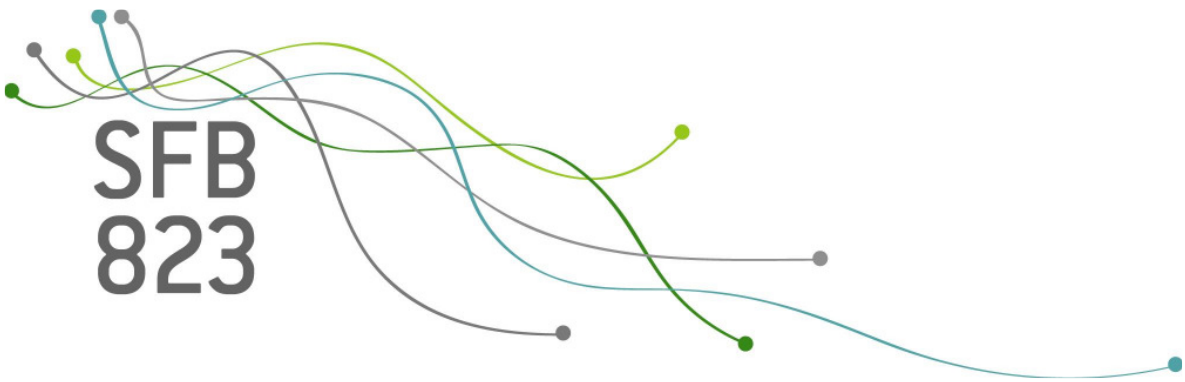
SFB 823

NPUA: A new approach for the analysis of computer experiments

Holger Dette, Andrey Pepelyshev

Nr. 1/2010

Discussion Paper



NPUA: A new approach for the analysis of computer experiments

Holger Dette
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum, Germany
e-mail: holger.dette@rub.de

Andrey Pepelyshev
Sheffield University
Department of Probability & Statistics
Sheffield, UK
email: a.pepelyshev@sheffield.ac.uk

11 Jan 2010

Abstract

The main issue in the analysis of computer experiments is an uncertainty of prediction and related inferences. To address the uncertainty analysis, the Bayesian analysis of deterministic computer models has been actively developed in the last decade. In the Bayesian approach, the uncertainty is expressed through a Gaussian process model. As a consequence, the resulting analysis is rather sensitive with respect to these prior assumptions. Moreover, for high dimensional data this approach leads to time consuming computations.

In the present paper we introduce a new approach for deriving the uncertainty in the analysis of computer experiments, where the distribution of uncertainty is obtained in a general nonparametric form. The proposed approach is called N(on) P(arametric) U(ncertainty) A(nalysis) and is based on a combination of sampling and regression techniques. In particular, it is computationally very simple. We compare NPUA with the Bayesian and Kriging method and investigate its performance for finding points for the next runs by re-analyzing the ASET model.

Keywords and Phrases: Computer experiment, uncertainty analysis, important sampling, regression, Jack-knife, sequential designs.

1 Introduction

In modern scientific studies, complex processes are described by mathematical computer models. These models serve as a replacement for natural (physical, chemical,

biological) experiments which are too time consuming or too costly. Computer models are used for modelling processes in engineering, fluid dynamics and thermodynamics, epidemiology, health and environmental sciences. In particular, mathematical models may describe phenomena which cannot be reproduced, for example, weather modelling and climate change.

A computer experiment consists of several runs of a computer program under different input conditions. Each run of the model is typically time consuming, since the computer program is based on a solution of a large system of sophisticated mathematical equations. As a result, the number of runs which are available for the analysis of the model is limited. One of the aims of the analysis is to construct a meta-model for predicting the output of the model at untried inputs. However, such a prediction is uncertain, and it is important to quantify this uncertainty. In the Bayesian approach (O'Hagan et al., 1999) the conception of a so-called emulator is introduced to describe the uncertainty. The emulator is a stochastic process that represents the unknown output of the computer model. The meta-model is defined by the mean of the emulator and the uncertainty is characterized by the variance of the emulator. The validity of the emulator is determined through diagnostics (Bastos and O'Hagan, 2009).

In general, the analysis of deterministic computer experiments is versatile. In addition to three basic objectives (prediction, uncertainty and diagnostics), there are more specific objectives such as calibration, data assimilation and sensitivity analysis (see Sacks et al., 1989, Kennedy and O'Hagan, 2001, Politis and Robertson, 2004, Oakley and O'Hagan, 2004) among many others). For the purpose of prediction, there are several techniques including adaptive spline interpolation (Friedman, 1991), Kriging (Cressie, 1993), Bayes linear approach (Goldstein and Wooff, 1995), Bayesian analysis (Kennedy and O'Hagan, 2001), neural networks (Smith, 1993), radial basis function approximation (Powell, 1987), and wavelet modeling (Mallet, 1998).

On the other hand – to the knowledge of the authors – the uncertainty analysis has been developed only in the Bayesian framework by employing stochastic processes of a premeditated class. The key feature of this approach is the handling of the dependence between outputs for different inputs as the correlation dependence. As a result, the meta-model does not provide easily accessible information on the shape of the model output since the meta-model is a posterior Gaussian process. Moreover, the computational complexity of the Bayesian approach is substantial for high dimensional data. The complexity is smaller for the Bayes linear approach (Goldstein and Wooff, 1995) which is less popular than the Bayesian approach. In addition, the diagnostics of the meta-model (see Bastos and O'Hagan, 2009) is based on fine rules as the verification of the normality of correlated differences between the outputs of the meta-model and the computer model at several points.

In the present paper, we propose a new approach for these three basic parts of the

analysis of computer experiments. Some intermediate steps of the suggested method are well-known standard statistical tools, but, to the best of our knowledge, have not been utilized in the proposed way for analyzing computer experiments. The idea of the approach is motivated by the important sampling method, the theory of regression experiments and cross-validation. The meta-model is defined as the sum of a mean term and a residual term. We treat the mean term as a major part of the model output similarly to the selection of an essential part in the important sampling method (Ripley, 1987). The regression theory is utilized to construct the mean term using the known basis functions with unknown parameters. We propose to construct the residual term, the unexplained behaviour in the model output, by exploiting a simple interpolation technique. The uncertainty is quantified using the Jack-knife principle. As a result, the distribution of uncertainty is defined through an empirical distribution. Finally, we propose to perform the diagnostics on the basis of the set of deleted residuals. Such a diagnostics is harmonized with the uncertainty analysis and purely based on the data used for the construction of the meta-model. Throughout this paper the proposed approach is called N(on) P(arametric) U(ncertainty) A(nalysis). Note that NPUA is rather simple, uses standard statistical tools from various statistical fields, and can easily be implemented.

The remaining part of the paper is organized as follows. In Section 2, NPUA is outlined and the basic algorithms of the Bayesian and Kriging method are reviewed for the sake of a clear comparison. In Section 3, the difference between the three methods are explained by an illustrative example. In Section 4, the performance of new methodology is demonstrated re-analyzing the ASET model developed by Cooper and Stroup (1985).

2 Outline of three approaches

Let y_i denote the output of a computer program for an input x_i, that is $y_i = \eta(x_i)$, $i = 1, \ldots, n$, where $\eta(x)$ is the computer model, which is called a simulator in the Bayesian approach. We assume that the model has a scalar output and d-dimensional inputs. In the following subsections, we describe three different ways of analyzing computer experiments of this type. In Section 2.1, a new method, NPUA, for the analysis of computer experiments is introduced. In Sections 2.2 and 2.3, we briefly describe the Bayesian and Kriging approaches which have been widely used in the literature (see Santner et al., 2003, Fang et al., 2006, O'Hagan and West, 2010).

2.1 The principle of NPUA

For the analysis of experiments of a deterministic computer model, we propose to build the meta-model in the form

$$\hat{\eta}(x) = \hat{\beta}^T f(x) + Z(x), \tag{1}$$

where the mean term $\beta^T f(x)$ is constructed by a stepwise regression technique, $\hat{\beta}$ denotes an appropriate estimate and the residual term $Z(x)$ is defined by an interpolation method on the basis of the residuals $z_i = y_i - \hat{\beta}^T f(x_i)$, $i = 1, \ldots, n$. We impose that the mean term $\beta^T f(x)$ has to reflect a major behaviour of the model $\eta(x)$ and, hence, the term $Z(x)$ remains to describe a residual behaviour. Although the term $\beta^T f(x)$ is called a mean term, indeed, this term is not the mean of some random variable, since there is no randomness in the proposed procedure.

The specific form of the meta-model (1) can be justified by the following arguments. From a practical point of view, the model $\eta(x)$ is unknown but it belongs to a given class of functions $\mathcal{F}$; $\eta \in \mathcal{F}$. Consequently, we can choose some basis functions $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_q$ such that the value

$$\min_{\beta} \operatorname{dist}\Big(\eta, \sum_{i=1}^{q} \beta_i \mathcal{V}_i\Big)$$

is rather small for some $q \in \mathbb{N}$, where $\operatorname{dist}(\cdot, \cdot)$ is an appropriate distance measure between the elements of the class $\mathcal{F}$. This concept is similar to a principle of the important sampling method, where a reduction of the variance is achieved by extracting a simple and essential part (Ripley, 1987). In our context, a linear combination of the basis functions represents a major behaviour of the model. Therefore, the uncertainty is smaller if the basis functions are more appropriate. This rule is a key tool for the qualitative reduction of the uncertainty.

To fix ideas, we propose to build the vector $f(x)$ in the model (1) using the stepwise regression procedure. Alternatively, any other more sophisticated method, like LASSO, could be used as well, see Hastie et al., (2009). The (forward) stepwise procedure means to include the new term $\mathcal{V}_r(x)$, that maximizes a coefficient of determination R^2 or minimizes some discrepancy measure, for example, the sum of squared residuals. The procedure is stopped at step q^* when, for example, the coefficient of determination R^2 is greater than 0.95. Note that q^* is usually small if the set of basis functions is appropriate. By our experience, we can say that usually $q^* < 15$. At the end of the stepwise procedure, we define

$$f(x) = (\mathcal{V}_1(x), \ldots, \mathcal{V}_{q^*}(x))^T,$$

and estimate $\beta \in \mathbb{R}^{q^*}$ in the model (1) by ordinary least squares , that is

$$\hat{\beta} = (F^T F)^{-1} F^T Y$$

where $Y = (y_1, \ldots, y_n)^T$ denotes the vector of outputs at input conditions $x_1, \ldots, x_n$ and $F = (f(x_1), \ldots, f(x_n))^T$ is the design matrix in the constructed linear regression model. To finalize the determination of the meta-model, we construct the residual term $Z(x)$ by a generalized inverse distance weighted interpolation of residuals between the output values and the estimated mean term at the corresponding input values

$$z_i = y_i - \hat{\beta}^T f(x_i),$$

$i = 1, \ldots, n$. Any other interpolation method could be used alternatively, see Cressie (1993, Sect. 5.9), Fang et al., (2006, Ch. 5), Lu and Wong (2008). To be precise, define

$$Z(x) = \frac{\sum_{i=1}^{n} z_i \, \dfrac{\kappa(x - x_i)}{||x - x_i||_2^p}}{\sum_{i=1}^{n} \dfrac{\kappa(x - x_i)}{||x - x_i||_2^p}}, \tag{2}$$

where $||x||_2 = \left(\sum_{s=1}^{d} x_s^2 \right)^{1/2}$, $\kappa(\cdot)$ is a positive symmetric unimodal function with $\kappa(0) = 1$ and p is a positive number. The parameter p and the function $\kappa(\cdot)$ can be varied and correspond to different forms of interpolation of the residuals $z_1, \ldots, z_n$. In geostatistics, the case $p = 2$ and $\kappa(\cdot) \equiv 1$ is known as the ordinary inverse distance weighted interpolation (Cressie, 1993, Sect. 5.9.2). Finally, the meta-model $\hat{\eta}(x)$ for the data set $(x_i, y_i)_{i=1}^n$ is a sum of $\hat{\beta}^T f(x)$ and $Z(x)$, which interpolates the values of the given dataset, that is $\hat{\eta}(x_i) = y_i$, $i = 1, \ldots, n$. Note that similar procedures for constructing the mean term have been used in Friedman and Stuetzle (1981), Koehler and Owen (1996), Fang and Lin (2003, p. 157).

To derive a distribution of uncertainty, we propose to use the Jack-knife technique (Efron and Tibshirani, 1993). To be precise, we construct meta-models $\hat{\eta}_1(x), \ldots, \hat{\eta}_n(x)$, where the meta-model $\hat{\eta}_j(x)$ based on the data excluding the jth point and the same vector $f(x)$, function $\kappa(\cdot)$ and parameter p. Then, the sample $(\hat{\eta}_1(x), \ldots, \hat{\eta}_n(x))$ yields an empirical distribution of uncertainty of the model output $\eta(x)$ for any x. This empirical distribution can be used for making probabilistic judgments for the output of the model $\eta(x)$, provided that the sample size n is sufficiently large. In general, one may derive a continuous distribution from the empirical distribution, however, this is not necessary. We believe that the consideration of the empirical distribution itself is enough for the analysis of the uncertainty. We note that the empirical distribution is not symmetric and the values $\max_i \hat{\eta}_i(x) - \hat{\eta}(x)$ and $\min_i \hat{\eta}_i(x) - \hat{\eta}(x)$ characterize the tails of the uncertainty distribution for any given x.

For the diagnostics of the constructed meta-model, we propose to compute the set

$$\mathcal{S} = \{\hat{\eta}(x_1) - \hat{\eta}_1(x_1), \ldots, \hat{\eta}(x_n) - \hat{\eta}_n(x_n)\} \tag{3}$$

which can be interpreted as the set of deleted residuals (Efron and Tibshirani, 1993). If the set $\mathcal{S}$ has an outlier or contains values which are larger than a given threshold, then the meta-model has to be considered as not accurate enough and additional runs of the model have to be performed. Note that, in contrast to Bastos and O'Hagan (2009), this diagnostics does not require additional runs of the computer model.

In the following paragraphs we present several details of the proposed approach.

2.1.1 Choice of the basis functions

The problem of choosing appropriate basis functions for the space $\mathcal{F}$ is of particular importance and can be solved in the following ways. One way is the identification of a set of candidate functions from the space $\mathcal{F}$ on the basis of scientific background and experts' knowledge. Another way is to draw scatterplots of the output versus each input. From these figures, one may guess reasonable functions describing the relation between the output and the input. The basis functions for high dimensional input can be constructed in the following manner.

Let $\{g_i(t)\}_i$ be a set of scalar functions of scalar variable t. Functions $g_i(t)$ may be monomials, exponentials, rational, trigonometric or wavelet-type functions, for example,

$$1, t, t^2, t^3, e^{-t}, te^{-5t}, t^2e^{-5t}, t/(0.05+t^2), t^2/(0.05+t^2), \cos(\pi t), \cos(2\pi t).$$

Similarly to An and Owen (2001), define the terms $\mathcal{V}_r(x)$ as products of the form

$$\mathcal{V}_r(x) = \prod_{s=1}^{d} g_{j_{r,s}}(x_s)$$

for some $j_{r,1}, \ldots, j_{r,d}$, $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. In general, the basis functions may be discontinuous in order to analyze a computer model with discontinuities in the output. One may also employ nonlinear regression models, but a higher complexity of parameter estimation should be taken into account in this case.

2.1.2 Choice of the meta-model via diagnostics

On the basis of the set $\mathcal{S}$ defined in (3), we introduce two measures of goodness-of-fit, that is

$$\begin{aligned} D_2 &= \left(\frac{1}{n} \sum_{i=1}^{n} \left(\hat{\eta}(x_i) - \hat{\eta}_i(x_i) \right)^2 \right)^{1/2}, \\ D_\infty &= \max_{i=1,\ldots,n} \left| \hat{\eta}(x_i) - \hat{\eta}_i(x_i) \right|. \end{aligned}$$

Similarly to Stone (1974), we can use these values for the choice of the vector $f(x)$ among several variants in the following manner. We consider a meta-model as more appropriate if the coefficient of determination R^2 is large and the value of the quantity

$$\alpha D_2 + (1-\alpha) D_\infty$$

is small, where α is a pre-specified constant. Consequently, we have the following procedure. We build several meta-models using different sets of basis functions and we choose a meta-model for which the value $\alpha D_2 + (1-\alpha) D_\infty$ is minimal provided that $R^2 > 0.95$. In the same manner, we can choose the parameter p and $\kappa(\cdot)$ and in the interpolation method for the construction of the residual process $Z(x)$.

2.1.3 Analysis of stochastic models

The proposed approach admits a generalization for the analysis of stochastic models (Kleijnen, 2005). Assume that the output of the model is disturbed by a random variable, that is

$$y_i = \eta(x_i) + \varepsilon_i.$$

A typical example of a "random" output occurs in the case where the mathematical model contains a stochastic differential equation, or the model is a simulation model.

Assume that the disturbances ε_i are independent identically distributed random values with zero mean and finite variance. Then the meta-model can be taken in the form (1) with

$$Z(x) = \frac{\sum\limits_{i=1}^{n} s(z_i, \tau)\kappa(x - x_i)||x - x_i||_2^{-p}}{\sum\limits_{i=1}^{n} \kappa(x - x_i)||x - x_i||_2^{-p}},$$

where $s(z, \tau)$ is given by

$$s(t, \tau) = \begin{cases} t + \tau & t < -\tau, \\ 0 & |t| \leq \tau, \\ t - \tau & t > \tau. \end{cases}$$

The value of the shrinkage parameter τ is determined from experts' knowledge. In general, the estimation of τ from the data is a hard problem, which requires observations at several points closely located to each other. Note that the threshold for R^2 in the stepwise procedure should be reduced as the variance of disturbances ε_i increases. For the case of heteroscedastic models, we should use $Z(x)$ with $s(z_i, \tau(x_i))$, where $\tau(x)$ is a value of the shrinkage parameter at the point x.

2.1.4 Sequential experiments

The amount of information extracted from the limited number of runs of the model increases by applying the sequential methodology, see e.g. (Santner et al. 2000, 2003, Sect. 6.3). It is clear that since each run of the program is time consuming, there is enough time for computing a 'good' point for the next run of the computer model, rather than to use an n-point Latin hypercube design, where n is the total number of runs.

We propose to start the investigation of the model with k runs according to a k-point space-filling design, where $n/3 \leq k \leq n/2$, and to compute the remaining design points sequentially. It is natural that a new point for the next run of the model should be a point at which the uncertainty of the meta-model output is maximal. Assume that the model output is obtained at the input conditions $x_1, \ldots, x_m$. Then we determine the point for the next run by

$$x^* = \text{argmax} \left\{ \max_{j=1,\ldots,m} \psi(x) \big| \hat{\eta}(x) - \hat{\eta}_j(x) \big| \;\Big|\; x \notin \bigcup_{i=1}^{m} S_i \right\}, \tag{4}$$

where

$$S_i = \left\{ x \,:\, ||x - x_i||_2 < \frac{1}{2} \min_{j \neq i} ||x_j - x_i||_2 \right\}$$

denotes a neighbourhood of the point x_i and $\psi(x)$ is a prior preference function such that $0 \leq \psi(x) \leq 1$. The maximization over the set $\bigcup_{i=1}^{m} S_i$ guarantees that the new point x^* is not close to the inputs $x_1, \ldots, x_m$. The function ψ is a weight function, reflecting the interest at different regions in the design space. If a specific subdomain Ω of the design space is of particular importance, one can put a larger weight at points of Ω. If only points in Ω are of equally interest, one can define $\psi(x) = 1_\Omega(x)$.

The performance of the proposed sequential design methodology will be demonstrated in Section 4.

2.2 Bayesian approach

In this subsection, we briefly describe the basic algorithm of the Bayesian approach for the analysis of computer experiments (Kennedy and O'Hagan, 2001). This approach is focusing on the Bayesian analysis of the emulator, which is a Gaussian process with mean $m_e(x) = \beta^T h(x)$ and covariance function $V_e(x, \tilde{x}) = \sigma^2 r(x, \tilde{x}|\psi)$, where $x \in \mathbb{R}^d$ is the d-dimensional input of the model, $h(x)$ is the vector of known functions, $\beta \in \mathbb{R}^q$ is the vector of unknown parameters, the output is one-dimensional, $r(x, \tilde{x}) = r(x, \tilde{x}|\psi)$ is the known correlation function and σ and ψ are unknown parameters. It is often assumed that $h(x)$ and $r(x, \tilde{x})$ are of the form $h(x) = (1, x^T)^T$ and

$$r(x, \tilde{x}|\psi) = \exp\Big(- \sum\nolimits_{s=1}^{d} (x_s - \tilde{x}_s)^\delta / \psi_s \Big) \tag{5}$$

with $\delta = 2$. The meta-model is the mean of the posterior emulator which is the conditional (on data) Gaussian process with mean $m_p(x)$ and covariance function $V_p(x, \tilde{x})$.

For the computation of the posterior emulator according to the Bayesian analysis, one has to specify a prior distribution for the parameters β, σ^2, ψ and integrate out to obtain the estimates. It is often assumed that the prior densities for the parameters are given by non-informative priors, for example, $p_a(\beta, \sigma^2) \propto \sigma^{-2}$, $p_a(\psi) \propto 1$. Note that these priors considerably simplify the calculations in the Bayesian approach.

The basic algorithm for computing Gaussian posterior emulator is described in the following paragraph. For details and modifications of this algorithm, we refer to the work of Kennedy and O'Hagan (2001), Bayarri et al. (2007), Rougier (2008), Bastos and O'Hagan (2009) and Liu and West (2009) among others.

Firstly, we define the posterior density for the correlation length parameters ψ

$$p_p(\psi|y) \propto |R|^{-1/2}|H^T R^{-1} H|^{-1/2} \left(\hat{\sigma}^2\right)^{-(n-q)/2}, \tag{6}$$

where

$$\hat{\sigma}^2 = \frac{1}{n-q-2} y^T \left(R^{-1} - R^{-1}H(H^T R^{-1} H)^{-1} H^T R^{-1}\right) y, \tag{7}$$

$H = (h(x_1), \ldots, h(x_n))^T \in \mathbb{R}^{n \times q}$ is the design matrix and $R = (R(x_i, x_j|\psi))_{i,j=1}^n$ is the covariance matrix. Next, according to the plug-in method (Kennedy and O'Hagan, 2001), calculate the estimate $\hat{\psi}$ of the correlation parameter by maximizing the posterior density defined in (6) and compute the matrix $\hat{R} = (R(x_i, x_j|\hat{\psi}))_{i,j=1}^n$. After that, calculate the Bayesian estimate of β by

$$\hat{\beta} = (H^T \hat{R}^{-1} H)^{-1} H^T \hat{R}^{-1} y. \tag{8}$$

Next, we compute the Bayesian estimate of σ^2 by (7) with the replacement R by $\hat{R}$. Finally, we define the mean of the posterior emulator by

$$m_p(x) = \hat{\beta}^T h(x) + t^T(x) \hat{R}^{-1} (y - H\hat{\beta}), \tag{9}$$

and the covariance function of the posterior emulator by

$$\hat{V}_p(x, \tilde{x}) = \hat{\sigma}^2 \left(r(x, \tilde{x}|\hat{\psi}) - t^T(x) \hat{R}^{-1} t(\tilde{x}) + s^T(x)(H^T \hat{R}^{-1} H)^{-1} s(\tilde{x})\right), \tag{10}$$

where $t(x) = (r(x, x_1|\hat{\psi}), \ldots, r(x, x_n|\hat{\psi}))^T$ and $s(x) = h(x) - H^T \hat{R}^{-1} t(x)$.

For a graphical representation of the posterior emulator, the mean and the 95%-confidence interval

$$(m_p(x) - \gamma\sqrt{\hat{V}_p(x, x)}, m_p(x) + \gamma\sqrt{\hat{V}_p(x, x)}) \tag{11}$$

are drawn, where γ is a 0.975-quantile of the Student distribution with $(n - q)$ degrees of freedom.

The diagnostics of the Gaussian process emulators has recently been developed by Bastos and O'Hagan (2009). New runs of the model at points of a validation set are required for this method, which is based on the verification of the normality of the residuals between the output of the meta-model at the points of the validation set and the true output of the model.

2.3 Kriging approach

In this subsection, we outline the approach based on the Kriging technique, which has found considerable attention in the field of spatial statistics (Cressie, 1993). This method is based on the model $\eta(x) = m(x) + Z(x)$, where $Z(x)$ is a stationary process with zero mean and known covariance function. Let the mean $m(x) = \beta^T f(x)$ be constructed by the stepwise regression technique (in geostatistics, the mean term $m(x)$ is constant) and $Z(x)$ is an isotropic Gaussian process with Gaussian covariance function

$$V(x, \tilde{x}|\psi_0) = \sigma^2 \exp\Big(-\sum\nolimits_{s=1}^{d} (x_s - \tilde{x}_s)^2 \Big/ \psi_0 \Big)$$

and unknown parameter $\psi_0 > 0$. Let the parameter β be estimated by the ordinary least squares, $\hat{\beta} = (F^T F)^{-1} F^T Y$, where $Y = (y_1, \ldots, y_n)^T$ and $F = (f(x_1), \ldots, f(x_n))^T$, and the correlation parameter ψ_0 is estimated by the variogram method for the residuals $Z_i = y_i - \hat{\beta}^T f(x_i)$ [see formula (2.4.12) and (2.6.12) in Cressie (1993)].[1] Finally, the mean of the posterior emulator is given by

$$m_p(x) = \hat{\beta}^T f(x) + t^T(x) \hat{V}^{-1} (y - F\hat{\beta})$$

where $\hat{V} = (V(x_i, x_j|\hat{\psi}))_{i,j=1}^n$ is the covariance matrix and

$$t(x) = (V(x, x_1|\hat{\psi}_0), \ldots, V(x, x_n|\hat{\psi}_0))^T.$$

The covariance function of the posterior emulator is given by

$$\hat{V}_p(x, \tilde{x}) = V(x, \tilde{x}|\hat{\psi}_0) - t^T(x) \hat{V}^{-1} t(\tilde{x}).$$

Finally, the uncertainty is defined in the same way as in Section 2.2.

[1] To be precise, the estimate of the variogram is given by

$$2\hat{\gamma}(h) = \frac{\Big(\frac{1}{|N(h)|} \sum_{(i,j) \in N(h)} |Z_i - Z_j|^{1/2} \Big)^4}{0.457 + 0.494/|N(h)|}$$

where $|N(h)|$ means the number of elements in the set $N(h) = \{(i,j) : ||x_i - x_j|| = h,\ i,j = 1, \ldots, n\}$, and the parameter ψ_0 is estimated by minimizing the sum

$$\sum_{j=1}^{K} |N(h_j)| \left(\frac{\hat{\gamma}(h_j)}{\gamma(h_j, \psi_0)} - 1 \right)^2,$$

where $h_1, \ldots, h_K$ are the distinct elements of the set $\{||x_i - x_j||\}_{i,j}$ and $\gamma(h, \psi_0) = \sigma^2(1 - \exp(-h^2/\psi_0))$.

3 Comparison of three approaches

In this section we present the conceptual differences and perform a numerical comparison of the three approaches.

3.1 Some general remarks

In NPUA there is no randomness and no correlation in the structure of the meta-model. In contrast, in the Bayesian and Kriging method, the dependence between outputs for different inputs is modeled by a stationary Gaussian process (specifying its correlation structure) and the meta-model is assumed to be the mean of a stochastic process. As a result, the structure of uncertainty is the consequence of these specific assumptions. In the Bayesian and Kriging approaches, the uncertainty for any input has a Student distribution and is as a consequence symmetric (Bastos and O'Hagan, 2009). In NPUA there is no specific form of the distribution for the uncertainty, because – similar to the bootstrap method – the distribution of the uncertainty is produced from the data.

Note that in the proposed and Kriging approaches, the mean term is a nonlinear function constructed in the same manner, but the the residual term $Z(x)$ is determined in different ways. On the other hand, in the Bayesian approach, the mean term of the Gaussian process is typically constant or a linear function (Bastos and O'Hagan, 2009). In the Kriging approach, the covariance function is isotropic. In particular, the relation between the correlation parameters in the Bayesian and Kriging approaches is $\psi = (\psi_0, \ldots, \psi_0)$.

Finally, we discuss the computational complexity of the three methods. The proposed and Kriging approaches have a similar complexity. They consist of k inversions of matrices of sizes from 1 to q^* where $q^* \approx 15$ and k depends on the number of basis functions. After that, in the proposed approach, n inversions of matrices of size q^* are performed to compute the distribution of uncertainty. In the Kriging approach, the empirical variogram is calculated and the variogram estimate of ψ_0 is computed through the minimization of a functional of one variable, and - in the final step - the inversion of the correlation matrix of size n is performed. In the Bayesian approach, the maximum likelihood estimate of the parameter ψ is computed through the maximization of a functional of $d + 1$ variables. The finding of the maximum requires numerous (about $100(d+1)$) inversions of the correlation matrices of size n. Moreover, the correlation matrices of size n may be ill-conditioned (see Neal, 1997).

3.2 Two illustrative examples

For clarity, we investigate two examples in the one-dimensional case, where it is easy to visualize the meta-model and the corresponding uncertainty (other examples are

available from the authors).

Let us consider the first dataset

x_i	0	0.25	0.5	0.75	1
y_i	0.28	0.37	0.5	0.19	0.07

(12)

and the second dataset

x_i	0	0.25	0.5	0.75	1
y_i	0	0.03	0.30	0.50	0.59

. (13)

Define $h(x) = (1, x)^T$ in the Bayesian approach and $f(x) = (1, x, x^2)^T$ in the Kriging method and the method proposed in this paper. Taking $p = 2$ and $\kappa(\cdot) \equiv 1$ in (2) in the proposed approach, we obtain that $R^2 = 0.9$, $D_\infty = 0.27$, $D_2 = 0.16$ for the dataset (12) and $R^2 = 0.976$, $D_\infty = 0.34$, $D_2 = 0.20$ for the dataset (13).

The meta-models and their corresponding uncertainty are depicted in Figure 1. Note that in the Bayesian and Kriging approaches the uncertainty is specified by 95% confidence intervals, which results in two dashed lines on the top and the middle panel of Figure 1, respectively. On the other hand, the uncertainty in the proposed approach is specified by n curves $\hat{\eta}_1(x), \ldots, \hat{\eta}_n(x)$ obtained by the Jack-knife technique. The true model is not given in order to stress the uncertainty of the prediction problem.

The conceptual differences of the three approaches can be observed in Figure 1. In the Bayesian approach, the uncertainty of the meta-model for the first dataset grows as a point goes outside of the interval $[0, 1]$. In the Kriging approach, the uncertainty of the meta-model is proportional to the variance $\hat{\sigma}$ for points which lie outside the neighborhood of the training data. In the Bayesian and Kriging approaches, the uncertainty has a symmetric quasi-periodic shape. In contrast, in NPUA, the uncertainty at some point strongly depends on the joint placement of the neighbourhood training points.

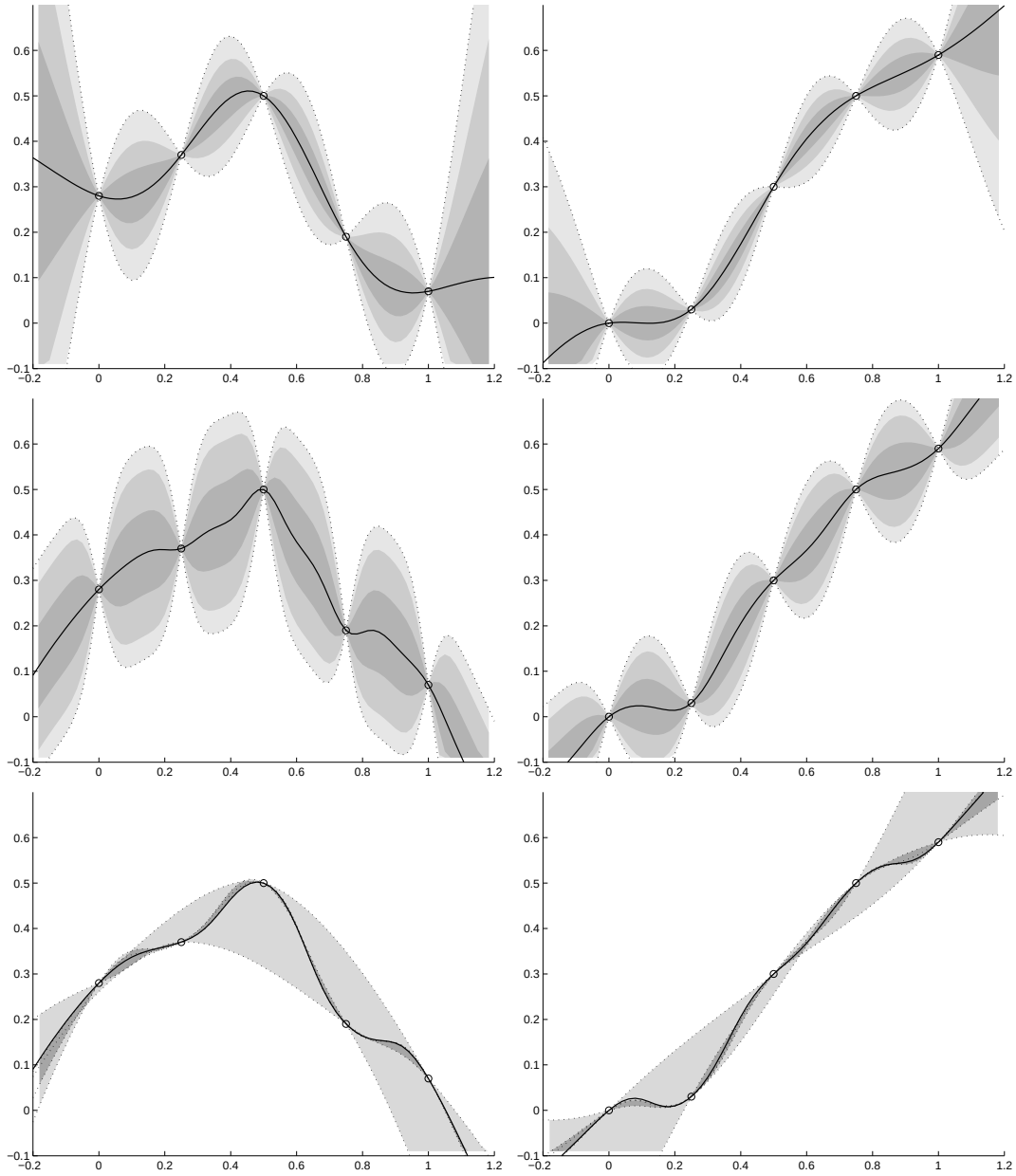


Figure 1: Meta-models (solid lines) and their uncertainty (dashed lines) for the dataset (12) (left column) and dataset (13) (right column) for the Bayesian approach (upper panel), the Kriging approach (middle panel) and the proposed approach (bottom panel).

4 Evolution of fires in enclosed areas

In this section we use NPUA for the analysis of the ASET model developed by Cooper and Stroup (1985). In our study, we used the ASET-B program implemented in BASIC by Walton (1985). In particular, this model has also been studied in Santner et al. (2003). This program describes the fire environment in a single room with closed windows and doors with a small leak at floor level. This leak prevents the pressure from increasing in the room. A fire starts at some point below the ceiling and releases energy and products of combustion. The hot products of combustion form a plume which rises towards the ceiling. When the plume reaches the ceiling it spreads out and forms a hot gas layer. There is a relatively sharp interface between the hot upper layer and the air in the lower part of a room. The program predicts the thickness and the temperature of the hot smoke layer as a function of time by solving a system of differential equations. The program has four inputs: the heat loss fraction for the room ($L \in [0.7, 0.9]$) the height of the fire source above the floor ($F \in [0.1, 4]$), the room height ($H \in [6, 12]$), the room floor area ($A \in [100, 250]$). The model output is the time it takes for the low bound of the hot smoke layer to reach five feet above the floor.

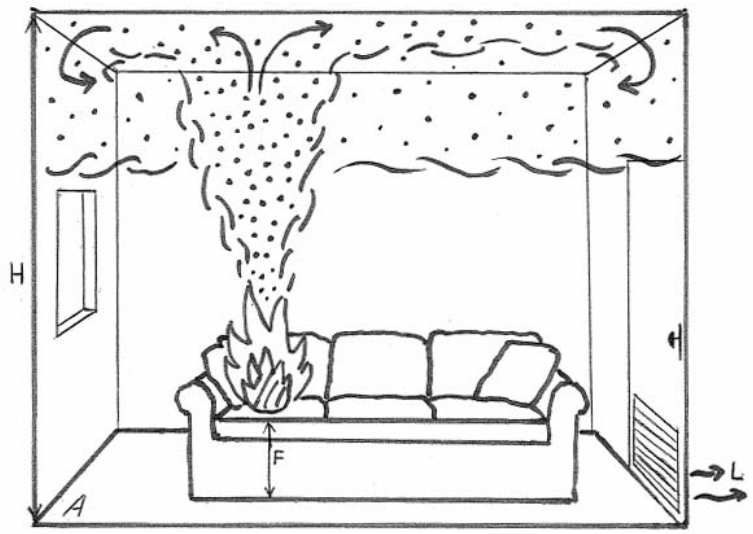


Figure 2: The room with the fire source. The thickness of the hot smoke layer increases from zero to the room height.

After re-parameterization, we consider the model in the following form

$$\eta(x) = \mathrm{ASET}\big(L = 0.7 + 0.2x_1, F = 0.1 + 3.9x_2, H = 6 + 6x_3, A = 100 + 150x_4\big)$$

defined on the cube $[0, 1]^4$. We run the model on the basis of a 30-point maximin Latin hypercube design (see e.g. Santner et al. 2003). Following Subsection 2.1.1, we define the set of scalar functions in the form $\{1, t, t^{2/3}, t^{3/2}\}$. Applying the stepwise regression

technique, we obtain the mean term of the meta-model in the form

$$m(x) = 18.7 + 42.1x_2^{2/3}x_3^{3/2}x_4^{3/2} + 15.8x_1^{3/2}x_2^{2/3} + 38.6x_3^{3/2}x_4^{3/2} + 25.5x_2^{2/3}$$

with the coefficient of determination $R^2 = 0.989$. Performing the diagnostics, we calculate $D_\infty = 8.21$ and $D_2 = 3.58$.

Let us pretend that the meta-model fails the diagnostics. Consequently, we would perform several new runs of the computer model and rebuild the meta-model. According to the 20 iterations of the algorithm proposed in Subsection 2.1.4, we obtain 20 new input conditions for which the computer model is evaluated. The total 50-point design is depicted in Figure 3. We observe that the additional design points are mostly located near the boundary of the design space.

Using 50 runs of the model, we obtain the mean term of the meta-model in the form

$$m(x) = 20.5 + 44.0x_2^{2/3}x_3^{3/2}x_4^{3/2} + 15.3x_1^{3/2}x_2^{2/3} + 35.7x_3^{3/2}x_4^{3/2} + 23.8x_2^{2/3}.$$

The diagnostics of the meta-model yields $D_\infty = 4.81$ and $D_2 = 2.49$.

Let us now compare the characteristics of the meta-model obtained by this sequential strategy with the characteristics of meta-models obtained by three alternative 50-point designs. In particular we consider a design ξ_u, where we add 20 uniformly chosen random points to the 30-point maximin LHD, and a 50-point maximin LHD. A further candidate for the comparison is the design ξ_d, where we add 20 additional points maximizing the minimal distance to the points of the 30-point LHD and between each other. The results of the comparison are summarized in Table 1.

Table 1: The diagnostic of the meta-models for different designs.

	type of design	D_∞	D_2
ξ_s	30-point maximin LHD and 20 seq. optimal points	4.8	2.5
ξ_u	30-point maximin LHD and 20 uniform random points	7.8	2.9
ξ_m	50-point maximin LHD	9.1	3.5
ξ_d	30-point maximin LHD and 20 maximin distance points	9.7	3.7

We observe that the 20 points, which are constructed by the sequential algorithm proposed in Subsection 2.1.4, significantly improve the goodness-of-fit of the meta-model in comparison with the other strategies. Note that a design ξ_d, in which 20 points are chosen to maximize the minimal distance to the previous points, yields a meta-model with slightly larger values for the D_2- and D_∞-criteria than the meta-model obtained for the 50-point maximin LHD ξ_m. This observation can be explained by a

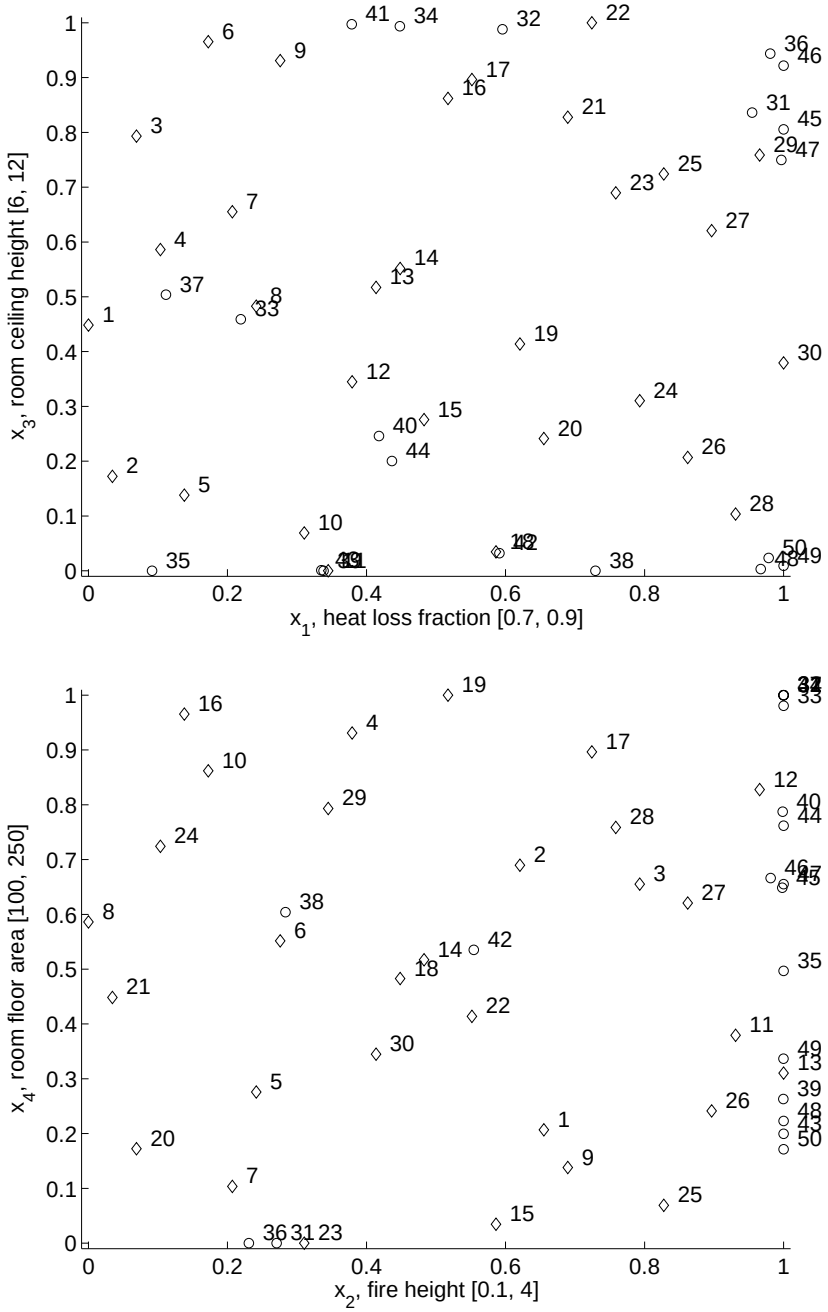


Figure 3: The projections of the 50-point design obtained as the union of a 30-point maximin Latin hypercube design with 20 sequentially chosen points. The points of 30-point design are numbered 1-30 and are marked by a diamond. The 20 points are numbered 31-50 and are marked by a ball. The upper part is the projection of design points onto the 1st and 3rd coordinates. The right part is the projection of the design points onto the 2nd and 4th coordinates.

less structured location of points. On the other hand, the meta-model for the design ξ_u is better than the meta-model for the design ξ_d since some points of the design ξ_u are close to each other [this artificially decreases the values of the criteria]. Summarizing these observations, we conclude that the proposed sequential design is the best among the designs considered in the study.

Sensitivity analysis can be performed by studying the expression of the mean term. This expression indicates the presence of interactions between the different input variables. Also, we see that the loss heat fraction L has the smallest effect on the output. The height F of the fire source above the floor and the room height H have a medium effect. The room floor area A has the largest effect on the output. These observations are consistent with results obtained in Santner et al. (2003, Section 7.1).

Finally, we perform the brute-force validation of the meta-model (corresponding to the sequential 50-point design) by running the model for 300 uniformly distributed random points. The errors of prediction are displayed in Figure 4 with respect to the distance of the points to the boundary of the design space $[0, 1]^4$. We observe that the absolute values of the errors are mostly smaller than D_2. It is remarkable that the errors are larger for the points located near the boundary of design space. This observation explains why the algorithm proposed in Subsection 2.4.1 yields mostly runs of the computer experiment in a neighbourhood of the boundary.

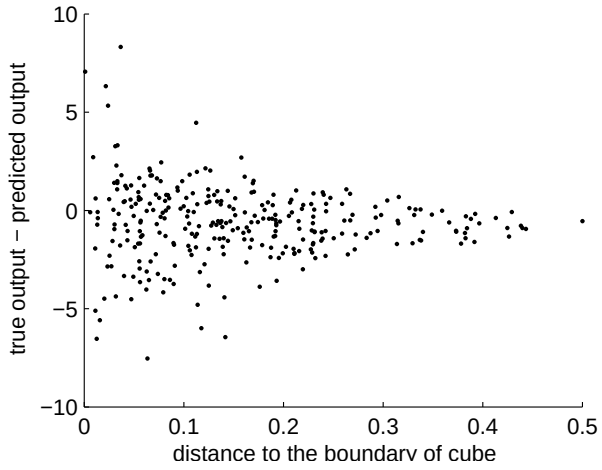


Figure 4: Errors of prediction with respect to the distance of a point to the boundary of design space for 300 uniform random points.

5 Conclusion

In the present paper, we have proposed a nonparametric approach for deriving the uncertainty in the analysis of computer experiments (NPUA), which is based on the

combination of several powerful statistical techniques. For this reason, we believe that it is of particular interest for practitioners. The conceptual idea of the proposed approach is that the structure of the mean term plays a primary role in the meta-model. This structure should be guessed as the extraction of an essential part in the important sampling method. The parameters of the mean term are determined by the stepwise regression technique, but other methods for variable selection could be used as well. The distribution of uncertainty is derived by employing the Jack-knife technique. As a result, we obtain a "nonparametric" uncertainty analysis. We recommend to run the computer experiment with a part of possible runs and then define the remaining inputs sequentially, such that the uncertainty is large and such that the new inputs are not too close to the points, which have already been used in the experiment. The differences between NPUA and the Bayesian and the Kriging approach are discussed in Section 3, where we demonstrate that NPUA yields an uncertainty which depends on the joint placement of the neighbourhood points.

The performance of NPUA is also illustrated by the re-analysis of the ASET model. In particular we have demonstrated that the rule for finding the points for the next runs of the computer model is efficient. Also, we have shown that the diagnostics of the meta-model provides reliable information about the accuracy of the meta-model. Moreover, the NPUA is not computationally demanding and is particularly suitable for large data sets, while the Bayesian and Kriging approaches may be infeasible in such cases.

Acknowledgments. This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823) of the German Research Foundation (DFG). Andrey Pepelyshev acknowledges the financial support provided by the MUCM project (EPSRC grant EP/D048893/1, http://mucm.group.shef.ac.uk). We would like to thank Anatoly Zhigljavsky, Anthony O'Hagan, Jeremy Oakley and Jonathan Cumming for their very perceptive comments and Martina Stein for typing parts of this manuscript with considerable technical expertise.

References

An J., Owen A. (2001) Quasi-regression. J. Complexity 17, no. 4, 588–607.

Bastos L., O'Hagan A. (2009) Diagnostics for gaussian process emulators. Technometrics 51, 425–438.

Bayarri M.J., Berger J.O., Cafeo J., Garcia-Donato G., Liu F., Palomo J., Parthasarathy R.J., Paulo R., Sacks J., Walsh D. (2007) Computer model validation with functional output. Ann. Statist. 35, no. 5, 1874–1906.

Cressie N.A.C. (1993) Statistics for Spatial Data, John Wiley & Sons, New York.

Cooper L.Y., Stroup D.W. (1985) Calculating Safe Egress Time (ASET) - A Computer Program and User's Guide. Fire Safety Jour. 9, 29–45.

Efron B., Tibshirani R.J. (1993) An Introduction to the Bootstrap. New York: Chapman & Hall.

Fang K.-T., Lin D. K. J. (2003) Uniform experimental designs and their applications in industry. Statistics in Industry, Handbook of Statist., 22, North-Holland, Amsterdam, 131–170.

Fang K.-T. Li R., Sudjianto A. (2006) Design and Modeling for Computer Experiments. Chapman & Hall/CRC.

Friedman J. H., (1991) Multivariate Adaptive Regression Splines. The Annals of Statistics 19, No. 1, 1–67.

Friedman J.H., Stuetzle W. (1981) Projection Pursuit Regression. J. of the Amer. Stat. Assoc., Vol. 76, No. 376, 817–823.

Goldstein M, Wooff DA (1995) Bayes linear computation: concepts, implementation and programs. Statistics and Computing 5, 327–341.

Hastie T., Tibshirani R., Friedman J. (2009) The elements of statistical learning. Data mining, inference, and prediction. Springer Series in Statistics. Springer-Verlag, New York.

Kennedy M.C., O'Hagan A. (2001) Bayesian calibration of computer models. J. R. Stat. Soc. Ser. B 63, no. 3, 425–464.

Kleijnen J.P.C. (2005) An overview of the design and analysis of simulation experiments for sensitivity analysis. European J. Oper. Res. 164, no. 3, 287–300.

Koehler J.R., Owen A.B. (1996) Computer Experiments. In Handbook of Statistics, 261–308.

Liu F., West M. (2009) A Dynamic Modelling Strategy for Bayesian Computer Model Emulation. Bayesian Analysis 4, 393–412.

Lu, G.Y., Wong, D.W. (2008) An adaptive inverse-distance weighting spatial interpolation technique. Computers & Geosciences 34, 1044–1055.

Mallet C.G. (1998) A Wavelet Tour of Signal Processing, Academic Press, Boston, MA.

Neal R. (1997) Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Tech. Rep. CRGTR972, Dept. of Computer Science, University of Toronto.

Oakley J.E., O'Hagan A. (2004) Probabilistic sensitivity analysis of complex models: a Bayesian approach. J. R. Stat. Soc. Ser. B Stat. Methodol. 66, no. 3, 751–769.

O'Hagan A., West M. (2010) The Oxford Handbook of Applied Bayesian Analysis. Oxford University Press.

O'Hagan A., Kennedy. M. C. and Oakley, J. E. (1999). Uncertainty analysis and other inference tools for complex computer codes (with discussion). In Bayesian Statistics 6, J. M. Bernardo et al. (eds.). Oxford University Press, 503–524.

Powell M. J. D., (1987) Radial Basis Functions for Multivariable Interpolation: A Review. In Algorithms for Approximation (Mason, J. C. and Cox, M. G., eds.), Oxford University Press, London.

Politis K., Robertson L. (2004) Bayesian updating of atmospheric dispersion after a nuclear accident. J. Roy. Statist. Soc. Ser. C 53, no. 4, 583–600.

Ripley B.D. (1987) Stochastic Simulation. Wiley & Sons.

Rougier J. (2008) Efficient emulators for multivariate deterministic functions. Journal of Computational and Graphical Statistics 17, 827–843.

Sacks J., Welch W.J., Mitchell T.J., Wynn H.P. (1989) Design and analysis of computer experiments. Statist. Sci. 4, no. 4, 409–435.

Santner T.J., Williams B.J., Notz W. (2000) Sequential design of computer experiments to minimize integrated response functions. Stat. Sinica 10, 1133–1152.

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

Smith M. (1993) Neural Networks for Statistical Modeling. von Nostrand Reinhold, New York.

Stone M. (1974) Cross-validatory choice and assessment of statistical predictions. J. Roy. Statist. Soc. Ser. B 36, 111–147.

Walton W. (1985) ASET-B: A room fire program for personal computers. Fire Technology, Vol. 21, No. 4, 293–309