Research Article

Quality and Reliability Engineering International

(wileyonlinelibrary.com) DOI: 10.1002/qre.1213

Published online 22 June 2011 in Wiley Online Library

Parameter Calibration and Inadequacy Modeling in a Computer Experiment

Wenny Chandra[a] and Dennis K. J. Lin[b]*†

Deterministic simulation is a popular tool used to numerically solve complex mathematical models in engineering applications. These models often involve parameters in the form of numerical values that can be calibrated when real-life observations are available. This paper presents a systematic approach in parameter calibration using Response Surface Methodology (RSM). Additional modeling by considering correlation in error structure is suggested to compensate the inadequacy of the computer model and improve prediction at untried points. Computational Fluid Dynamics (CFD) model for manure storage ventilation is used for illustration. A simulation study shows that in comparison to likelihood-based parameter calibration, the proposed parameter calibration method performs better in accuracy and consistency of the calibrated parameter value. The result from sensitivity analysis leads to a guideline in setting up factorial distance in relation to initial parameter values. The proposed calibration method extends RSM beyond its conventional use of process yield improvement and can also be applied widely to calibrate other types of models when real-life observations are available. Moreover, the proposed inadequacy modeling is useful to improve the accuracy of simulation output, especially when a computer model is too expensive to run at its finest level of detail.

Keywords: computational fluid dynamics; deterministic simulation; parameter calibration; response surface methodology; steepest descent

1. Introduction

Computer simulation has been used widely to mimic the real-world process or system over time. A simulation model can be utilized to study a system before it is built or to predict the impact of changes on the performance of an existing one. Deterministic simulation, in particular, is a popular tool to numerically solve complex mathematical models for engineering applications. It has been used extensively in various disciplines including finite element analysis to determine the best product design, e.g. aerospike nozzle design[1] and Computational Fluid Dynamics (CFD) to model indoor or outdoor airflow, e.g. manure storage ventilation[2].

In Zhao *et al.*[2], simulation based on CFD is used to determine the time needed until toxic gas in manure storages reaches a level below the permissible standard. The simulation is performed using the standard $k-\varepsilon$ model with two default numerical values in PHOENICS (Parabolic, Hyperbolic Or Elliptic Numerical Integration Code Series) software. Despite efforts to validate the simulation model, the discrepancy between measured and simulated values remains apparent.

Motivated by the problems above, the objective of this paper is twofold: (i) to acquire a systematic parameter calibration procedure to achieve the minimum difference between observed and simulated values and (ii) to model the remaining bias of simulation results after calibration is performed in order to improve future prediction.

Conventionally, parameter calibration is often an *ad hoc* process of changing parameter values until the best match between simulated and observed values are obtained[3, 4]. Moreover, when more than one parameter is involved, the search for the optimal parameter value is sometimes done by tuning one parameter at a time[5]. Cox *et al.*[6] provided a more analytical approach to estimate the unknown parameters through non-linear least squares and likelihood-based models. Kennedy and O'Hagan[7] introduced a Bayesian calibration technique based on the Gaussian process model with correlated errors—derive the mathematical form of fitted parameters and then correct the discrepancy between model predictions and observed data. Higdon *et al.*[8] overcame the complexities of Kennedy and O'Hagan's technique in multivariate settings by applying principal component analysis to reduce the dimensionality of the problem. Note that Cox *et al.*[6] calibrated tuning constants while Kennedy and O'Hagan[7] and Higdon *et al.*[8] calibrated parameters that have physical meaning. Han *et al.*[9] differentiated between tuning parameter (constant) and

[a]*Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA, U.S.A.*
[b]*Department of Statistics, The Pennsylvania State University, University Park, PA, U.S.A.*
**Correspondence to: Dennis K. J. Lin, Department of Statistics, The Pennsylvania State University, University Park, PA, U.S.A.*
†*E-mail: dkl5@psu.edu*

calibration parameter (with physical meaning) and determined both by using the discrepancy measure and hierarchical Bayesian model.

These methods require sophisticated mathematical models and high computational effort for parameter estimation in addition to the deterministic simulation model. For example, the Bayesian calibration approach requires prior knowledge and Gaussian distributional assumption is often chosen to ease mathematical derivation although it may not be appropriate in reality. A more practical yet effective calibration method is needed.

In this paper, a calibration technique based on Response Surface Methodology (RSM) is proposed. The procedure involves sequential runs of simulation, enables fine-tuning of multiple parameters, and determines the best parameter value based on available observations. Recognizing that the computer simulator may not be run at the highest level of detail due to time constraint and difficulty in convergence, an inadequacy model with correlated errors is proposed to compensate this problem and improve the prediction ability of the simulator.

This paper is organized as follows: Section 2 proposes a step-by-step calibration and inadequacy modeling methodology, followed by an illustration using a manure storage ventilation case in Section 3. Section 4 evaluates the performance of proposed calibration method by using theoretical functions found in the literature in comparison to maximum likelihood based calibration[6]. Sensitivity analysis is also performed to investigate whether some settings in the proposed method have significant impacts on the results obtained. A discussion and the future work are given in Section 5.

2. Proposed methodology

The focus here is on cases where deterministic simulation is used to solve a complicated mathematical model which represents a real system. There are often some parameters in the mathematical model where initial or default values are used in the simulation. However, fine-tuning is necessary to obtain the best value at which computer outputs closely match the observations. We denote this numerical value as parameter (**T**) and call the fine-tuning of its value as parameter calibration. Given the values of **T**, the observation ($\mathbf{Y_{obs}}$) is a function of design variable (**X**).

To calibrate the parameter used in the simulator and to improve the future prediction of this simulator, we propose a two-part methodology:

Part 1: *Parameter Calibration* aims to find the best parameter values that minimize the difference between observations and computer outputs.

Part 2: *Inadequacy Modeling* aims to improve the accuracy of simulator by modeling the difference between calibrated simulation output and field observations.

2.1. Part 1: Parameter calibration

RSM first originated in chemical experimentation as a sequential procedure to achieve a more desirable response[10]. The collection of statistical tools in RSM have since been developed further[11] and used widely in industrial processes[12]. While RSM was first intended to be used in hands-on experiments, recent development has seen the popularity of response surface approximation as the metamodel representing complex computer experiments[13–15].

Our methodology uses RSM to find the best parameter value in a deterministic simulation. The desired response is the minimum difference between simulated and observed values, represented by root mean-squared errors (RMSE). Given available observations $\mathbf{Y_{obs}}$ and the computer simulator, the following procedure is proposed:

1. Set up the values of design matrix, **X**, that correspond to $\mathbf{Y_{obs}}$, initial parameter values $\mathbf{T_0}$, and minimum percentage of RMSE improvement ($RMSE_{stop}$) for iteration rule.
2. Using **X** and $\mathbf{T_0}$ setup in step 1, obtain output $\mathbf{Y_{comp}}$ by running the computer simulator.
3. Testing **T** as the experimental factor (with **X** fixed), form a small factorial experiment region surrounding $\mathbf{T_0}$ as the center point. Figure 1 shows the region in coded values for two-parameter case, initial parameter values, $\mathbf{T_0}$, are at level 0. At $T_{-1,1}$ the first parameter is at low level and second parameter at high level. The number of runs needed in this step is 2^kplus 1

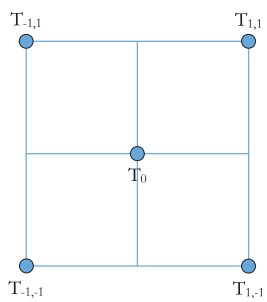

Figure 1. Factorial region with center point with T as factor (*X* is fixed as given)

center point where k is the number of parameters. If the number of simulations needed is a concern, a fractional factorial design can be used instead.

4. Run the simulator at design points to obtain $\mathbf{Y_{obs}}$ (a typical $n \times 1$ vector output) and calculate RMSE at each point by the following formula:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (Y_{\text{obs},i} - Y_{\text{comp},i})^2}{n}}. \tag{1}$$

5. Response surface model building.

   5.1. With RMSE as the response, build a first-order response surface model with the main effects and interaction, two-parameter model is illustrated as follows:

$$\text{RMSE} = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_{12} T_1 T_2 + \varepsilon.$$

   5.2. Test whether interaction term or curvature is significant—basically, the interaction and quadratic curvature. For two-factor example, this is

$$\text{Ratio of sum of squares of interaction: } \text{SS}_{T_1 T_2} = \frac{[t_{12} + (1) - t_1 - t_2]^2}{4}$$

   where (1), t_1, t_2, and t_{12} are RMSE at points T_{-1-1}, T_{1-1}, T_{-11}, T_{11}, respectively and

$$\text{Sum of squares of curvature: } \text{SS}_{\text{pure quadratic}} = \frac{n_F n_C (\bar{y}_F - \bar{y}_C)^2}{n_F + n_C},$$

   where n_F and n_C are the number of factorial and center points, respectively

   $\bar{y}_F$: the average of RMSE from four factorial points and
   $\bar{y}_C$: RMSE of the center point (for one center point, this is simply y_c).

      5.2.1. If either one or both are significant, run a central composite design by collecting additional axial points, shown in Figure 2 for a two-parameter case. To maintain orthogonality, the optimal choice of distance from the center point is $\sqrt{k}$, where k is the number of parameters[12].
      For this design, a second-order model can be written in general as

$$\text{RMSE} = \beta_0 + \sum_{i=1}^{k} \beta_i T_i + \sum_{i=1}^{k} \beta_{ii} T_i^2 + \sum \sum_{i<j=2}^{k} \beta_{ij} T_i T_j + \varepsilon,$$

      and the fitted model as $\text{RMSE} = b_0 + \mathbf{T}'\mathbf{b} + \mathbf{T}'\hat{\mathbf{B}}\mathbf{T}$, where b_0 is the estimate of the intercept

$$T = [T_1, T_2, \ldots, T_k]$$

$$\mathbf{b}' = [b_1, b_2, \ldots b_k]$$

$$\hat{\mathbf{B}} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1k} \\ & b_{22} & \cdots & b_{2k} \\ & & \ddots & \vdots \\ \text{sym.} & & & b_{kk} \end{bmatrix}.$$

      Setting $\partial \hat{y} / \partial \mathbf{T} = \mathbf{b} + 2\hat{\mathbf{B}}\mathbf{T} = 0$, we obtain the stationary point, and in this case the optimal parameter value is: $\mathbf{T}^* = -(1/2)\hat{\mathbf{B}}^{-1}\mathbf{b}$. This concludes the algorithm.

      5.2.2. If neither is significant, then fit a first-order model with just the main effects: $\text{RMSE} = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \varepsilon$.

6. Determine the Path of Steepest Descent as $-\mathbf{b}$, where $\mathbf{b} = [\hat{\beta}_1, \hat{\beta}_2]$ is the coefficient of the first-order model fitted in step 5.2.2.
7. Starting with $j = 1$, new points along the Path of Steepest Descent are determined by the following formula:

$$\text{The } j\text{th new point is } \mathbf{T_{new,j}} = s_j \times (-\mathbf{b}),$$

   where the jth step size is $s_j = j / \|\mathbf{b}\|$.

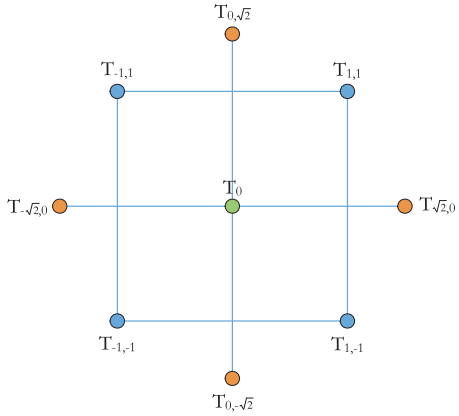

Figure 2. Central composite design for a two-parameter case

8. Run the simulator and compute RMSE at *j*th new point.
   8.1. If the percentage of RMSE improvement is more than $RMSE_{stop}$, increase *j* by 1 and return to step 7 to compute the next new point.
   8.2. If the percentage of RMSE improvement is not more than $RMSE_{stop}$, return to step 3 to form another factorial design with center point.

2.2. *Part 2: Inadequacy modeling*

Using the calibrated parameter ($\mathbf{T}^*$) obtained in part 1, the difference between calibrated simulation output and observations may still be apparent. This difference is caused by systematic inadequacy of the simulator and could be modeled using kriging model—a linear model with correlated errors—given the consideration that the response may have a spatial relationship.

9. Assuming additive relationship, inadequacy of the computer model ($\mathbf{I}$) is defined as the difference between observation ($\mathbf{Y_{obs}}$) and computer output obtained using calibrated parameter ($\mathbf{Y^*_{comp}}$): $\mathbf{I} = \mathbf{Y_{obs}} - \mathbf{Y^*_{comp}}(\mathbf{X}, \mathbf{T}^*)$.
10. Model $\mathbf{I}$ in the form of a linear function of X: $\mathbf{I}(\mathbf{X}) = \sum_{\mathbf{j=0}}^{\mathbf{p}} \beta_{\mathbf{j}} \mathbf{X_j} + \varepsilon$, with ε: error with mean 0 and variance σ^2.
11. Prediction at untried points is then given by the calibrated computer output plus the prediction of the inadequacy, $\hat{\mathbf{I}}(\mathbf{X}) = \sum_{\mathbf{j=0}}^{\mathbf{p}} \hat{\boldsymbol{\beta}}_{\mathbf{j}} \mathbf{X_j}$, i.e.:

$$\hat{\mathbf{Y}} = \mathbf{Y^*_{comp}}(\mathbf{X}, \mathbf{T}^*) + \hat{\mathbf{I}}(\mathbf{X}).$$

The proposed two-part method provides a separate solution to parameter calibration and refinement of the simulation model. Each part can be used independently or both can be used together in the order specified, i.e. parameter calibration followed by inadequacy modeling. In part 1 parameter calibration, the sequential approach enables the user to understand and visualize the iterative process of searching the best parameter. This approach works well when the RMSE function is relatively smooth and does not have multiple local optima. Part 2 provides a quick modeling of the remaining bias between calibrated simulation results and observed values. The basis of the proposed correlated linear model is that there may be some relationship among design variables, such as when the output is spatially correlated.

3. Case study: parameter calibration for manure storage ventilation

Confined manure storage entry is a major safety issue. There has been deaths of farm workers entering the storage, caused by toxic and asphyxiating gases produced by the fermentation process. According to Occupational Safety and Health Administration (OSHA) Standards, forced ventilation in a confined space is imperative prior to entry, and effective ventilation strategies can potentially reduce the risks associated with confined manure storage entry. H_2S is one of the primary toxic gases, OSHA has set Permissible Exposure Limit (PEL) of 10 ppm for this gas.

Using the previously determined effective ventilation strategies for prismatic confined-space manure storages, Zhao *et al.*[2] validated a simulation-based CFD model. CFD modeling protocol is useful to predict air movement in ventilated spaces, including spatial variations in temperature and pollutant concentrations. The time taken to reach PEL (T_{PEL}) is simulated using PHOENICS software. The standard $k-\varepsilon$ model consists of six sets of differential equations and includes two important parameters, $C_{\varepsilon 1}$ and $C_{\varepsilon 2}$. Note that these two parameters influence the computer output but they are independent of the observation.

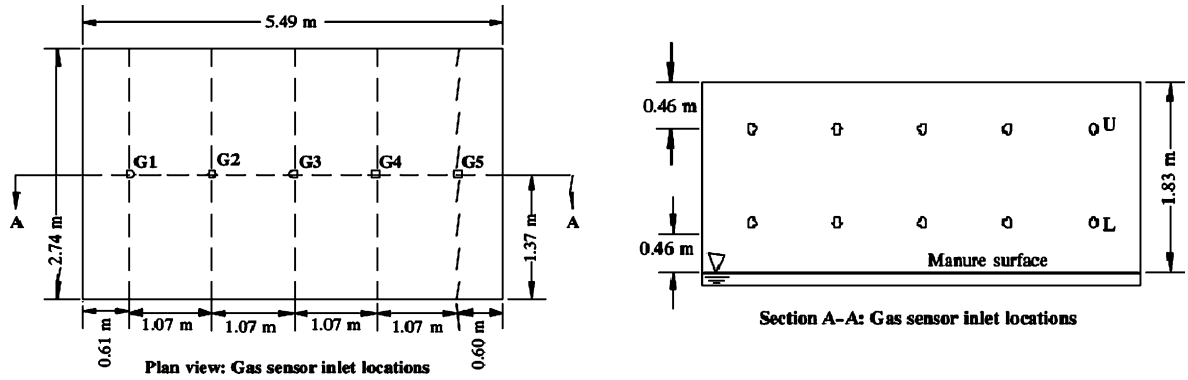

Figure 3. Sampling location of H_2S measurements

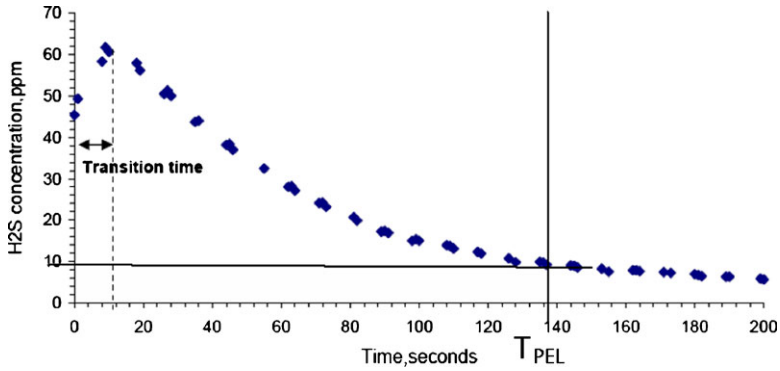

Figure 4. H_2S decay curve during forced ventilation

PHOENICS (Parabolic, Hyperbolic Or Elliptic Numerical Integration Code Series) is a general-purpose CFD software package for quantitative prediction of fluid flow, heat transfer, changes of chemical and physical compositions via computer-based simulation. This software has been used in a wide array of applications, ranging from wind energy simulation, environmental concern of city square car park exhaust, lecture hall ventilation, to airflow analysis in pharmaceutical clean room. CFD users rely on validated models before applying them to their particular cases.

For this case study we use a simulator built for a solid floor storage type and high velocity fan that creates high air exchange rate[16]. Computer output and field measurements are real-time H_2S concentration which are then used to compute T_{PEL} at those locations. The plan view of Figure 3 shows five grid points (G1 through G5) of sampling location seen from above the manure storage. The cross section A-A shows the upper (U) and lower (L) locations at each grid point. From these 10 sampling locations, H_2S concentration is measured at 10-s intervals.

Figure 4 shows a typical curve of H_2S decay where the transition time is the period right after manure agitation ceases until H_2S concentration is stable and begins the exponential decay. T_{PEL} is calculated as the time when H_2S concentration reaches the PEL of 10 ppm, shown in Figure 4 at the 136th second.

Note that these 10 T_{PEL} values ($\mathbf{Y_{obs}}$) have identical parameter setups ($\mathbf{T_1} = C_{\varepsilon 1}$ and $\mathbf{T_2} = C_{\varepsilon 2}$), but with different locations (the *XZ*-coordinates). Such a location difference will be treated as random errors in the response surface model building. We next follow the steps stated in the previous section to fine-tune the parameters and improve the future prediction:

3.1. Part 1: Parameter Calibration

1. Set of variable input **X** associated with the given observations are *XZ*-coordinates of 10 location measurements. We illustrate the procedure by choosing initial values of parameters as $\mathbf{T_1} = 1.3$ and $\mathbf{T_2} = 1.6$.
2. Run simulator (Phoenics) and obtain output $\mathbf{Y_{comp}}$, which is a 10×1 vector of T_{PEL} values.
3. Run factorial design with center point.
   One unit of coded value is chosen to be 5% of the original parameter value. This value is chosen since it is considered small enough for linearity assumption to hold (first-order model is used) but also large enough to cause meaningful changes in the output. Figure 5 shows the complete values used for the factorial design plus the center point.
4. Simulation in Steps 2 and 3 produces 10 simulation outputs at each factorial point and the center point. Using Equation (1) and the same set of 10 Y_{obs} values, RMSE is computed for each point. The result is shown in Figure 6. The value of 6.38 for $T_{-1,1}$ point, for example, is obtained by computing the difference between Y_{obs} and Y_{comp} obtained from simulation in step 3 run at low level of T_1 and high level of T_2.

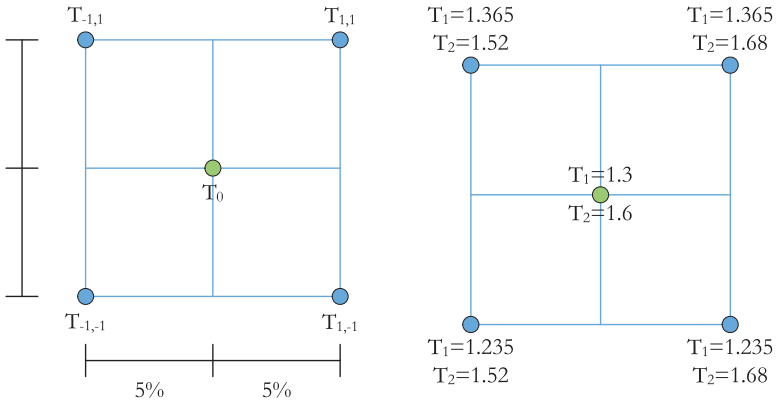

Figure 5. First four factorial points around the original parameter

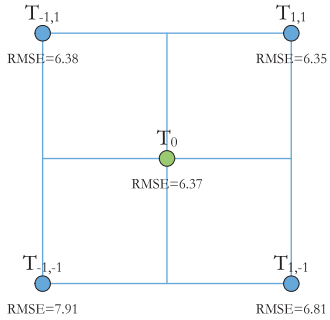

Figure 6. RMSE of first four factorial points around the original parameter point

5. Response surface model building

   5.1. First-order response surface model for RMSE (y) with interaction term is as follows:

$$\text{RMSE}(y) = 6.76 - 0.28T_1 - 0.49T_2 + 0.27T_1T_2 + \varepsilon.$$

   5.2. The residual error has only 1 degree of freedom, therefore the p-value from the F test of significance of terms may not be reliable. Instead, the percentage contribution of the sum of squares total (1.78) is used to decide the significance. Since the contribution of the interaction and curvature term is only 16% and 11%, respectively, it is concluded that the first-order main effect model is sufficient and the final model obtained is

$$\text{RMSE} = 6.76 - 0.28T_1 - 0.49T_2 + \varepsilon.$$

6. Using the coefficients obtained in the main effect model, the Path of Steepest Descent is determined as $-\mathbf{b} = [0.280.49]$.
7. Determine new points along the Path of Steepest Descent. The first new point ($j = 1$) is

$$\mathbf{T_{new,j}} = \frac{j}{\|\mathbf{b}\|} \times (-\mathbf{b})$$

$$\mathbf{T_{new,1}} = \frac{1}{\sqrt{0.28^2 + 0.49^2}} \times [0.28 \quad 0.49] = [0.49 \quad 0.87].$$

   As this point is still within the factorial points region (see Figure 7), it will not improve the RMSE value, therefore the second new point is determined in the same manner:

$$\mathbf{T_{new,2}} = \frac{2}{\sqrt{0.28^2 + 0.49^2}} \times [0.28 \quad 0.49] = [0.99 \quad 1.74]$$

   in coded value, which is translated into a real value of $\mathbf{T_{new,2}} = [1.33 \quad 1.67]$.
8. RMSE at $\mathbf{T_{new,2}}$ is 6.32. Since this is an improvement compared to RMSE = 6.37 at T_0, return to step 7 to calculate and simulate another new point. Figure 7 shows new points simulated along the steepest descent path. RMSE is increasing at $\mathbf{T_{new,6}}$, so $\mathbf{T_{new,5}} = [1.43; 1.88]$ is set as new $\mathbf{T_0}$ and the procedure is repeated from step 3.

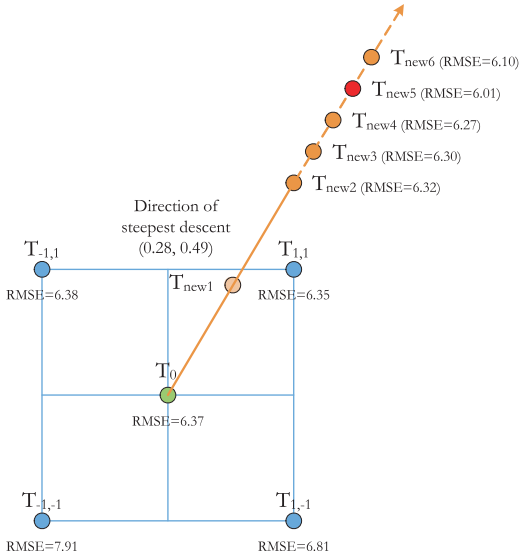

Figure 7. First round steepest descent path

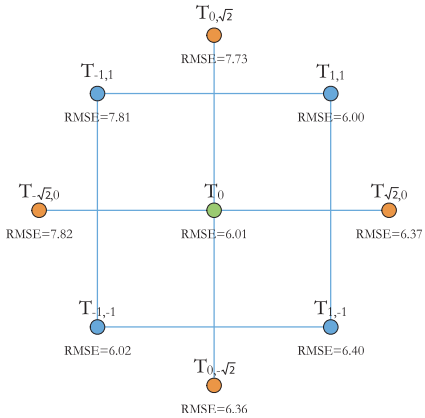

Figure 8. Second round central composite design with star points

In this second round of response surface model building, the sequential sum of squares of interaction term contributes 60% towards the sum of squares total in the ANOVA table, thus 4 star points at $\sqrt{2}$ distance from the center point are added in order to fit a second-order model (see Figure 8).

The second-order model is fitted as follows:

$$\text{RMSE} = 6.01 - 0.44T_1 + 0.42T_2 - 0.55T_1T_2 + 0.41T_1^2 + 0.39T_2^2,$$

$$(s.e.) \quad (0.45) \quad (0.16) \quad (0.16) \quad (0.23) \quad (0.26) \quad (0.26)$$

resulting in the optimal parameter value of $T^* = [0.2113; -0.2381]$ in the second round region, which translates into $T^* = [1.48; 1.93]$. The Final RMSE at this parameter value is predicted to be 5.89, which is a 7% improvement compared to the initial RMSE of 6.37. The confirmation run at this optimal point gives an RMSE of 5.84.

3.2. *Part 2: Inadequacy modeling*

9. Inadequacy is computed as the difference between observed values and calibrated computer outputs: $\mathbf{I} = \mathbf{Y_{obs}} - \mathbf{Y^*_{comp}}$.
10. Fit **I** as a linear function of **X**. The following calculation is done by using `geoR` and `sp` packages in R. For exploratory purposes, first a variogram of the data is calculated and plotted in Figure 9. This sample variogram can be compared with a theoretical variogram to determine the type of function suitable to describe the covariance. However, no known pattern is apparent from Figure 9, therefore the exponential form of covariance is chosen for parameter estimation. Using maximum likelihood estimation, the following model is then obtained:

$$I(X) = 15.20 - 0.60X_1 - 11.92X_2 + \varepsilon,$$

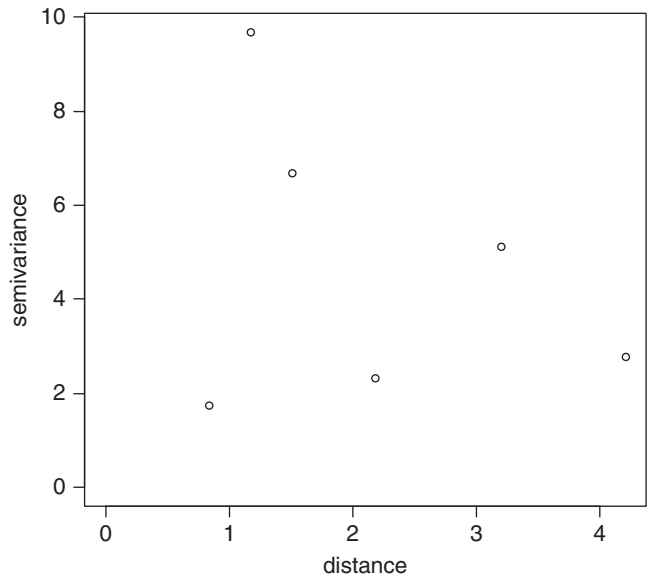

Figure 9. Variogram of H_2S measurement of case study

where X_1: X-coordinate of measurement location; X_2: Z-coordinate of measurement location; and ε: error with mean 0 and variance $\sigma^2 = 4.18$.

11. Prediction at untried point $\mathbf{X_0}$ is given by $\hat{\mathbf{Y}} = \mathbf{Y^*_{comp}}(\mathbf{X_0}, \mathbf{T}^*) + \hat{\mathbf{I}}(\mathbf{X_0})$.

In step 10, a linear function with correlated errors is chosen because the output variable is measured in several locations in the $X-Z$ plane. This is the reason to believe that the output maybe correlated spatially, hence the use of kriging model. Inference on the model coefficients is done by maximum likelihood.

Basically, kriging is a spline smoothing, it derives a smooth curve that connects original data points, it is not possible to obtain prediction at these points. Therefore, to evaluate the predictive ability, the estimation is done by leaving the original data point out (cross validation). For each data point i, linear function $I_{(i)}$ (subscript (i) means all data points used except the ith data) is fitted and used to give prediction at i. Thus, prediction at all original data points is obtained and then compared with the observed values to compute RMSE. RMSE after inadequacy modeling is 2.94 compared to calibrated simulation without inadequacy modeling which yields RMSE of 5.89. Applying the two-step proposed method to another set of 10 observations collected on a different occasion, inadequacy modeling improves RMSE from 6.09 to 4.14.

4. Empirical performance evaluation

To investigate whether the proposed method works well in a wide range of scenarios, simulations are performed with four simple functions acting as simulators. We use the same functions as in Cox *et al.*[6] to enable a comparison between the proposed method and the existing work on parameter calibration. Cox *et al.*[6] introduce a likelihood-based method to estimate the unknown tuning parameters in a computer model. Because Cox *et al.*[6] use iid normal error to generate computer output and observation, the same elements are also added to the proposed method in order to compare both methods fairly.

Table I shows the functions used in the simulation study: Function 1 represents a simple function with a small number of variable and parameters while Function 3, on the other hand, with 6 variables and 4 parameters; Function 2 represents a more complicated relation among the variables and parameters; Function 4 includes a parameter inside the sin function.

In addition to the true parameter values (T_{true}) used in Cox *et al.*[6], we use initial parameter values (T_0) randomly generated from uniform distribution $U(a,b)$ with $a = 0.8*$(true value) and $b = 1.2*$(true value). We choose a 5% distance between factorial points to this initial parameter. The procedure ends when the percentage of RMSE improvement compared to the previous run is less than 5%.

Table II shows the criteria used to evaluate the simulation results. The true values of each function are shown in brackets in the header row. The mean of the calibrated parameter is simply the average of the calibrated parameter values (T^*) divided by 100 000, which is the number of simulation replicates for each function. We can see that the calibrated parameters are very close to the true values. The consistency of the result is also good, as shown by the small standard errors, calculated as the standard deviation of T^* divided by the square root of 100 000. The worst case of RMSE improvement is 30% for function 2.

The accuracy of the calibrated parameters is also judged by measures of distance to the true values. The average absolute distance measures the absolute values of the distance averaged over 100 000 simulations. Looking at the average percentage distance to true value, the proposed method is accurate within 11% of the true value. Another measure of the accuracy is the norm distance to the true value which is the norm of average of $|T^* - T_{\text{true}}|$.

Table I. Functions used as deterministic simulator for performance evaluation (taken from Cox *et al.*[6])

	Function	Number of Variables	Number of Parameters
1.	$f(X,T)=T_1\times\exp(-T_2X)$	1	2
2.	$f(X,T)=\frac{X_1}{T_1}[\sqrt{T_2+(X_2+X_3^2)T_3X_4/X_1^2}-T_2]+X_1+3X_4$	4	3
3.	$f(X,T)=T_1\exp(\lvert X_1+X_2\rvert)+T_2(X_4+1.2X_5+1)/2.5$ $+T_3(T_2+2X_3+X_4)+T_4\exp(X_1+X_3-X_5-X_6)$ $+3\cos[6(X_2+X_3+X_4)]$	6	4
4.	$f(X,T)=T_1\times\exp[T_2(X_1+X_2)]/3+T_3X_4\sin(T_4X_3)+T_2T_4X_3$	3	4

Table II. Proposed method performance evaluation

Criteria	Function 1 (1;1)*	Function 2 (2;4;1)*	Function 3 (2;1;4;4)*	Function 4 (2;1;-1;1)*
Mean of calibrated parameters	0.995; 0.998	2.0056; 3.9832; 1.0078	2.003; 1.0001; 3.9938; 3.999	2.005; 0.9999; −1.0009; 0.9993
Standard error of mean	0.00007; 0.00014	0.0008; 0.0016; 0.0004	0.0001;0 .0004; 0.0014; 0.0015	0.00075; 0.00007; 0.00038; 0.00038
Average RMSE improvement	62.14%	30.26%	72.94%	77.37%
Average absolute distance to true value	0.0175; 0.0329	0.1987; 0.4169; 0.1084	0.0245; 0.1017; 0.3806; 0.4045	0.2024; 0.00175; 0.1024; 0.1018
Average % distance to true value	1.75; 3.29	9.93; 10.42; 10.84	1.23; 10.17; 9.52; 10.11	10.12; 1.75; 10.24; 10.18
Norm distance to true value	0.0373	0.4743	0.5652	0.2492
% no improvement	4.58	11.39	11.96	10.65
Average number of runs	15.13	20.81	33.16	33.40

*True parameter values.

Despite the good performance, in some cases the full factorial points do not produce better RMSE than the original point (initial parameter values), this results in no change of parameter values. In the worst case (function 3), this happens 12% of the time. This is likely because either the region of the factorial points is too narrow, or the initial parameters are too close to the true parameter values.

Table III shows a comparison between the proposed RSM and Cox's PMLE method Cox *et al.*[6] proposed several methods: most based on maximum likelihood estimation and one based on non-linear least squares estimation. The final recommendation is the so-called partially separated MLE (PMLE for short). PMLE samples form the whole region of possible X values, and then estimate the parameters based on an assumed model.

Using three criteria—mean of calibrated parameters, norm distance (which measure the accuracy), and standard error (which measures the consistency),the RSM method outperforms the PMLE method. The RSM method consistently gives more accurate results than PMLE. For instance, the norm distance for function 2 for RSM (0.47) is only slightly less than PMLE (0.56), but for function 1 RMS (0.037) is three times smaller than PMLE (0.101). In terms of the consistency, RSM also outperforms PMLE, especially for function 2 where RSM's standard error ranges in the hundredth while PMLE's ranges from 1.8 to 4.5.

It is of interest to study how robust the performance of the proposed procedure is to the current simulation setup:

1. Initial parameter values are generated from a uniform distribution which ranges 20% below and above the true value.
2. Factorial distance is 5% from the original parameter values.
3. The procedure is stopped when the percentage of RMSE improvement is below 5% compared to the previous run.

Table III. Comparison between the proposed (RSM) and Cox's (PMLE) calibration method

	Function 1 (1;1)*		Function 2 (2;4;1)*	
Criteria	RSM	PMLE	RSM	PMLE
Mean of calibrated parameters	0.995; 0.998	1.082; 1.059	2.0056; 3.9832; 1.0078	2.434; 3.927; 1.351
Standard error of mean	0.00007; 0.00014	0.116; 0.167	0.0008; 0.0016; 0.0004	4.562; 3.28; 1.801
Norm distance to true value	0.037	0.101	0.474	0.563
Criteria	Function 3 (2;1;4;4)*		Function 4 (2;1;−1;1)*	
	RSM	PMLE	RSM	PMLE
Mean of calibrated parameters	2.003; 1.0001; 3.9938; 3.999	1.712; 1.242; 3.08; 4.071	2.005; 0.9999; −1.0009; 0.9993	2.313; 0.926; −0.748; 0.829
Standard error of mean	0.0001; 0.0004; 0.0014; 0.0015	0.602; 0.77; 1.086; 1.159	0.00075; 0.00007; 0.00038; 0.00038	0.748; 0.643; 1.629; 1.79
Norm distance to true value	0.565	0.996	0.249	0.444

*True parameter values.

Table IV. Simulation settings for sensitivity analysis

	Low	Base	High
Initial parameter	±5%	±20%	±50%
Factorial distance	±1%	±5%	±10%
$RMSE_{stop}$	±1%	±5%	±10%

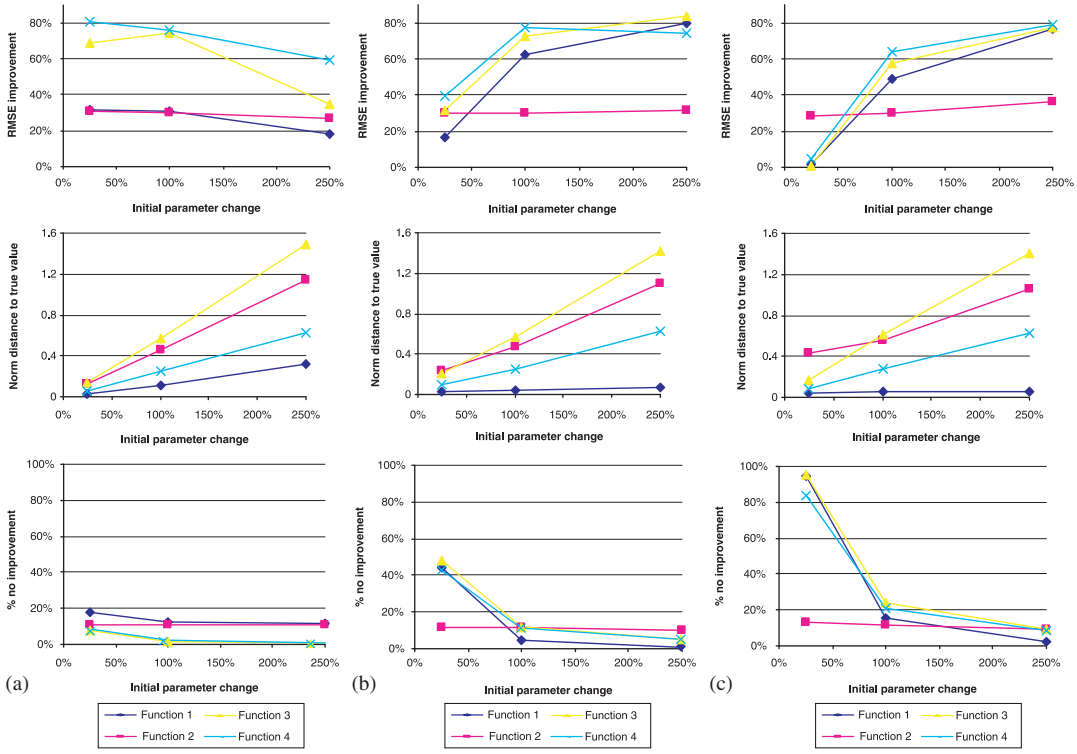

Figure 10. Sensitivity analysis of initial parameter and factorial distance: (a) low factorial distance; (b) base factorial distance; and (c) high factorial distance

Table IV shows the original values (as the base) of these three factors, plus low and high values which provide sufficient large variation to the performance. Preliminary analysis based on the setting in Table IV shows that there is no effect by changing $RMSE_{stop}$. Therefore, we proceed with all nine possible combinations of low-base-high setting to analyze the possible interaction between initial parameter and factorial distance.

Figure 10 shows the effect of different initial parameter values for different values of factorial distance. The three columns from left to right correspond to low, base, and high factorial distance. The three rows represent three performance criteria evaluated, from top: RMSE improvement, norm distance to true value, and percentage of no improvement.

The first row in Figure 10 indicates that when initial parameters are far from true values, the RMSE improvement achieved will be larger. However, when the factorial distance is small, the improvement in RMSE gets smaller when initial parameters get further away. This suggests that if we believe that the initial value is far from the true values, it is advisable to have a larger factorial distance in order to facilitate a wider and quicker search towards the true value.

The middle row of Figure 10 shows the effect of initial parameters' variation on the norm distance of the calibrated parameters. The three graphs from left to right (correspond to low, base, and high factorial distance) show a similar pattern that if the procedure starts from parameter values close to the true value, then the final calibrated values will be more accurate. However, when the factorial distance is larger, the norm distance is slightly smaller. This implies that it is better to have a larger factorial distance if we are unsure of the initial parameter values.

The last row in Figure 10 reveals that when initial parameter is too close to the true value, the calibration often fails to further improve RMSE. This condition becomes worse when the factorial distance is too high. That is, if we believe that we have good initial values for the parameters, we can set a small factorial distance to concentrate the search to a smaller area.

5. Discussion and conclusion

Parameter calibration in a computer experiment has been investigated by Kennedy and O'Hagan[7] and Cox *et al.*[6]. The former utilize Bayesian approach and the latter use likelihood-based modeling to obtain parameter estimates. Both require strong model assumptions and sophisticated computational effort. This paper proposes using the RSM approach to calibrate the parameters. Instead of optimizing the response function as in the traditional RSM, the objective of calibration is to minimize the difference between observations and computer outputs, therefore RMSE is used as the function to be minimized. The proposed method extends RSM beyond its conventional use of process yield improvement.

As the proposed approach is a search method, it is particularly suitable when an initial estimate of parameter value is available but needs to be fine-tuned to minimize the bias of simulation output against the available observations. This situation is common in the case where a general simulation model is available but in order for it to be useful for a specific application, more precise parameter values are needed. Not limited to calibrating a computer model, this method can be applied widely to calibrate other types of models when real-life observations are available.

The proposed method also deals with inadequacy modeling as a means to compensate some level of details sacrificed to achieve faster computer runtime. Correlation among variables is considered by having correlated errors in the model. This approach works well whenever variables are correlated, as in the case study where response is measured spatially. In general, the proposed inadequacy modeling can be used as a complement to improve the accuracy of simulation output, especially when a computer model is too expensive to be run at its finest level of detail.

There are certain assumptions that need to be satisfied in order for the proposed method to work. Observation is a function of design variables but not of parameters. Calibrating the parameters will bring the computer output closer to observation but does not influence the future observation. We also assume that parameters (T) are independent from design variables (X), i.e. calibrated parameters (T^*) are the best value for the model regardless of the values of X.

Furthermore, the response function in RSM is represented by a first-order linear model, this necessitates a smooth function that can be locally reduced to polynomial form of first degree assuming that higher order terms are negligible[12, 17]. If the first-order approximation holds, steepest ascent/descent direction will be given by the regression coefficients[12], which will lead to a new set of parameters with improved RMSE.

In our PHOENICS study, there are only two parameters involved. In general, however, there could be much more parameters in which some of them may not be relevant. In this case, the classical RSM to remove insignificant factors (the screening process in first stage) can be applied (see, for example, Lin[18]).

As a conclusion, we have shown that the RSM methodology is effective in calibrating parameters in a computer experiment. A recommendation is also made to determine the factorial distance depending on how accurate we believe the initial parameter values are.

Acknowledgements

We thank the reviewer's constructive comments which have led to a much more improved version of this paper.

References

1. Simpson TW, Mauery TM, Korte JJ, Mistree F. Kriging models for global approximation in simulation-based multidisciplinary design optimization. *American Institute of Aeronautics and Astronautics Journal* 2001; **39**(12):2233–2241.
2. Zhao J, Manbeck HB, Murphy DJ. Computational fluid dynamics simulation and validation of H_2S removal from fan-ventilated confined-space manure storage. *Transactions of the ASABE* 2007; **50**(6):2231–2246.
3. Banks J, Nelson BL, Nicol DM. *Discrete-Event System Simulation*. Prentice-Hall: New Delhi, 2005.
4. Wolf J, Evans LG, Semenov MA, Eckersten H, Iglesias A. Comparison of wheat simulation models under climate change. I. Model calibration and sensitivity analyses. *Climate Research* 1996; **7**:253–270.
5. Abdulhai B, Sheu JB, Recker W. Simulation of ITS on the Irvine FOT area using 'Paramics 1.5' scalable microscopic traffic simulator—Phase I: Model calibration and validation, 1999. California PATH program, Institute of Transportation Studies, University of California, Berkeley, CA.
6. Cox DD, Park JS, Singer CE. A statistical method for tuning a computer code to a data base. *Computational Statistics & Data Analysis* 2001; **37**:77–92.
7. Kennedy MC, O'Hagan A. Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B* (*Statistical Methodology*) 2001; **63**(3):425–464.
8. Higdon D, Gattiker J, Williams B, Rightley M. Computer model calibration using high dimensional output. *Journal of the American Statistical Association* 2008; **103**(482):570–583.
9. Han G, Santer TJ, Rawlinson JJ. Simultaneous determination of tuning and calibration parameters for computer experiments. *Technometrics* 2009; **51**(4):464–474.
10. Box GEP, Wilson KB. On the experiment attainment of optimum conditions. *Journal of the Royal Statistical Society. Series B* (*Statistical Methodology*) 1951; **XIII**(1):1–45.
11. Box GEP, Draper NR. *Response Surfaces, Mixtures, and Ridge Analyses*. Wiley: New Jersey, 2007.
12. Myers RH, Montgomery DC, Anderson-Cook CM. *Response Surface Methodology*: *Process and Product Optimization Using Designed Experiments*. Wiley: New York, 2009.
13. Barton RR. Simulation metamodels. *Proceedings of the 1998 Winter Simulation Conference*, Washington, DC, 167–174.
14. Simpson TW, Peplinski JD, Koch PN, Allen JK. Metamodels for computer-based engineering design: survey and recommendations. *Engineering with Computers* 2001; **17**:129–150.
15. Chen VCP, Tsui KL, Barton RR, Allen JK. *A Review of Design and Modeling in Computer Experiments*. Elsevier Science: Amsterdam, 2003.
16. Zhao J. Simulation and validation of hydrogen sulfide removal from fan ventilated confined-space manure storages. *Dissertation*, The Pennsylvania State University, 2006.
17. Khuri AI, Cornell JA. *Response Surfaces*: *Designs and Analyses*. Marcel Dekker: New York, 1996.
18. Lin DKJ. Discussion on papers by Box and Liu, Box, and Myers. *Journal of Quality Technology* 1999; **31**:61–66.

Author's biography

Dr Dennis K. J. Lin is a University Distinguished Professor of Statistics and Supply Chain Management at Penn State University. His research interests are quality engineering, industrial statistics, data mining and response surface. He has published over 150 papers in various journals. Dr Lin is an elected fellow of ASA and ASQ, an elected member of ISI, a lifetime member of ICSA, and a fellow of RSS.

He is an honorary chair professor for various universities, including a Chang-Jiang Scholar of China at Renmin University, National Chengchi University (Taiwan), Fudan University, and XiAn Statistical Institute (China).

Dr Lin presents several distinguished lectures, including the 2010 Youden Address (FTC) and the 2011 Loutit Address (SSC). He is also the recipient of the 2004 Faculty Scholar Medal Award at Penn State University.