Proceedings of the 11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Portsmouth, Virginia, 6-8 September, U.S.A.

Assessing the Value of Another Cycle in Surrogate-Based Optimization

Nestor V. Queipo*, Alexander Verde†, and Salvador Pintos‡
Applied Computing Institute, University of Zulia, Venezuela.
{nqueipo, averde, spintos} @ ica.luz.ve

and

Raphael T. Haftka§
Aerospace and Mechanical Engineering, University of Florida, USA
haftka @ ufl.edu

Surrogate-based optimization (SBO) for engineering design has become popular in the optimization of engineering systems (e.g., aerospace, automotive, oil industries) requiring expensive computer simulations. SBO proceed in design cycles, each cycle consisting of gathering input/output data using computer simulations, construction of a surrogate based on these data, estimation of the optimum using the surrogate, and a simulation at that optimum (PBS). However, due to time and cost constraints, the design optimization is limited to a small number of cycles (short cycle SBO) and rarely allowed to proceed to convergence. The current frontier of surrogate-based engineering design lacks statistically rigorous procedures for assessing the merit of investing in another cycle of analysis versus accepting the PBS.

This paper presents a methodology to address this issue. The proposed methodology establishes an estimate of the probability of improving a specified target if another cycle (with a given set of points) is undertaken. It relies on three components: i) a covariance model (structure and parameters) obtained from available input/output data, ii) a surrogate model such as those provided by polynomial regression, kriging, and support vector regression, and iii) the assumption that the points in the next cycle are a realization of a Gaussian process with a covariance matrix and mean specified using i) and ii). Gaussian processes are frequently used for problems of regression and classification and are closely related to a variety of surrogate modeling approaches including neural networks, kriging, and generalized radial basis functions. In this study, a particular form of kriging is used to evaluate the proposed methodology considering that capturing a covariance model is at the core of this surrogate modeling approach. Validation results obtained using elements of statistical inference in the context of the SBO of the Branin-Hoo test function, and its application in the optimization of alkali-surfactant-polymer flooding of petroleum reservoirs is also discussed.

Nomenclature

a	=	statistical significance level
B	=	statistic - number of prediction points with function values below the target
B	=	random variable - number of prediction points with function values below the target
Cov	=	covariance function
E	=	expected value
H_o	=	null hypothesis
N	=	multivariate normal distribution
$N_{Zp\|Zt}$	=	normal conditional probability distribution of Z_p given Z_t
p_i	=	probability of improvement corresponding to prediction point i

* Professor and Director, Applied Computing Institute, University of Zulia, Maracaibo, ZU, 4011, Venezuela
† Research Engineer, Applied Computing Institute, University of Zulia, Maracaibo, ZU, 4011, Venezuela
‡ Professor, Applied Computing Institute, University of Zulia, Maracaibo, ZU, 4011, Venezuela
§ Distinguished Professor, University of Florida, Gainesville, FL, 32611, Fellow AIAA

pb_k = probability of obtaining k prediction points with function values below target
PBS = present best solution
R = correlation function
SD = standard deviation
T = target
V = variance
x = model input
Z_p = prediction data set
Z_t = training data set
m = mean vector
s_z = scale factor
q = correlation parameters

I. Introduction

Surrogate based optimization (SBO) of computationally demanding simulation-based models has become very popular over the last decade[1-3]. A typical SBO constructs the surrogate based on a number of simulations, estimates the optimum design based on the surrogate, and then performs an exact simulation at that estimated position (*checking phase*). This constitutes one *cycle*. The process is then repeated until resources run out or convergence is established. There has been much progress recently in developing SBO methods with proven convergence[4,5], and in the SBO under uncertainty for robust design and reliability-based design optimization as evidenced in the DAKOTA and i-SIGHT optimization frameworks[6-9]. However, in many applications, time and resources limit the approach to a small number of cycles[10,11].

The current frontier of surrogate-based engineering design lacks statistically rigorous procedures for assessing the merit of investing in another cycle of analysis versus accepting the present best solution (PBS). More precisely, the designer faces a question whose answer has received limited attention: what is the probability that the present best solution (PBS) can be improved at least a certain amount? There is available, however, the Gaussian processes (GP) perspective to surrogate modeling which has a long history in the field of statistics and will prove to be useful in this context. Just as a Gaussian distribution is specified by its mean and a covariance matrix, a Gaussian process is specified by a mean and a covariance model; here, the mean is a function of the location in the model input space, and the covariance is a function expressing how correlated the model output values are at two locations. GP are frequently used for problems of regression (e.g., kriging) and classification and are closely related to a variety of surrogate modeling approaches including neural networks[12], kriging[13,14], generalized radial basis functions[15], and kernel methods[16]. Rasmussen[17] conducted a comparison of GP regression with several other state of the art methods on a number of problems and, in general, found its performance comparable or superior to most methods. A more recent comparison of GP modeling versus the response surface method is available in Hollingsworth and Mavris[18].

This paper presents a methodology to address the problem of interest which relies on three components: i) a covariance model (structure and parameters) obtained from available input/output data, ii) a surrogate model such as those provided by polynomial regression, Kriging, and support vector regression, and iii) the assumption that the points in the next cycle are a realization of a Gaussian distribution with a covariance matrix and mean specified using i) and ii). The methodology is validated (through elements of statistical inference) using a well-known analytical test function (i.e., Branin and Hoo), and evaluated in the surrogate-based modeling of a field scale alkali-surfactant-polymer (ASP) enhanced oil recovery (EOR) process. ASP flooding is the most promising EOR solution for one of the greatest challenges facing the oil industry worldwide: after conventional water flooding the residual oil (drops trapped by capillary forces) in reservoirs around the world is likely to be around 70 % of the original oil in place[19,20].

The remainder of the paper is structured as follows: problem statement (Section II), solution approach (Section III), case studies (Section IV), results, and discussion (Section V), and conclusions (Section VI).

II. Problem Statement

In the context of surrogate-based optimization, given a surrogate model (built from a set of training points) and a sample of prediction locations, what is the probability of improving a particular target at one or more of the prediction locations if another cycle is undertaken?. As an illustration of the problem of interest, Fig. 1 shows a kriging-based model of the Branin and Hoo test function, a set of prediction points, the present best solution-PBS, and a target-T.

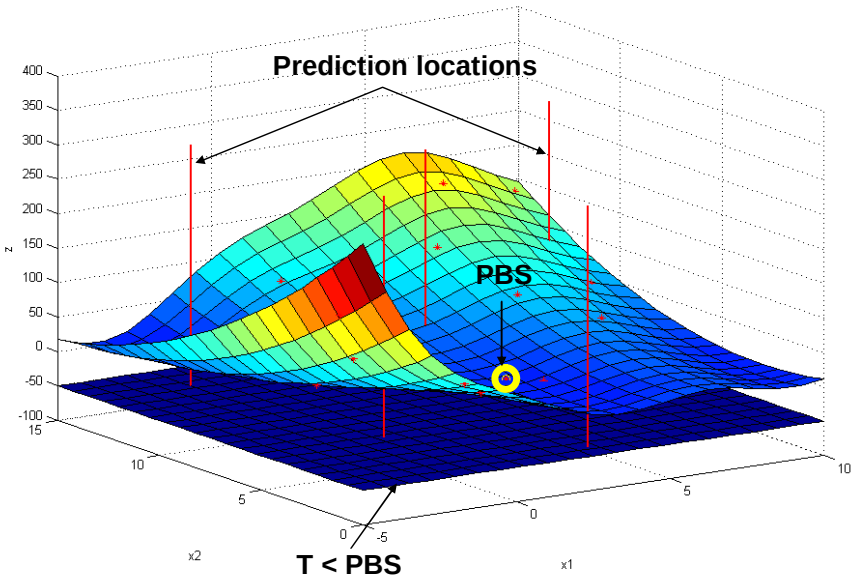

Figure 1. A Kriging-based model of the Branin-Hoo function with a set of five (5) prediction locations where the probability of improving a specified target is sought.

Under a GP perspective[13,21,22], the problem of interest can be mathematically formulated as follows. Considering, the vectors Z_t, and Z_p denote training and prediction data (output) sets, respectively, the points in the next cycle in surrogate-based optimization can be seen as a realization of the following Gaussian distribution[23] (with a mean value equal to zero for the Z_t output data)[¶]:

$$N_{Zp|Zt}\left\{\sum_{pt}\sum_{tt}^{-1}Z_t\,,\ \sum_{pp}-\sum_{pt}\sum_{tt}^{-1}\sum_{tp}\right\} \tag{1}$$

where $N_{Zp|Zt}$ is a multivariate normal distribution representing the conditional probability distribution of Z_p given Z_t and the matrices denoted by Σ specify the variances and covariances of the components in vectors Z_p and Z_t. Note that the terms in brackets represent the mean of the prediction at the prediction locations and the conditional covariance matrix of Z_p given Z_t, respectively. The components in the variance and covariance matrices (denoted by Σ) can be calculated by identifying a covariance function Cov(z, z); the general form of the covariance function expresses the idea that nearby inputs will have highly correlated outputs and some parameters allow a different distance measure for each input dimension.

In this context, the problem of interest is then to calculate the probability that a target T can be met or surpassed by at least one of the components of Z_p given a set of training points in Z_t.

III. Solution Approach

Given the previously cited GP perspective, the solution approach relies on three components: i) a covariance model (structure and parameters) obtained from available input/output data (this issue is discussed at the end of this section), ii) a surrogate model such as those provided by polynomial regression, kriging, and support vector regression, and iii) the assumption that the points in the next cycle are a realization of a Gaussian process (GP) with a covariance matrix and mean specified using i) and ii). Once the Gaussian process is specified (i.e., through its mean and covariance matrix), the probability of interest can be calculated as:

$$\Pr ob\left(at\ least\ Z_{pj} < T \middle| Z_t\right) = 1 - \Pr ob\left(Z_p > W \middle| Z_t\right) \tag{2}$$

[¶] The expression for a non-zero mean just would involve a little extra complexity

where W is a vector with dimension equal to the number of prediction points and with all its values equal to the target T, that is, $W = [T \quad T \quad . \quad . \quad T]^t$.

Furthermore, a GP is a stochastic process for which any finite set of outputs (e.g., predictions) has a joint multivariate Gaussian distribution. Hence, considering the symmetry of the multivariate Gaussian density function with respect to its mean value μ (Fig. 2), we can write:

$$\Pr ob(Z_p > W | Z_t) = \Pr ob(Z_p \le 2m - W | Z_t) \tag{3}$$

This transformation is required because we can compute the right hand side of the previous equation using well known algorithms for evaluating Gaussian cumulative probability distributions in high dimensions[24].

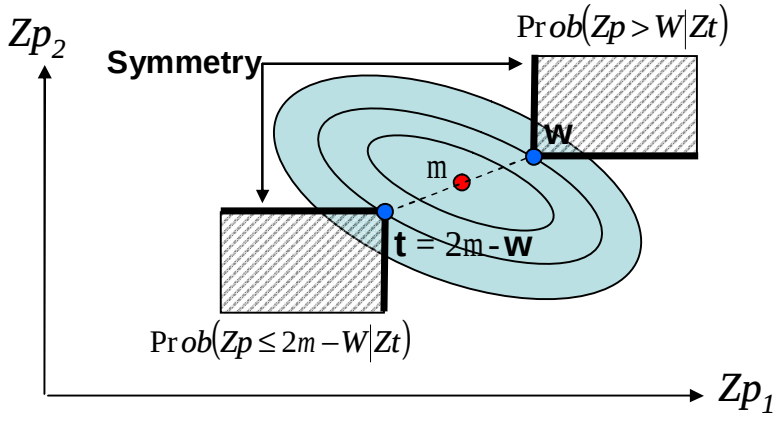

Figure 2. An illustration of how to estimate Prob (Zp>W|Zt) in Eq. (2) using multivariate normal cumulative distributions.

More precisely, the solution approach includes the following steps:

1. Construct vector Z_t from the available data. It includes the output values in the data used to construct the current surrogate model.
2. Identify a covariance model for the output values in the training (Z_t) and prediction Z_p sets. This issue will be fully discussed later in this section.
3. Using the covariance model identified in Step 2, calculate the covariance matrices denoted as: $\sum_{tt}, \sum_{tp}, \sum_{pt}, \sum_{pp}$
4. Compute the conditional covariance matrix of Z_p given Z_t, that is:

$$\Sigma_{pp} | Z_t = \sum_{pp} - \sum_{pt} \cdot \sum_{tt}^{-1} \cdot \sum_{tp} \tag{4}$$

5. Create a mean vector m equal to the surrogate model predictions at the prediction points; for example, in the case of kriging (assuming a trend equal to zero) the mean vector can be expressed as:

$$m = \sum_{pt} \cdot \sum_{tt}^{-1} \cdot Z_t \tag{5}$$

In other modeling approaches (polynomial regression, support vector regression, etc.) the mean at any prediction point would be the surrogate modeling prediction at that location.

6. Establish a desired target T whose probability of improving in the next cycle is sought and construct a vector W of dimension equal to the number of prediction points and components with values equal to T

7. Compute the symmetric vector *s* with respect to *W as:*

$$s = 2m - W \tag{6}$$

8. Compute the value of the multivariate normal cumulative distribution function-CDF corresponding to vector s, namely: $\Pr ob\left(Z_p \le s \mid Z_t\right)$
9. Calculate the probability of interest as:

$$\Pr ob\left(at\ least\ Z_{pj} < T \mid Z_t\right) = 1 - \Pr ob\left(Z_p \le s \mid Z_t\right) \tag{7}$$

A. Covariance Model Identification

Identification in this context means to establish the ***structure*** and ***parameters*** of the covariance function. The covariance function C(z, z´) is a function of the model inputs (x, x´), returns the covariance of the outputs corresponding to two inputs, and it encodes our assumptions about the problem (for example, that is smooth and continuous). Formally, we are required to specify a function which will generate a non-negative definite covariance matrix for any set of input points. From a modeling point of view, we wish to specify covariance models with a ***structure*** that contain our previous beliefs about the function we are modeling. The general form of the covariance function expresses the idea that nearby inputs will have highly correlated outputs with some ***parameters*** (θ) allowing a different distance measure for each input dimension.

A frequent assumption in the context of surrogate modeling using GP is to have the covariance function to be stationary; that is, Cov(z, z´) is a function of x-x´; if additionally, Cov only depends on the magnitude of the distance between x and x´ then the covariance function is said to be isotropic. The use of stationary covariance functions is appealing since it makes the prediction invariant under shifts of the origin in the input space, and greatly simplifies the covariance model identification. One commonly used covariance function for inputs in R^n is:

$$\operatorname{cov}(z, z') = s_z^2 \cdot R(q, x, x') \tag{8}$$

where s_z^2 is a scale factor and $R(q, x, x')$ is a correlation function:

$$R(q, x, x') = \prod_{j=1}^{n} R_j\left(q, x_j - x_j'\right) \tag{9}$$

This is simply the product of *n* correlation functions with a set of parameters θ. Table 1 shows commonly used correlation functions; note that the correlation function does not have to be ¨Gaussian¨ and that the parameters model different length scales in each dimension.

Table 1 – Commonly used correlation functions

Name	$R\left(q, x_j - x_j'\right)$
Exponential	$\exp\left(-q_j \cdot \left\|x_j - x_j'\right\|\right)$
Gaussian	$\exp\left(-q_j \cdot \left\|x_j - x_j'\right\|^2\right)$
Exponential - Gaussian	$\exp\left(-q_j \cdot \left\|x_j - x_j'\right\|^{q_{n+1}}\right), \quad 0 < q_{n+1} \le 2$

Once the covariance function structure has been set, the parameters can be estimated using the training data. There are several approaches for achieving this purpose: i) maximum likelihood estimates-MLE[25], ii) cross validation (CV), and general cross validation (GCV) methods, as discussed in Wahba[26], and iii) through variogram modeling[13]. In particular, the MLE approach consists of maximizing the log likelihood of the training vector Zt under a Gaussian process with known mean and covariance matrix calculated using a previously specified covariance function. Given the likelihood and its derivatives with respect to the parameters θ, the

maximum of the likelihood can be estimated using standard optimization routines. The evaluation of the likelihood and its partial derivatives takes time $O(n^3)$ unless a special structure in the problem can be exploited and can be a difficult problem in high dimensions; approximate methods such as that proposed by Vecchia[27] have been shown to be useful in such scenarios. A more robust approach for covariance function identification can be made through the so called variogram modeling process from geostatistics. This approach has been limited to low dimensional problems, and extensions to high dimensional problems are not obvious. In any event, there is empirical evidence that even somewhat crude MLEs can lead to useful predictions and quantifications of uncertainty[21].

IV. Case Studies

This section describes the evaluation approach, and the analytical (Branin and Hoo function[28]) and the industrial (Alkali-Surfactant-Polymer enhanced oil recovery optimization[11]) case studies used to test the proposed solution methodology.

A. Validation Approach

The validation approach is limited to a particular case of the proposed solution methodology (uncorrelated prediction points), and will include as modeling approach ordinary kriging[13]. It includes the following steps:

1. Given a set of *N* (prediction) locations where the probabilities of improvement p_i´s over a specified target T can be calculated, evaluate the computationally expensive model and check at each location whether an improvement was observed (success=1) or not (failure=0). This will be called an experiment.

   The p_i´s can be computed since a prediction under the GP approach is also normally distributed with mean, and variance (i.e., $\Sigma_{pp}|Z_t$) as specified in steps 5, and 4 of the solution approach, respectively. Note that for calculating each of the p_i´s step 4 is conducted one prediction point at a time. Fig. 3 illustrates this calculation.

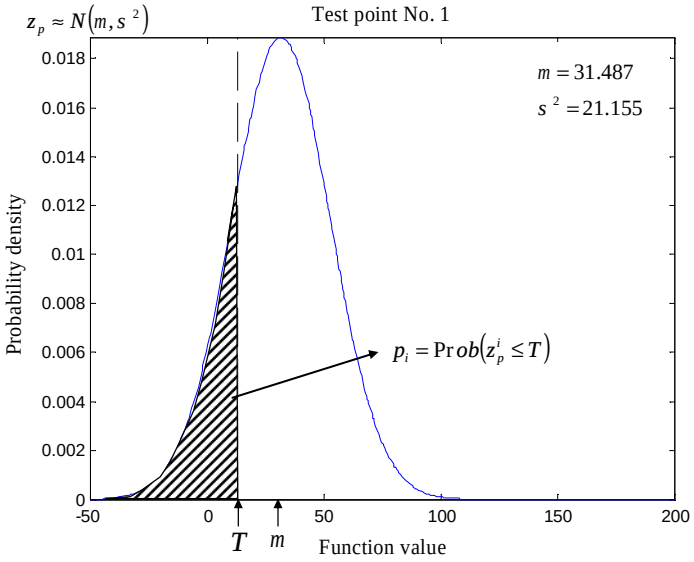

Figure 3. Illustration of the probability of improvement.

2. Define a random variable B equal to the number of prediction points with function values below the target, and recognize that, assuming independence, this variable should follow a generalized binomial distribution (GBD) with expected value $E(B) = \sum_{i=1}^{n} p_i$ and variance $V(B) = \sum_{i=1}^{n} p_i.(1-p_i)$. Note that this is just the well-known binomial distribution but with variable probabilities for the trials.

3. Compute a statistic *b* equal to the number of prediction points with function values below the target for a particular experiment.
4. Conduct a hypothesis test (statistical inference) using as null hypothesis H_o that *B* actually follows a GBD with *E(B)* and *V(B)* as specified in step 2, and compute the p-value associated with the statistic *b* calculated in step 3. In this context the p-value of B is the probability that B will assume a value "at least as extreme" as an observed value *b* given that the null hypothesis is true. If the p-value is high (e.g., greater than 0.05) the null hypothesis can not be rejected and hence the approach can be considered statistically consistent. Figure 4 illustrates the p-value calculation. The calculation procedure differs depending on whether the statistic *b* is lower (Eq. 11) or higher (Eq. 12) than E(B)

Case b<E(B)

$$p-value = 2\cdot\sum_{k=0}^{b} pb_k \tag{11}$$

Case b>E(B)

$$p-value = 2\cdot\left(1-\sum_{k=0}^{b-1} pb_k\right) \tag{12}$$

where $\sum_{k=0}^{b_{\max}} pb_k = 1$, and pb_k is the probability of obtaining *k* prediction points with function values below target.

As previously stated, the probabilities of improvement at each of the test locations are calculated assuming independence and a normal distribution with mean equal to the surrogate model prediction and an estimated variance (analytical). In the case of kriging, empirical estimates of these variances are also considered. The analytical variance neglects the fact that the kriging correlation parameters are estimated from a sample and there is uncertainty about their values. As discussed by den Hertog[29], the analytical variance underestimates the true one, and more accurate estimates can be obtained through parametric bootstrapping[30]. The kriging modeling was conducted using the Matlab toolbox developed by Lophaven et al.[31]; in particular, it was implemented ordinary kriging with a Gaussian correlation function and parameters identified using maximum likelihood principles.

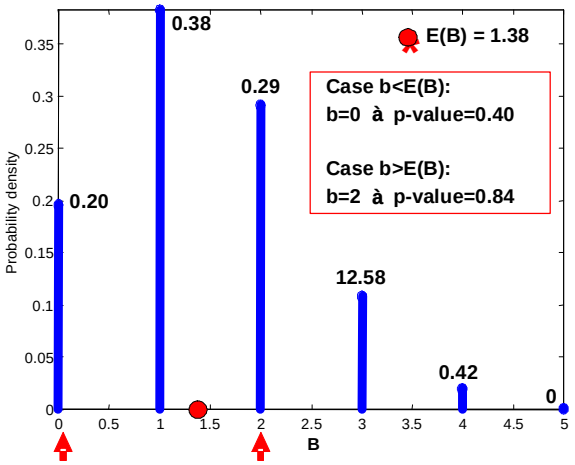

Figure 4. An illustration of the p-value calculation. Two different

In each of the following case studies, after the above cited evaluation is conducted on the analytical test function (Branin and Hoo), the results of the proposed solution methodology for the more general scenario of *correlated* prediction points will be calculated.

B. Branin and Hoo Test Function

This test function is expressed as:

$$f(x,y)=\left(y-\frac{5.1x^2}{4p^2}+\frac{5x}{p}+6\right)^2+10\left(1-\frac{1}{8p}\right)\cos(x)+10 \tag{13}$$

Input space range: x ϵ [-5, 10] & y ϵ [0, 15]

The statistic *b* will be calculated using ten (10) different Latin hypercube designs with sixteen (16) training points, and five (5) prediction points. The target is made equal to the present best solution in each training set, minus ten (10) percent of the segment between the present best solution and the global optimum of the function (0.3978).

C. Alkali-Surfactant-Polymer (ASP) Optimization

The problem of interest is to find the values of the design variables, namely, concentration of alkaline, surfactant and polymer, and ASP slug size (expressed in the form of the injection time) that maximize the cumulative oil production. The ranges of the design variables are presented in Table 2. The cumulative oil production is calculated at 487 days expressed as percentage of the original oil in place (OOIP).

Table 2 – Design variable restrictions- ASP optimization

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

As illustrated in Fig. 5, 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 4150 ft., has an average initial pressure of 1770 psi, and the porosity is assumed to be constant throughout the reservoir and equal to 0.3. The numerical grid is composed of 19x19x3 blocks in the x, y and z directions. The OOIP is 395,427 bbls, the crude oil viscosity is 40 cp, the initial brine 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 program.

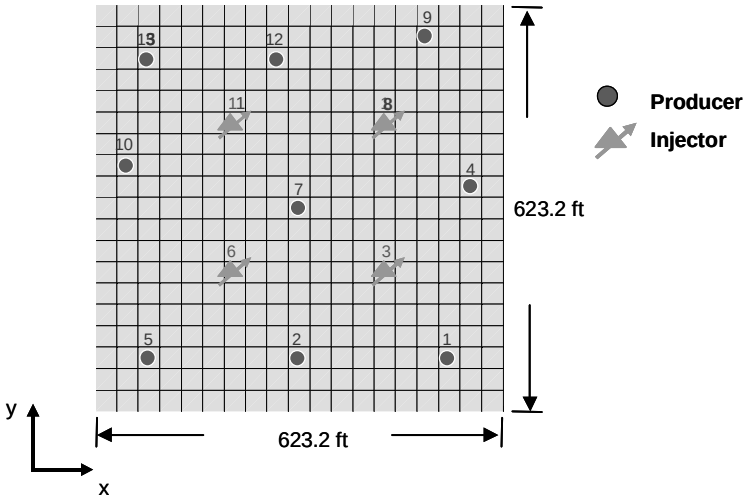

Figure 5. Well pattern illustration (ASP modeling case study).

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.

The statistic b will be calculated using one (1) Latin hypercube design with forty one (41) training points, and five (5) sets of prediction points with ten (10) points each. The target is made equal to the best solution in the training set (30.13%).

V. Results and Discussion

A. Branin & Hoo Case Study

Table 4 shows the estimated kriging parameters for the evaluated configurations, namely, the mean μ and the correlation parameters.

Table 4 – Estimated kriging parameters for the evaluated configurations (Branin & Hoo case study)

Config.	q_1	q_2	m	$s^2 \times 10^3$
1	0.0386	1.0865	74.4185	2.5067
2	0.0245	0.0234	76.2245	8.0430
3	0.0232	0.0184	109.6764	6.0829
4	0.0388	0.0146	74.0953	3.0256
5	0.0554	0.0086	82.5646	4.9569
6	0.0003	5.1336	53.5383	4.1119
7	0.0203	0.0357	64.8435	4.0426
8	0.0536	0.0122	43.6384	1.5753
9	0.0438	0.0041	151.6764	13.070
10	0.0376	0.0043	155.5163	15.472

Table 5 shows the probabilities of improvement over the target for each of the ten (10) configurations assuming the test points correlated, and uncorrelated (independent) using both analytical and empirical prediction variance estimates. An illustration of the probabilities of improvement is shown in Fig. 6. In general, when the output values at the prediction locations were assumed to be *independent*, the probabilities of improvement were found to be smaller when the analytical variances were used. When the correlation between the output values at the prediction locations were accounted for the probabilities of improvement were even smaller.

Table 5 - Probability of improvement for the evaluated configurations (Branin & Hoo case study)

Config.	Independent (empirical)	Independent (analytical)	Dependent
1	0.5575	0.4821	0.4541
2	0.5003	0.4514	0.3546
3	0.9737	0.9933	0.9927
4	0.3956	0.3663	0.3248
5	0.5208	0.5272	0.4859
6	0.8038	0.6809	0.6576
7	0.0889	0.0722	0.0727
8	0.5697	0.5965	0.5673
9	0.9064	0.9807	0.9694
10	0.4707	0.3945	0.3158

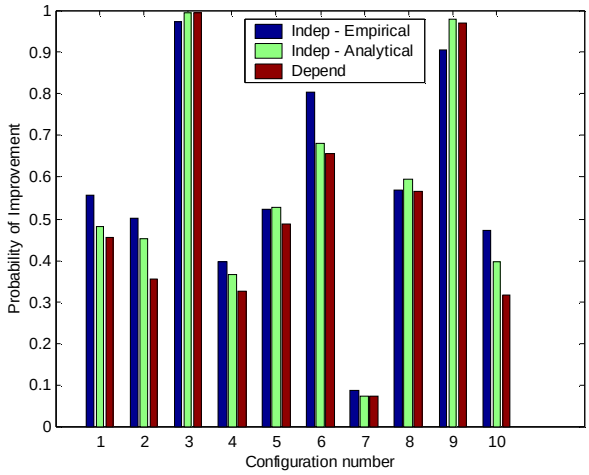

Figure 6 - Probability of improvement for the evaluated configurations (Branin & Hoo case study)

Table 6 shows for each of the configurations, the target, the statistic *b* (i.e., the number of prediction points below the target), the expected value of B, its standard deviation, and p-values, considering both analytical and empirical prediction variance estimates. Note the high p-values obtained for all the configurations, which shows that the computed probabilities of improvement are consistent with the experimental results. Furthermore, while the sample of configurations is not big enough to make a definite statement, the average probability of improvement is in excellent agreement with the observed results (i.e., 5 out of 10 configurations have prediction points below the target).

Table 6 – p-values comparison for the evaluated configurations (Branin & Hoo case study)

Config.	T	b	$E_a - E_e$	$SD_a - SD_e$	$p\text{-}value_a - p\text{-}value_e$
1	7.0366	2	0.5997 - 0.7334	0.7123 - 0.7763	0.2161 - 0.3102
2	2.5250	0	0.4992 - 0.5700	0.6031 - 0.6330	1.0 - 1.0
3	1.0355	0	0.9930 - 0.9769	0.0919 - 0.1842	0.0155 - 0.0576
4	1.7308	1	0.3257 - 0.3631	0.4686 - 0.4809	0.6514 - 0.7261
5	2.6180	0	0.4985 - 0.4988	0.5000 - 0.5000	1.0 - 1.0
6	0.9219	0	0.9797 - 1.3628	0.8650 - 0.9875	0.6532 - 0.4009
7	2.6751	1	0.0698 - 0.0863	0.2548 - 0.2807	0.1396 - 0.1725
8	0.4168	0	0.5649 - 0.5531	0.4966 - 0.5076	0.8711 - 0.9044
9	11.8487	2	1.4585 - 1.2922	0.5528 - 0.6499	0.9743 - 0.8000
10	15.3256	1	0.4050 - 0.5013	0.5676 - 0.6133	0.7288 - 0.8767

B. Alkali-Surfactant-Polymer (ASP) Modeling

Table 7 shows the estimated kriging parameters for the evaluated configurations, namely, the mean μ and the correlation parameters.

Table 7 – Estimated kriging parameters for the evaluated configurations (ASP case study)

Config.	q_1	q_2	q_3	q_4	m	s^2
1	3.5691×10^{5}	0.9562×10^{5}	0.00016×10^{-4}	0.9572×10^{-4}	22.278	18.020

Table 8 shows the probabilities of improvement for each of the five (5) configurations assuming the test points correlated, and uncorrelated (independent) using both analytical and empirical prediction variance estimates. The probabilities of improvement over the target followed the same trends observed in the analytical case study.

Table 8 - Probability of improvement for the evaluated configurations (ASP case study)

Config.	Independent (empirical)	Independent (analytical)	Dependent
1	0.7365	0.5806	0.5800
2	0.5279	0.0308	0.0303
3	0.5200	0.1044	0.1027
4	0.6971	0.3251	0.3248
5	0.4554	0.0532	0.0523

Table 9 shows for each of the configurations, the target, the statistic b (the number of prediction points below the target), the expected value of B, standard deviation, and p-values, considering both analytical and empirical prediction variance estimates. Note the high p-values obtained for all the configurations, which shows that the computed probabilities of improvement are again consistent with the experimental results. An illustration of the p-values for each of the configurations is shown in Fig. 5.

Table 9 - p-values comparison for the evaluated configurations (ASP case study)

Config.	b	$E_a - E_e$	$SD_a - SD_e$	$p\text{-value}_a - p\text{-value}_e$
1	0	0.728 – 1.065	0.716 – 1.065	0.839 – 0.527
2	1	0.031 – 0.676	0.174 – 0.743	0.06 – 1.0
3	1	0.105 – 0.637	0.311 – 0.694	0.20 – 1.0
4	0	0.337 – 1.047	0.501 – 0.901	1.0 – 0.60
5	0	0.054 – 0.566	0.229 – 0.701	1.0 – 1.0

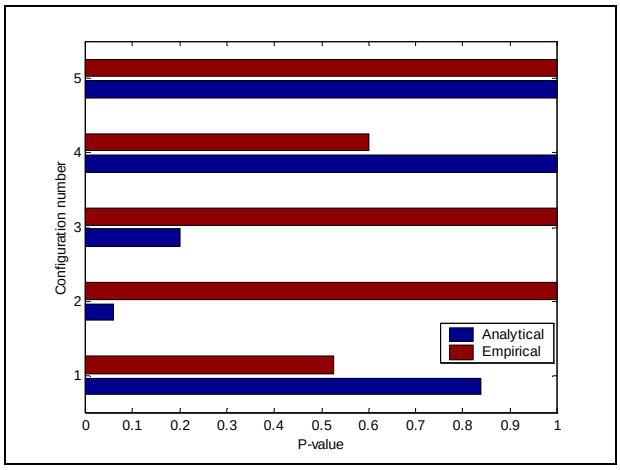

Figure 9 - p-values comparison for the evaluated configurations (ASP case study)

VI. Summary and Conclusions

- This paper presented an approach based on the Gaussian process perspective for assessing the merit of investing in another cycle of analysis (probability of improvement) versus accepting the present best solution (PBS). The Gaussian processes (GP) perspective to surrogate modeling has a long history in

the field of statistics, are frequently used for problems of regression (e.g., kriging) and classification and are closely related to a variety of surrogate modeling approaches including neural networks, kriging, generalized radial basis functions, and kernel methods.

- The proposed approach relies on three components: i) a covariance model (structure and parameters) obtained from available input/output data, ii) a surrogate model such as those provided by polynomial regression, kriging, and support vector regression, and iii) the assumption that the points in the next cycle are a realization of a Gaussian process with a covariance matrix and mean specified using i) and ii).
- Using ordinary kriging as modeling approach, considering estimates of both analytical and empirical variances, the proposed approach gave results statistically consistent when applied to an analytical case study (Branin-Hoo) and to the surrogate-based optimization of an Alkali-Surfactant-Polymer process, and holds promise to be effective in broader contexts.
- Current work focuses on evaluating the proposed approach using popular alternative surrogate modeling schemes such as polynomial regression, and support vector regression, and, on developing strategies for setting reasonable targets in another cycle in surrogate-based optimization, by using, for example, concepts from extreme value theory from statistics.

Acknowledgments

This material is based upon work supported by National Science Foundation under Grant DDM-423280, and the Fondo Nacional de Ciencia, Tecnología e Innovación (FONACIT), Venezuela under Grant F-2005000210. The authors also thank the Center for Petroleum and Geosystems Engineering of The University of Texas at Austin for providing the UTCHEM compositional simulator.

.

VII. References

1. Simpson, T.W., Booker, A.J., Ghosh, D., Giunta, A.A., Koch, P.N., and Yang, R.-J., "Approximation Methods in Multidisciplinary Analysis and Optimization: A Panel Discussion," *3rd ISSMO/AIAA Internet Conference on Approximations in Optimization*, Oct. 14-18, 2002.

2. Li, W., and S. Padula, "Approximation Methods for Conceptual Design of Complex Systems," *in Eleventh International Conference on Approximation Theory*, Edited by Chui C, Neaumtu M, Schumaker L, May 2004.

3. Queipo, N., R. Haftka, W. Shyy, T. Goel, R. Vaidyanathan and P. Kevin Tucker, "Surrogate-Based Analysis and Optimization," *Journal of Progress in Aerospace Sciences*, vol. 41, 2005, pp. 1-28.

4. Rodriguez, J.F, J. E. Renaud, and L. T. Watson. "Trust Region Augmented Lagrangian Methods for Sequential Response Surface Approximation and Optimization", *ASME J. Mech. Design*, Vol. 120, 1998, pp. 58-66.

5. Alexandrov, N., "A Trust Region Framework For Managing the Use of Approximation Models in Optimization," *Structural Optimization*, Vol. 15, No. 1, 1998, pp. 16-23.

6. Wojtkiewicz, S.F., M.S. Eldred, R.V. Field, Jr., A. Urbina, and J.R. Red-Horse. "A Toolkit for Uncertainty Quantification in Large Computational Engineering Models," *Proceedings of the 42nd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, Seattle, WA, AIAA-2001-1455, April 16-19, 2001.

7. Eldred, M. S., Giunta, A. A., Wojtkiewicz, S. F., Jr. and Trucano, T. G., "Formulations for Surrogate-Based Optimization Under Uncertainty," *9th AIAA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Atlanta, GA, AIAA-2002-5585, September 4-6, 2002.

8. Padula, S.L., Korte, J.J., Dunn, H.J., Salas, A., "Muldisciplinary Optimization Branch Experience Using iSIGHT Software", NASA/TM-1999-209714, Langley Research Center, Hampton, Virginia, November 1999.

9. Koch, P. N. and Gu, L., 2001, "Addressing Uncertainty using the iSIGHT Probabilistic Design Environment," *First Annual Probabilistic Methods Conference*, Newport Beach, CA, June 18-19, 2001.

10. Kageyama, Y., Q. Yu, and M. Shiratori, "Robust and Optimal Parameter Design Using SDSS and Reliability Design and Application to Hydraulic Servo Valve", *Computer Aided Optimum Design of Structures VII*, S. Hernandez and C. A. Brebbia, eds., WIT Press, Southampton, UK, 2001, pp. 13-22.

11. Zerpa, L., N. Queipo, S. Pintos and J. Salager, "An Optimization Methodology of Alkaline-Surfactant-Polymer Flooding Processes Using Field Scale Numerical Simulation and Multiple Surrogates," *Journal of Petroleum Science and Engineering*, 47, 2005, pp. 197-208.

12. Neal, R., *Bayesian Learning for Neural Networks*, Springer, New York, Lecture Notes in Statistics, 118, 1996.

13. Cressie, N., *Statistics for Spatial Data*, Wiley, New York, 1993.

14. Chiles, J.P., Delfiner, P., *Geostatistics: Modeling Spatial Uncertainty*, Wiley-Interscience, 1999.

15. Poggio, T., and Girosi, F., "A Theory of Networks for Approximation and Learning", Technical report A.I., 1140, M.I.T.

16. Lowe, D., *Similarity Metric Learning for a Variable Kernel Classifier, Neural Computation*, 7, 72-85.

17. Rasmussen, C., "Evaluation of Gaussian Processes and Other Methods for Non-Linear Regression", Ph.D. Thesis, Department of Computer Science, University of Toronto, 1996.

18. Hollingsworth, P., Mavris, D., "A Technique for Use of Gaussian Processes in Advanced Meta-Modeling," *Proceedings of the SAE Aerospace Congress and Exhibition*, Montreal, Canada, September, 2003.

19. Dosher, T.M., and F. A. Wise, "Enhanced Oil Recovery Potential. An Estimate," *Paper SPE 5800, J. Petroleum Technology*, May 1976, p. 575.
20. Lake, L., *Enhanced Oil Recovery*, Englewood Cliffs, NJ: Prentice Hall, p. 408, 1989.
21. Sacks, J., Welch, W., Mitchell, T., Wynn, H., "Design and Analysis of Computer Experiments", *Statistical Science*, 4, Nov. 1989, pp. 409-423.
22. Williams, C., "Prediction with Gaussian Processes: From Linear Regression to Linear Prediction and Beyond". Ed. M. Jordan, Learning in Graphical Models, Kluwer Academic, 1998, pp. 599-621.
23. Rao, C.R., *Linear Statistical Inference and Its Applications*, Second Edition, New York: John Wiley & Sons, Inc., 2002, pp. 516-533.
24. Genz, A. (1993), "A Comparison of Methods for Numerical Computation of Multivariate Normal Probabilities", *Computing Science and Statistics* 25, pp. 400-405.
25. Williams, C., Gaussian processes, *The Handbook of Brain Theory and Neural Networks*, Second Edition, Ed. M. Arbib, MIT press, 2002.
26. Wahba, G., *Spline Models for Observational Data*, SIAM, CBMS-NSF Regional Conference Series in Applied Mathematics, 1990.
27. Vecchia, A., "Estimation and Model Identification for Continuous Spatial Processes," *J. Roy. Statist. Soc. Ser. B*, 50, 1998, pp. 297-312.
28. Dixon, L.C.W. and Szeg, G.P., *The Global Optimization Problem: An Introduction,* North-Holland, Amsterdam 1978.
29. Den Hertog, D., Kleijnen, J., Siem, A., "The Correct Kriging Variance Estimated by Bootstrapping", *Journal of Operational Research Society*, 57, 4, April 2006, pp. 400-409.
30. Efron, B., and Tibshirani, R., *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.
31. Lophaven, S., Nielsen, H., and Sondergaard, J. (2002), "DACE: A Matlab Kriging Toolbox, Version 2.0". Informatics and Mathematical Modeling, Technical University of Denmark. http://www.imm.dtu.dk/~hbn/dace/