J Glob Optim
DOI 10.1007/s10898-011-9837-4

Setting targets for surrogate-based optimization

Nestor V. Queipo · Salvador Pintos · Efrain Nava

Received: 16 April 2011 / Accepted: 26 December 2011

Abstract In the context of surrogate-based optimization (SBO), most designers have still very little guidance on when to stop and how to use infill measures with target requirements (e.g., one-stage approach for goal seeking and optimization); the reason: optimum estimates independent of the surrogate and optimization strategy are seldom available. Hence, optimization cycles are typically stopped when resources run out (e.g., number of objective function evaluations/time) or convergence is perceived, and targets are empirically set which may affect the effectiveness and efficiency of the SBO approach. This work presents an approach for estimating the minimum (target) of the objective function using concepts from extreme order statistics which relies only on the training data (sample) outputs. It is assumed that the sample inputs are randomly distributed so the outputs can be considered a random variable, whose density function is bounded (a, b), with the minimum (a) as its lower bound. Specifically, an estimate of the minimum (a) is obtained by: (i) computing the bounds (using training data and the moment matching method) of a selected set of analytical density functions (catalog), and (ii) identifying the density function in the catalog with the best match to the sample outputs distribution and corresponding minimum estimate (a). The proposed approach makes no assumption about the nature of the objective functions, and can be used with any surrogate, and optimization strategy even with high dimensional problems. The effectiveness of the proposed approach was evaluated using a compact catalog of *Generalized Beta* density functions and well-known analytical optimization test functions, i.e., F2, Hartmann 6D, and Griewangk 10D and in the optimization of a field scale alkali-surfactant-polymer enhanced oil recovery process. The results revealed that: (a) the density function (from a catalog) with the best match to a function outputs distribution, was the same for both large and reduced samples, (b) the true optimum value was always within a 95% confidence interval of the estimated

N. V. Queipo (✉) · S. Pintos · E. Nava
Applied Computing Institute, University of Zulia, Maracaibo 4011, ZU, Venezuela
e-mail: nqueipo@ica.luz.edu.ve

S. Pintos
e-mail: spintos@ica.luz.edu.ve

E. Nava
e-mail: enava@ica.luz.edu.ve

Published online: 07 January 2012

minimum distribution, and (c) the estimated minimum represents a significant improvement over the present best solution and an excellent approximation of the true optimum value.

Keywords Minimum estimates from sample data · Extreme value theory · Setting targets · Surrogate-based optimization

1 Introduction

Assessing the merit of another cycle in surrogate-based optimization for engineering design versus accepting the present best solution [1] is an issue of considerable interest in the optimization of complex engineering systems in the aerospace [2–5], automotive ([6,7]) and oil industries ([8,9]); review papers on the subject of surrogate-based optimization are those of Li and Padula [4], Queipo et al. [5], Wang and Shan [10], and Forrester and Keane [11]. Each cycle consists of the analysis of a number of designs, the fitting of a surrogate, optimization based on the surrogate, and exact analysis at the design obtained by the optimization. The cycles are typically stopped when resources run out (e.g., number of objective function evaluations/time) or convergence is perceived, such as when the latest improvement represents a particular fraction of the range of the objective function evaluations. On the other hand, promising infill measures such as one-stage approach for goal-seeking and optimization [12,13] have limited their application because the targets (goals) are empirically set with significant uncertainty. Hence, considering optimum estimates are seldom available, most designers have still very little guidance on when to stop and how to set up the targets (goals).

Jones et al. [14] using the so called expected improvement (EI) as infill measure stopped the search when the maximum EI was less than 1% of the present best solution. Sasena et al. [15] compared alternative infill sampling plans using a generalized EI measure while stopping the cycles after a fixed number of objective function evaluations. Sobester et al. [16] used a weighted EI criterion and also limited the cycles to a fixed number of objective function evaluations. Huang et al. [17] presented a so called augmented EI to address stochastic black box systems and used as stopping criterion a tolerance for the ratio between the maximal EI and the active span of the responses. Alternatively, Apley et al. [18] in the context of robust design gave guidelines for additional cycles depending on whether or not the analytical prediction intervals for potential designs overlapped. Forrester and Jones [19] proposed an EI measure with no user defined parameters and stopped the cycles after a particular target is reached. These works consider the deployment of a single point in each additional cycle. In contrast, clustered approaches for the deployment of multiple points in additional cycles were conducted for probability of improvement (PI) as infill measure [12] and generalized EI [20]; the former did not specify a stopping criterion while the latter used a fixed number of objective function evaluations. Using a fixed number of cycles [21] gave results for both EI and PI as infill sampling criteria also allowing for multiple points in each additional cycle; two heuristics were used for the EI calculations. Note that, in general, the stopping criteria were not based on optimum estimates.

In the one-stage approach for goal-seeking the covariance parameters in Gaussian Process (GP) modeling are jointly estimated (maximum likelihood) with the location at which a particular goal (target) may be achieved. An extension of this idea for optimization essentially establishes as infill criterion the goal seeking approach for multiple targets. Note that, in contrast to traditional GP-based optimization, the covariance model estimation is coupled with the optimization process (one-stage) and can be more effective when the training data is sparse and misleading. A discussion of these approaches for GP and radial basis functions, can be found in [12] and [13], respectively; in all instances, targets where empirically set which may affect the effectiveness and efficiency of the surrogate-based optimization.

On the other hand, in surrogate-based optimization, assuming the initial design of experiment (inputs) is a random sample, the training data outputs can be considered a random variable. In that context, there are asymptotic distributions for the extremes (minimum/maximum) of a random variable (with unknown distribution) [22–24], but estimating their parameters require observations corresponding to the minimum/maximum of several samples, which are rarely available in surrogate-based optimization. In contrast, this work estimates the minimum (target) of an objective function using a single sample, i.e., training data outputs, assuming a particular density function. An estimate of the minimum (a density function bound) is then obtained through the moment matching method. The assumption imposed on the sample inputs is not restrictive since the initial design of experiment (DOE) for surrogate modeling and optimization aims to distribute the sample points uniformly in the design space (to reduce bias errors). The uniformity property in designs is sought by, for example, maximizing the minimum distances among design points [25], or by minimizing correlation measures among the sample data [26,27]. Practical implementation of these strategies includes Latin Hypercube sampling (LHS, e.g., [28]) and OA-based LHS [29,30] and other optimal LHS schemes [31,32]. Note that in the proposed approach: (i) the minimum of the objective function of interest is estimated at the beginning of the surrogate-based optimization process, and (ii) except for continuity, no assumption is made about the nature of the objective functions, and can be used with any surrogate, and optimization strategy even with high dimensional problems.

The remainder of the paper is structured as follows: problem statement (Sect. 2), solution approach (Sect. 3), case studies (Sect. 4), results and discussion (Sect. 5), and summary and conclusions (Sect. 6).

2 Problem definition

The problem of interest can be stated as: given a sample of model input/output pairs, estimate the minimum of the model output (objective function) which relies only on the training data (sample) outputs. It is assumed that the model output is a scalar, and the sample inputs are randomly distributed so the outputs can be considered a random variable.

3 Proposed approach

Given a sample of points, an estimate for the minimum of a function is obtained through the following three steps:

A. Generate a *catalog* with a variety of bounded (a, b) *analytical* density functions that may provide a good fit to training data (outputs) *empirical* density function,
B. For each of the density functions in the catalog, estimate its bounds (a, b) using the moment matching method, and,
C. Identify the density function from the above-referenced catalog with the best match to the sample outputs distribution; the lower bound (a) for the identified density function is the minimum estimate sought.

Table 1 Density and cumulative density functions for the *Beta* (p, q) distribution defined in the interval

Probability density function	$f_{Beta}(x\|p, q) = \frac{x^{p-1}(1-x)^{q-1}}{B(p,q)}$
Cumulative distribution function	$F_{Beta}(x\|p, q) = \frac{1}{B(p,q)} \int_0^x t^{p-1}(1-t)^{q-1}dt$
With $B(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1}dt$	

Details of each of these steps are given below.

3.1 Generate a catalog with a variety of bounded (a, b) analytical density functions that may provide a good fit to training data (outputs) empirical density function

Ideally, the catalog should provide a compact description of a whole range of density functions. This can be accomplished using a *Generalized Beta* (p, q, a, b) density function for a random variable X (objective function values) defined in the interval (a, b); where X is considered a *Generalized Beta* (p, q, a, b) if $Z = \left(\frac{X-a}{b-a}\right)$ is a *Beta* (p, q). More specifically, different shapes of the density function can be identified modifying the p and q shape parameters in a *Beta* (p, q) density function for random variable $Z = \left(\frac{X-a}{b-a}\right)$ defined in the interval (0,1); see Table 1. Note that since $Z = \left(\frac{X-a}{b-a}\right)$ is a linear transformation, the *Beta* (p, q, a, b) and *Beta* (p, q) density functions share the same shape, hence the latter can be used to select proper p, q parameters (without knowing the bounds a, b). Figure 1 shows a catalog with nine (9) different compactly specified density functions that are expected to match a variety of empirical density functions for the objective function values (X). Note that shape parameters (p, q) equal to [1,1] resemble a uniform distribution, simultaneously increasing the values of p, q modifies the shape to a Gaussian-like density function (i.e., [3, 3], [5, 5]), and a biased density function to the left or right can be obtained with $q > p$ (i.e., [1, 2.5], [1, 5], [2.5, 5]) and $p > q$ (i.e., [2.5, 1], [5, 1], [5, 2.5]), respectively.

3.2 For each of the density functions in the catalog, estimate its bounds (a, b) using the moment matching method

It includes solving a system of equations for the bounds (a, b) of each of the analytical density functions in the catalog (specified in step A). The system is obtained equating the expected value for the minimum and maximum (B.1), with the sample outputs minimum and maximum, respectively. Note that selecting *Generalized Beta* (p, q, a, b) density functions for the catalog makes the above-referenced system of equations linear for the bounds (a, b) with analytical solutions (B.2).

3.2.1 Expected value for the minimum and maximum

Given a random sample $X_1, \ldots, X_N$ of a random variable $\boldsymbol{X}$ with density and cumulative distributions $f(x)$ and $F(x)$, respectively, the distribution of the maximum, i.e., $\max(X) = max(X_1, \ldots, X_N)$ and minimum values, i.e., $\min(X) = min(X_1, \ldots, X_N)$ is expressed as $F_{\max}(x) = F(x)^N$ and $F_{\min}(x) = 1 - (1 - F(x))^N$, respectively (See [33]). The corresponding density functions are:

$$f_{\max}(x) = NF(x)^{N-1} f(x)$$
$$f_{\min}(x) = N(1 - F(x))^{N-1} f(x)$$

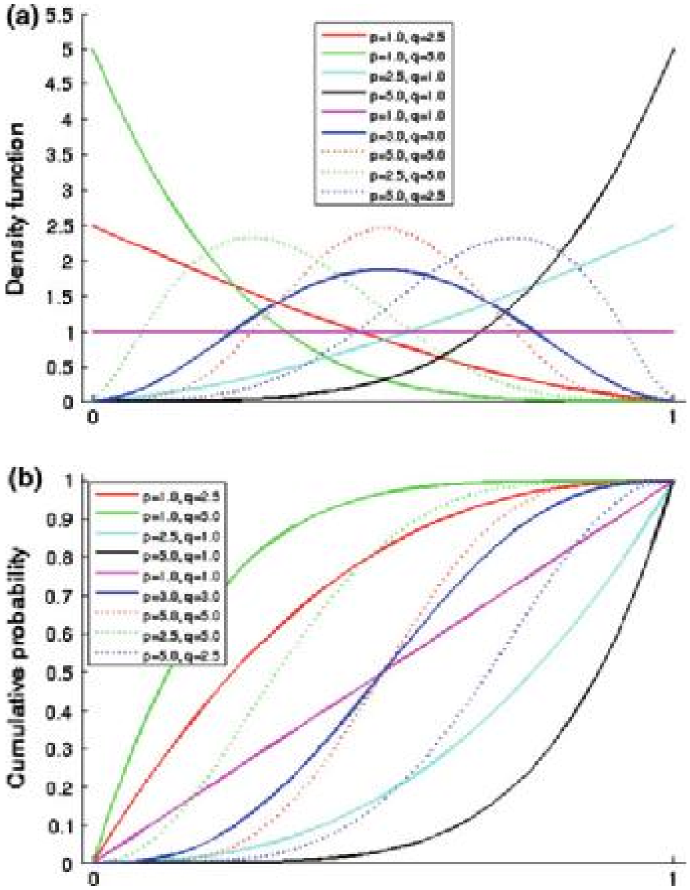


Fig. 1 Catalog of bounded analytical density functions (**a**) and corresponding cumulative distributions (**b**) using *Beta* (p, q) distributions with selected values for the shape parameters p and q. Note the variety of density functions available

For bounded function outputs $[a, b]$, hence, the expected value for the maximum can be calculated as:

$$E\left(\max(x)\right) = \int_a^b x f_{\max}(x)dx = xF_{\max}(x)|_a^b - \int_a^b F_{\max}(x)dx = b - \int_a^b F(x)^N dx$$

And the expected value for the minimum can be shown to be:

$$E\left(\min(x)\right) = a + \int_a^b \left(1 - F(x)\right)^N dx$$

On the other hand, for *GeneralizedBeta* (p, q, a, b), its cumulative distribution, $F_{GBeta}(x|p, q, a, b)$, can be shown to be equal to $F_{Beta}(z|p, q)$ when $z = \left(\frac{x-a}{b-a}\right)$; hence, the expected values of interest, i.e., $E(\max(x))$ and $E(\min(x))$, can be expressed as:

$$E(\max(x)) = b - \int_a^b F_{GBeta}(x|p,q,a,b)^N dx = b - (b-a)d_N$$

where $d_N = \int_0^1 F_{Beta}(z|p,q)^N dz$

$$E\left(\min(x)\right) = a + \int_a^b \left(1 - F_{GBeta}(x|p,q,a,b)\right)^N dx = a + (b-a)c_N$$

where $c_N = \int_0^1 \left(1 - F_{Beta}(z|p,q)\right)^N dz$.
Note that c_N and d_N are one dimensional integrals (numerically easy to solve) depending only on the shape parameters (p, q) of the *Beta* (p, q) distribution under consideration.

3.2.2 *System of equations for the density function bounds and its analytical solution*

Equating the expected value for the minimum/maximum (B.1) and the sample output minimum ($x_{\min}$) and maximum ($x_{\max}$), the following linear system of equations is obtained:

$$\left\{\begin{array}{l} x_{\max} = \max\left(\left\{x_1, ..x_j, \ldots, x_n\right\}\right) = b - (b-a)d_N \\ x_{\min} = \min\left(\left\{x_1, ..x_j, \ldots, x_n\right\}\right) = a + (b-a)c_N \end{array}\right\}$$

Solving the above system of equations for the bounds (a, b), the solution can be found to be:

$$b = x_{\max} + d_N \frac{x_{\max} - x_{\min}}{1 - c_N - d_N}$$
$$a = x_{\min} - c_N \frac{x_{\max} - x_{\min}}{1 - c_N - d_N}$$

Note that the bounds (a, b) are essentially estimates for the model output (objective function) minimum and maximum, respectively, obtained using: (i) the minimum and maximum of the sample outputs ($\mathrm{x}_{\min}$ and $\mathrm{x}_{\max}$), and (ii) the density function, i.e., *Beta* (p, q) distribution (through c_N and d_N) under consideration.

3.3 Identify the density function in the catalog with the best match to the sample outputs distribution and corresponding minimum estimate (a)

The best match refers to the analytical density function (*Generalized Beta* (p, q, a, b)) with the lowest maximum absolute difference[1] ($\mathrm{D}_{\max}$) between its cumulative distribution and the corresponding to the sample outputs. The density function of interest is selected from the catalog of *Generalized Beta* (p, q, a, b) distributions, with p, q specified in step A (Fig. 1), and bounds a, b obtained in step B. The minimum estimate of interest (a) is the one associated with the density function with the best match in the catalog.

4 Case studies

The proposed approach for solving the problem of interest, is evaluated using three well-known analytical optimization test functions [34–36]: F2 (Fig. 2), Hartmann 6D, Griewangk

[1] In the spirit of the statistic used in the Kolmogorov-Smirnov (nonparametric) test for the equality of continuous, one-dimensional probability distributions.

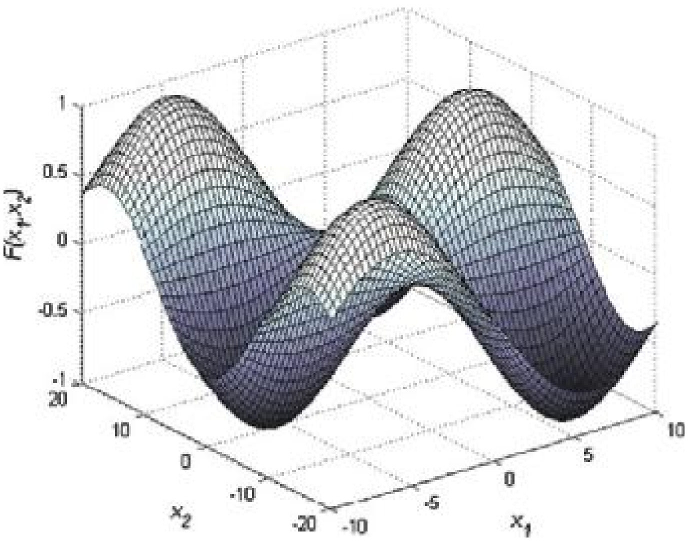


Fig. 2 F2 test function

10D (Fig. 3) and in the optimization of a field scale alkali-surfactant-polymer (ASP) enhanced oil recovery (EOR) process [37–40].

Three sample sizes in the order of $10 \cdot k$ [14] samples are considered, with k being number of input dimensions. For each case study and sample size, the effect of: (i) DOE, i.e., a hundred (100) latin-hypercubes, and (ii) noise, uniformly distributed with two levels ($\alpha_1 = 0.1, \alpha_2 = 0.2$); is also evaluated. The noisy test functions are specified by the following expression: $F(x) \cdot [1 + \alpha \boldsymbol{U} - 1/2)]$, where F is the test function under consideration and $\boldsymbol{U}$ is a uniform distribution in the interval (0, 1).

4.1 F2 [34]

$$f(x_1, x_2) = \sin\left(\frac{\pi \cdot x_1}{12}\right) \cdot \cos\left(\frac{\pi \cdot x_2}{16}\right) \quad \begin{array}{l} -10 \le x_1 \le 10 \\ -20 \le x_2 \le 20 \end{array} \quad \begin{array}{l} \text{Range} = [-1, 1] \\ f_{\text{opt}} = [-1, -1, -1] \\ x_{\text{opt}} = [(-6, 0), (6, -16), (6, 16)] \end{array}$$

4.2 Hartmann 6D [35]

$$f(x) = -\sum_{i=1}^{4} c_i \exp\left(-\sum_{j=1}^{6} a_{ij}(x_j - p_{ij})^2\right) \quad \begin{array}{l} 0 \le x_j \le 1 \quad \text{Range} = [-3.3224, 0] \\ \text{for } j = 1, 2, \ldots, 6 \end{array}$$

$$A = \begin{pmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{pmatrix}, c = \begin{pmatrix} 1 \\ 1.2 \\ 3 \\ 3.2 \end{pmatrix}$$

$$P = \begin{pmatrix} 0.1312 & 0.1696 & 0.5569 & 0.0124 & 0.8283 & 0.5886 \\ 0.2329 & 0.4135 & 0.8307 & 0.3736 & 0.1004 & 0.9991 \\ 0.2348 & 0.1451 & 0.3522 & 0.2883 & 0.3047 & 0.6650 \\ 0.4047 & 0.8828 & 0.8732 & 0.5743 & 0.1091 & 0.0381 \end{pmatrix}$$

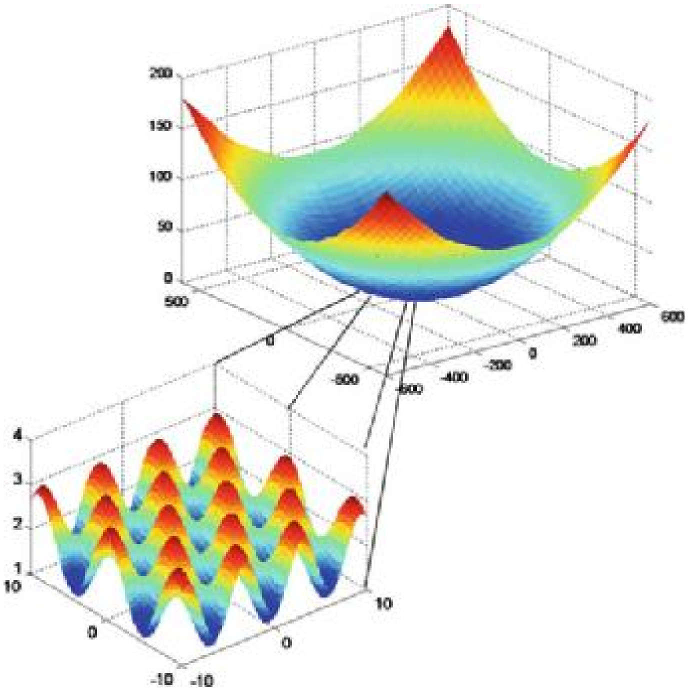


Fig. 3 Two-dimensional representation of the Griewangk 10D test function

$$f_{\mathrm{opt}} = -3.3224 \qquad x_{\mathrm{opt}} = [0.2017, 0.1500, 0.4769, 0.2753, 0.3117, 0.6573]$$

4.3 Griewangk 10D [36]

$$f(x) = 1 + \sum_{i=1}^{10} \frac{x_i^2}{4000} - \prod_{i=1}^{10} \cos\left(\frac{x_i}{\sqrt{i}}\right) \quad -600 \le x_j \le 600 \quad \text{for} \quad j = 1, 2, \ldots, 10$$

$$\text{Range} = [0, 900], \quad f_{\mathrm{opt}} = 0, \quad x_{\mathrm{opt}} = 0 \,\text{in every dimension}$$

4.4 Alkali-Surfactant-Polymer (ASP) EOR[2] process optimization [37–40]

The problem of interest is to find optimum estimates for cumulative oil production in a ASP flooding pilot given a range of values (Table 2) for the following design variables: concentration of alkaline, surfactant, polymer, and ASP slug size (expressed in the form of the injection time). The cumulative oil production is calculated at 487 days expressed as a percentage of the original oil in place (OOIP).

As illustrated in Fig. 4, the ASP flooding pilot has an inverted five-spot pattern and a total of 13 vertical wells, 9 producers and 4 injectors. The reservoir is at a depth of 4,150 ft., has an average initial pressure of 1,770 psi, and the porosity is assumed to be constant throughout the reservoir and equal to 0.3. The numerical grid is composed of 19 × 19 × 3 blocks in the x, y and z directions. The OOIP is 395,427 bbls, the crude oil viscosity is 40 cp, the initial brine

[2] Enhanced Oil Recovery.

Table 2 Design variable restrictions—ASP-EOR case study

Design variable	Range		Units
	Min.	Max.	
Alkaline concentration (Na_2CO_3)	0	0.5898	meq/ml
Surfactant concentration	0.001815	0.01	Vol. fract.
Polymer concentration	0.0487	0.1461	wt%
Injection time	111	326	Days

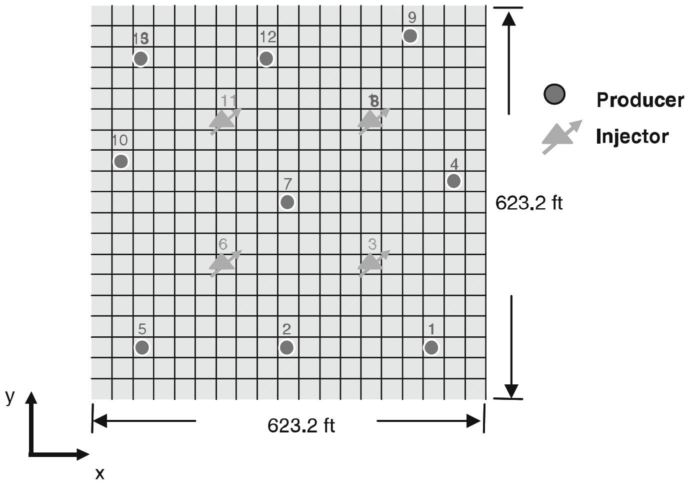


Fig. 4 Well pattern illustration—ASP-EOR case study

salinity is 0.0583 meq/ml and the initial brine divalent cation concentration is 0.0025 meq/ml. This is the reference configuration whose details can be found in the sample data archives of the UTCHEM [41] program.

Three flowing phases and eleven components are considered in the numerical simulations. The phases are water, oil and microemulsion, while the components are water, oil, surfactant, polymer, chloride anions, divalent cations (Ca++, Mg++), carbonate, sodium, hydrogen ion, and oil acid. The ASP interactions are modeled using the reactions: in situ generated surfactant, precipitation and dissolution of minerals, cation exchange with clay and micelle, and chemical adsorption. Note the detailed chemical reaction modeling, and the heterogeneous and multiphase petroleum reservoir under consideration.

4.5 Performance criteria

For a given case study, these are: (i) the *Beta* (p, q, a, b) analytical density function with the minimum D_{max} (best match) should have the same shape parameters (p, q) for both reduced and larger sample sizes, with reasonable dispersion for the D_{max} distribution. In addition, it is desirable that performance (relative error) does not deteriorate significantly when selecting an analytical density function with D_{max} distribution similar to the one exhibited by the best matching density function (*robustness*); relative error is calculated as the difference

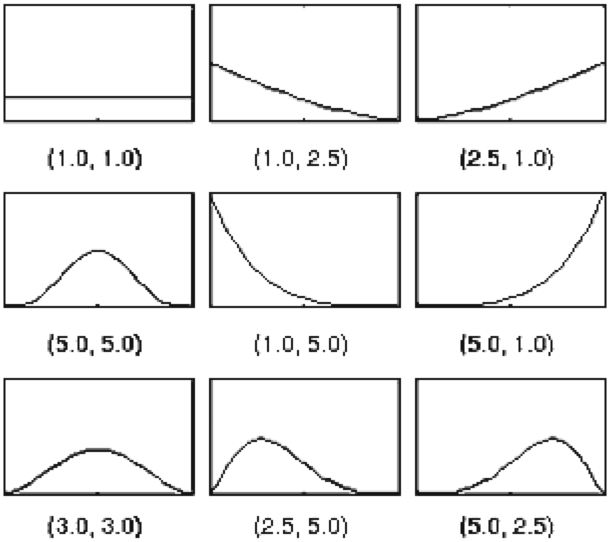


Fig. 5 Catalog of $Beta(p, q)$ density functions used for obtaining the optimum estimates of the case studies. The shape parameters (p, q) are also shown

between the estimated minimum and the true optimum value as a fraction of the function range, (ii) the true optimum should be within a 95% confidence interval of the estimated minimum; the bounds of the confidence intervals correspond to the 0.025 and 0.975 quantiles of the estimated minimum empirical distribution (*statistical soundness*), (iii) the estimated minimum should be a good approximation (low relative error) for the true optimum, and a meaningful improvement over the sample outputs minimum (present best solution) even for modest sample sizes (*accuracy*), and iv) the minimum estimates should exhibit statistically significant improvements (median and dispersion of the relative error) for larger sample sizes (*consistency*).

4.6 Catalog of density functions (*Beta* (p, q)) to match training data (outputs) empirical density function

Figure 5 illustrates the catalog with nine (9) density functions used for obtaining the optimum estimates of the case studies. Note the whole range of density functions available and the compact representation through the shape parameters $[p, q]$, i.e., [1, 2.5], [1, 5], [2.5, 1], [5, 1], [1, 1], [3, 3], [5, 5], [2.5, 5], [5, 2.5].

5 Results and discussion

The performance of the proposed approach for optimum estimation is discussed considering the performance criteria specified in the previous section:

- *Robustness*: Figs. 6, 7, 8 and 9 show, for each case study, the boxplots[3] of the $D_{\max}$ empirical distribution for a hundred LHS with three increasingly bigger sample sizes.

[3] In a boxplot, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. Points are drawn as outliers if they are larger than q3 + 1.5(q3 – q1) or smaller than q1 – 1.5(q3 – q1), where

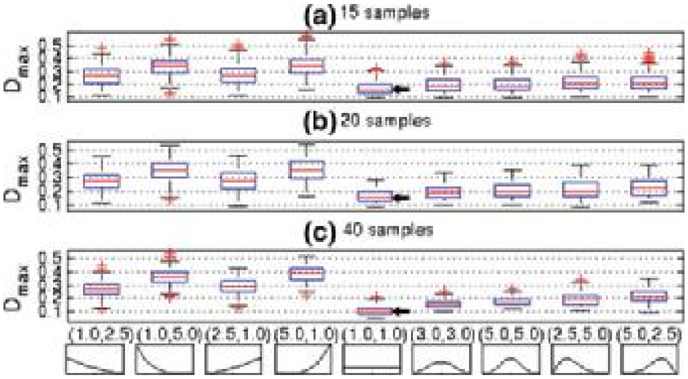


Fig. 6 Boxplots of the D_{max} corresponding to each of the *Beta* distributions in the catalog for a 100 LHS with samples of size (**a**) 15, (**b**) 20 and (**c**) 40. An *arrow* points to the Beta distribution with the best match to the sample outputs distribution. The parameters (p, q) and shape of the *Beta* distributions in the catalog are depicted below each of the columns of boxplots—F2 case study

The D_{max} distributions exhibit significant differences depending on the shape parameters (p, q). Specifically, the best matching *Generalized Beta* (p, q) density function (minimum median D_{max}) in the catalog:

i. Had the same shape parameters (p, q) for all sample sizes, but different ones depending on the case study (Table 3), and,
ii. Exhibited a coefficient of variation (robust estimates[4]) for D_{max} in the order of 20–30% and 30–40% for the case studies with up to six dimensions and ten dimensions (Griewangk 10D), respectively; the median and coefficient of variation for D_{max}, was relatively insensitive to the sample sizes under consideration, and,
iii. In those instances where the second best matching density function had the value for the median of D_{max} close to the best matching one (i.e., ASP-EOR and Griewangk 10D), the corresponding relative errors for minimum estimates deteriorate, but these errors remained lower or equal to 10% for all sample sizes, e.g., 3% versus 8% (ASP-EOR, sample size 40), and 2% versus 7% (Griewangk 10D, sample size 150).

All of the above confirms the effectiveness of D_{max} for selecting best matching density functions, and justifies the wide variety of functions available in the catalog.

- *Statistical soundness*: In all case studies, even in the ten dimensional one, the true optimum value was always within a 95% confidence interval of the estimated minimum distribution corresponding to a hundred LHS designs (even for reduced sample sizes); in fact, p-values for the minimum were, in general, well above (greater than 0.5 in all but three instances) the 0.05 significance level used as a threshold to reject the null hypothesis

Footnote 3 continued
q1 and q3 are the 25th and 75th percentiles, respectively. The default of 1.5 corresponds to approximately $^{+}/-2.7\,\sigma$ and 99.3% coverage if the data are normally distributed. The plotted whisker extends to the adjacent value, which is the most extreme data value that is not an outlier.

[4] A robust estimate of the coefficient of variation, i.e., $(\mu - (\mu - \sigma))/\mu$, can also be written as $[\phi^{-1}(p_2) - \phi^{-1}(p_1)]/\phi^{-1}(p_2)$, if ϕ^{-1} denotes the inverse cdf of a Normal probability distribution, and p_1 and p_2 represent $\phi(\mu - \sigma) \approx 0.159$ and $\phi(\mu) = 0.5$, respectively.

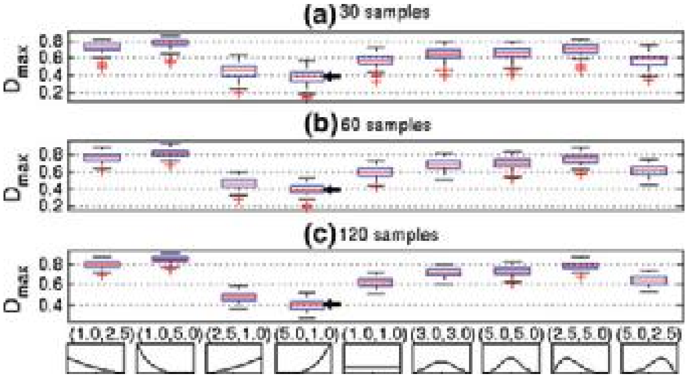


Fig. 7 Boxplots of the D_{max} corresponding to each of the *Beta* distributions in the catalog for a 100 LHS with samples of size (**a**) 30, (**b**) 60 and (**c**) 120. An *arrow* points to the Beta distribution with the best match to the sample outputs distribution. The parameters (p, q) and shape of the *Beta* distributions in the catalog are depicted below each of the columns of boxplots—Hartmann 6D case study

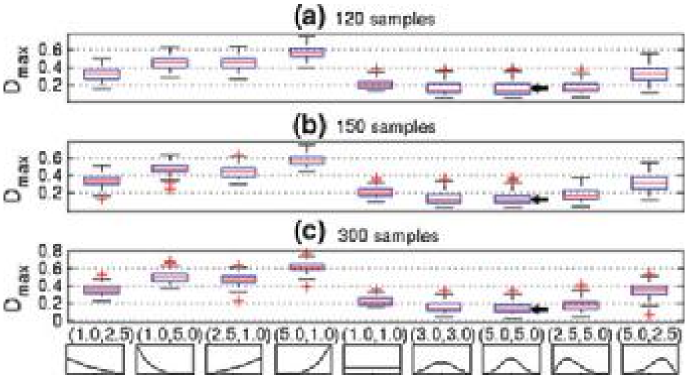


Fig. 8 Boxplots of the D_{max} corresponding to each of the *Beta* distributions in the catalog for a 100 LHS with samples of size (**a**) 120, (**b**) 150 and (**c**) 300. An *arrow* points to the Beta distribution with the best match to the sample outputs distribution. The parameters (p, q) and shape of the *Beta* distributions in the catalog are depicted below each of the columns of boxplots—Giewangk 10D case study

(i.e., true minimum is a sample of the estimated minimum distribution) (Figs. 10, 11, 12 and 13).

- *Accuracy and consistency*: In each of the case studies, the estimated minima obtained from a hundred LHS designs (Figs. 14, 15, 16 and 17):

  i. Represented an excellent approximation of the true optimum value considering the median of the relative error was lower than seven percent (7%), even for reduced sample sizes and high dimensional problems (Table 4),
  ii. Showed to be significantly closer to the minimum (maximum relative error of 6.5%) than the present best solution (maximum relative error of 49.9%),
  iii. In general, included the zero relative error within its 25th and 75th percentiles, and,
  iv. The errors (median and dispersion) were, in general, reduced with larger sample sizes.

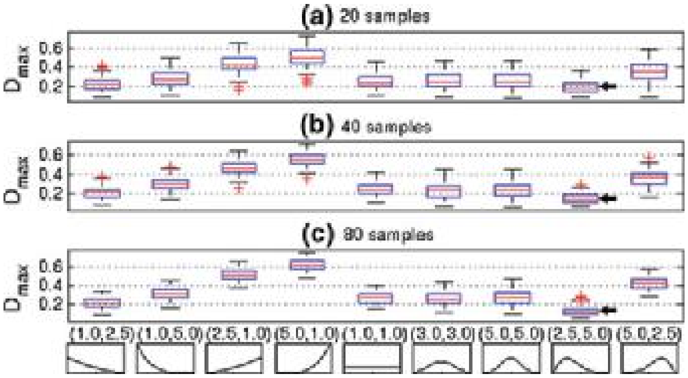
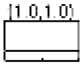


Fig. 9 Boxplots of the $\mathrm{D}_{\max}$ corresponding to each of the *Beta* distributions in the catalog for a 100 LHS with samples of size (**a**) 20 (**b**) 40 and (**c**) 80. An *arrow* points to the Beta distribution with the best match to the sample outputs distribution. The parameters (p, q) and shape of the *Beta* distributions in the catalog are depicted below each of the columns of boxplots—ASP-EOR case study

Table 3 Best matching *Beta* (p, q) density function for each case study and sample sizes

Case study	Best matching *Beta* (p, q) density function	Sample sizes
F2		15/20/40
Hartmann 6D		30/60/120
Griewangk 10D		120/150/300
ASP-EOR		20/40/80

For all case studies and sample sizes, the best matching density function remained unaltered for the noisy version of the test functions, and the differences in the relative error of the minimum estimation, with (noise 10%, α_1; 20%, α_2) and without noise, were statistically insignificant; see, for example, the results corresponding to Griewangk 10D with a sample size of 150 (Fig. 18).

6 Conclusions

This work presents an approach for estimating the expected value for the minimum (target) of an objective function at a given cycle using concepts from extreme order statistics. It is assumed that the sample inputs are randomly distributed so the outputs can be considered a random variable, whose density function is bounded (a, b), with the minimum being its lower bound. Specifically, an estimate of the minimum (a) is obtained by: (i) computing the bounds (using training data and the moment matching method) of a selected set of analytical density functions (catalog), and (ii) identifying the density function in the catalog with the best match to the sample outputs distribution and corresponding minimum estimate (a).

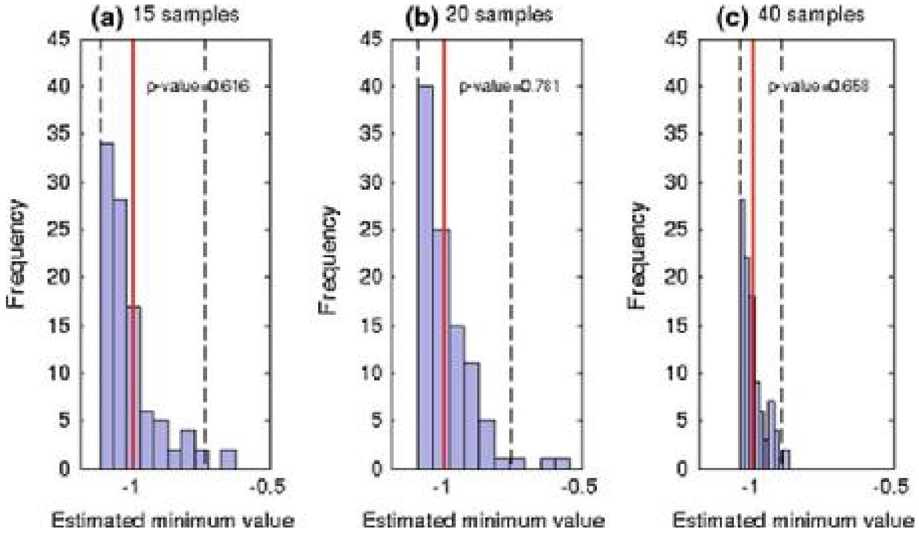


Fig. 10 Empirical distribution of estimated minimum values with an indication of the true minimum value for samples of size (**a**) 15, (**b**) 20 and (**c**) 40. Ninety five percent confidence intervals and p-values are also shown—F2 case study

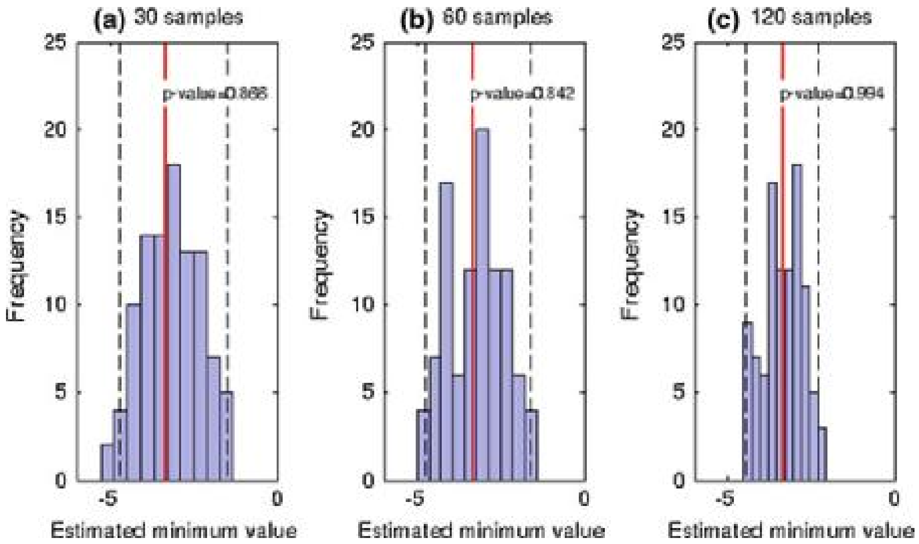


Fig. 11 Empirical distribution of estimated minimum values with an indication of the true minimum value for samples of size (**a**) 30, (**b**) 60 and (**c**) 120. Ninety five percent confidence intervals and p-values are also shown—Hartmann 6D case study

The effectiveness of the proposed approach was evaluated using a compact catalog of *Generalized Beta* density functions and well-known analytical optimization test functions, i.e., F2, Hartmann 6D, and Griewangk 10D and in the optimization of a field scale alkali-surfactant-polymer (ASP) enhanced oil recovery (EOR) process. In this context:

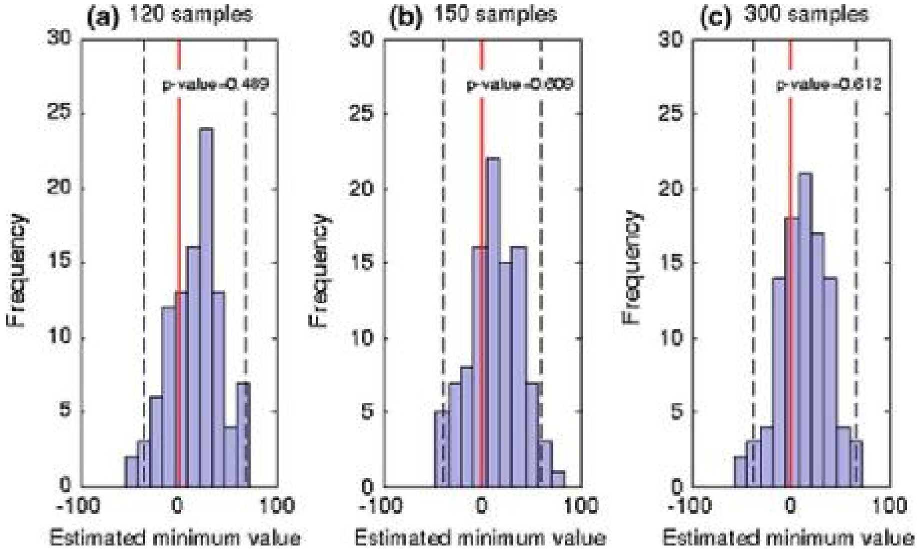


Fig. 12 Empirical distribution of estimated minimum values with an indication of the true minimum value for samples of size (**a**) 120, (**b**) 150 and (**c**) 300. Ninety five percent confidence intervals and p-values are also shown—Griewangk 10D case study

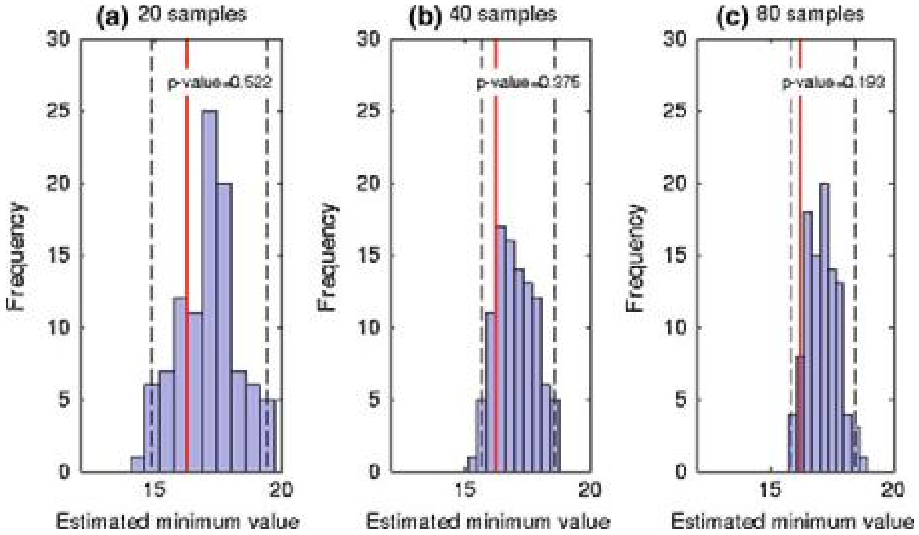


Fig. 13 Empirical distribution of estimated minimum values with an indication of the true minimum value for samples of size (**a**) 20, (**b**) 40 and (**c**) 80. Ninety five percent confidence intervals and p-values are also shown—ASP-EOR case study

- It was possible to setup a compact catalog with a variety of density functions (nine) by modifying the shape parameters (p, q) of a *Beta* (p, q) distribution. While the cited catalog includes the most anticipated density function shapes (e.g., uniform, Gaussian like, biased to the left or right), the catalog can evolve to meet the requirements of empirical sample outputs distribution related to particular problems.

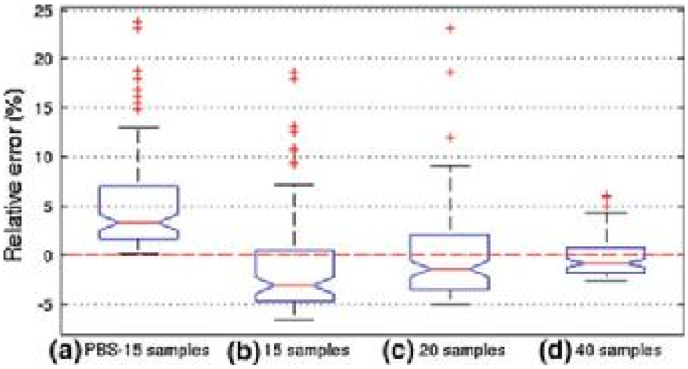


Fig. 14 Boxplots of the relative error for: (*a*) sample minimum value (PBS) and estimated minimum value using a sample of size (*b*) 15, (*c*) 20 and (*d*) 40—F2 case study

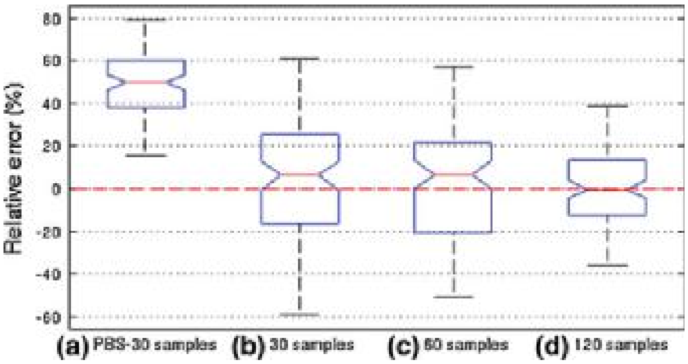


Fig. 15 Boxplots of the relative error for: (*a*) sample minimum value (PBS) and estimated minimum value using a sample of size (*b*) 30, (*c*) 60 and (*d*) 120—Hartmann 6D case study

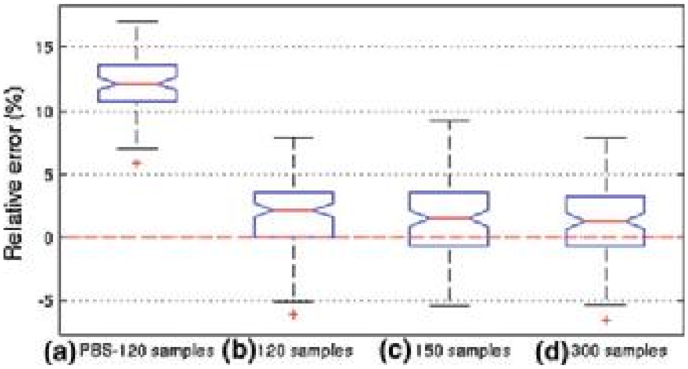


Fig. 16 Boxplots of the relative error for: (*a*) sample minimum value (PBS) and estimated minimum value using a sample of size (*b*) 120, (*c*) 150 and (*d*) 300—Griewangk 10D case study

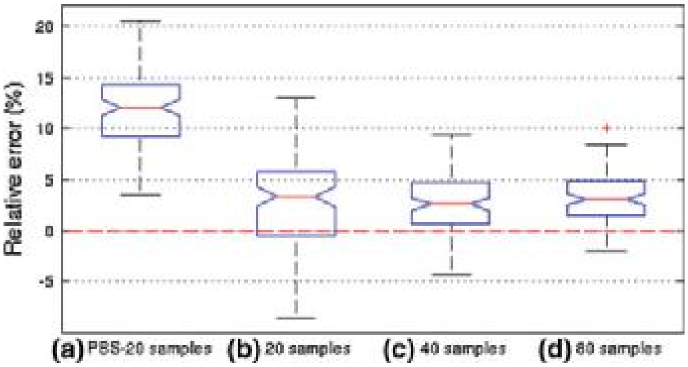


Fig. 17 Boxplots of the relative error for: (*a*) sample minimum value (PBS) and estimated minimum value using a sample of size (*b*) 20, (*c*) 40 and (*d*) 80—ASP-EOR case study

Table 4 Relative errors for case studies and sample sizes

Case study	Sample size			Relative error (median) [%]			
				Proposed approach			Present best solution (sample size)
F2	15	20	40	3.1	1.3	0.8	3.4 (15)
Hartmann 6D	30	60	120	6.2	6.5	0.3	49.9 (30)
Griewangk 10D	120	150	300	2.2	1.5	1.3	12.2 (120)
ASP-EOR	20	40	80	3.4	2.6	3.1	12.0 (20)

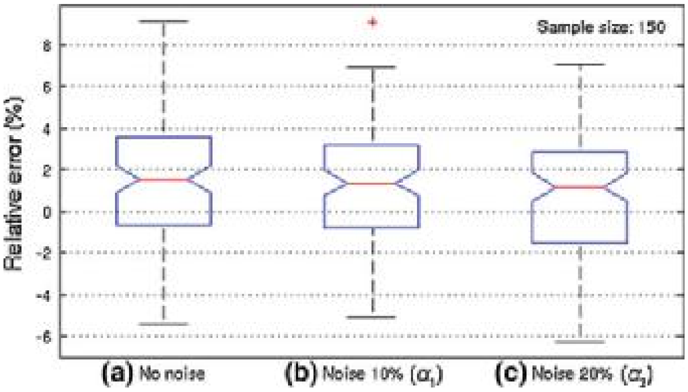


Fig. 18 Boxplots of the relative error of the minimum estimation (this work) considering: (*a*) no noise, (*b*) noise 10% (α_1), and (*c*) noise 20% (α_2), for a sample size of 150—Griewangk 10D case study

- The process for selecting the density function that best matches the function outputs distribution was shown to be robust, i.e., the best matching density function in the catalog was the same for all sample sizes, but different depending on the case study.
- The true optimum value was always within a 95% confidence interval of the estimated minimum distribution with p-values, in general, greater than 50% for all sample sizes, even for high dimensional problems (up to 10D).

- The estimated minimum represented an excellent approximation of the true optimum value even for reduced sample sizes with significant improvements over the present best solution, and did not show to be significantly affected by the curse of dimensionality.

The proposed approach is independent of the surrogate and optimization strategies, can be tailored to fit a variety of risk attitudes and design environments, and holds promise to be useful in setting targets and assessing the value of another cycle in surrogate-based optimization. Future work should focus on strategies for updating targets throughout surrogate-based optimization cycles where the samples of the inputs may no longer be considered random.

Acknowledgments One of us (NQ) acknowledges this material is based upon work supported by National Science Foundation under Grant DDM-0856431.

References

1. Queipo, N., Pintos, S., Verde, A., Haftka, R.: Assessing the value of another cycle in Gaussian process surrogate-based optimization. Struct. Multidiscip. Optim. **39**(5), 459–475 (2009). doi:10.1007/s00158-008-0346-0
2. Giunta, A.A., Balabanov, V., Haim, D., Grossman, B., Mason, W.H., Watson, L.T., Haftka, R.T.: Multidisciplinary optimization of a supersonic transport using design of experiments, theory and responsive surface modeling. Aeronaut. J. **101**, 347–356 (1997)
3. Balabanov, V., Haftka, R., Grossman, B., Mason, W., Watson, L.: Multidisciplinary response model for HSCT wing bending material weight. In: Proc 7th AIAA/USAF/NASA/ISSMO Symp Multidiscip Anal Optim, AIAA paper 98-4804, St. Louis, MO, pp. 778—788 (1998)
4. Li, W., Padula, S.: Approximation methods for conceptual design of complex systems. In: Chui, C., Neamtu, M., Schumaker, L. (eds.) 11th Int Conf Approx Theory. Gatlinburg, TN, May (2004)
5. Queipo, N., Haftka, R., Shyy, W., Goel, T., Vaidyanathan, R., Tucker, P.K.: Surrogate-based analysis and optimization. J. Prog. Aerosp. Sci. **41**, 1–28 (2005)
6. Craig, K., Stander, N., Dooge, A., Varadappa, S.: MDO of automotive vehicles for crashworthiness using response surface methods. In: 9th AIAA/ISSMO Symp Multidiscip Anal Optim, AIAA paper 2002-5607, Atlanta, 4–6 September (2002)
7. Kurtaran, H., Eskamdarian, A., Marzougui, D., Bedewi, N.: Crashworthiness design optimization using successive response surface approximations. Comput. Mech. **29**, 409–421 (2002)
8. Queipo, N., Goicochea, J., Pintos, S.: Surrogate modeling-based optimization of SAGD processes. J. Pet. Sci. Eng. **35**(1–2), 83–93 (2002)
9. Queipo, N., Verde, A., Canelon, J., Pintos, S.: Efficient global optimization of hydraulic fractuing designs. J. Pet. Sci. Eng. **35**(3–4), 151–166 (2002)
10. Wang, G., Shan, S.: Review of metamodeling techniques in support of engineering design optimization. ASME Trans. J. Mech. Des. **129**(4), 370 (2007)
11. Forrester, A., Keane, A.: Recent advances in surrogate-based optimization. Prog. Aerosp. Sci. **45**(1–9), 50–79 (2009)
12. Jones, D.: A taxonomy of global optimization methods based on response surfaces. J. Glob. Optim. **21**, 345–383 (2001)
13. Gutmann, H.: A radial basis function method for global optimization. J. Glob. Optim. **19**, 201–227 (2001)
14. Jones, D., Schonlau, M., Welch, W.: Efficient global optimization of expensive black-box functions. J. Glob. Optim. **13**, 455–492 (1998)
15. Sasena, M., Papalambros, P., Goovaerts, P.: Exploration of metamodeling sampling criteria for constrained global optimization. Eng. Optim. **34**(3), 263–278 (2002)
16. Sobester, A., Leary, S., Keane, A.: On the design of optimization strategies based on global response surface approximation models. J. Glob. Optim. **33**, 31–59 (2005)
17. Huang, D., Allen, T., Notz, W., Zeng, N.: Global optimization of stochastic black-box systems via sequential kriging meta-models. J. Glob. Optim. **34**, 441–466 (2006)
18. Apley, D., Liu, J., Chen, W.: Understanding the effects of model uncertainty in robust design with computer experiments. J. Mech. Des. **128**(4), 945–958 (2006)
19. Forrester, A., Jones, D.: Global optimization of deceptive functions with spare sampling. In: Proc 12th AIAA/ISSMO Multidiscip Anal Optim Conf, BC, Canada, 10–12 September (2008)

20. Ponweiser, W., Wagner, T., Vincze, M.: Clustered multiple generalized expected improvement: a novel infill sampling criterion for surrogate models. In: Michalewicz, Z. (ed.) IEEE Congr on Evol Comp, pp. 3514—3521. IEEE Computer Society. (2008)
21. Ginsbourger, D., Le Riche, R., Carraro, L.: Multi-points criterion for deterministic parallel global optimization based on Kriging. In: Intl Conf Nonconvex Progr, Rouen, France (2007)
22. Gumbel, E.J.: Statistics of Extremes. pp. 375 Columbia University Press, New York (1958)
23. Finkenstadt, B., Rootzéen, H. (eds.): Extreme Values in Finance, Telecommunications and the Environment. Chapman and Hall/CRC Press, London (2003)
24. Coles, S.G.: An Introduction to Statistical Modeling of Extreme Values. Springer, New York (2001)
25. Johnson, M., Moore, L., Ylvisaker, D.: Minimax and maximin distance designs. J. Stat. Plan. Inference **26**, 131–148 (1990)
26. Iman, R., Conover, W.: A distribution-free approach to inducing rank correlation among input variables. Commun. Stat. Part B Simulat. Comput. **11**, 311–334 (1982)
27. Owen, A.B.: Controlling correlations in latin hypercube samples. J. Stat. Assoc. **89**, 1517–1522 (1994)
28. McKay, M., Conover, W., Beckman, R.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics **21**, 239–245 (1979)
29. Tang, B.: Orthogonal array-based latin hypercubes. J. Am. Stat. Assoc. **88**, 1392–1397 (1993)
30. Ye, K.: Orthogonal column latin hypercubes and their application in computer experiments. J. Am. Stat. Assoc. **93**, 1430–1439 (1998)
31. Palmer, K., Tsui, K.: A minimum bias latin hypercube design. IIE Trans. **33**, 793–808 (2001)
32. Leary, S., Bhaskar, A., Keane, A.: Optimal orthogonal array-based latin hypercubes. J. Appl. Statist. **30**, 585–598 (2003)
33. Cramer, H.: Mathematical Methods of Statistics. Princeton University Press, Princeton (1999)
34. Jin, R., Chen, W., Simpson, T.: Comparative studies of metamodeling techniques under multiple modeling criteria. Struct. Multidisc. Optim. **23**, 1–13 (2000)
35. Dixon, L., Szegö, G.: The global optimization problem: an introduction. In: Dixon, L., Szegö, G. (eds.) Towards Global Optimization, 2, North-Holland, Amsterdam (1978)
36. Digalakis, J., Margaritis, G.: On benchmarking functions for genetic algorithms. Int. J. Comp. Math. **77**(4), 481–506 (2001)
37. Zerpa, L., Queipo, N., Pintos, S., Salager, S.: An optimization methodology for alkaline-surfactant-polymer flooding processes using field scale numerical simulations and multiple surrogates. J. Pet. Sci. Eng. **47**(3–4), 197–208 (2005)
38. Carrero, E., Zerpa, L., Queipo, N., Pintos, S.: Global sensitivity analysis of alkali-surfactant-polymer enhanced oil recovery processes. J. Pet. Sci. Eng. **50**(1–2), 30–42 (2007)
39. Sanchez, E., Queipo, N., Pintos, S.: Toward an optimal ensemble of kernel-based approximations with engineering applications. Struct. Multidiscip. Optim. **36**(3), 247–267 (2008)
40. Nava, E., Pintos, S., Queipo, N.: A geostatistical perspective for the surrogate-based integration of variable fidelity models. J. Pet. Sci. Eng. **51**, 56–66 (2010)
41. UTCHEM: Utchem-9.0 a three-dimensional chemical flood simulator (2000). http://www.cpge.utexas.edu/utchem/