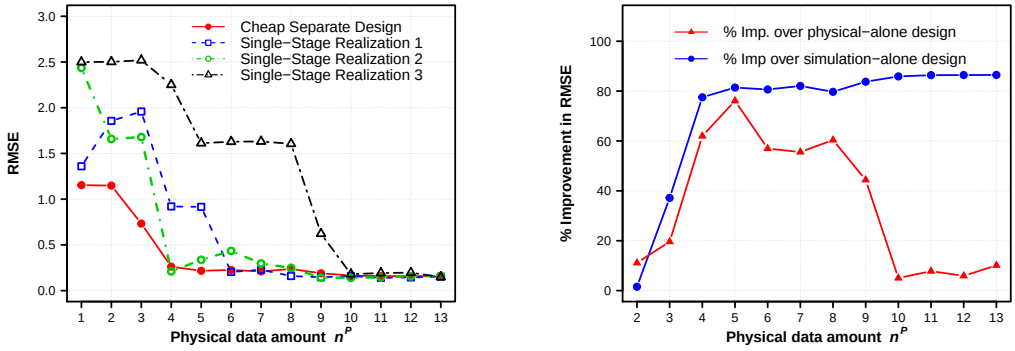

Figure 6: Buckypaper fabrication case study: comparison of the cheap separate design with the single-stage designs (left panel) and comparison with the physical-alone and simulation-alone designs (right panel).

We would also like to show the benefit of a sequential design for calibration over designs that only use the physical or the simulation data alone. Please note that when using only one source of data without the other, the calibration parameter does not play a role. When using the simulation data without the physical data, for instance, the calibration parameter can no longer be estimated and becomes a nuisance factor in prediction. For the simulation-alone design, all the 149 simulation data points are used to develop the simulation model, while in both the cheaper separate design and the physical-alone design, a sequential procedure is adopted. The right panel of Figure 6 presents the percentage improvements of the cheap separate design over the other two approaches in terms of prediction accuracy. The x-axis again represents the physical data amount employed in the designs at each run.

When the physical data amount is small, the performance of the cheap separate design converges to that of the simulation-alone design because the predictive model is dominated by the simulation data. The limited information provided by a small amount of physical data is not enough to allow a quality estimation of $\theta(x)$. As the amount of physical data increases, cheap separate design improves upon the simulation-alone design steadily. As more and more physical data points become available, using the simulation data alone is simply not a good option anymore. The improvement made by cheap separate design over the physical-alone design initially increases, reaches a peak when there are around five physical data points, and then gradually declines afterwards. When there are enough

physical data points, using the physical data alone gradually converges to the performance of the calibrated model. This is much expected because as physical data become denser, the under-sampling error is gradually alleviated and then using the physical data solely becomes self-sufficient. A similar finding is observed in Pourhabib et al. (2015).

5.2 Resistance spot-welding experiment

The second case study is a resistance spot-welding experiment that was initially studied by Bayarri et al. (2007). The observable input variables are x_1, the applied load that compresses the water-cooled copper electrodes against the two metal sheets to be welded and x_2, the electric current intensity applied to melt a local area of the sheet metals and thus weld them together. The welding action produces a weld nugget and the size of the nugget is the response y. The calibration parameter θ in this process is the resistance at the interface, known to be a function of the localized temperature, which is in turn dependent on the combination of the applied load and the current intensity. But the resistance cannot be directly observed and its functional relation with the temperature is unknown. For this experiment, $p = 2$, $\mathbf{x} = (\text{applied load}, \text{current intensity})^T$, $c = 1$, $\theta = (\text{resistance})$, $d = p + c = 3$. Furthermore, θ is a function of both inputs, denoted by $\theta(x_1, x_2)$.

The available dataset consists of 12 physical experiment data points and 35 simulated data points from a finite element analysis. As in the buckypaper fabrication example, the cheap separate design approach is compared to a single-stage design, in which the data points are randomly sampled, all at once, from the pool of available data. Four data points are reserved for testing. The RMSE is compared in the left panel of Figure 7 as the physical sample sizes increases in an increment of one data point per run. The right panel of Figure 7 presents the percentage improvements of cheap separate design over both physical-alone and simulation-alone designs. This case study suggests a similar message to that resulting from the buckypaper fabrication case study.

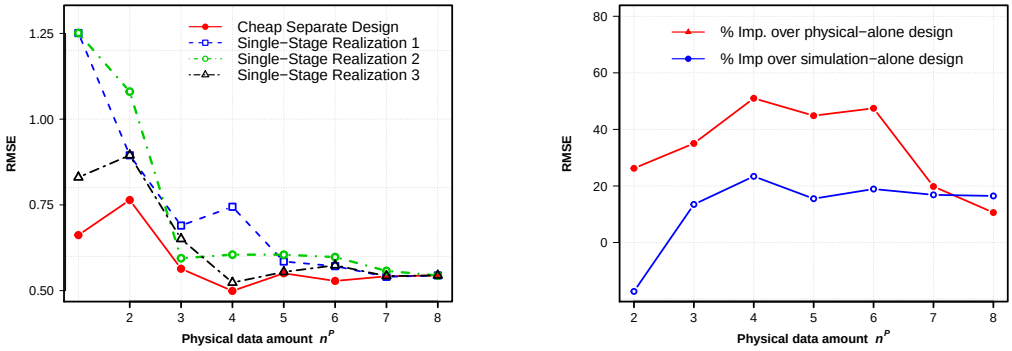

Figure 7: Resistance spot welding case study: comparison of the cheap separate design with the single-stage designs (left panel) and comparison with the physical-alone and simulation-alone designs (right panel).

6. SUMMARY AND CONCLUDING REMARKS

Since the emergence of the calibration framework, the associated design issues have not received enough attention. Recently, several engineering applications have been reported where calibration parameters are of a functional form and as such, their functional estimation heavily relies on the quality of designs of the inputs to the calibration. Our analysis suggests that handling the sequential designs separately for the physical experiments and for the computer experiments can offer tempting advantages to the experimenters. This is particularly true when the computer experiments are cheap enough to be executed in a large number of runs. Moreover, given the different nature of physical and computer experiments, the separate design strategy provides experimenters the flexibility to choose different design mechanisms, more fitting to the needs of the respective experiments.

Our study can be considered as a guideline for researchers, practitioners and experimenters to design their data collection procedures when combining physical and computer experimental data in the presence of functional calibration parameters. This research also calls upon both the research community and practitioners in the area of calibration and multi-fidelity analysis to look for possibly more sophisticated solutions beyond the separate design approach or our simple extension of the nested design for dealing with the different sized design spaces and the existence of functional calibration parameters.

7. ACKNOWLEDGEMENTS

The authors would like to acknowledge the generous support from their sponsors. Aziz Ezzat and Ding are partially supported by NSF under grant no. CMMI-1000088 and CMMI-1545038.

REFERENCES

Arendt, P., Apley, D., and Chen, W. (2012), "Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability," *Transactions of ASME, Journal of Mechanical Design*, 134.

Atamturktur, S., Hegenderfer, J., Williams, B., Egeberg, M., Lebensohn, R. A., and Unal, C. (2015), "A Resource Allocation Framework for Experiment-based Validation of Numerical Models," *Mechanics of Advanced Materials and Structures*, 22, 641–654.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C., and Tu, J. (2007), "A Framework for Validation of Computer Models." *Technometrics*, 49, 138–154.

Brown, D. A. and Atamturktur, S. (2018), "Nonparametric Functional Calibration of Computer Models," *Statistica Sinica*, *To appear*, DOI:10.5705/ss.202015.0344.

Cohn, D. A., Ghahramani, Z., and Jordan, M. (1996), "Active Learning with Statistical Models," *Journal of Applied Intelligent Research*, 4, 129–145.

Fang, K. (1979), "The Uniform Design: Application of Number-Theoretic Methods in Experimental Design," *Acta Mathematicae Applicatae Sinica*, 3, 363–372.

Goh, J., Bingham, D., Holloway, J., Grosskopf, M. J., Kuranz, C., and Rutter, E. (2013), "Prediction and Computer Model Calibration Using Outputs From Multifidelity Simulators," *Technometrics*, 55, 501–512.

Gramacy, R. B. and Lee, H. K. (2009), "Adaptive Design and Analysis of Supercomputer Experiments," *Technometrics*, 51, 130–145.

Johnson, M. (1990), "Minimax and Maximin Distance Designs," *Journal of Statistical Planning and Inference*, 26, 131–148.

Jones, D. R., Shonlau, M., and Welch, W. J. (1998), "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, 13, 455–492.

Joseph, V. R. and Melkote, S. (2009), "Statistical Adjustments to Engineering Models," *Journal of Quality Technology*, 41, 362–375.

Kennedy, M. and O'Hagan, A. (2000), "Predicting The Output From a Complex Computer Code When Fast Approximations are Available," *Biometrika*, 87, 1–13.

— (2001), "Bayesian Calibration of Computer Models," *Journal of The Royal Statistical Society, Series B (Statistical Methodology)*, 63, 425–464.

Le Gratiet, L. and Cannamela, C. (2015), "Co-kriging Based Sequential Design Strategies Using Fast Cross-Validation Techniques for Multi-Fidelity Computer Codes," *Technometrics*, 57, 418–427.

Le Gratiet, L. and Garnier, J. (2014), "Recursive Co-kriging Model for Design of Computer Experiments With Multiple Levels of Fidelity," *International Journal for Uncertainty Quantification*, 4, 365–386.

Li, W., Chen, S., Jiang, Z., Apley, D., Lu, Z., and Chen, W. (2016), "Integrating Bayesian Calibration, Bias Correction, and Machine Learning for the 2014 Sandia Verification and Validation Challenge Problem," *Transactions of ASME, Journal of Verification, Validation and Uncertainty Quantification*, 1, 1–12.

Loeppky, J., Bingham, D., and Welch, W. (2006), "Computer Model Calibration or Tuning in Practice," Tech. rep., University of British Columbia, Vancouver, BC, CA.

Mackay, D. J. (1992), "Information-based Objective Functions for Active Data Selection," *Neural Computation*, 4, 589–603.

Mckay, M., Beckman, R., and Conover, W. (1979), "A comparison of Three Methods for Selecting Values of Input Variables in The Analysis of Output from a Computer Code," *Technometrics*, 21, 239–245.

Picheny, V., Ginsbourger, D., Richet, Y., and Caplin, G. (2013), "Quantile-Based Optimization of Noisy Computer Experiments With Tunable Precision," *Technometrics*, 55, 2–13.

Plumlee, M., Joseph, V., and Yang, H. (2016), "Calibrating Functional Parameters in The Ion Channel Models of Cardiac Cells," *Journal of The American Statistical Association*, 111, 500–509.

Pourhabib, A. and Balasundaram, B. (2015), "Non-isometric Curve to Surface Matching with Incomplete Data for Functional Calibration," *arXiv preprint*, arXiv:1508.01240 [stat.ML].

Pourhabib, A., Huang, J. Z., Wang, K., Wang, B., and Ding, Y. (2015), "Modulus Prediction of Buckypaper Based on Multi-Fidelity Analysis Involving Latent Variables," *IIE Transactions*, 47, 141–152.

Pourhabib, A., Tuo, R., Ding, Y., and Huang, J. Z. (2016), "Local Calibration In Computer Models," *Journal of The American Statistical Association*, under Review.

Qian, P. Z. (2009), "Nested Latin Hypercube Designs," *Biometrika*, 96, 957–970.

Qian, P. Z. and Wu, J. C. F. (2008), "Bayesian Hierarchical Modeling for Integrating Low-accuracy and High-accuracy Experiments," *Technometrics*, 50, 192–204.

Qian, Z., Seepersad, C. C., Joseph, V. R., Allen, J. K., and Wu, J. C. (2006), "Building surrogate models based on detailed and approximate simulations," *Journal of Mechanical Design*, 128, 668–677.

Rasumussen, C. and Williams, K. (2006), *Gaussian Processes for Machine Learning*, The MIT Press.

Reese, C., Wilson, A., Hamada, M., Martz, H., and Ryan, K. (2004), "Integrated Analysis of Computer and Physical Experiments," *Technometrics*, 46, 153–164.

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

Shewry, M. and Wynn, H. (1987), "Maximum Entropy Sampling," *Journal of Applied Statistics*, 14, 165–170.

Tuo, R., Qian, P. Z. G., and Wu, J. C. F. (2013), "Comment: A Brownian Motion Model for Stochastic Simulation With Tunable Precision," *Technometrics*, 55, 29–31.

Tuo, R. and Wu, J. C. (2015), "Efficient calibration for imperfect computer models," *The Annals of Statistics*, 43, 2331–2352.

Tuo, R., Wu, J. C. F., and Yu, D. (2014), "Surrogate Modeling of Computer Experiments With Different Mesh Densities," *Technometrics*, 56, 372–380.

Wang, K. (2013), "Statistics-enhanced Multistage Process Models for Integrated Design and Manufacturing of Poly (vinyl Alcohol) Treated Buckypaper," Ph.D. thesis, Florida State University, Tallahassee, FL.

Wang, K., Vanli, A., Zhang, C., and Wang, B. (2017), "Calibration and adjustment of mechanical property prediction model for poly(vinyl alcohol)-enhanced carbon nanotube buckypaper manufacturing," *The International Journal of Advanced Manufacturing Technology*, 88, 1889–1901.

Wang, Z., Liang, Z., Wang, B., Zhang, C., and Kramer, L. (2004), "Processing and property investigation of single-walled carbon nanotube (SWNT) buckypaper/epoxy resin matrix nanocomposites," *Composites Part A: Applied Science and Manufacturing*, 35, 1225–1232.

Williams, B. J., Santner, T. J., and Notz, Williams, I. (2000), "Sequential Design of Computer Experiments To Minimize Integrated Response Functions," *Statistica Sinica*, 10, 1133–1152.

Xia, H., Ding, Y., and Mallick, B. (2011), "Bayesian Hierarchical Model for Combining Misaligned Two-resolution Metrology Data," *IIE Transactions*, 43, 242–258.

Xiong, S., Qian, P. Z. G., and Wu, J. C. F. (2013), "Sequential Design and Analysis of High-accuracy and Low-accuracy Computer Codes," *Technometrics*, 55, 37–46.