Computational Statistics and Data Analysis 113 (2017) 346–362

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Designing combined physical and computer experiments to maximize prediction accuracy

Erin R. Leatherman [a,*], Angela M. Dean [b], Thomas J. Santner [b]

[a] *West Virginia University, United States*
[b] *The Ohio State University, United States*

ARTICLE INFO

Article history:
Received 1 March 2016
Received in revised form 18 July 2016
Accepted 19 July 2016
Available online 25 July 2016

Keywords:
Augmented nested LHD
Calibration
Gaussian process Kriging interpolator
IMSPE
Latin hypercube design
Simulator

ABSTRACT

Combined designs for experiments involving a physical system and a simulator of the physical system are evaluated in terms of their accuracy of predicting the mean of the physical system. Comparisons are made among designs that are (1) locally optimal under the minimum integrated mean squared prediction error criterion for the combined physical system and simulator experiments, (2) locally optimal for the physical or simulator experiments, with a fixed design for the component not being optimized, (3) maximin augmented nested Latin hypercube, and (4) I-optimal for the physical system experiment and maximin Latin hypercube for the simulator experiment. Computational methods are proposed for constructing the designs of interest. For a large test bed of examples, the empirical mean squared prediction errors are compared at a grid of inputs for each test surface using a statistically calibrated Bayesian predictor based on the data from each design. The prediction errors are also studied for a test bed that varies only the calibration parameter of the test surface. Design recommendations are given.

1. Introduction

Combinations of observations from a physical system and a deterministic computer simulator of that system have been used, for example, to calibrate the simulator statistically, to optimize the physical system, and to achieve other objectives (see, for example, Higdon et al., 2004, 2008; Leatherman et al., 2014b). This paper considers settings where such data come from a *physical experiment* that varies the values of the inputs to the physical system and a *computer experiment* that varies the input values to the simulator code. The goal is to determine the *initial* design of a combined simulator and physical experiment with the objective of most accurately predicting the *mean of the physical system* using a statistically calibrated simulator.

For both physical and computer experiments there has been much research on the optimal design of experiments using intuitively defined criteria. For example the *I-optimality criterion*, which minimizes the integrated mean squared prediction error (MSPE) for a specified regression model, is a design metric for physical experiments that emphasizes prediction accuracy (Studden, 1977; Hardin and Sloane, 1993).

Designs that minimize the integrated MSPE (IMSPE) have also been proposed for computer experiments. However, because the analysis of simulator output ordinarily is based on non-parametric Kriging predictors, most simulator designs constructed in the literature are locally optimal, corresponding to a specification of correlation and other model parameters

[*] Correspondence to: Department of Statistics, PO Box 6330, Morgantown, WV 26506, United States. Fax: +1 304 293 2272.
E-mail address: erleatherman@mail.wvu.edu (E.R. Leatherman).

http://dx.doi.org/10.1016/j.csda.2016.07.013
0167-9473/© 2016 Elsevier B.V. All rights reserved.

(although the designs of Leatherman et al., 2016, are constructed for a weighting of the parameter values). More often, the initial design of a computer experiment uses a "space-filling" criterion resulting in, for example, *minimax* designs (Johnson et al., 1990), *minimum average reciprocal distance* designs (Audze and Eglais, 1977; Welch, 1985; Liefvendahl and Stocki, 2006), and *lattices, nets, and uniform* designs (Niederreiter, 1978, 1992; Fang and Wang, 1994; Owen, 1995).

Previous work on the design of combined physical and simulator experiments includes studies of how to take follow-up runs. For example, Ranjan et al. (2011) and Williams et al. (2011) focus on batch sequential design optimization but use standard space-filling designs as the initial physical and simulator designs. The initial observations are used to estimate model parameters, and additional design points are added to improve the design's measure of goodness, in particular to provide the maximum IMSPE reduction in Ranjan et al. (2011) and the maximum generalized expected improvement for global fit in Williams et al. (2011).

Using a calibrated Bayesian predictor for the mean of a physical system, this paper compares the accuracy of *local IMSPE-optimal designs* for combined physical and simulator experiments with *maximin augmented nested Latin hypercube designs (MmANLHD)* and other designs. The comparisons are based on the empirical mean squared prediction error (EMSPE) in a large test bed of examples. Section 2 describes the model used to relate the simulator experiment output and the physical system output. Section 3.1 gives the formulas for the MSPE and IMSPE and defines local IMSPE-optimal designs while Section 3.2 defines MmANLHDs. Sections 4.1 and 4.2 give algorithms for constructing local minimum IMSPE designs and MmANLHDs, respectively. Section 5 presents a study of the prediction accuracy of the initial combined designs. Fourteen designs are selected in Section 5.1 to compare across 18 corresponding physical and simulator test-bed families, where the test beds are described in Section 5.2. Section 5.4 compares the designs' prediction accuracy across the surfaces using the EMSPE criterion defined in Section 5.3. An additional comparison of prediction accuracy is made in Section 6 where test beds are formed from stationary GP draws with $\boldsymbol{\theta} \neq 0.5 \times \mathbf{1}_{d_t}$ when the design used to collect training data for prediction is locally optimal for $\boldsymbol{\theta} = 0.5 \times \mathbf{1}_{d_t}$. A brief summary of the conclusions is given in Section 7.

2. Modeling combined simulator and physical outputs

In the simulator code let $\boldsymbol{x}^s$ and $\boldsymbol{t}$ denote a $d_x \times 1$ vector of control inputs and a $d_t \times 1$ vector of calibration inputs, respectively. *Control inputs* can be 'set' by the researcher in both the physical experiment as well as in simulator runs. *Calibration inputs* can be varied in the simulator runs but are fixed and unknown in the associated physical experiment; for example, while the material properties of meniscal tissue are fixed values in a biomechanical cadaver study of stresses in the knee, a finite element simulator may regard these values as inputs. Let $\boldsymbol{x}^p$ be a $d_x \times 1$ vector of control inputs in the physical experiment and $\boldsymbol{\theta}$ be the true, but unknown, $d_t \times 1$ vector of calibration parameters. Assume the input space of the control variables is rectangular but transformed so that $\boldsymbol{x}^s, \boldsymbol{x}^p \in [0, 1]^{d_x}$, while the input space of the calibration variables is also rectangular but transformed so that $\boldsymbol{t}, \boldsymbol{\theta} \in [0, 1]^{d_t}$. Finally let $y^s(\boldsymbol{x}^s, \boldsymbol{t})$ and $y^p(\boldsymbol{x}^p)$ denote the outputs from the simulator and physical experiments when run at $(\boldsymbol{x}^s, \boldsymbol{t})$ and $\boldsymbol{x}^p$, respectively.

Adopting the model of Kennedy and O'Hagan (2001), denoted KO hereafter, this paper regards the simulator output $y^s(\boldsymbol{x}^s, \boldsymbol{t})$ as a draw from the Gaussian Process (GP)

$$Y^s\left(\boldsymbol{x}^s, \boldsymbol{t}\right) = \sum_{\ell=1}^{k} f_\ell(\boldsymbol{x}^s, \boldsymbol{t})\beta_\ell + Z(\boldsymbol{x}^s, \boldsymbol{t}) = \boldsymbol{f}^T(\boldsymbol{x}^s, \boldsymbol{t})\boldsymbol{\beta} + Z(\boldsymbol{x}^s, \boldsymbol{t}), \tag{1}$$

where $\boldsymbol{f}(\boldsymbol{x}^s, \boldsymbol{t}) = (f_1(\boldsymbol{x}^s, \boldsymbol{t}), f_2(\boldsymbol{x}^s, \boldsymbol{t}), \ldots, f_k(\boldsymbol{x}^s, \boldsymbol{t}))^T$ are known regression functions, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)^T$ is a vector of unknown regression coefficients, $\boldsymbol{w}^T$ denotes the transpose of $\boldsymbol{w}$, and $Z(\cdot, \cdot)$ is a zero-mean, stationary GP over $[0, 1]^{d_x+d_t}$ with process variance σ_Z^2 and separable Gaussian correlation function:

$$\begin{aligned} Cor\left(Y(\boldsymbol{x}_1^s, \boldsymbol{t}_1), Y(\boldsymbol{x}_2^s, \boldsymbol{t}_2)\right) &= R_Z\left((\boldsymbol{x}_1^s, \boldsymbol{t}_1) - (\boldsymbol{x}_2^s, \boldsymbol{t}_2) \mid \boldsymbol{\rho}_Z\right) \\ &= \prod_{j=1}^{d_x} \rho_{Z,j}^{4\,(x_{1,j}^s - x_{2,j}^s)^2} \prod_{j=1}^{d_t} \rho_{Z,d_x+j}^{4\,(t_{1,j}-t_{2,j})^2}, \end{aligned} \tag{2}$$

where $x_{i,j}^s$ and $t_{i,j}$ are the jth elements of inputs $\boldsymbol{x}_i^s$ and $\boldsymbol{t}_i$, respectively, $i = 1, 2$. The parameter $\rho_{Z,j} \in [0, 1]$ is the correlation between outputs at inputs $(\boldsymbol{x}_1^s, \boldsymbol{t}_1)$ and $(\boldsymbol{x}_2^s, \boldsymbol{t}_2)$ that differ *only* in the jth input by half the range of this input.

Let $\zeta(\boldsymbol{x}^p) \equiv \zeta(\boldsymbol{x}^p, \boldsymbol{\theta})$ denote the mean of the physical system run at input $\boldsymbol{x}^p$. The output $y^p(\boldsymbol{x}^p)$ is modeled as the sum of $\zeta(\boldsymbol{x}^p)$ and a zero mean measurement error $\epsilon(\boldsymbol{x}^p)$, so that $y^p(\boldsymbol{x}^p)$ is a realization of $Y^p(\boldsymbol{x}^p) = \zeta(\boldsymbol{x}^p) + \epsilon(\boldsymbol{x}^p)$ where additional assumptions regarding $\epsilon(\boldsymbol{x}^p)$ are stated below. The KO model assumes that the simulator, even when run at the true value of $\boldsymbol{\theta}$, need not perfectly represent the underlying physical process because the mathematical model uses simplified physics or biology. Following KO, denote the simulator model bias (discrepancy) as

$$\delta(\boldsymbol{x}^p) \equiv \zeta(\boldsymbol{x}^p) - y^s(\boldsymbol{x}^p, \boldsymbol{\theta}),$$

and assume that $\delta(\boldsymbol{x}^p)$ can be regarded as a realization of $\Delta\left(\boldsymbol{x}^p\right)$ which is a stationary, zero-mean GP over $[0, 1]^{d_x}$ with process variance σ_δ^2 and separable Gaussian correlation function

$$R_\delta\left(\boldsymbol{x}_1^p - \boldsymbol{x}_2^p \mid \boldsymbol{\rho}_\delta\right) = \prod_{j=1}^{d_x} \rho_{\delta,j}^{\;4\left(x_{1,j}^p - x_{2,j}^p\right)^2}, \tag{3}$$

where $x_{i,j}^p$ is the jth element of input $\boldsymbol{x}_i^p$, $i = 1, 2$, and $\rho_{\delta,j} \in [0, 1]$ is interpreted analogously to $\rho_{Z,j}$. The zero-mean assumption for $\Delta\left(\boldsymbol{x}^p\right)$ is interpreted as saying that there is no global trend in $\delta\left(\boldsymbol{x}^p\right)$ requiring regression modeling.

As in KO, assume $y^p\left(\boldsymbol{x}^p\right)$ can be regarded as a realization of

$$Y^p\left(\boldsymbol{x}^p\right) = Y^s\left(\boldsymbol{x}^p, \boldsymbol{\theta}\right) + \Delta\left(\boldsymbol{x}^p\right) + \epsilon\left(\boldsymbol{x}^p\right), \tag{4}$$

where $Y^s\left(\boldsymbol{x}^p, \boldsymbol{\theta}\right)$, $\Delta\left(\boldsymbol{x}^p\right)$, and $\epsilon\left(\boldsymbol{x}^p\right)$ are mutually independent with the distribution of $Y^s\left(\boldsymbol{x}^p, \boldsymbol{\theta}\right)$ defined through (1) and (2), $\Delta\left(\boldsymbol{x}^p\right)$ has distribution described in the previous paragraph, and $\epsilon\left(\boldsymbol{x}^p\right)$ is a white noise process with variance σ_ϵ^2.

Suppose that n_s simulator observations $\boldsymbol{y}^s = \left(y^s(\boldsymbol{x}_1^s, \boldsymbol{t}_1), y^s(\boldsymbol{x}_2^s, \boldsymbol{t}_2), \ldots, y^s(\boldsymbol{x}_{n_s}^s, \boldsymbol{t}_{n_s})\right)^T$ result from running a set of inputs specified by the rows of the $n_s \times (d_x + d_t)$ simulator design matrix

$$\boldsymbol{X}^s = \begin{bmatrix} \boldsymbol{x}_1^s & \boldsymbol{x}_2^s & \cdots & \boldsymbol{x}_{n_s}^s \\ \boldsymbol{t}_1 & \boldsymbol{t}_2 & \cdots & \boldsymbol{t}_{n_s} \end{bmatrix}^T.$$

Additionally, suppose that n_p physical observations $\boldsymbol{y}^p = \left(y^p(\boldsymbol{x}_1^p), y^p(\boldsymbol{x}_2^p), \ldots, y^p(\boldsymbol{x}_{n_p}^p)\right)^T$ are to be collected at inputs which are the rows of the $n_p \times d_x$ physical design matrix $\boldsymbol{X}^p = \left[\boldsymbol{x}_1^p, \boldsymbol{x}_2^p, \ldots, \boldsymbol{x}_{n_p}^p\right]^T$. The combined physical and simulator observations are denoted by the $(n_p + n_s) \times 1$ vector $\boldsymbol{y} = \left[(\boldsymbol{y}^p)^T, (\boldsymbol{y}^s)^T\right]^T$.

In the experimental setting of this paper, the goal is to predict the mean of the *physical* system output $\zeta\left(\boldsymbol{x}_0\right)$ at $\boldsymbol{x}_0$ based on the physical and simulator training data. Focusing on the prediction of $\zeta\left(\boldsymbol{x}_0\right) = y^s(\boldsymbol{x}_0, \boldsymbol{\theta}) + \delta(\boldsymbol{x}_0)$ eliminates the problem of unidentifiability of predicting separately $\delta(\cdot)$ and $\boldsymbol{\theta}$. When $\boldsymbol{\beta}$ in (1) is unknown, while $\boldsymbol{\Omega} = \left(\boldsymbol{\theta}, \sigma_Z^2, \boldsymbol{\rho}_Z, \sigma_\delta^2, \boldsymbol{\rho}_\delta, \sigma_\epsilon^2\right)$ in (1)–(3) are known, Sec 3.3 of Santner et al. (2003) presents the argument that shows that the best linear unbiased predictor (BLUP) of $\zeta\left(\boldsymbol{x}_0\right)$ is

$$\widehat{\zeta}_{\text{blup}}(\boldsymbol{x}_0) = \boldsymbol{f}_0^T\widehat{\boldsymbol{\beta}} + \boldsymbol{v}_0^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\left(\boldsymbol{y} - \boldsymbol{F}\widehat{\boldsymbol{\beta}}\right), \tag{5}$$

where $\boldsymbol{f}_0 = \boldsymbol{f}(\boldsymbol{x}_0, \boldsymbol{\theta})$ is the $k \times 1$ vector of known regressors at control input $\boldsymbol{x}_0$ and calibration parameter $\boldsymbol{\theta}$; $\widehat{\boldsymbol{\beta}} = (\boldsymbol{F}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{y}$ is the $k \times 1$ general least squares estimator of $\boldsymbol{\beta}$; $\boldsymbol{F}$ is a $(n_p + n_s) \times k$ matrix of known regressors with the first n_p rows defined by $f_j(\boldsymbol{x}_i^p, \boldsymbol{\theta})$ for $1 \le i \le n_p$ and $1 \le j \le k$, and the last n_s rows defined by $f_j(\boldsymbol{x}_i^s, \boldsymbol{t}_i)$ for $1 \le i \le n_s$ and $1 \le j \le k$; and $\boldsymbol{v}_0 = [(\boldsymbol{v}_0^p)^T, (\boldsymbol{v}_0^s)^T]^T$ is the $(n_p + n_s) \times 1$ vector of covariances, with the ith element of $\boldsymbol{v}_0^p$ being

$$\begin{aligned}&\text{Cov}(Y^s(\boldsymbol{x}_0, \boldsymbol{\theta}) + \Delta(\boldsymbol{x}_0),\ Y^s(\boldsymbol{x}_i^p, \boldsymbol{\theta}) + \Delta(\boldsymbol{x}_i^p))\\ &\quad = \sigma_Z^2 R_Z\left((\boldsymbol{x}_0, \boldsymbol{\theta}) - (\boldsymbol{x}_i^p, \boldsymbol{\theta}) \mid \boldsymbol{\rho}_Z\right) + \sigma_\delta^2 R_\delta\left(\boldsymbol{x}_0 - \boldsymbol{x}_i^p \mid \boldsymbol{\rho}_\delta\right), \quad \text{for } i = 1, \ldots, n_p,\end{aligned}$$

while the jth element of $\boldsymbol{v}_0^s$ is

$$\text{Cov}(Y^s(\boldsymbol{x}_0, \boldsymbol{\theta}) + \Delta(\boldsymbol{x}_0),\ Y^s(\boldsymbol{x}_j^s, \boldsymbol{t}_j)) = \sigma_Z^2 R_Z\left((\boldsymbol{x}_0, \boldsymbol{\theta}) - (\boldsymbol{x}_j^s, \boldsymbol{t}_j) \mid \boldsymbol{\rho}_Z\right), \quad \text{for } j = 1, \ldots, n_s.$$

Also, $\boldsymbol{\Sigma}_{\boldsymbol{y}}$ is the $(n_p + n_s) \times (n_p + n_s)$ covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{Z}}{}^{pp} & \boldsymbol{\Sigma}_{\boldsymbol{Z}}{}^{ps} \\ (\boldsymbol{\Sigma}_{\boldsymbol{Z}}{}^{ps})^T & \boldsymbol{\Sigma}_{\boldsymbol{Z}}{}^{ss} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Sigma}_\delta + \boldsymbol{\Sigma}_\epsilon & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \equiv \boldsymbol{\Sigma}_{\boldsymbol{Z}} + \begin{pmatrix} \boldsymbol{\Sigma}_\delta + \boldsymbol{\Sigma}_\epsilon & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}, \tag{6}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{Z}}{}^{pp} = \left(\sigma_Z^2 R_Z\left((\boldsymbol{x}_i^s, \boldsymbol{\theta}) - (\boldsymbol{x}_j^s, \boldsymbol{\theta}) \mid \boldsymbol{\rho}_Z\right)\right)$ is $n_p \times n_p$, $\boldsymbol{\Sigma}_{\boldsymbol{Z}}{}^{ps} = \left(\sigma_Z^2 R_Z\left((\boldsymbol{x}_i^p, \boldsymbol{\theta}) - (\boldsymbol{x}_j^s, \boldsymbol{t}_j) \mid \boldsymbol{\rho}_Z\right)\right)$ is $n_p \times n_s$, $\boldsymbol{\Sigma}_{\boldsymbol{Z}}{}^{ss} = \left(\sigma_Z^2 R_Z\left((\boldsymbol{x}_i^s, \boldsymbol{t}_i) - (\boldsymbol{x}_j^s, \boldsymbol{t}_j) \mid \boldsymbol{\rho}_Z\right)\right)$ is $n_s \times n_s$, $\boldsymbol{\Sigma}_\delta = \left(\sigma_\delta^2 R_\delta\left(\boldsymbol{x}_i^p - \boldsymbol{x}_j^p \mid \boldsymbol{\rho}_\delta\right)\right)$ is $n_p \times n_p$, and $\boldsymbol{\Sigma}_\epsilon = \sigma_\epsilon^2 I_{n_p}$. In Section 3.1, the BLUP of $\zeta\left(\boldsymbol{x}_0\right)$ in (5) will be used to define IMSPE and the designs that are locally optimal for the IMSPE measure.

The predictors used in the simulation comparisons of Sections 5 and 6 do not assume that $[\boldsymbol{\beta}, \boldsymbol{\Omega}]$ is known. Instead, the (fully) Bayesian predictor

$$\begin{aligned}\widehat{\zeta}_{\text{FB}}(\boldsymbol{x}_0) &= E\left[Y^s\left(\boldsymbol{x}^p, \boldsymbol{\theta}\right) + \Delta\left(\boldsymbol{x}^p\right) | \boldsymbol{y}\right]\\ &= E_{[\boldsymbol{\beta}, \boldsymbol{\Omega} | \boldsymbol{y}]}\left(E\left(Y^s\left(\boldsymbol{x}^p, \boldsymbol{\theta}\right) + \Delta\left(\boldsymbol{x}^p\right) | \boldsymbol{\beta}, \boldsymbol{\Omega}, \boldsymbol{y}\right)\right)\\ &= E_{[\boldsymbol{\beta}, \boldsymbol{\Omega} | \boldsymbol{y}]}\left(\boldsymbol{f}_0^T\boldsymbol{\beta} + \boldsymbol{v}_0^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\left(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\beta}\right)\right),\end{aligned} \tag{7}$$

is used where a prior for $[\boldsymbol{\beta}, \boldsymbol{\Omega}]$ is assumed. Unfortunately, the predictor (7) can be analytically demanding and, therefore, to make predictions in Sections 5 and 6, (7) is computed based on draws from $[\boldsymbol{\beta}, \boldsymbol{\Omega} \mid \boldsymbol{y}]$ that are constructed using the Markov Chain Monte Carlo algorithm that is implemented in the GPM/SA software of Gattiker (2008).

3. Designs for combined physical and simulator experiments

In their initial research on calibration, KO speculated on the design of combined physical and simulator experiments. They noted that the physical design is often "not a matter of choice". Next, they suggested that the simulator design should cover both the *control* input space and *calibration* input space well, and that a sequential design is a good way to ensure the coverage of the calibration input space. Their final observation is that there should be simulator design points that are 'close' to physical design points in order to determine the model bias.

While physical designs and simulator designs have been studied extensively in their own right, the combination of these designs has received much less attention in the literature. Ranjan et al. (2011) and Williams et al. (2011) studied combined follow-up designs. However, to the authors' knowledge, the effect of the *initial combined design* has not yet been presented in the literature.

The following subsections define two design criteria for initial combined designs. The first is the *local minimum IMSPE* criterion described in Section 3.1. This criterion is prediction-based; i.e., it is defined using the physical and simulator output models and the BLUP from Section 2. The second is the *MmANLHD* criterion that is defined in Section 3.2. This geometrically-based criterion yields space-filling simulator designs with corresponding physical designs whose points replicate some of the control input values from the simulator design. Thus, this criterion follows the KO suggestion to align simulator and physical design points. A third design is used in Section 5 for comparison; this commonly-used design consists of an I-optimal design for the physical experiment paired with an MmLHD for the simulator experiment.

3.1. Locally optimal designs for minimizing the integrated mean squared prediction error

Designs constructed using the local minimum IMSPE criterion are meant to provide small expected prediction errors of the mean of the physical system on average across the control input space. The predictor $\widehat{\zeta}_{\text{blup}}(\boldsymbol{x}_0)$ in (5) depends on $\boldsymbol{y}$, the physical and simulator designs $\boldsymbol{X}^p$ and $\boldsymbol{X}^s$ (through $\boldsymbol{F}$, $R_Z(\cdot \mid \boldsymbol{\rho}_Z)$, and $R_\delta(\cdot \mid \boldsymbol{\rho}_\delta)$), and on $\boldsymbol{\Omega} = \left(\boldsymbol{\theta}, \sigma_Z^2, \boldsymbol{\rho}_Z, \sigma_\delta^2, \boldsymbol{\rho}_\delta, \sigma_\epsilon^2\right)$. For a given $\boldsymbol{X}^p$, $\boldsymbol{X}^s$, and $\boldsymbol{\Omega}$, the MSPE of $\widehat{\zeta}_{\text{blup}}(\cdot)$ at $\boldsymbol{x}_0 \in [0,1]^{d_x}$ is

$$\begin{aligned}\text{MSPE}\left(\boldsymbol{x}_0, \boldsymbol{X}^p, \boldsymbol{X}^s \mid \boldsymbol{\Omega}\right) &= E\left(\left(\widehat{\zeta}_{\text{blup}}(\boldsymbol{x}_0) - \left(Y^s(\boldsymbol{x}_0, \boldsymbol{\theta}) + \Delta(\boldsymbol{x}_0)\right)\right)^2 \middle| \boldsymbol{\Omega}\right) \\ &= \sigma_Z^2 + \sigma_\delta^2 - \begin{bmatrix}\boldsymbol{f}_0^T & \boldsymbol{v}_0^T\end{bmatrix} \begin{bmatrix}\boldsymbol{0} & \boldsymbol{F}^T \\ \boldsymbol{F} & \boldsymbol{\Sigma}_{\boldsymbol{y}}\end{bmatrix}^{-1} \begin{bmatrix}\boldsymbol{f}_0 \\ \boldsymbol{v}_0\end{bmatrix},\end{aligned} \tag{8}$$

where $\boldsymbol{f}_0$, $\boldsymbol{F}$, $\boldsymbol{v}_0$, and $\boldsymbol{\Sigma}_{\boldsymbol{y}}$ are defined below (5) and the expectation (8) is taken with respect to $\left(Y^s(\boldsymbol{x}_0, \boldsymbol{\theta}) + \Delta(\boldsymbol{x}_0), \left(\boldsymbol{Y}^p\right)^T, \left(\boldsymbol{Y}^s\right)^T\right)$, where $\boldsymbol{Y}^p$ and $\boldsymbol{Y}^s$ are the GP model analogs of $\boldsymbol{y}^p$ and $\boldsymbol{y}^s$.

To avoid the numerical non-invertibility of $\boldsymbol{\Sigma}_{\boldsymbol{y}}$ when simulator design points $\{x_i^s\}$ are "too close" together, simulator designs that are constructed using (8) are restricted to be an element of the set of all possible n_s-run simulator designs in $d_x + d_t$ inputs where no two design rows are within an ϵ-ball of diameter $b = 10^{-3}$ of each other, denoted by $\mathcal{D}^s_{n_s, d_x+d_t, b}$ hereafter. Note that the replication of physical design points does not pose a similar problem because of the measurement error term $\epsilon(\cdot)$ in the physical model (4). Thus, physical designs will be selected from $\mathcal{D}^p_{n_p, d_x}$, the set of all possible n_p-run physical designs in d_x inputs.

Given parameters $\boldsymbol{\Omega} = \left(\boldsymbol{\theta}, \sigma_Z^2, \boldsymbol{\rho}_Z, \sigma_\delta^2, \boldsymbol{\rho}_\delta, \sigma_\epsilon^2\right)$, the IMSPE of the predictor $\widehat{\zeta}_{\text{blup}}(\cdot)$ using design $(\boldsymbol{X}^p, \boldsymbol{X}^s)$ is

$$\begin{aligned}\text{IMSPE}\left(\boldsymbol{X}^p, \boldsymbol{X}^s \mid \boldsymbol{\Omega}\right) &= \int_{[0,1]^{d_x}} \text{MSPE}\left(\boldsymbol{x}_0, \boldsymbol{X}^p, \boldsymbol{X}^s \mid \boldsymbol{\Omega}\right) \mathrm{d}\boldsymbol{x}_0 \\ &= \sigma_Z^2 + \sigma_\delta^2 - \text{trace}\left[\begin{bmatrix}\boldsymbol{0} & \boldsymbol{F}^T \\ \boldsymbol{F} & \boldsymbol{\Sigma}_{\boldsymbol{y}}\end{bmatrix}^{-1} \int_{[0,1]^{d_x}} \begin{pmatrix}\boldsymbol{f}_0\boldsymbol{f}_0^T & \boldsymbol{f}_0\boldsymbol{v}_0^T \\ \boldsymbol{v}_0\boldsymbol{f}_0^T & \boldsymbol{v}_0\boldsymbol{v}_0^T\end{pmatrix} \mathrm{d}\boldsymbol{x}_0\right].\end{aligned} \tag{9}$$

Given $\boldsymbol{\Omega}$, an $(\boldsymbol{X}^p, \boldsymbol{X}^s)$ that minimizes Eq. (9) over $\left\{\mathcal{D}^p_{n_p, d_x}, \mathcal{D}^s_{n_s, d_x+d_t, b}\right\}$ is said to be a *local IMSPE-optimal combined design* (w.r.t. $\boldsymbol{\Omega}$). Equivalently, factoring (9) into σ_Z^2 times

$$\text{IMSPE}^\star\left(\boldsymbol{X}^p, \boldsymbol{X}^s \mid \boldsymbol{\Omega}^\star\right) \equiv \frac{1}{\sigma_Z^2}\text{IMSPE}\left(\boldsymbol{X}^p, \boldsymbol{X}^s \mid \boldsymbol{\Omega}\right), \tag{10}$$

shows that a local IMSPE-optimal combined design minimizes $\text{IMSPE}^\star$ $(\boldsymbol{X}^p, \boldsymbol{X}^s \mid \boldsymbol{\Omega}^\star)$ and depends only on $\boldsymbol{\Omega}^\star = \left(\boldsymbol{\theta}, \boldsymbol{\rho}_Z, \sigma_\delta^2/\sigma_Z^2, \boldsymbol{\rho}_\delta, \sigma_\epsilon^2/\sigma_Z^2\right)$. Notice that a local IMSPE-optimal design is independent of $\boldsymbol{\beta}$ but depends on the values of the regression functions at the training data inputs, through $\boldsymbol{F}$, and at the point to be predicted, through $\boldsymbol{f}_0$.

The parameters $\boldsymbol{\Omega}^\star$ needed to calculate $\text{IMSPE}^\star$ typically are not known in advance of the experiment. The simulation study in Section 5.4 examines the prediction accuracy of a range of local IMSPE-optimal designs to determine whether there is a choice of $\boldsymbol{\Omega}^\star$ that allows for accurate empirical predictions for a variety of test-bed surfaces. The choices of $\boldsymbol{\Omega}^\star$ that are used to construct locally optimal designs for the simulation study of Section 5 are based on the results of Leatherman et al. (2016) for the simulator-only setting.

Table 1
A local IMSPE-optimal combined 10 × 2 physical and 15 × 3 simulator design constructed using $\mathbf{\Omega}^\star = (\theta = 0.5, \boldsymbol{\rho}_Z = 0.25 \times \mathbf{1}_3, \sigma_\delta^2/\sigma_Z^2 = 0.1, \boldsymbol{\rho}_\delta = 0.25 \times \mathbf{1}_2, \sigma_\epsilon^2/\sigma_Z^2 = 0.01)$ and the constant mean $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t}) = 1$.

Physical design		Simulator design		
x_1^p	x_2^p	x_1^s	x_2^s	t_1
0.4682	0.7144	1.0000	0.4026	0.4954
0.7852	0.3414	1.0000	1.0000	0.5004
0.3766	0.0795	1.0000	1.0000	0.0000
0.1381	0.2439	0.3070	0.5737	0.4987
0.7092	0.9092	0.0000	1.0000	0.5001
0.9286	0.0883	0.0000	1.0000	1.0000
0.0480	0.5500	0.3881	1.0000	0.4997
0.1570	0.8472	1.0000	1.0000	1.0000
0.4502	0.3779	0.0000	0.0000	0.4954
0.8898	0.6930	0.5411	0.0000	0.0000
		0.6703	0.0010	0.4932
		0.6330	0.5643	0.4962
		0.5411	0.0004	0.0019
		0.0000	1.0000	0.0000
		0.6702	0.0000	0.4932

3.2. Maximin augmented nested latin hypercube designs

A second design criterion that will be considered in the simulation study of Section 5 is the MmANLHD criterion. Recall that the projections of an $(n_s \times d_x)$ LHD onto every one of the d_x axes has one design value on each grid point $\{0,\ 1/(n_s - 1), \dots, 1\}$. The MmANLHDs described below are constructed from nested LHDs (NLHDs). NLHDs have the property that the smaller (physical) design points must coincide with a subset of the *control inputs* for the larger (simulator) design. Because the NLHDs are marginally non-collapsing and are selected to be space-filling, the MmANLHDs constructed from these designs inherit these properties.

First, an NLHD that maximizes the minimum inter-point distance over pairs of rows is selected. Assuming $n_p \le n_s$, the full NLHD is used for the control variables in the simulator experiment and the smaller (nested) design is used for the physical design. Then, augmentation is performed by adding columns to the simulator design matrix for the calibration inputs. Formally, let $\mathcal{L}_{n_p,d_x,n_s,d_x+d_t}$ denote the set of all $n_s \times (d_x + d_t)$ LHDs whose first d_x columns form the specified maximin NLHD with designs sizes (n_p, d_x) and (n_s, d_x). Any design $\boldsymbol{X} \in \mathcal{L}_{n_p,d_x,n_s,d_x+d_t}$ that maximizes

$$\min_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in \boldsymbol{X}} \sqrt{\sum_{i=1}^{d_x+d_t} \left(x_{1,i} - x_{2,i}\right)^2}, \tag{11}$$

is said to be a *maximin augmented nested Latin hypercube design*. The MmANLHDs used for comparison in the simulation study of Section 5.4 were constructed starting with the maximin NLHDs posted on the website of van Dam et al. (2013).

4. Design construction algorithms

4.1. A global/local algorithm for constructing local IMSPE-optimal designs

To find the combined designs $(\boldsymbol{X}^p \in \mathcal{D}^p_{n_p,d_x}, \boldsymbol{X}^s \in \mathcal{D}^s_{n_s,d_x+d_t,b})$ that minimize IMSPE* in (10) for $b = 10^{-3}$ at a specific set of parameter values $\mathbf{\Omega}^\star$, this paper uses particle swarm optimization (PSO) to identify a design that serves as the starting point for a gradient-based quasi-Newton (QN) search for the best design. A detailed description of this heuristic approach and an illustrative example is presented in Leatherman et al. (2014a).

Briefly, PSO begins with a set of N_{des} starting combined designs $(\boldsymbol{X}^p, \boldsymbol{X}^s)$ spread over the design space $\left\{\mathcal{D}^p_{n_p,d_x}, \mathcal{D}^s_{n_s,d_x+d_t,b}\right\}$. Each design is iterated N_{its} times. At each iteration, a design is updated to a new design that is "between" the global best design among all combined designs generated thus far and the best design among those having started at the same $(\boldsymbol{X}^p, \boldsymbol{X}^s)$. For the examples in this paper, the PSO parameter settings followed the recommendations of Kennedy and Eberhart (1995) and Yang (2010), and the PSO algorithm was run with $N_{\text{des}} = 4\left(n_p d_x + n_s(d_x + d_t)\right)$ starting designs and $N_{\text{its}} = 2N_{\text{des}}$ iterations. The best design constructed by PSO was used as the starting design for a single run of a QN algorithm (as implemented in the MATLAB (MATLAB, 2015) code `fmincon.m`) to produce the final (approximate) local IMSPE-optimal design.

An example of a local IMSPE-optimal combined 10-run × 2-d physical design and 15-run × 3-d simulator design w.r.t. $\mathbf{\Omega}^\star = \left(\theta = 0.5, \boldsymbol{\rho}_Z = 0.25 \times \mathbf{1}_3, \sigma_\delta^2/\sigma_Z^2 = 0.1, \boldsymbol{\rho}_\delta = 0.25 \times \mathbf{1}_2, \sigma_\epsilon^2/\sigma_Z^2 = 0.01\right)$ and the constant mean, $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t}) = 1$, is listed in Table 1. The physical design and the 2-d projection of the simulator design onto (x_1, x_2) space is shown in the left

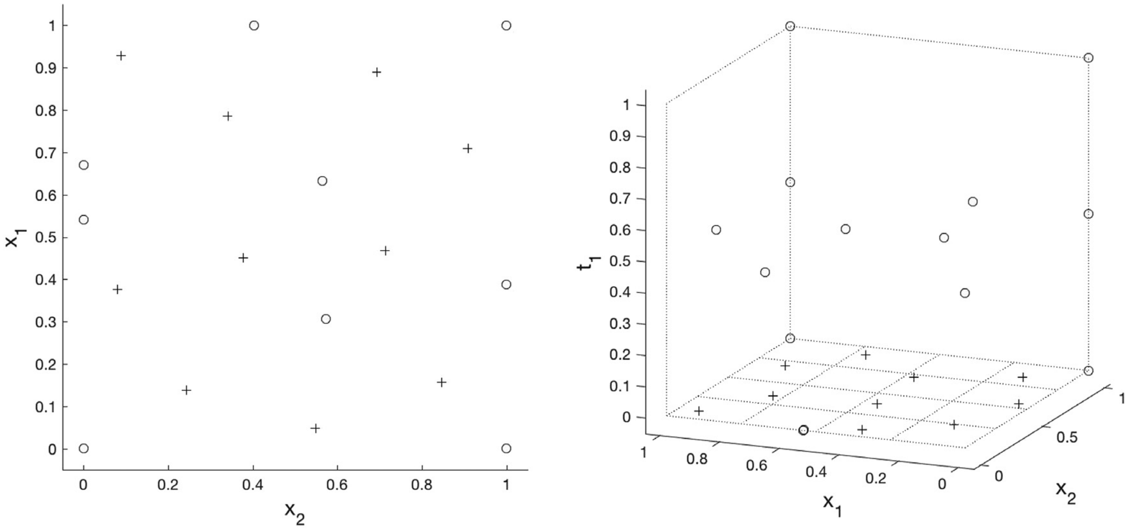

Fig. 1. Projections of the local IMSPE-optimal combined design in Table 1. Left panel: projection of the simulator design onto (x_1, x_2) space. Right panel: scatterplot of the 3-d simulator design, with 2-d physical design shown on plane $t_1 = 0$. The physical design shown as '+' symbols; the simulator design shown as open circles.

panel of Fig. 1; this projection is visually space-filling. The right panel of Fig. 1 shows the full 3-d simulator design which includes the calibration parameter t_1 on the vertical axis. The right panel also shows the 2-d physical design in the plane $t_1 = 0$. Note that the eleventh and fifteenth point of the simulator design in Table 1 have a Euclidean distance of only 0.001005 between them. This distance is very near the simulator design minimum distance restriction that is permitted by the ϵ-ball with radius 10^{-3}. Other authors who have found this near-replication of simulator points are Crary (2002) and Crary and Stormann (2015).

4.2. A Smart Swap algorithm for constructing MmANLHDs

Each MmANLHD used in the simulation study of Section 5.4 was constructed by appending columns to the design matrix of a fixed maximin NLHD with d_x inputs, using a modified version of the Smart Swap algorithm described by Moon et al. (2011). (This modification also allows the algorithm to be used to add inputs to any fixed LHD having points on a grid.) Specifically, for this paper, d_t additional columns are required for the input settings of the calibration parameters in the n_s-run simulator design (with nested n_p-run physical design). For the given NLHD, it is required that the resulting augmented design is an MmANLHD, i.e., it should have a maximum value of the minimum inter-point Euclidean distance (11).

The Smart Swap algorithm of Moon et al. (2011) iteratively improves a candidate design as follows: at each iteration, one of the design points involved in the minimum inter-point distance is selected at random and, if it improves the design, a coordinate of this design point is swapped with the corresponding coordinate of another point randomly selected from the design. Additionally, this coordinate swap occurs with a probability specified by the user when the swap produces an equivalent value of (11); this probability was set to 0.05 for the designs constructed for this paper. For the MmANLHDs in this paper, the swap is only applied to the last d_t columns of the candidate design matrix, since the first d_x columns are fixed.

One other minor modification must be applied to the original Smart Swap algorithm when the design is to be constructed on a grid (rather than design points randomly selected within the "cells" that are formed by the grid). Since, in this situation, multiple pairs of design points are likely to have the same value of minimum inter-point distance, the algorithm is modified to identify and list all such pairs.

Table 2 lists an MmANLHD of size 10×2 for the physical experiment and 15×3 for the simulator experiment. The design was constructed using the modified Smart Swap algorithm to augment a maximin NLHD from the website created by van Dam et al. (2013). The left panel of Fig. 2 shows the physical design with the 2-d projection of the simulator design onto (x_1, x_2) space; this panel visually demonstrates the space-filling property of the d_x control inputs for the physical and simulator designs, separately, and the coinciding property of the two designs in these dimensions. The right panel of Fig. 2 shows the 3-d simulator design with the physical design on the plane at $t_1 = 0$.

5. Comparison of design prediction accuracy

This section examines the prediction accuracy of specific physical and simulator designs when used to predict for a test bed of surfaces. The EMSPE, defined below in Section 5.3 for the Bayesian predictor $\hat{\zeta}_{FB}(\boldsymbol{x}_0)$ in (7), will be used to compare the prediction accuracy of the designs described in Section 5.1 for the surfaces described in Section 5.2.

Table 2
An MmANLHD with 10 × 2 physical and 15 × 3 simulator component designs that was constructed using the modified Smart Swap algorithm.

Physical design		Simulator design		
x_1^p	x_2^p	x_1^s	x_2^s	t_1
0.0000	0.1429	0.0000	0.1429	0.0714
0.0714	0.6429	0.0714	0.6429	0.5000
0.2143	0.3571	0.2143	0.3571	0.7143
0.2857	0.0000	0.2857	0.0000	0.5714
0.3571	0.7857	0.3571	0.7857	0.8571
0.5000	0.2143	0.5000	0.2143	0.9286
0.5714	1.0000	0.5714	1.0000	0.4286
0.7143	0.4286	0.7143	0.4286	0.6429
0.8571	0.8571	0.8571	0.8571	0.7857
1.0000	0.5714	1.0000	0.5714	0.3571
		0.1429	0.9286	0.1429
		0.4286	0.5000	0.2857
		0.6429	0.7143	0.0000
		0.7857	0.0714	0.2143
		0.9286	0.2857	1.0000

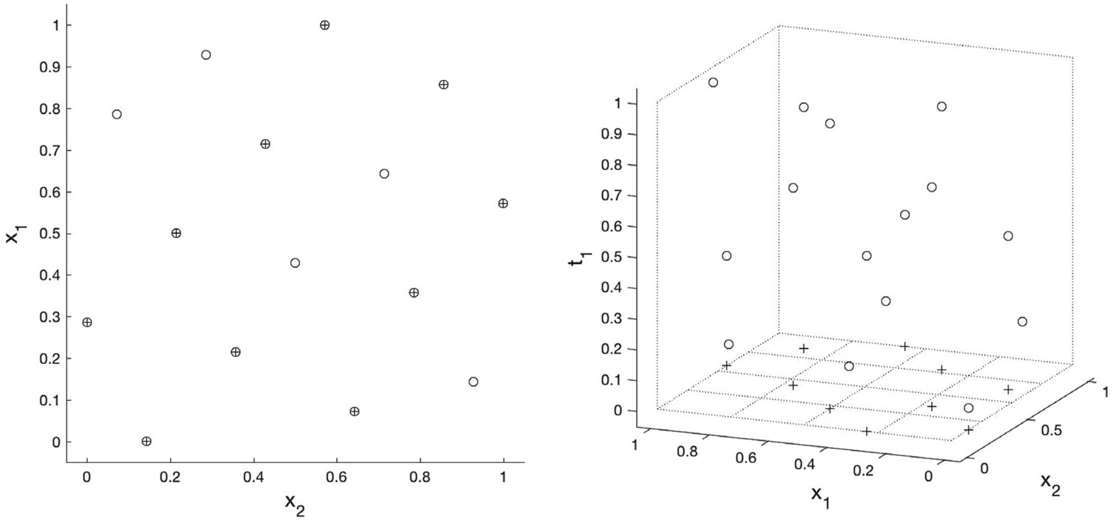

Fig. 2. An MmANLHD with 10 × 2 physical and 15 × 3 simulator component designs. Left panel: projection of the simulator design onto (x_1, x_2) space. Right panel: scatterplot of the 3-d simulator design, with 2-d physical design shown on plane $t_1 = 0$. The physical design shown as '+' symbols; the simulator design shown as open circles.

5.1. Combined physical and simulator designs studied

The prediction accuracy of eight design sizes $(n_p, d_x, n_s, d_x + d_t)$ for each of fourteen $(\boldsymbol{X}^p, \boldsymbol{X}^s)$ designs are compared in Section 5.4 in terms of EMSPE. The designs are described in the following paragraphs and are summarized in Table 3.

Four of the designs were constructed by minimizing IMSPE* $(\boldsymbol{X}^p, \boldsymbol{X}^s \mid \boldsymbol{\Omega}^\star)$ in (10) for a specific $\boldsymbol{\Omega}^\star = \left(\boldsymbol{\theta}, \boldsymbol{\rho}_Z, \sigma_\delta^2/\sigma_Z^2, \boldsymbol{\rho}_\delta, \sigma_\epsilon^2/\sigma_Z^2\right)$. In the simpler problem of predicting simulator output based on a set of simulator training data, Leatherman et al. (2016) found that local IMSPE-optimal designs constructed using "small" correlation parameter values yielded smaller empirical prediction errors than designs based on "larger" correlation parameter values. Guided by these observations, the correlation vectors $\boldsymbol{\rho}_Z$ and $\boldsymbol{\rho}_\delta$ were chosen to have a small common correlation value ρ for the local IMSPE-optimal designs used in the study of Section 5.4. Four $\left(\boldsymbol{\rho}_Z, \boldsymbol{\rho}_\delta\right)$ vector combinations were used in total: $\left(0.25 \times \mathbf{1}_{d_x+d_t}, 0.25 \times \mathbf{1}_{d_x}\right)$, $\left(0.25 \times \mathbf{1}_{d_x+d_t}, 0.5 \times \mathbf{1}_{d_x}\right)$, $\left(0.5 \times \mathbf{1}_{d_x+d_t}, 0.25 \times \mathbf{1}_{d_x}\right)$, and $\left(0.5 \times \mathbf{1}_{d_x+d_t}, 0.5 \times \mathbf{1}_{d_x}\right)$. The calibration parameter vector $\boldsymbol{\theta}$ is an element of $[0, 1]^{d_t}$ and, as a "naïve" selection, was set to be $\boldsymbol{\theta} = 0.5 \times \mathbf{1}_{d_t}$. Additionally, the variance ratios were selected to be $\sigma_\delta^2/\sigma_Z^2 = 0.1$ and $\sigma_\epsilon^2/\sigma_Z^2 = 0.01$, and the constant mean, $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t}) = 1$, was used for the GP in (1).

For each of the four $\boldsymbol{\Omega}^\star$ specified in the previous paragraph, the local IMSPE-optimal combined design $(\boldsymbol{X}^p, \boldsymbol{X}^s)$ was constructed using the PSO plus QN optimization algorithm described in Section 4.1. In the sections that follow, these four designs are denoted $D_{0.25,0.25}^{PS}$, $D_{0.25,0.5}^{PS}$, $D_{0.5,0.25}^{PS}$, and $D_{0.5,0.5}^{PS}$, where the superscript "PS" denotes that the optimality criterion

Table 3
Labels used to denote the designs compared in the simulation study of Section 5.4. The superscripts denote the portion of the combined design that was optimized using the local minimum IMSPE criterion. The subscripts denote the common correlation parameter values ρ that were used in the IMSPE optimization, where $\boldsymbol{\rho}_Z = \rho \times \mathbf{1}_{d_x+d_t}$ and $\boldsymbol{\rho}_\delta = \rho \times \mathbf{1}_{d_x}$. The subscripts are listed in the order $\boldsymbol{\rho}_Z, \boldsymbol{\rho}_\delta$.

Design type		Design label	Common correlation	
Physical	Simulator		For $\boldsymbol{\rho}_Z$	For $\boldsymbol{\rho}_\delta$
Combined	IMSPE-opt	$D^{PS}_{0.25,0.25}$	0.25	0.25
		$D^{PS}_{0.25,0.5}$	0.25	0.5
		$D^{PS}_{0.5,0.25}$	0.5	0.25
		$D^{PS}_{0.5,0.5}$	0.5	0.5
I-opt	IMSPE-opt	$D^{S}_{0.25,0.25}$	0.25	0.25
		$D^{S}_{0.25,0.5}$	0.25	0.5
		$D^{S}_{0.5,0.25}$	0.5	0.25
		$D^{S}_{0.5,0.5}$	0.5	0.5
IMSPE-opt	MmLHD	$D^{P}_{0.25,0.25}$	0.25	0.25
		$D^{P}_{0.25,0.5}$	0.25	0.5
		$D^{P}_{0.5,0.25}$	0.5	0.25
		$D^{P}_{0.5,0.5}$	0.5	0.5
I-opt	MmLHD	I-opt + MmLHD	–	–
MmANLHD		ANLHD	–	–

is applied to the combined (physical and simulator) design. The subscripts denote the common values of the simulator and the discrepancy correlation parameters, respectively (see also Table 3).

A second class of designs fixes the physical design $\boldsymbol{X}^p$ and constructs a local IMSPE-optimal simulator design $\boldsymbol{X}^s$ to minimize IMSPE*. Four of the designs included in this paper are such local IMSPE-optimal simulator designs used in conjunction with a *given* I-optimal physical design. The parameter values used to optimize IMSPE* are identical to those used to construct the combined local IMSPE-optimal designs $D^{PS}_{\_,\_}$. The fixed I-optimal designs were constructed with JMP® (JMP, 1989–2007) assuming a cubic mean model that also included quadratic interaction terms, as this was the largest polynomial model that could be fit for all n_x and d_x studied. In the sections that follow, the local IMSPE-optimal simulator designs combined with the fixed I-optimal physical design are denoted $D^S_{0.25,0.25}$, $D^S_{0.25,0.5}$, $D^S_{0.5,0.25}$, and $D^S_{0.5,0.5}$, where “S” in the superscript indicates that only the simulator design is IMSPE optimal. As for the case of combined designs, the subscripts state the common value of the $\boldsymbol{\rho}_Z$ and $\boldsymbol{\rho}_\delta$ parameters, respectively (see also Table 3).

Analogously, a third class of designs fixes the simulator design $\boldsymbol{X}^s$ and constructs a local IMSPE-optimal physical design $\boldsymbol{X}^p$ to minimize IMSPE*. Again, four designs were included in the EMSPE studies. These $(\boldsymbol{X}^p, \boldsymbol{X}^s)$ were determined by combining a *given* maximin LHD (MmLHD) for $\boldsymbol{X}^s$ with a local IMSPE-optimal $\boldsymbol{X}^p$. The parameter values used to optimize the physical design are identical to those for obtaining $D^{PS}_{\_,\_}$. The MmLHDs were obtained from van Dam et al. (2013). In the sections that follow, these designs are denoted $D^P_{0.25,0.25}$, $D^P_{0.25,0.5}$, $D^P_{0.5,0.25}$, and $D^P_{0.5,0.5}$, where the superscript “P” indicates that only the physical design is IMSPE optimal. Again, the subscripts state the common value of the $\boldsymbol{\rho}_Z$ and $\boldsymbol{\rho}_\delta$ parameters, respectively (see also Table 3).

In the two single design optimization scenarios above, the PSO plus QN algorithm was used to optimize the simulator design $\boldsymbol{X}^s$ with a fixed physical design $\boldsymbol{X}^p$, and vice versa. The PSO algorithm was initiated with N_{des} starting designs taken from the appropriate design space, where $N_{\text{des}} = 4n_s\,(d_x + d_t)$ when optimizing $\boldsymbol{X}^s$ with a fixed $\boldsymbol{X}^p$, and $N_{\text{des}} = 4n_p d_x$ when optimizing $\boldsymbol{X}^p$ with a fixed $\boldsymbol{X}^s$. In both cases, the PSO employed $N_{\text{its}} = 2N_{\text{des}}$ iterations.

The final designs studied in the following sections are space-filling. The first design is an “off-the-shelf” design that combines an MmLHD for the simulator runs and an I-optimal design for the physical experiment observations. These MmLHD and I-optimal designs are the same designs used in the fixed simulator and fixed physical design settings, respectively, from above. This combined design is denoted I-opt + MmLHD in Table 3. The second design is an MmANLHD which is an intuitively more sophisticated version of the off-the-shelf design; the MmANLHD uses common inputs for the physical experiment and the control portion of the simulator inputs. This design was computed using the Smart Swap algorithm in Section 4.2 by augmenting the (n_2-grid) maximin NLHD from van Dam et al. (2013) having n_p and n_s points in d_x dimensions. This design is denoted ANLHD in Table 3.

The designs described above were constructed using Linux compute machines having Dual Eight Core Xeon 2.7 E5-2680 processors with 384 GB RAM. The exception was the $(n_s, d_x, n_p, d_x + d_t) = (10, 2, 15, 3)$ combined local IMSPE-optimal design, which was constructed using a Linux compute machine with Dual Quad Core Xeon 2.66 E5430 processors with 32 GB RAM. For local IMSPE-optimal designs, Table 4 lists the average hours of computation time, over the four correlation

Table 4
The average computation time (in hours), over the four correlation scenarios, needed to construct each design type and size. The computation time listed for ANLHD is for a single MmANLHD, where the starting maximin NLHD was given.

$(n_s, d_x, n_p, d_x + d_t)$	Design type			
	$D^{PS}_{-,-}$	$D^{P}_{-,-}$	$D^{S}_{-,-}$	ANLHD
(10, 2, 15, 3)	1.62[a]	0.03	0.12	0.0009
(20, 2, 30, 3)	3.54	0.11	0.53	0.0057
(20, 4, 25, 5)	6.78	0.45	1.53	0.0032
(20, 4, 30, 6)	9.83	0.56	2.98	0.0252
(30, 3, 50, 5)	19.30	1.57	10.34	0.1069
(40, 4, 50, 5)	40.24	4.37	13.00	0.0253
(40, 4, 60, 6)	52.25	4.82	40.10	0.1934

[a] Constructed using a different type of compute machine than for the other design types and sizes (see text).

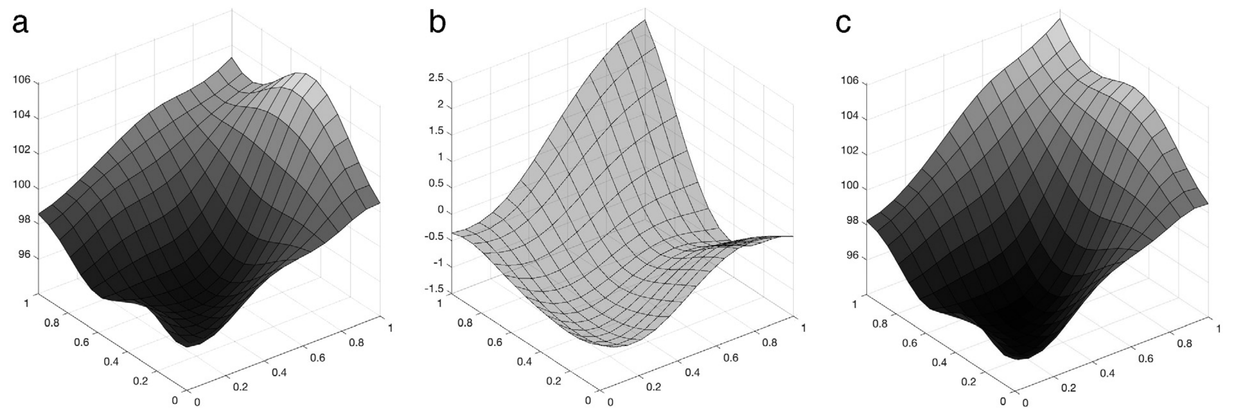

Fig. 3. A surface $\zeta_{\text{test}}(\boldsymbol{w})$ when $d_x = 2$, $d_t = 1$, and $\theta = 0.25$: (a) a 2-d simulator surface $S^{Krig}_{0.25}$ evaluated at $\theta = 0.25$, (b) a Kriging discrepancy surface $B^{Krig}_{0.5}$, (c) the mean surface $\zeta_{\text{test}}(\boldsymbol{w})$ obtained by summing (a) and (b).

scenarios, that were needed to construct each design type and size. The times listed in Table 4 for ANLHD are the hours of computation time needed to augment each starting maximin NLHD obtained from van Dam et al. (2013).

5.2. The test bed of physical and simulator surfaces

To compare the prediction accuracy of the designs of Section 5.1, a test bed of non-linear physical and simulator surfaces was created. Each mean surface of the physical observations $\zeta(\boldsymbol{x}^p)$ is constructed as the summation of a corresponding simulator response surface and a discrepancy (bias) response surface. When physical "observations" are made, an additional observation error is added to the mean physical response.

5.2.1. Simulator surfaces

The test beds of simulator surfaces used in the simulation study of Section 5.4 can be categorized in three groups. The first group of simulator surfaces uses the Kriging surfaces of Trosset (1999). These surfaces have the form

$$y^s_{\text{test}}(\boldsymbol{w}) = \hat{\beta}_Z + \boldsymbol{r}_Z(\boldsymbol{w})^T \boldsymbol{R}_Z^{-1}\left(\boldsymbol{Y}^{500} - \boldsymbol{1}_{500}\hat{\beta}_Z\right), \quad \boldsymbol{w} \in [0, 1]^{d_x+d_t}, \tag{12}$$

where $\boldsymbol{Y}^{500}$ is a 500×1 vector drawn from a stationary GP based on inputs that form a $500 \times (d_x + d_t)$ approximate MmLHD, denoted by $\boldsymbol{L}_{\boldsymbol{d_x}+\boldsymbol{d_t}}$.

The sampled GP has mean $\beta_Z = 100$, variance $\sigma^2_Z = 10$, and Gaussian correlation function (2), where $\boldsymbol{\rho}_Z$ is either $0.25 \times \boldsymbol{1}_{d_x+d_t}$ or $0.5 \times \boldsymbol{1}_{d_x+d_t}$. For numerical stability, a nugget 10^{-6} was added to the diagonal of $\sigma^2_Z \boldsymbol{R}_Z$. In (12), $\widehat{\beta}_Z = \left(\boldsymbol{1}^T_{500}\boldsymbol{R}_Z^{-1}\boldsymbol{1}_{500}\right)^{-1} \boldsymbol{1}^T_{500}\boldsymbol{R}_Z^{-1}\boldsymbol{Y}^{500}$, $\boldsymbol{r}_Z(\boldsymbol{w})$ is the 500×1 vector of correlations having the ith component $R_Z(\boldsymbol{x}_i, \boldsymbol{w})$ for $\boldsymbol{x}^T_i \in \boldsymbol{L}_{\boldsymbol{d_x}+\boldsymbol{d_t}}$, and $\boldsymbol{R}_Z$ is the 500×500 matrix of correlations having (i, j)th element $R_Z(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for $\boldsymbol{x}^T_i, \boldsymbol{x}^T_j \in \boldsymbol{L}_{\boldsymbol{d_x}+\boldsymbol{d_t}}$. For each $\boldsymbol{\rho}_Z$, 30 surfaces $y^s_{\text{test}}(\boldsymbol{w})$ were constructed. The collections of Kriging simulator surfaces are denoted $S^{Krig}_{0.25}$ and $S^{Krig}_{0.5}$, where the subscript specifies the common correlation value used to construct the surface. A representative simulator surface from $S^{Krig}_{0.25}$ is plotted in Panel (a) of Fig. 3.

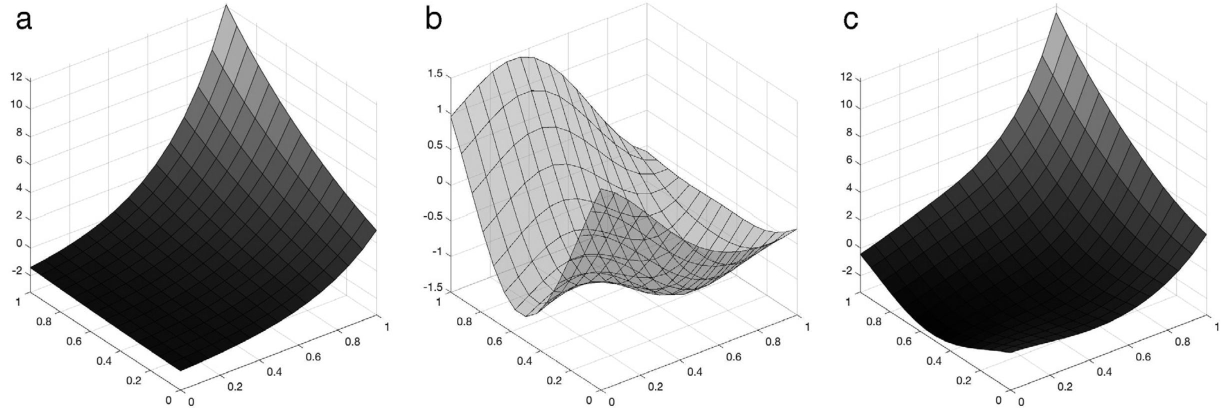

Fig. 4. A surface $\zeta_{\text{test}}(\boldsymbol{w})$ when $d_x = 2$, $d_t = 1$, and $\theta = 0.25$: (a) a 2-d simulator surface S_U^{SL} evaluated at $\theta = 0.25$, (b) a Kriging discrepancy surface $B_{0.5}^{Krig}$, (c) the mean surface $\zeta_{\text{test}}(\boldsymbol{w})$ obtained by summing (a) and (b).

The second group of simulator surfaces uses the exponential model of Sobol´ and Levitan (1999)

$$y^s_{\text{test}}(\boldsymbol{w}) = \exp\left(\boldsymbol{b}^T\boldsymbol{w}\right) - I_{d_x+d_t}, \quad \boldsymbol{w} \in [0, 1]^{d_x+d_t}, \tag{13}$$

where $I_{d_x+d_t} = \prod_{i=1}^{d_x+d_t} \frac{e^{b_i}-1}{b_i}$, and b_i is the ith element of $\boldsymbol{b}$. The b_i parameters were chosen based on results from Loeppky et al. (unpublished). Thirty sets of $\boldsymbol{b}$ (and surfaces (13)) were constructed around two central $\boldsymbol{b}$ values. The central value of the first 30 $\boldsymbol{b}$ draws produced surfaces with inputs having equal effects, whose main effects contributed 50% of the overall variance, and whose calibration inputs contributed 75% of the main effect variance. The central value of the second set of 30 $\boldsymbol{b}$ draws produced surfaces whose inputs had unequal effects, whose main effects contributed 95% of the overall variance, and whose calibration inputs contributed 75% of the main effect variance. The simulator exponential surface families are denoted S_E^{SL} and S_U^{SL}, where the subscript specifies whether the inputs have "E"qual or "U"nequal effects. A representative simulator test-bed surface from S_U^{SL} is plotted in Panel (a) of Fig. 4.

The third group of simulator surfaces uses a modification of the non-stationary function described by Xiong et al. (2007) (see also Ba and Joseph, 2012). Two test beds of non-stationary test surfaces are formed from

$$y^s_{\text{test}}(\boldsymbol{w}) = 10^{(d_x+d_t)/2} \prod_{i=1}^{d_x+d_t} \left\{ \sin\left(a_i\left(w_i - b_i\right)^4\right) \cos\left(2\left(w_i - b_i\right)\right) + \frac{w_i - b_i}{2} \right\}, \quad \boldsymbol{w} \in [0, 1]^{d_x+d_t} \tag{14}$$

where a_i and b_i are varied to form each surface. Two families of test beds are formed. In the first test bed, the base family is defined by (14) and has non-stationary activity that occurs near the edges of the input domain $[0, 1]^{d_x+d_t}$; in the second test bed, a change of variable is made in the base function to $y^s_{\text{test}}(|w_1 - 0.5|, \ldots, \left|w_{d_x+d_t} - 0.5\right|)$ which results in non-stationary activity near the center of $[0, 1]^{d_x+d_t}$. Using these two formulations of (14), two non-stationary test-bed families of 30 surfaces were formed, each by taking i.i.d. Uniform(20, 35) draws $a_1, a_2, \ldots, a_{d_x+d_t}$ and i.i.d. Uniform(0.5, 0.9) draws $b_1, b_2, \ldots, b_{d_x+d_t}$. The families of such surfaces are denoted S_{edge}^{MXB} and S_{mid}^{MXB}, respectively. A representative simulator test-bed surface from S_{mid}^{MXB} is shown in panel (a) of Fig. 5.

5.2.2. Discrepancy (bias) surfaces

All discrepancy (bias) surfaces are of the form

$$\delta_{\text{test}}(\boldsymbol{w}) = \hat{\beta}_\delta + \boldsymbol{r}_\delta(\boldsymbol{w})^T \boldsymbol{R}_\delta^{-1} \left(\boldsymbol{V}^{500} - \boldsymbol{1}_{500}\hat{\beta}_\delta\right), \quad \boldsymbol{w} \in [0, 1]^{d_x}. \tag{15}$$

Here $\boldsymbol{V}^{500}$ is a 500×1 vector drawn from a stationary GP based on a $500 \times d_x$ approximate MmLHD, denoted by $\boldsymbol{L}_{\boldsymbol{d_x}}$. The sampled GP has mean $\beta_\delta = 0$, variance $\sigma_\delta^2 = 1$, and Gaussian correlation function (3), where $\boldsymbol{\rho}_\delta$ is either $0.5 \times \boldsymbol{1}_{d_x}$ or $0.75 \times \boldsymbol{1}_{d_x}$. For numerical stability, a nugget 10^{-6} was added to the diagonal of the covariance matrix, $\sigma_\delta^2 \boldsymbol{R}_\delta$. In (15), $\widehat{\beta_\delta} = \left(\boldsymbol{1}_{500}^T \boldsymbol{R}_\delta^{-1} \boldsymbol{1}_{500}\right)^{-1} \boldsymbol{1}_{500}^T \boldsymbol{R}_\delta^{-1} \boldsymbol{V}^{500}$, $\boldsymbol{r}_\delta(\boldsymbol{w})$ is the 500×1 vector of correlations with ith component $R_\delta(\boldsymbol{x}_i, \boldsymbol{w})$ for $\boldsymbol{x}_i^T \in \boldsymbol{L}_{d_x}$, and $\boldsymbol{R}_\delta$ is the 500×500 matrix of correlations having (i, j)th element $R_\delta(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for $\boldsymbol{x}_i^T, \boldsymbol{x}_j^T \in \boldsymbol{L}_{\boldsymbol{d_x}}$. For each $\boldsymbol{\rho}_\delta$, 30 surfaces $\delta_{test}(\boldsymbol{w})$ were constructed. The collections of bias surfaces are denoted $B_{0.5}^{Krig}$ and $B_{0.75}^{Krig}$, where the subscript specifies the common correlation value used to construct the surface. Three representative bias surfaces from $B_{0.5}^{Krig}$ can be seen in panel (b) of Figs. 3–5. When no bias is required, $\delta_{\text{test}}(\boldsymbol{w})$ is set to zero for $\boldsymbol{w} \in [0, 1]^{d_x}$.

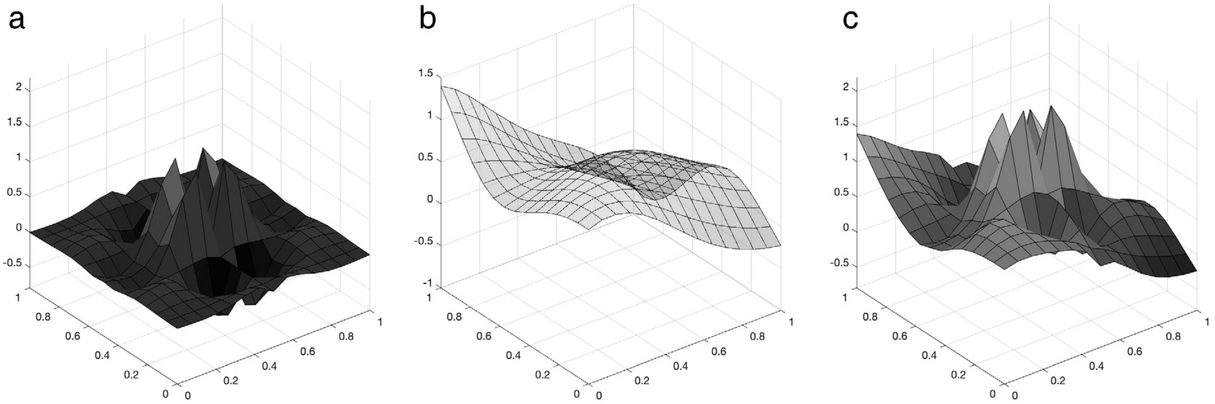

Fig. 5. A surface $\zeta_{\text{test}}(\boldsymbol{w})$ when $d_x = 2$, $d_t = 1$, and $\theta = 0.25$: (a) a 2-d simulator surface $S^{MXB}_{0.25}$ evaluated at $\theta = 0.25$, (b) a Kriging discrepancy surface $B^{Krig}_{0.5}$, (c) the mean surface $\zeta_{\text{test}}(\boldsymbol{w})$ obtained by summing (a) and (b).

Table 5
Labels for the 18 mean surfaces $\zeta_{\text{test}}(\boldsymbol{w})$ defined in (16). These numeric labels are used to identify test-bed surfaces in Figs. 6–9 and in the related figures in the online Supplementary material (see Appendix A).

		$\delta_{\text{test}}(\boldsymbol{x}_0)$		
		0	$B^{Krig}_{0.5}$	$B^{Krig}_{0.75}$
$y^s_{\text{test}}(\boldsymbol{x}_0, \boldsymbol{\theta})$	$S^{Krig}_{0.25}$	1	2	3
	$S^{Krig}_{0.5}$	4	5	6
	S^{SL}_{E}	7	8	9
	S^{SL}_{U}	10	11	12
	S^{MXB}_{edge}	13	14	15
	S^{MXB}_{mid}	16	17	18

5.2.3. Mean of the physical system

For the simulation study of Section 5, the physical surface means are constructed using $\boldsymbol{\theta} = 0.25 \times \mathbf{1}_{d_t}$ for the true value of the calibration parameter. Notice that this true parameter value is different from the "naïve" selection of $\boldsymbol{\theta} = 0.5 \times \mathbf{1}_{d_t}$ that was used to construct the local IMSPE-optimal designs described in Section 5.1. Alternative values of the calibration parameter used for test-bed generation are investigated in Section 6.

In more detail, each mean physical response surface for Section 5 is the summation of a simulator surface (12), (13) or (14) and a discrepancy surface (15); that is,

$$\zeta_{\text{test}}(\boldsymbol{w}) \equiv \zeta_{\text{test}}(\boldsymbol{w}, \boldsymbol{\theta} = 0.25 \times \mathbf{1}_{d_t}) = y^s_{\text{test}}(\boldsymbol{w}, 0.25 \times \mathbf{1}_{d_t}) + \delta_{\text{test}}(\boldsymbol{w}), \quad \boldsymbol{w} \in [0, 1]^{d_x}, \tag{16}$$

or, for no bias, $\delta_{\text{test}}(\boldsymbol{w}) = 0$. Each "observable" physical response is the summation of a mean physical response (16) plus i.i.d. observation error:

$$y^p_{\text{test}}(\boldsymbol{w}, 0.25 \times \mathbf{1}_{d_t}) = \zeta_{\text{test}}(\boldsymbol{w}) + \epsilon, \quad \boldsymbol{w} \in [0, 1]^{d_x}, \tag{17}$$

where ϵ is Normal$(0, \sigma^2_\epsilon = 1)$. The 18 settings used to construct the physical response surfaces $\zeta_{\text{test}}(\boldsymbol{w})$ are listed in Table 5. For each of these parameter settings, 30 physical surfaces were drawn along with their corresponding simulator surfaces as described above. A different set of 18 × 30 surfaces was drawn for each of the (physical, simulator) dimensions $(d_x, d_x + d_t) \in \{(2, 3), (3, 5), (4, 5), (4, 6)\}$.

Three representative $\zeta_{\text{test}}(\boldsymbol{w}, \boldsymbol{\theta})$ in (16) are shown in panels (c) of Figs. 3–5. In Fig. 3, panels (a) and (b) plot a representative simulator surface $S^{Krig}_{0.25}$ and discrepancy surface $B^{Krig}_{0.5}$, respectively, while $\zeta_{\text{test}}(\boldsymbol{w}, \theta = 0.25)$ in panel (c) is the sum of panels (a) and (b). Similarly panels (a) and (b) of Fig. 4, show representative simulator and discrepancy surfaces S^{SL}_{U} and $B^{Krig}_{0.5}$, respectively, that were summed to construct panel (c) of the same figure. Lastly, panels (a) and (b) of Fig. 5, show representative simulator and discrepancy surface S^{MXB}_{U} and $B^{Krig}_{0.5}$, respectively, that were summed to construct panel (c) of the same figure.

5.3. Measuring design prediction accuracy

In order to quantify the prediction accuracy of a combined physical and simulator design $(\boldsymbol{X}^p, \boldsymbol{X}^s)$ across the physical system design space $[0, 1]^{d_x}$, the EMSPE was calculated using an equally-spaced and computationally-feasible grid of g

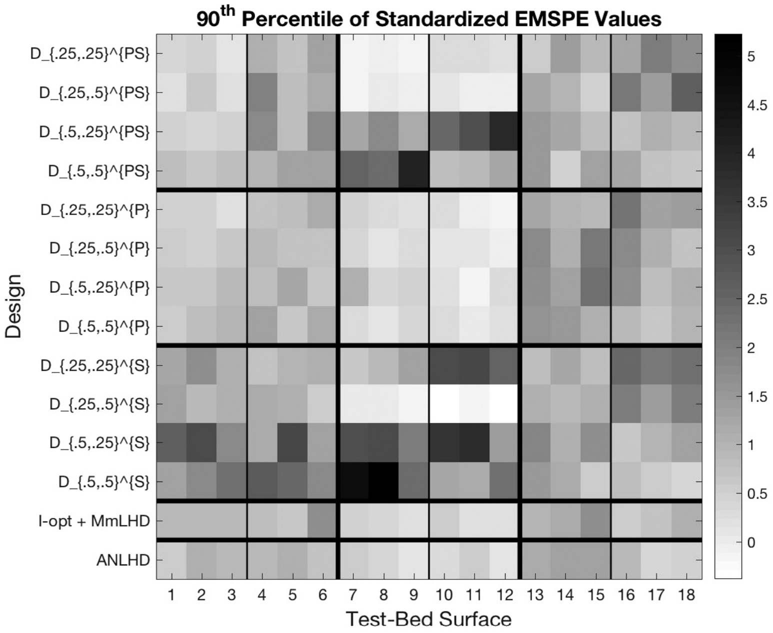

Fig. 6. $(n_p, d_x, n_s, d_x + d_t) = (10, 2, 15, 3)$: A grayscale heatmap of the 90th percentile of the standardized EMSPE values for the 14 designs listed in Table 3 and the 18 test-bed surface types listed in Table 5. Test bed surfaces 1–3, 4–6, 7–9, 10–12, 13–15 and 16–18 use $S^{Krig}_{0.25}$, $S^{Krig}_{0.5}$, S^{SL}_{E}, S^{SL}_{U}, S^{MXB}_{edge}, and S^{MXB}_{mid}, respectively, as $y^s(\boldsymbol{x}, t)$. Within each grouping of three simulator surfaces, the $\zeta_{\text{test}}(\boldsymbol{x})$ to be estimated is the sum of $y^s_{\text{test}}(\boldsymbol{x}, 0.25)$ and $\delta_{\text{test}}(\boldsymbol{x})$ which is: $\equiv 0$, $B^{Krig}_{0.5}$, $B^{Krig}_{0.75}$, in order.

points taken from $[0, 1]^{d_x}$. Given the mean of a physical test-bed surface $\zeta_{\text{test}}(\boldsymbol{x}) \equiv \zeta_{\text{test}}(\boldsymbol{x}, \boldsymbol{\theta})$ the EMSPE is defined by

$$\text{EMSPE}(\boldsymbol{X}^p) = \frac{1}{g}\sum_{i=1}^{g}\left(\widehat{\zeta}_{\text{FB}}(\boldsymbol{x}_i) - \zeta_{\text{test}}(\boldsymbol{x}_i)\right)^2, \tag{18}$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_g$ are the g grid points. In the examples below, $g = 50^2$ for $d_x = 2$, $g = 20^3$ for $d_x = 3$, and $g = 10^4$ for $d_x = 4$. For the comparisons made in Sections 5.4 and 6, the predictor $\widehat{\zeta}_{\text{FB}}(\cdot)$ in (7) was calculated using the Markov Chain Monte Carlo posterior distribution $[\boldsymbol{\beta}, \boldsymbol{\Omega} \mid \boldsymbol{y}]$ that is constructed using the methodology and GPM/SA software described in Higdon et al. (2008) and Gattiker (2008). Alternatively, one could make predictions with the BLUP $\widehat{\zeta}_{\text{blup}}(\cdot)$ in (5) with REML estimates of the model parameters $\boldsymbol{\Omega}$. No matter the predictor used, designs with small EMSPE values are desirable as they allow for more accurate predictions on average across the control inputs.

5.4. Design comparisons and recommendations

Thirty representative surfaces were drawn from each of the 18 test beds listed in Table 5. Training data were collected from each of the 30 × 18 surfaces using each of the 14 designs in Table 3 and for each of eight (n_p, d_x, n_s, d_x+d_t) design sizes. For each of the 30 × 18 × 14 × 8 sets of training data, predictions were made using $\widehat{\zeta}_{\text{FB}}(\cdot)$ in Eq. (7) at a grid of inputs and the EMSPE in Eq. (18) was calculated. The EMSPE values were standardized within each test bed and design size combination because the surfaces vary in complexity across test beds and the prediction accuracy differs depending on the amount of training data available as specified by the $(n_p, d_x, n_s, d_x + d_t)$ design size. For each test bed and design size combination, the $420 = 30 \times 14$ EMSPE values were standardized by subtracting the mean (taken over the 30 surface realizations and 14 designs) and dividing by the standard deviation. Figs. 6–9 are four comparative plots of the 90th percentiles of the 30 standardized EMSPE values for the 18 × 14 test bed by design combinations, for the four design sizes

$$(n_p, d_x, n_s, d_x + d_t) \in \{(10, 2, 15, 3), (20, 4, 25, 5), (30, 3, 50, 5), (40, 4, 50, 5)\}. \tag{19}$$

Corresponding plots of the 90th percentiles of standardized EMSPE values for the additional 4 design sizes (15, 3, 25, 5), (20, 2, 30, 3), (20, 4, 30, 6), and (40, 4, 60, 6) are included in the Supplementary material (see Appendix A), as are tables of the 90th percentiles of the 30 non-standardized EMSPE values for all 8 × 18 × 14 cases.

Starting with the tables in the Supplementary material (see Appendix A), the effect of *increasing the sample size* from 5 runs per dimension to 10 runs per dimension can be quantified by comparing the entries in the pairs of tables with common (d_x, d_t); there are four such pairs of tables. The 90th percentile of the non-standardized EMSPE values is reduced in 91% of

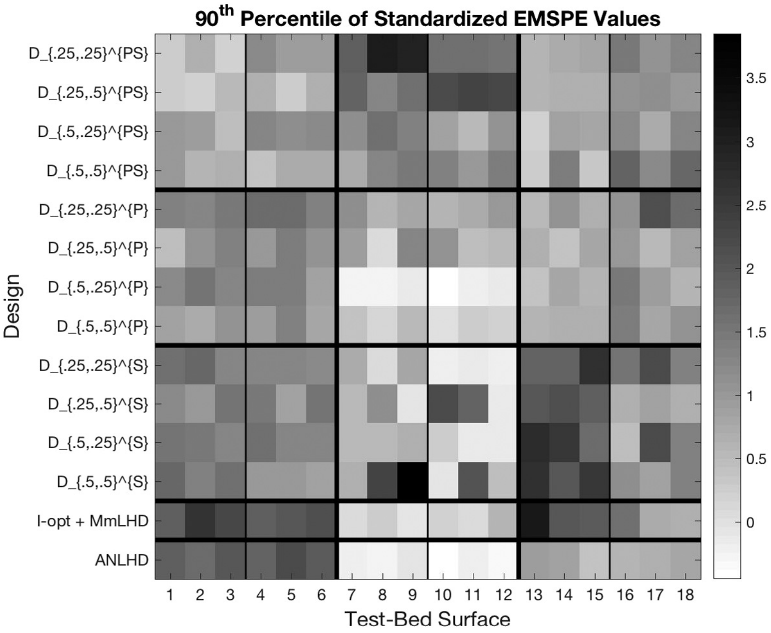

Fig. 7. $(n_p, d_x, n_s, d_x + d_t) = (20, 4, 25, 5)$: A grayscale heatmap of the 90th percentile of the standardized EMSPE values for the 14 designs listed in Table 3 and the 18 test-bed surface types listed in Table 5. Test bed surfaces 1–3, 4–6, 7–9, 10–12, 13–15 and 16–18 use $S^{Krig}_{0.25}$, $S^{Krig}_{0.5}$, S^{SL}_{E}, S^{SL}_{U}, S^{MXB}_{edge}, and S^{MXB}_{mid}, respectively, as $y^s(\boldsymbol{x}, t)$. Within each grouping of three simulator surfaces, the $\zeta_{\text{test}}(\boldsymbol{x})$ to be estimated is the sum of $y^s_{\text{test}}(\boldsymbol{x}, 0.25)$ and $\delta_{\text{test}}(\boldsymbol{x})$ which is: $\equiv 0$, $B^{Krig}_{0.5}$, $B^{Krig}_{0.75}$, in order.

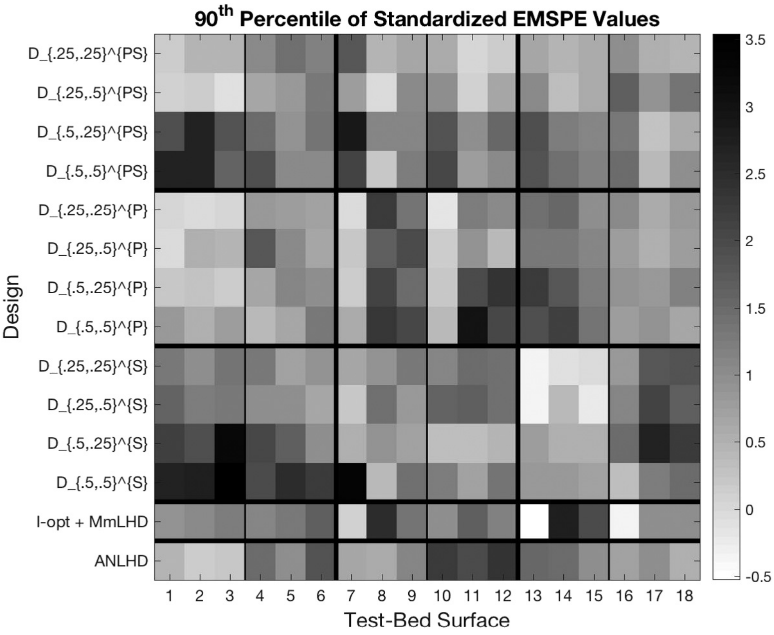

Fig. 8. $(n_p, d_x, n_s, d_x + d_t) = (30, 3, 50, 5)$: A grayscale heatmap of the 90th percentile of the standardized EMSPE values for the 14 designs listed in Table 3 and the 18 test-bed surface types listed in Table 5. Test bed surfaces 1–3, 4–6, 7–9, 10–12, 13–15 and 16–18 use $S^{Krig}_{0.25}$, $S^{Krig}_{0.5}$, S^{SL}_{E}, S^{SL}_{U}, S^{MXB}_{edge}, and S^{MXB}_{mid}, respectively, as $y^s(\boldsymbol{x}, \boldsymbol{t})$. Within each grouping of three simulator surfaces, the $\zeta_{\text{test}}(\boldsymbol{x})$ to be estimated is the sum of $y^s_{\text{test}}(\boldsymbol{x}, 0.25 \times \boldsymbol{1}_2)$ and $\delta_{\text{test}}(\boldsymbol{x})$ which is: $\equiv 0$, $B^{Krig}_{0.5}$, $B^{Krig}_{0.75}$, in order.

the 1008 (design × test-bed × design size pair) combinations when using 10 runs per input compared with 5 runs per input. The 6 test-beds formed using the Sobol´–Levitan simulator in Eq. (13) were most likely to have the largest reduction, with one design × test bed combination having a 94% reduction in the 90th percentile of the non-standardized EMSPE values.

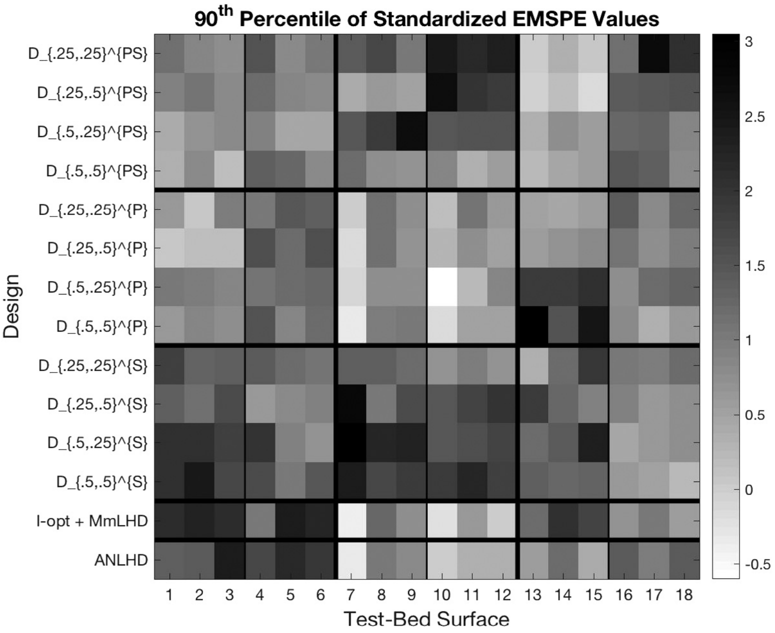

Fig. 9. $(n_p, d_x, n_s, d_x + d_t) = (40, 4, 50, 5)$: A grayscale heatmap of the 90th percentile of the standardized EMSPE values for the 14 designs listed in Table 3 and the 18 test-bed surface types listed in Table 5. Test bed surfaces 1–3, 4–6, 7–9, 10–12, 13–15 and 16–18 use $S^{Krig}_{0.25}$, $S^{Krig}_{0.5}$, S^{SL}_{E}, S^{SL}_{U}, S^{MXB}_{edge}, and S^{MXB}_{mid}, respectively, as $y^s(\boldsymbol{x}, t)$. Within each grouping of three simulator surfaces, the $\zeta_{\text{test}}(\boldsymbol{x})$ to be estimated is the sum of $y^s_{\text{test}}(\boldsymbol{x}, 0.25)$ and $\delta_{\text{test}}(\boldsymbol{x})$ which is: $\equiv 0$, $B^{Krig}_{0.5}$, $B^{Krig}_{0.75}$, in order.

The 6 test-beds using the non-stationary modified Xiong/Ba simulator surface in Eq. (14) occasionally predicted worse when using 10 runs per input compared with 5 runs per input; with the largest observed increase in the 90th percentile of the non-standardized EMSPE values being 250%. Similar increases were observed for the *simulator-only* prediction of modified Xiong/Ba surfaces in Leatherman et al. (2016).

Figs. 6–9 and the corresponding figures in the Supplementary material tell a complicated story about the designs' prediction accuracy (see Appendix A). First, there is no single design that predicts better than *all* competing designs for *all* test-beds and *all* design sizes, i.e., no (design) row is uniformly "lighter" than all other rows for all of Figs. 6–9. However, some designs are clearly inferior to others for sufficiently many design sizes, such that one should use one of the better alternatives. First, the I-opt + MmLHD should be avoided because it is often inferior to ANLHD. Second, the four designs that add simulator runs to the I-opt physical experiment design, the $D^S_{\_,\_}$ designs, are each inferior to the corresponding $D^P_{\_,\_}$ designs. Third, the $D^P_{0.5,\_}$ designs are inferior to the $D^P_{0.25,\_}$ designs. Fourth, the $D^{PS}_{0.5,\_}$ designs are inferior to the $D^{PS}_{0.25,\_}$ designs.

This leaves five designs that are candidates to provide the most accurate predictions: $D^{PS}_{0.25,0.25}$, $D^{PS}_{0.25,0.5}$, $D^P_{0.25,0.25}$, $D^P_{0.25,0.5}$, and ANLHD. ANLHD is particularly effective for predicting non-stationary surfaces, such as the Xiong/Ba surfaces (#13–#18) with the center displaying different behavior than the edge or vice-versa. The same is true for the Sobol´–Levitan surfaces (#7–#12) which have a large spike in one corner of the input space. In both the Xiong/Ba and the Sobol´–Levitan surfaces, not having a design point near the non-stationary activity leads to large prediction errors.

Among the local IMSPE-optimal designs $D^{PS}_{\_,\_}$ and $D^P_{\_,\_}$, those with $\boldsymbol{\rho}_Z = 0.25$ produce more accurate predictions than designs with $\boldsymbol{\rho}_Z = 0.5$. Thus the designs producing the most accurate predictions appear robust to the choice of $\boldsymbol{\rho}_\delta$ but depend heavily on the choice of $\boldsymbol{\rho}_Z$. Designs $D^{PS}_{0.25,\_}$ and $D^P_{0.25,\_}$ predict well for the Kriging surfaces (#1–#6), which is not surprising as these are the surfaces for which the designs were constructed. Additionally, the locally optimal design $D^{PS}_{0.25,0.5}$ has many points near the edges of the design space and therefore predicts well for the test-beds with surfaces that have non-stationary activity near the edges of the input space (#7–#15). Similarly, designs with inputs near the middle of the design space predict well for the modified Xiong/Ba surfaces that have non-stationary activity near the middle of the input space (#16–#18); designs ANLHD, $D^P_{0.25,0.25}$, and $D^P_{0.25,0.5}$ all predict well for these surfaces.

Recalling that $D^P_{\_,\_}$ designs require roughly 10% of the computing time to construct compared to the corresponding $D^{PS}_{\_,\_}$ designs, combining a MmLHD for the simulator experiment with a design that optimizes the IMSPE criterion for the physical experiment is an attractive design choice. While the prediction criterion for this paper is different than that of Ranjan et al. (2011), our conclusions agree qualitatively with their assessment that designs that are locally IMSPE-optimal in the physical element ($D^P_{\_,\_}$) predict best, followed by the designs that are combined locally IMSPE-optimal ($D^{PS}_{\_,\_}$), and lastly followed by designs that are locally IMSPE-optimal in the simulator element ($D^S_{\_,\_}$).

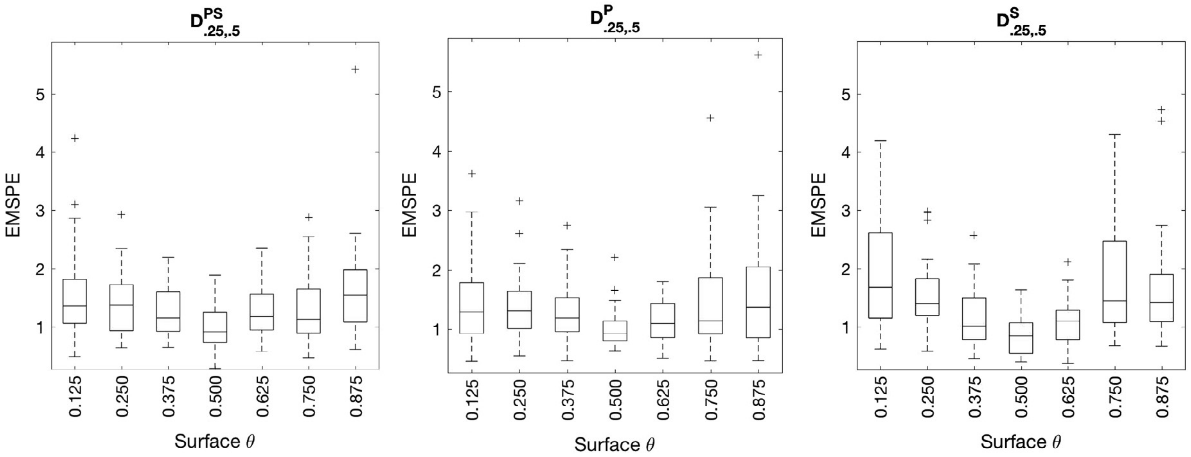

Fig. 10. $(n_p, d_x, n_s, d_x + d_t) = (10, 2, 15, 3)$: Boxplots of (non-standardized) EMSPE values when predicting 30 realizations of the surface $\zeta_{\text{test}}(\boldsymbol{x}) = y^s_{\text{test}}(\boldsymbol{x}, \theta) + \delta_{\text{test}}(\boldsymbol{x})$ for $\theta \in \{0.125, 0.25, \ldots, 0.875\}$ when $y^s_{\text{test}}(\boldsymbol{x}, t)$ is $S^{Krig}_{0.25}$ and $\delta_{\text{test}}(\boldsymbol{x})$ is $B^{Krig}_{0.5}$. Panels (from left to right) correspond to designs $D^{PS}_{0.25,0.5}$, $D^{P}_{0.25,0.5}$, and $D^{S}_{0.25,0.5}$, which were constructed under the assumption $\theta = 0.5$.

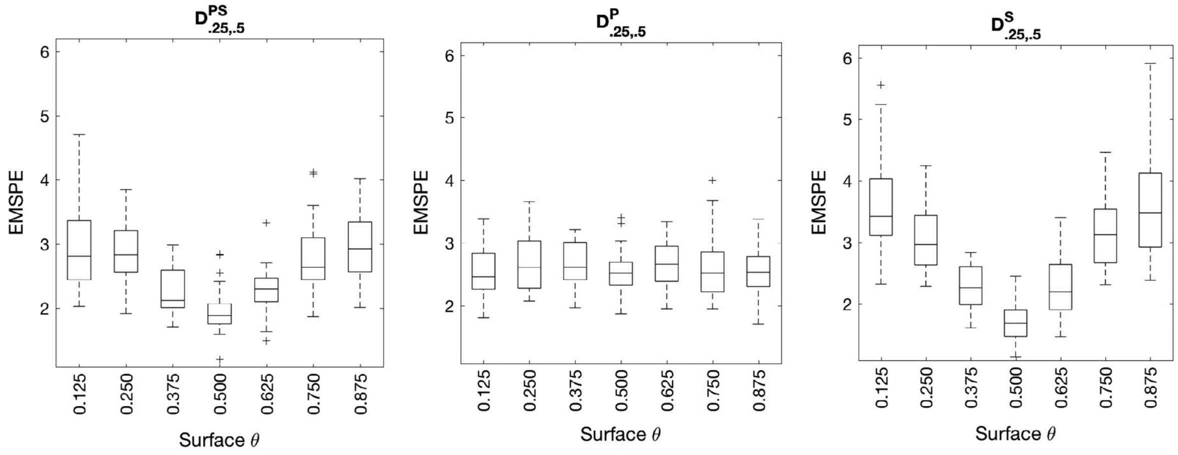

Fig. 11. $(n_p, d_x, n_s, d_x + d_t) = (40, 4, 50, 5)$: Boxplots of (non-standardized) EMSPE values when predicting 30 realizations of the surface $\zeta_{\text{test}}(\boldsymbol{x}) = y^s_{\text{test}}(\boldsymbol{x}, \theta) + \delta_{\text{test}}(\boldsymbol{x})$ for $\theta \in \{0.125, 0.25, \ldots, 0.875\}$ when $y^s_{\text{test}}(\boldsymbol{x}, t)$ is $S^{Krig}_{0.25}$ and $\delta_{\text{test}}(\boldsymbol{x})$ is $B^{Krig}_{0.5}$. Panels (from left to right) correspond to designs $D^{PS}_{0.25,0.5}$, $D^{P}_{0.25,0.5}$, and $D^{S}_{0.25,0.5}$, which were constructed under the assumption $\theta = 0.5$.

6. The dependence of prediction accuracy on the assumed $\boldsymbol{\theta}$

All local IMSPE-optimal designs used for the simulation study in Section 5 were constructed with $\boldsymbol{\theta} = 0.5 \times \mathbf{1}_{d_t}$, $\sigma^2_\delta/\sigma^2_Z = 0.1$, and $\sigma^2_\epsilon/\sigma^2_Z = 0.01$. The Kriging test-bed surfaces in the simulation study of Section 5 were used to assess prediction accuracy when the test-bed correlation parameters $\boldsymbol{\rho}_Z$ and $\boldsymbol{\rho}_\delta$ are different than their assumed values for design construction. The current section focuses on prediction accuracy in test beds formed from stationary GP draws with $\boldsymbol{\theta} \neq 0.5 \times \mathbf{1}_{d_t}$. Each of the $D^{PS}_{0.25,0.5}$, $D^{S}_{0.25,0.5}$, and $D^{P}_{0.25,0.5}$ designs was used to predict 30 test surfaces $\zeta_{\text{test}}(\boldsymbol{x}) = y^s_{\text{test}}(\boldsymbol{x}, \boldsymbol{\theta}) + \delta_{\text{test}}(\boldsymbol{x})$ for a given $\boldsymbol{\theta} \in \left\{0.125 \times \mathbf{1}_{d_t}, 0.25 \times \mathbf{1}_{d_t}, \ldots, 0.875 \times \mathbf{1}_{d_t}\right\}$, where $y^s_{\text{test}}(\boldsymbol{x}, \boldsymbol{t})$ is a draw from $S^{Krig}_{0.25}$ and $\delta_{\text{test}}(\boldsymbol{x})$ is a draw from $B^{Krig}_{0.5}$.

For the eight design sizes listed in and directly below Eq. (19), the EMSPE in (18) was calculated; because all test surfaces are based on the same stationary GP model, the EMSPE was not standardized. Boxplots of the EMSPE values for the 30 surface realizations are shown in Figs. 10 and 11 for design sizes (10, 2, 15, 3) and (40, 4, 50, 5), respectively, and in the Supplementary material for the remaining six design sizes (see Appendix A). Each individual boxplot corresponds to a test bed of surfaces constructed for a specific value of $\boldsymbol{\theta}$.

Most panels in Figs. 10 and 11 (and in the related figures in the Supplementary material, Appendix A) provide intuitive results. The EMSPE values are smallest for test-bed surfaces constructed using $\boldsymbol{\theta} = 0.5 \times \mathbf{1}_{d_t}$, the calibration parameter used to construct the design. For many panels (across design types and sizes), the EMSPE increases as the absolute value of the distance between $\boldsymbol{\theta} = 0.5 \times \mathbf{1}_{d_t}$ used in the design construction and the $\boldsymbol{\theta}$ used to construct the test bed increases. Also the range of the EMSPEs appears to increase as the absolute distance between the design $\boldsymbol{\theta}$ and surface $\boldsymbol{\theta}$ increases. Focusing on Fig. 11, where $(n_p, d_x, n_s, d_x + d_t) = (40, 4, 50, 5)$, the middle panel shows that $D^{P}_{0.25,0.5}$, is robust to the calibration parameter used to construct the stationary GP. However, the EMSPE values for $D^{P}_{0.25,0.5}$ never reach the best EMSPE values achieved by $D^{S}_{0.25,0.5}$ when the surface is constructed using $\theta = 0.5$. On the other hand, the EMSPE values for $D^{P}_{0.25,0.5}$ also do not reach the worst EMSPE values achieved by $D^{S}_{0.25,0.5}$ when the surface is constructed using small or large θ.

7. Summary

This paper compares a variety of criteria to select the *initial design* of a physical system experiment combined with that of a deterministic simulator of the physical system when the goal is accurate prediction of the mean of the physical system. Design criteria and construction algorithms are described for local IMSPE-optimal designs and maximin augmented nested LHDs (MmANLHDs). A local minimum IMSPE design for one of the experiments, simulator or physical system, is constructed when the other experiment uses an "off-the-shelf" design, either an MmLHD for the simulator experiment or an I-optimal design for the physical experiment. These designs are compared with the frequently-used combination of an I-optimal design with an MmLHD. For a large test bed of stationary and non-stationary surfaces and for each design studied, a Bayesian calibrated predictor is used to estimate the mean of the physical system. EMSPE values are calculated for each test surface at a grid of inputs.

This simulation study shows that no single design predicts better than all competing designs for all test-beds and all design sizes. However certain designs should be avoided, general recommendations can be made, and particular designs can be recommended for specific surface types. The I-optimal physical design + MmLHD for the simulator is inferior to the MmANLHD. The I-optimal physical design + local IMSPE-optimal simulator designs are inferior to both the local IMSPE-optimal combined designs and to the MmLHD for the simulator + local IMSPE-optimal physical system experimental designs. Local IMSPE-optimal combined designs or local IMSPE-optimal physical system experimental designs that use $\boldsymbol{\rho}_Z = 0.5$ are inferior to those that use $\boldsymbol{\rho}_Z = 0.25$.

The five recommended designs, ANLHD, $D^{PS}_{0.25,0.25}$, $D^{PS}_{0.25,0.5}$, $D^{P}_{0.25,0.25}$, and $D^{P}_{0.25,0.5}$, have different strengths. ANLHD is least expensive to compute, and $D^{PS}_{0.25,-}$ is the most expensive to compute. Among the local IMSPE-optimal designs, $D^{P}_{0.25,-}$ requires the least computational effort. For stationary GP draws, $D^{P}_{0.25,-}$ and $D^{PS}_{0.25,-}$ predict well, with $D^{P}_{0.25,-}$ often being superior to the correlation comparable $D^{PS}_{0.25,-}$. For non-stationary surfaces, designs with inputs near the non-stationary activity predict well. For example, $D^{PS}_{0.25,0.5}$ and ANLHD have design points near the spike in the Sobol´–Levitan simulator surface and consistently predict well for this surface. ANLHD predicts well for the non-stationary modified Xiong/Ba simulator surfaces that have activity near the edges of the input space. Finally, many designs, including ANLHD, $D^{P}_{0.25,0.25}$, and $D^{P}_{0.25,0.5}$ have points near the middle of the design space and thus predict well for the non-stationary simulator surfaces with activity near the middle of the input space.

A second simulation study shows that local IMSPE-optimal designs perform best for draws from stationary GPs having the same $\boldsymbol{\theta}$ used to construct the design. These studies show both an increase in the median EMSPE and the range of EMSPE values as the physical surface draws have $\boldsymbol{\theta}$ values that are further from the $\boldsymbol{\theta}$ used to construct the design.

Acknowledgments

The authors wish to thank the reviewers and the Editor for suggestions which led to the improvement of this paper. The authors also wish to thank Bradley Jones for helpful discussions. This research was sponsored, in part, by an allocation of computing time from the Ohio Supercomputer Center, and by the National Science Foundation under Agreements DMS-0806134 and DMS-1310294 (awarded to The Ohio State University). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Appendix A. Supplementary material

The designs and construction code discussed in this manuscript can be found on the first author's website http://stat.wvu.edu/~erl/CombinedDesigns/. Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2016.07.013.

References

Audze, P., Eglais, V., 1977. New approach for planning out of experiments. Probl Dynam. Strengths 35, 104–107.
Ba, S., Joseph, V.R., 2012. Composite gaussian process models for emulating expensive functions. Ann. Appl. Stat. 6 (4), 1838–1860.
Crary, S., 2002. Design of computer experiments for metamodel generation. Analog Integr. Circuits Signal Process. 32.
Crary, S., Stormann, J., 2015. Four-Point, 2D, Free-Ranging, IMSPE-Optimal, Twin-Point Designs. ArXiv e-prints.

Fang, K., Wang, Y., 1994. Number Theoretic Methods in Statistics. Chapman & Hall.

Gattiker, J.R., 2008. Gaussian Process Models for Simulation Analysis (GPM/SA) Command, Function, and Data Structure Reference. Tech. Rep. LA-UR-08-08057. Los Alamos National Laboratory.

Hardin, R.H., Sloane, N.J.A., 1993. A new approach to the construction of optimal designs. J. Statist. Plann. Inference 37, 339–369.

Higdon, D., Gattiker, J., Williams, B., Rightley, M., 2008. Computer model calibration using high dimensional output. J. Amer. Statist. Assoc. 103, 570–583.

Higdon, D., Kennedy, M., Cavendish, J., Cafeo, J., Ryne, R., 2004. Combining field data and computer simulations for calibration and prediction. SIAM J. Sci. Comput. 26, 448–466.

JMP, 1989–2007. Version 11. SAS Institute Inc., Cary, NC.

Johnson, M.E., Moore, L.M., Ylvisaker, D., 1990. Minimax and maximin distance designs. J. Statist. Plann. Inference 26, 131–148.

Kennedy, J., Eberhart, R., 1995. Particle swarm optimization. In: IEEE International Conference on Neural Networks, 1995. Proceedings. Vol. 4. pp. 1942–1948.

Kennedy, M.C., O'Hagan, A., 2001. Bayesian calibration of computer models (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 63, 425–464.

Leatherman, E.R., Dean, A.M., Santner, T.J., 2014a. Computer experiment designs via particle swarm optimization. In: Melas, V., Mignani, S., Monari, P., Salmaso, L. (Eds.), Topics in Statistical Simulation: Research Papers from the 7th International Workshop on Statistical Simulation. Vol. 114. Springer, pp. 309–317.

Leatherman, E.R., Guo, H., Gilbert, S.L., Hutchinson, I., Maher, S.A., Santner, T.J., 2014b. Using a statistically calibrated biphasic finite element model of the human knee joint to identify robust designs for a mensical substitute. J. Biomech. Eng. 136 (7), 071007.

Leatherman, E.R., Santner, T.J., Dean, A.M., 2016. Computer experiment designs for prediction. Tech. Rep. 895. The Ohio State University, Department of Statistics.

Liefvendahl, M., Stocki, R., 2006. A study on algorithms for optimization of latin hypercubes. J. Statist. Plann. Inference 136, 3231–3247.

Loeppky, J.L., Williams, B.J., Welch, W.J., 2010. A critical assessment of computer model calibration (unpublished manuscript).

MATLAB. 8.5.0.197613 (R2015a). The MathWorks Inc., Natick, Massachusetts, 2015.

Moon, H., Santner, T., Dean, A.M., 2011. Algorithms for generating maximin latin hypercube and orthogonal designs. J. Stat. Theory Pract. 5, 81–98.

Niederreiter, H., 1978. Quasi-monte carlo methods and pseudo-random numbers. Bull. Amer. Math. Soc. 84, 957–1041.

Niederreiter, H., 1992. Random Number Generation and Quasi-Monte Carlo Methods. SIAM, Philadelphia.

Owen, A.B., 1995. Randomly permuted (t, m, s)-nets and (t, s) sequences. In: Niederreiter, H., Shiue, P.J.-S. (Eds.), Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing. Springer-Verlag, New York, pp. 299–317.

Ranjan, P., Lu, W., Bingham, D., Reese, S., Williams, B.J., Chou, C.-C., Doss, F., Grosskopf, M., Holloway, J.P., 2011. Follow-up experimental designs for computer models and physical processes. J. Stat. Theory Pract. 5 (1), 119–136.

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

Sobol´, I., Levitan, Y., 1999. On the use of variance reducing multipliers in monte carlo computations of a global sensitivity index. Comput. Phys. Comm. 117 (1–2), 52–61.

Studden, W.J., 1977. Optimal designs for integrated variance in polynomial regression. In: Gupta, S.S., Moore, D.S. (Eds.), Statistical Decision Theory and Related Topics. Vol. II. Academic Press, New York, pp. 411–420.

Trosset, M.W., 1999. The krigifier: a procedure for generating pseudorandom nonlinear objective functions for computational experimentation. Tech. Rep. 35. Institute for Computer Applications in Science and Engineering, NASA Langley Research Center.

van Dam, E., den Hertog, D., Husslage, B., Rennen, G., 2013. Space-filling designs. http://www.spacefillingdesigns.nl/, (accessed: October–December 2015).

Welch, W.J., 1985. Aced: Algorithms for the construction of experimental designs. Amer. Statist. 39, 146.

Williams, B.J., Loeppky, J.L., Moore, L.M., Macklem, M.S., 2011. Batch sequential design to achieve predictive maturity with calibrated computer models. Reliab. Eng. Syst. Saf. 96 (9), 1208–1219.

Xiong, Y., Chen, W., Apley, D., Ding, X., 2007. A non-stationary covariance-based kriging method for metamodelling in engineering design. Internat. J. Numer. Methods Engrg. 71 (6), 733–756.

Yang, X.-S., 2010. Engineering Optimization: An Introduction with Metaheuristic Applications, first ed. Wiley Publishing.