11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference
6 - 8 September 2006, Portsmouth, Virginia

AIAA 2006-7047

Performance Estimate and Simultaneous Application of Multiple Surrogates

Tushar Goel[1*], Raphael T Haftka[1†], Nestor V Queipo[2‡], and Wei Shyy[3§]
[1]*University of Florida, Gainesville, US 32611*

[2]*University of Zulia, Maracaibo, Venezuela*

[3]*University of Michigan, Ann Arbor, US 48109*

Abstract

A typical approach in surrogate-based modeling is to assess the performance of alternative surrogate models and select the model that performs the best. In this paper, we extend the utility of an ensemble of surrogates to: i) identify regions of high uncertainties at locations where predictions of surrogates widely differ, and ii) provide a more robust approximation approach. We explore the possibility of using the best surrogate or a weighted average surrogate model instead of individual surrogate models. The weights associated with each surrogate model are determined based on the errors in surrogates. We demonstrate the advantages of an ensemble of surrogates using analytical problems and an engineering problem of radial turbine design for space launch vehicle. We show that for a single problem the choice of the surrogate can be substantially influenced by the design of experiments.

I. Introduction

Surrogate models have been extensively used in the design and optimization of computationally expensive problems. Different surrogate models have been shown to perform well in different conditions. Barthelemy and Haftka[1] reviewed the application of meta-modeling techniques in structural optimization. Sobeiszczanski-Sobieski and Haftka[2] reviewed different surrogate modeling applications in multi-disciplinary optimization. Giunta and Watson[3] compared polynomial response surface approximations and Kriging on analytical example problems of varying dimensions. Simpson et al.[4] reviewed different surrogates and gave recommendations on the usage of different surrogates for different problems. Jin et al.[5] compared different surrogate models based on multiple performance criteria such as accuracy, robustness, efficiency, transparency and conceptual simplicity. They recommended using radial basis function for high-order nonlinear problems, Kriging for low-order nonlinear problems in high dimension spaces and polynomial response surfaces for low-order nonlinear problems. They also noted difficulties in constructing different surrogate models. Li and Padula[6] and Queipo et al.[7] recently reviewed different surrogate models used in the aerospace industry.

There are also a number of studies comparing different surrogates for specific applications. Papila et al.[8], Shyy et al.[9], Vaidyanathan et al.[10], Mack et al.[11] presented studies comparing radial basis neural networks and response surfaces while designing the liquid rocket injector, supersonic turbines, and the shape of bluff body for mixing enhancement. For crashworthiness optimization, Stander et al.[12] compared polynomial response surface approximation, Kriging and neural networks while Fang et al.[13] compared polynomial response surface approximation and radial basis functions. As expected, no single surrogate model is superior in general.

While most researchers have primarily been concerned with the choice among different surrogates, there has been relatively little work about use of an ensemble of surrogates. Zerpa et al.[14] presented one application of using an ensemble of surrogates to construct weighted average surrogate model for the optimization of an alkali-surfactant-polymer flooding process. They suggested that the weighted average surrogate model has better modeling capabilities than individual surrogates.

Typically the cost of obtaining data required for developing surrogate models is high, and it is desired to extract as much information as possible from the data. Using an ensemble of surrogates, which can be constructed without a significant expense compared to the cost of acquiring data, can prove effective in distilling

[*] PhD Candidate, Student Member AIAA
[†] Distinguished Professor, Fellow AIAA
[‡] Professor
[§] Clarence L "Kelly" Johnson Collegiate Professor, Fellow AIAA

correct trends from the data and may protect against bad surrogate models. Averaging surrogates is one approach motivated by our inability to find a unique solution to the non-linear inverse problem of identifying the model from a limited set of data (Queipo et al.[7]). In this context, model averaging essentially serves as an approach to account for model uncertainty. In this work, we explore methods to exploit the potential of use of an ensemble of surrogates. Specifically, we present the following two aspects:

i. Ensemble of surrogates can be used to identify regions where we expect large uncertainties (contrast)
ii. Use of an ensemble of surrogates *via* weighted averaging (combination) or selection of best surrogate model based on error statistics for more robust approximation than individual surrogates

We demonstrate the advantages of an ensemble of surrogates using analytical problems and an engineering problem of radial turbine design for space launch vehicle. This paper is organized as follows: In the next section, we present a method to use an ensemble of surrogates to identify the regions with large uncertainty, and the conceptual framework of constructing weighted average surrogate models. Thereafter we discuss the test problems, numerical procedure and results supporting our claims. We close the paper by recapitulating salient points presented.

II. Conceptual Framework

A. Identification of Region of Large Uncertainty

Surrogate models are used to predict the response in unsampled regions. There is an uncertainty associated with the predictions. An ensemble of surrogates can be used to identify the regions of large uncertainty. The concept is described as follows: Let there be N_{SM} surrogate models. We compute the standard deviation of the predictions at a design point $\mathbf{x}$ as,

$$s_{resp}\left(\hat{y}(\mathbf{x})\right)=\sqrt{\frac{\sum_{i=1}^{N_{SM}}\left(\hat{y}_i(\mathbf{x})-\overline{y}_i(\mathbf{x})\right)^2}{N_{SM}-1}} \quad (1)$$
$$\text{where}\quad \overline{y}_i(\mathbf{x})=\frac{\sum_{i=1}^{N_{SM}}\hat{y}_i(\mathbf{x})}{N_{SM}}$$

The standard deviation of the predictions will be high in regions where the surrogates differ greatly. A high standard deviation may indicate a region of high uncertainty in the predictions of any of the surrogates, and additional sampling points in this region can reduce that uncertainty. Note that while high standard deviation indicates high uncertainty, low standard deviation does not guarantee high accuracy. It is possible for all surrogate models to predict similar response (yielding low standard deviation) yet perform poorly in a region.

B. Weighted Average Surrogate Model Concept

We develop a weighted average surrogate model as,

$$\hat{y}_{wt.avg.}(\mathbf{x})=\sum_{i}^{N_{SM}}w_i(\mathbf{x})\hat{y}_i(\mathbf{x}) \quad (2)$$

where, $\hat{y}_{wt.avg.}(\mathbf{x})$ is the predicted response by the weighted average of surrogate models, $\hat{y}_i(\mathbf{x})$ is the predicted response by the i^{th} surrogate model and $w_i(\mathbf{x})$ is the weight associated with the i^{th} surrogate model at design point $\mathbf{x}$. Furthermore, the sum of the weights must be one $\left(\sum_{i=1}^{N_{SM}}w_i=1\right)$ so that if all the surrogates agree, $\hat{y}_{wt.avg.}(\mathbf{x})$ will also be the same.

A surrogate model, deemed more accurate, should be assigned a large weight, and conversely, a less accurate model should have lower influence on the predictions. The confidence in surrogate models is given by different measures of "goodness" (quality of fit) which can be broadly characterized as (i) global versus local measures and (ii) measures based on surrogate models versus measures based on data. Weights associated with each surrogate based on the local measures of goodness are function of space $w_i=w_i(\mathbf{x})$; for example, weights based on the pointwise model error estimates like prediction variance, mean squared error (surrogate based), or weights based on the interpolated cross-validation errors (data based). When weights are selected based on the basis of global measures of goodness, they are fixed in design space $w_i(\mathbf{x}) = C_i, \forall \mathbf{x}$; for example, weights based on RMS error $\hat{\sigma}$ for polynomial response surface approximation, process variance for Kriging (surrogate based), or weights based on cross-validation error (data based). While variable weights may capture local behavior better than constant weights, reasonable selection of weight functions is a formidable task.

Zerpa et al.[14] constructed a local weighted average model from three surrogates (polynomial response surface approximation, Kriging and radial basis functions) for the optimization of an alkali surfactant-polymer flooding process. Their approach was based on the pointwise estimate of the variance predicted by the three surrogate models.

There are different strategies of selecting weights. A few can be enumerated as follows:

2.2.1 Non-parametric Surrogate Filter (NPSF)

Weights are a function of relative magnitude of (global data-based) errors. The weight associated with i^{th} surrogate is given as:

$$w_i = \frac{\sum_{j=1, j\neq i}^{N_{SM}} E_j}{\left(N_{SM}-1\right)\sum_{j=1}^{N_{SM}} E_j} \tag{3}$$

where E_j is the global data-based error measure for j^{th} surrogate model. This choice of weights gives only a small premium to the better surrogates when N_{SM} is large. For example, the best surrogate has a weight equal to or less than $1 \big/ \left(N_{SM}-1\right)$, which becomes unreasonably low when N_{SM} is large. On the positive side the weights selected this way protect against errors induced by the surrogate models which perform extremely well at the sampled data points but give poor predictions at unsampled locations.

2.2.2 Best PRESS for Exclusive Assignments

Traditional method of using an ensemble of surrogates is to select the best model among all considered surrogate models. However, once the choice is made, it is usually kept even as the design of experiment is refined. If the choice is revisited for each new design of experiment, we consider it as a weighting scheme where the model with least (global data-based) error is assigned a weight of one and all other models are assigned zero weight. In this study, we call this strategy the "*best PRESS* model".

2.2.3 Parametric Surrogate Filter (PSF)

As discussed above, there are two issues associated with the selection of weights: (i) weights should reflect our confidence in the surrogate model, and (ii) weights should filter out adverse effects of the model which represents the data well but performs poorly in unexplored regions. A strategy to select weights which addresses both issues can be formulated as follows:

$$\begin{aligned} w_i^* &= \left(E_i + \alpha E_{avg}\right)^{\beta}, \qquad w_i = w_i^* \Big/ \sum_i w_i^* \\ E_{avg} &= \sum_{i=1}^{N_{SM}} E_i \Big/ N_{SM}; \qquad \beta < 0, \alpha < 1 \end{aligned} \tag{4}$$

This weighting scheme requires the user to specify two parameters α and β which respectively control the importance of averaging and importance of individual surrogate. Small values of α and large negative values of β impart high weights to the best surrogate model. Large α values and small negative β values represent high confidence in the averaging scheme. In this study, we have used $\alpha = 0.05 \text{and} \beta = -1$. The sensitivity to these parameters is studied in a section on parameter sensitivity.

The above-mentioned formulation of weighting schemes is used with generalized mean square cross-validation error (GMSE) (leave-one-out cross validation or *PRESS* in polynomial response surface approximation terminology), defined in the Appendix, as global data-based error measure, by replacing E_j by

$\sqrt{GMSE_j}$. We have used three surrogate models, polynomial response surface approximation (PRS), Kriging (KRG) and radial basis neural networks (RBNN) (Orr[15]), to construct the weighted average surrogate model. The parametric weighted surrogate model (PWS) can then be given as follows:

$$\hat{y}_{pws} = w_{prs}\hat{y}_{prs} + w_{krg}\hat{y}_{krg} + w_{rbnn}\hat{y}_{rbnn} \tag{5}$$

where weights are selected according to the parametric surrogate filter PSF (Equation (4)). The rationale behind selecting these surrogate models to demonstrate the proposed approach was (i) these surrogate models are commonly used by practitioners and (ii) they represent different parametric and non-parametric approaches (Queipo et al.[7]).

The cost of constructing surrogate models is usually low compared to that of analysis. If this cost is not small (for example, when using a Kriging model and GMSE for large data sets), the user may want to explore surrogate models that provide a compromise solution between accuracy and construction cost. In general, the choice of surrogate models which are most amenable to averaging and uncertainty identification remains a question of future research (Sanchez et al.[16]).

Since global measures of error depend on the data and design of experiments, weights implicitly depend on the choice of the design of experiments. This dependence can be seen from Figure 1 where we show boxplots of weights obtained for 1000 instances of Latin hypercube sampling (LHS) design of experiments (DOEs) for Camelback function (described in next section). The center line of each boxplot shows the 50th-percentile (median) value and the box encompasses the 25th - and 75th -percentile of the data. The leader lines (horizontal lines) are plotted at a distance of 1.5 times the inter-quartile range in each direction or the limit of the data (if the limit of the data falls within 1.5 times inter-quartile range). The data points outside the horizontal lines are shown by placing a '+' sign for each point.

We can see that the weights for different surrogates vary over a wide range with DOEs. The weights also give an assessment of relative contribution of different surrogate models to the weighted average surrogate model. In this example polynomial response surface approximation had the highest weight most of the time (880 times) but not all the times (59 times Kriging had the highest weight and 61 times RBNN had the highest weight).

III. Test Problems, Numerical Procedure, and Prediction Metrics

A. Test Problems

To test the predictive capabilities of the proposed approach of using an ensemble of surrogates, we employ two types of problems: (i) analytical (Dixon-Szegö[17]) which are often used to test global optimization methods, and (ii) industrial: a radial turbine design problem (Mack et al.[18]) motivated by space launch. The details of each test problem are given as follows:

(i) Branin-Hoo Function

$$x \in [-5,10], \qquad y \in [0,15]$$

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

(ii) Camelback Function

$$x \in [-3,3], \qquad y \in [-2,2]$$

$$f(x,y) = \left(4 - 2.1x^2 + \frac{x^4}{3}\right)x^2 + xy + \left(-4 + 4y^2\right)y^2 \tag{7}$$

(iii) Goldstein-Price Function

$$x, y \in [-2,2]$$

$$f(x,y) = \left[1 + (x+y+1)^2\left(19 - 4x + 3x^2 - 14y + 6xy + 3y^2\right)\right] \times \tag{8}$$

$$\left[30 + (2x - 3y)^2\left(18 - 32x + 12x^2 + 48y - 36xy + 27y^2\right)\right]$$

Figure 2 depicts these two-variable test problems and shows zones of high gradients.

(iv) Hartman Functions

$$f(\mathbf{x}) = -\sum_{i=1}^{m} c_i \exp\left\{-\sum_{j=1}^{n} a_{ij}\left(x_j - p_{ij}\right)^2\right\}$$
$$where\,\mathbf{x} = \left(x_1, x_2, \ldots, x_n\right) \quad x_i \in [0,1] \tag{9}$$

Two instances of this problem are considered based on the number of design variables. For the chosen examples, $m = 4$.

(a) Hartman3

This problem has three variables. The choice of parameters is given in Table 1 (Dixon-Szegö[17]).

(b) Hartman6

This instance of the problem has six design variables and the parameters used in the function are tabulated in Table 2 (Dixon-Szegö[17]).

Figure 3 illustrates the complexity of the analytical problems. It shows the boxplots of function values at a uniform grid of points with 21 points in each direction (for Hartman problem with six variables we used 5 points in each direction); the mean, coefficient of variation and median are given in Table 3. We can see that for all the problems the coefficient of variation was close to one or more which indicates large variation in the function values. It is clear from Figure 3 that the function values follow non-uniform distribution which is also reflected by large differences in the mean and median. These conditions translate into high gradients in the functions and may pose difficulties in accurate modeling of the responses. Goldstein-Price and Hartman problem with six variables had a significant number of points which had higher function values than the inter-quartile range of the data. This is reflected in high coefficient of variation of these two functions.

(v) Radial Turbine Design for Space Launch

As described by Mack et al.[18], this six-variable problem is motivated by the design of compact radial turbine used to drive pumps that deliver liquid hydrogen and liquid oxygen to combustion chamber of a spacecraft. The objective of the design is to increase the work output of a turbine in the liquid rocket expander cycle engine while keeping the overall weight of the turbine low. If the turbine inlet temperature is held constant, the increase in turbine work is directly proportional to the increase in efficiency. Thus the design goal is to maximize the turbine efficiency while minimizing the turbine weight. Our interest in this problem is to develop accurate surrogate model(s) of the efficiency as a function of six design variables. The description of design variables and their corresponding ranges are given in Table 4 (Mack et al.[18]).

The objectives of the design were calculated using a one-dimensional flow analysis “Meanline” code (Huber[19]). Mack et al.[18] identified the appropriate region of interest by iteratively refining the design space. They also identified the most important variables using global sensitivity analysis.

B. Numerical Procedure

For all analytical problems, Latin hypercube sampling (LHS) was used to pick design points such that the minimum distance between the design points is maximized. We used Matlab[20]® routine *lhsdesign* with *maximin* criterion (maximize the minimum distance between points) and a maximum of *20 iterations* to obtain optimal configuration of points. For the radial turbine design problem, Mack et al.[18] sampled 323 designs in the six-dimensional region of interest, using LHS and a five level factorial design on the three most important design variables (identified by global sensitivity analysis). Out of these 323 designs, 13 designs were found infeasible. The remaining 310 design points were used to construct and test the surrogate models. For this study, we randomly select 56 points to construct the surrogate model and use the remaining 254 points to test the surrogate model. To reduce the effect of random sampling for both analytical and radial turbine design problems we present results based on 1000 instances of design of experiments for all the problems in low dimension spaces. However to keep computational cost low for six-variable problems, we used 100 design of experiments and then used 1000 bootstrap (Hesterberg et al.[21]) samples to estimate results.

The numerical settings used to fit different surrogate models for each problem are given in Table 5. The total number of test points (on a uniform grid) is p^{N_v} where N_v is the number of variables and p is the number of points along each direction (Table 5) except for the radial turbine problem where the number represents the total number of test points. We used reduced-quadratic or reduced-cubic polynomials for PRS. A Gaussian correlation function and a linear trend model were used in Kriging approximation of all test problems. Parameters “Spread” and “Goal” for radial basis neural network were selected according to problem characteristics (*spread* controls the decay rate of radial basis function and *goal* is the desired level of accuracy of the RBNN model on training points). It should be pointed out that no attempt was made to improve the predictions of any surrogate model.

C. Prediction Metrics

The following metrics were used to compare the prediction capabilities of different surrogate models:

(i) Correlation Coefficient

The correlation coefficient between actual and predicted response at the test points $R(y,\hat{y})$ is given as

$$R(y,\hat{y}) = \frac{\frac{1}{V}\int_V (y-\bar{y})(\hat{y}-\bar{\hat{y}})dv}{\sigma(y)\sigma(\hat{y})} \tag{10}$$

It is numerically evaluated from the data for test points by implementing quadrature[**] for integration (Ueberhuber[22]) as given in (11).

$$\begin{aligned} &\frac{1}{V}\int_V y\hat{y}dv = \sum_{i=1}^{N_{test}} \gamma_i y_i \hat{y}_i \Big/ N_{test} \quad ;\bar{y} = \sum_{i=1}^{N_{test}} \gamma_i y_i \Big/ N_{test} \\ &\sigma(y) = \sqrt{\frac{1}{V}\int_V (y-\bar{y})^2 dv} = \sqrt{\sum_{i=1}^{N_{test}} \gamma_i (y_i - \bar{y}_i)^2 \Big/ N_{test}} \end{aligned} \tag{11}$$

where $\bar{y}$ is the mean of actual response, $\bar{\hat{y}}$ is the mean of predicted response, N_{test} is the number of test points, and γ_i is the weight used for integration using trapezoidal rule. For radial turbine problem, we used a non-uniform set of data points so the correlation coefficient is obtained using (11) with weight $\gamma_i = 1$. For a high quality surrogate model, the correlation coefficient should be as high as possible. The maximum value of $R(y,\hat{y})$ is one which defines exact linear relationship between the predicted and the actual response.

(ii) RMS Error

For all the test problems the actual response at test points was known, which allowed us to compute error at all test points. The root mean square error (RMSE) in the design domain, as defined in (12), was used to assess the goodness of the predictions.

$$RMSE = \sqrt{\frac{1}{V}\int_V (y-\hat{y})^2 dv} \tag{12}$$

Equation (12) can be evaluated using trapezoidal rule as denoted in (13).

$$RMSE = \sqrt{\sum_{i=1}^{N_{test}} \gamma_i (y_i - \hat{y}_i)^2 \Big/ N_{test}} \tag{13}$$

For radial turbine problem, we used (13) with weight $\gamma_i = 1$ to get RMS error. Of course, a good surrogate model gives low RMS error.

(iii) Maximum Error

Another measure of the quality of prediction of a surrogate is the maximum absolute error at the test points. This is required to be low. A combination of high correlation coefficient and low RMS and maximum error would indicate a good prediction.

[**] Here we used trapezoidal rule for integration.

IV. Results and Discussion

In this section, we present some numerical results to demonstrate the capabilities of multiple surrogate models using the test problems discussed in Section 3.

A. Identification of Zones of High Uncertainty

We demonstrate the application of an ensemble of surrogates to identify region of high uncertainty with the help of different test problems. Results for a single instance of a DOE for Branin-Hoo example are presented in detail. Figure 4 shows the contour plots of absolute errors in prediction ($\left|y(\mathbf{x})-\hat{y}(\mathbf{x})\right|$) due to different surrogate models and the standard deviation of the responses.

Figure 4(A)-(C) shows contour plots of actual absolute errors in different surrogate models. It can be seen that the middle section of the design space was approximated very well (errors are low) but the left boundary was poorly represented by different surrogate models. The errors (and hence responses) from PRS, KRG and RBNN differed in the region close to the top-left corner. The contour plot of the standard deviation (Figure 4(D)) of predicted responses correctly indicated the region of high uncertainty near the top-left corner due to high standard deviation. It also appropriately identified good predictions in the central region of design space. The predictions in the region of high uncertainty can be improved by sampling additional points.

It is also noted that although all the surrogate models had high errors near the bottom-left corner of the design space (Figure 4(A)-(C)), the standard deviation of the predicted responses was not high. This means that we can use the standard deviation of surrogate models to identify regions of high uncertainty but we can-not use it to identify regions of high fidelity. This particular situation demands further investigation if the objective of using an ensemble of surrogates was to identify region of high error in the predictions.

To further show the independence of the result with respect to design of experiments, we simulated the Branin-Hoo function with 1000 DOEs. For each DOE, we computed the standard deviation of responses in design space. At the *location of maximum standard deviation* for each DOE we computed actual errors in the predictions of different surrogates. Similarly, we calculated actual errors in the predictions of different surrogates at the *location of minimum standard deviation*. Figure 5(A) shows the magnitude of maximum standard deviation and actual errors in predictions using different surrogates for 1000 DOEs and Figure 5(B) shows the magnitude of minimum standard deviation and actual errors in predictions using different surrogates from 1000 DOEs. By comparing Figure 5(A) and (B), it is clear that high standard deviation of responses corresponded to the regions with large uncertainties in predictions and low standard deviation corresponded to regions with low uncertainty and there was an order of magnitude difference.

To generalize the findings, we simulated all test problems and identified the actual errors at the *locations of maximum and minimum standard deviation* of responses. The results are summarized in Table 6 and Table 7. A one-to-one comparison of results for different test problems show that when the standard deviation of responses was highest, the actual errors in predictions were high and when the standard deviation of responses was lowest, the actual errors in predictions were low. We note that the results are more useful for a qualitative comparison than quantitative; i.e., identifying the regions where we expect large uncertainties in prediction rather than quantifying the magnitude of actual errors.

We also estimated the maximum (over the entire design space) errors due to each surrogate model for different test problems and compared with the maximum standard deviation of responses. The results are presented in Table 8. While the maximum standard deviation of responses was same order of magnitude as the maximum actual error for all surrogate models, it underestimated the maximum error by a factor of 2.5 – 4.0. When the number of data points to construct the surrogate model was increased (Branin-Hoo function was modeled with 31 points and Camelback function was modeled with 40 points, refer to Section 6.2.5 for details about modeling) the underestimation of the maximum actual error was reduced.

The main conclusions of the results presented in this section are: (i) dissimilar predictions of surrogate models (high standard deviation of responses) indicate regions of high errors, (ii) similar predictions of surrogate models (low standard deviation of responses) do not necessarily imply small errors and, (iii) the maximum standard deviation of responses underestimates the actual maximum error.

B. Robust Approximation *via* Ensemble of Surrogates

Next, we demonstrate the need of robust approximation with the help of Table 9 that enlists the number of times each surrogate yields the least PRESS error for all test problems. As can be seen, no surrogate model is universally the best for all problems. Besides, for any given problem, the choice of best surrogate model is affected by the design of experiment (except radial turbine design problem). The results presented in Table 9 clearly establish the need to search approximation models which are robust (i.e. the same surrogate model can be applied to different problems, and the results produced are not significantly influenced by the choice of DOE).

We present results to reflect the advantages of using an ensemble of surrogates. We compare the parametric weighted surrogate (PWS) model and the surrogate model corresponding to the best generalization error among the three surrogates (*best PRESS* model) with individual surrogate models (PRS, Kriging and RBNN). For each problem, the summary of the results based on 1000 DOEs is shown with the help of boxplots. A small size of the box suggests small variation in results with respect to the choice of design of experiment.

6.2.1. Correlations

Figure 6 shows the correlation coefficient (between actual and predicted responses) for different test problems. The results were statistically significant (*p*-value is smaller than *1e-4*) for all problems and DOEs. It is evident that no single surrogate worked the best for all problems and correlation coefficient for individual surrogates varied with DOE. Both the *best PRESS* and the PWS models were better than the worst surrogate model and at par with the corresponding best surrogate for most problems. The PWS model generally performed better than the *best PRESS* model. The variation in results with respect to design of experiments for both the PWS model and the *best PRESS* model was also comparable to best surrogate for all problems except Hartman problem with six variables.

For all the problems we observed that some of the design of experiments (DOEs) yielded very poor correlations. Analysis of the corresponding experiments revealed two scenarios:

i. Some times the DOE was not satisfactory and a large portion of the design space was unsampled. This led to poor performance of all the surrogate models.
ii. For a few poor correlation cases, despite a good DOE, one or more surrogates failed to capture the correct trends.

The PWS model and the *best PRESS* model were able to correct the anomalies in these scenarios to some extent. The tail of the boxplot corresponding to the PWS model and the *best PRESS* model was shorter compared to the worst surrogate (Figure 6).

Table 10 shows the mean and the coefficient of variation for different test problems to assess the performance of different surrogate models. It is clear that the average correlation coefficient for the PWS model was either the best or the second best for all the test problems. Also the low coefficient of variation underscored the relatively low sensitivity of the PWS model with respect to the choice of design of experiments. Performance of the *best PRESS* model was also comparable to the best surrogate model for each problem. The overall performance of all three surrogates was comparable. It can also be seen from Table 10 that the PWS model outperformed the *best PRESS* model for all cases but radial turbine design problem.

The mean of the correlation coefficient for different problems is reported based on one set of 1000 DOEs. Since the distribution of mean is approximately Gaussian, the coefficient of variation of the mean (of correlation coefficient) can be given as $COV/\sqrt{N_{DOE}}$ where COV is the coefficient of variation (of correlation coefficient) based on 1000 DOEs ($N_{DOE}=1000$), leading to a coefficient of variation of the mean that is about 30 times lower than the native coefficient of variation. The number of digits in the table is based on this estimate of the coefficient of variation.

We verified the results by performing the bootstrap analysis (Hesterberg et al.[21] 2005) by considering 1000 samples of 1000 DOEs each. The distribution of the mean for one representative case (mean correlation coefficient predicted using Kriging approximation for Branin-Hoo function) is plotted in Figure 7. The mean correlation coefficient evidently follows Gaussian distribution as the data falls on the straight line depicting the normal distribution. Similar results were observed for all other cases. Bootstrapping also confirmed that the coefficient of variation of the mean value followed the simple expression given above.

6.2.2. RMS Errors

Next we compared different surrogate models based on the RMS errors in predictions at test points. Figure 8 shows the results on different test problems. While no single surrogate performed the best on all problems, individual surrogate models approximated different problems better than others. The parametric weighted surrogate (PWS) model and the *best PRESS* model performed reasonably for all test problems. The results indicate that if we know that a particular surrogate performs the best for a given problem, it is best to use that surrogate model for approximation. However, for most problems the best surrogate model is not known a priori or the choice of best surrogate may get affected by choice of DOE (Table 9). Then an ensemble of surrogates (*via* the PWS or the *best PRESS* model) may prove beneficial to protect against the worst surrogate model.

The mean and coefficient of variation of RMS errors using different surrogates on different problems are tabulated in Table 11. Note that, Kriging (most often) had the lowest RMS errors compared to other surrogates. When the RMS errors due to all surrogates were comparable, as was the case for Branin-Hoo and Camelback functions, the predictions using the PWS model were more accurate (lower RMS error) than any individual surrogate. However when one or more surrogate models were much more inaccurate than others, the predictions using the PWS model were only reasonably close to the accurate surrogate model(s). We also observed that both the *best PRESS* model and the PWS model were able to significantly reduce the errors compared to the worst surrogate. This suggests that using an ensemble of surrogate models, we can protect against poor choice of a surrogate.

The PWS model generally yielded lower RMS errors than the *best PRESS* model. Relatively poor performance of the PWS model (compared to the *best PRESS* model) for six variables Hartman problem and radial turbine problem was attributed to accurate modeling of the response by one surrogate or inaccuracy in the representation of weights (see section on the role of generalized cross-validation errors).

6.2.3. Maximum Absolute Errors

Figure 9 shows the maximum absolute error for 1000 DOEs using different surrogate models on different test problems. As was observed for RMS errors, the PWS model and the *best PRESS* model performed reasonably for all test problems though individual surrogate models performed better for different test problems.

Numerical quantification of the results is given in Table 12. The maximum absolute error obtained using the PWS model and the *best PRESS* model were comparable to the maximum absolute error obtained using the best surrogate model for that test problem. For most cases, the PWS model also delivered a lower maximum absolute error than the *best PRESS* model. Relatively poor performance of the PWS model for Goldstein-Price test problem was attributed to the poor performance of one of the surrogate models (RBNN) on the prediction points.

The results presented in this section suggest that the strategy of using an ensemble of surrogate models potentially yields robust approximation (good correlation, low RMS and maximum errors) for problems of varying complexities and dimensions and the results are less sensitive to the choice of DOE. The PWS model may have an advantage compared to the *best PRESS* model.

6.2.4. Studying the Role of Generalized Cross-validation Errors

We observed that the PWS model did not perform well for Camelback and Goldstein-Price function where RBNN model noticeably yielded large variations. To investigate the underlying issue, we studied the weights and hence the role of *PRESS* error which is used to determine the weights. Our initial assumption was that the *PRESS* error is a good estimate of the actual RMS errors for all surrogate models. To validate this assumption, we computed the ratio of actual RMS errors and *PRESS* for different surrogate models over 1000 DOEs. The results are summarized in Figure 10 and corresponding mean and standard deviation (based on 1000 DOEs) are given in Table 13.

It is observed from the results that *PRESS* (generalized cross-validation error) on average underestimated actual RMS errors for polynomial response surface approximation but overestimated RMS error in Kriging and RBNN. For Goldstein-Price the mean was skewed for RBNN because of three simulations which gave very large ratio of RMS error and *PRESS* (the median is 0.42). The implication of this under/over estimate was that the weights associated with polynomial response surface model were overestimated and weights for Kriging and radial basis neural network were underestimated. Noticeably, there were a large number of instances for Camelback and Goldstein-Price functions where PRESS underestimated the RMS errors for RBNN (see long tail of points with RMS error to PRESS ratio greater than two). This indicated wrong emphasis of RBNN model for these models compared to other more accurate surrogates and hence relatively poor performance of the parametric weighted surrogate model was observed. This anomaly in accurately representing the actual errors or developing measures to correct the weight to account for the over-/under-estimation is a scope of future research.

6.2.5. Effect of Sampling Density

Often an initial DOE identifies regions of interest, and then the DOE is refined in these regions. At other times, the initial DOE is found insufficient for good approximation, so that it must be refined. The refinement of the DOE can be carried out in two ways: (i) increasing the number of points in the original design space, and (ii) reducing the size of design space. The refinement of the DOE may change the identity of the best surrogate model, so that even if a single surrogate model is used, it may be useful to switch surrogates. Additionally the choice between *best PRESS* and the PWS model may depend on sampling density. To investigate these issues, we study two representative problems: Branin-Hoo function and Camelback function which were not adequately approximated by different surrogate models (low correlations). Both problems are now modeled with increased number of points (31 points were used for Branin-Hoo function and Camelback function was modeled with 40 points) such that all regions were adequately modeled. We used a cubic polynomial to model Branin-Hoo function and a quartic polynomial to model Camelback function. All other parameters were kept the same. The results obtained for the increased number of points were compared with the previously presented results for smaller number of points in Table 14 and Table 15.

As can be seen from Table 14 and Table 15, the predictions improved with increasing number of points. The improvement in Kriging (which models the local behavior better) was significantly more than the other two surrogates. The performance of both the *best PRESS* model and the PWS model was comparable to the best individual surrogate model and significantly better than the worst surrogate model. For the problems considered here, the *best PRESS* model outperformed the PWS model. This result is expected because of much improved modeling of the objective function by one or more of the surrogates. The results corroborate our earlier findings: (i) if we a priori know the best surrogate model for a given problem, that surrogate should be used for approximation and, (ii) ensemble of surrogates protects us against the worst surrogate model. These results were evident irrespective of the number of points used to model the response. However, we also note that even if a single surrogate is used, its choice depends on sampling density. For Branin-Hoo function with 12 points, the polynomial response surface approximation had the best correlation and lowest maximum error. Its mean RMS error is slightly higher than Kriging but standard deviation is much better. With 31 points Kriging is the best surrogate.

6.2.6. Sensitivity Analysis of PSF Parameters

To study the effect of variation in the parameters α and β (see (4)), we constructed the PWS model for Goldstein-Price function with different values of α and β. This problem was selected because of significant differences in the performance of different surrogate models. All other parameters were kept the same. The comparison of correlation coefficient and errors based on 1000 DOE samples is given in Table 16. To eliminate the skewness of the data due to a few spurious results, we show median, 1st and 3rd quartile data for all cases.

When we increased α keeping β constant, we observed modest decrease in errors. This was expected because by increasing α we reduced the importance of individual surrogates and assigned more importance to the averaging, which helped in reducing the effect of bad surrogates. However, it is noteworthy that a few designs which gave poor performance of one surrogate deteriorated the performance of the PWS model for respective cases. By increasing β keeping α constant, we emphasized the importance of individual surrogates more than the averaging. For this case, the overall effect was the deterioration of correlation and increase in errors. The effect of variation in β on the results was more pronounced than the effect of variation in α. The above results indicated that the parameters α and β should be chosen according to the performance of the individual surrogates.

V. Conclusions

In this paper, we presented a case to simultaneously use multiple surrogates (i) to identify regions of high uncertainty in predictions, and (ii) to develop a robust approximation strategy. The main findings of the paper can be summarized as follows.

i. Regions of high standard deviation in the predicted response of the surrogates correspond to high errors in the predictions of the surrogates. However we caution the user not to interpret the regions of low standard deviation (uncertainty) as regions of low error.

ii. The magnitude of the standard deviation of responses usually underestimates the error.

iii. Simultaneous use of multiple surrogate models can improve robustness of the predictions by reducing the impact of a poor surrogate model (which may be an artifact of choice of design of experiment or the inherent unsuitability of the surrogate to the problem). Two suggested ways of using an ensemble of surrogates were to construct parametric weighted surrogate model or to select the surrogate model which has the least PRESS error among all considered surrogate models.

iv. The proposed PRESS error based selection of multiple surrogates performed at par with the *best* individual surrogate model for all test problems and showed relatively low sensitivity to the choice of DOE, sampling density, and dimensionality of the problem.

v. The parametric weighted surrogate model yielded best correlation between actual and predicted response for different test problems.

vi. While different surrogates performed the best for reducing error (RMS and maximum absolute error) in different test problems, the performance of surrogate models was influenced by the selection of DOE. Ensemble of surrogates (*via* the parametric weighted surrogate and the *best PRESS* model) performed at par with the corresponding best surrogate model for all test problems. The parametric weighted surrogate model in general outperformed the surrogate model with best PRESS error.

vii. It was also observed that PRESS in general underestimated the actual RMS error for polynomial response surface approximation and overestimated the actual RMS error for Kriging and radial basis neural network. The correction in weights to account for the under-/over- estimate of RMS errors by PRESS is a scope of future research.

viii. Though the best individual surrogate can change with increase in sampling density, the ensemble of surrogates performs comparably with the best surrogate.

We conclude that for most practical problems, where the best surrogate is not known beforehand, use of an ensemble of surrogates may prove a robust approximation method.

VI. Acknowledgements

The present efforts have been supported by the Institute for Future Space Transport (under the NASA Constellation University Institute Program (CUIP) with Ms. Claudia Meyer as program monitor) and National Science Foundation (Grant # DMI-423280). Authors also acknowledge the reviewers for their thoughtful suggestions.

VII. References

[1.] Barthelemy, J-F.M., Haftka, R.T., Approximation Concepts for Optimum Structural Design – A Review, Structural Optimization, Vol. 5, 1993, pp.129-144.

[2.] Sobieszczanski-Sobieski, J., Haftka, R.T., Multidisciplinary Aerospace Design Optimization: Survey of Recent Developments, Structural Optimization, Vol. 14, 1997, pp. 1-23.

[3.] Giunta, A.A., Watson, L.T., A Comparison of Approximation Modeling Techniques: Polynomial versus Interpolating Models, In: 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis & Optimization St. Louis, MO, Vol. 1, 1998, pp. 392–404, AIAA-98-4758.

[4.] Simpson, T.W., Peplinski, J.D., Koch, P.N., Allen, J.K., Meta-Models for Computer Based Engineering Design: Survey and Recommendations, Engineering with Computers, Vol. 17, 2001, pp. 129-150.

[5.] Jin, R., Chen, W., Simpson, T.W., Comparative Studies of Meta-modeling Techniques Under Multiple Modeling Criteria, Structural and Multi-disciplinary Optimization, Vol. 23, 2001, pp. 1-13.

[6.] Li,W., Padula, S., Approximation Methods for Conceptual Design of Complex Systems, In: Eleventh International Conference on Approximation Theory (eds. Chui C, Neaumtu M, Schumaker L), May 18-24, 2004.

[7.] Queipo, N.V., Haftka, R.T., Shyy, W., Goel, T., Vaidyanathan, R., Tucker, P.K., Surrogate-Based Analysis and Optimization, Progress in Aerospace Sciences, Vol. 41, 2005, pp.1-28.

[8.] Papila, N., Shyy, W., Griffin, L.W., Dorney, D.J., Shape Optimization of Supersonic Turbines using Response Surface and Neural Network Methods, Journal of Propulsion and Power, Vol. 18, 2001, pp. 509-518.

[9.] Shyy, W., Papila, N., Vaidyanathan, R., Tucker, P.K., Global Design Optimization for Aerodynamics and Rocket Propulsion Components, Progress in Aerospace Sciences, Vol. 37, 2001, pp. 59-118.

[10.] Vaidyanathan, R., Tucker, P.K., Papila, N., Shyy, W., CFD Based Design Optimization for a Single Element Rocket Injector, Journal of Propulsion and Power, Vol. 20 (4), 2004, pp. 705-717 (also presented in 41st Aerospace sciences meeting and exhibit, Reno, NV, 2003).

[11.] Mack, Y., Goel, T., Shyy, W., Haftka, R.T., Queipo, N.V., Multiple Surrogates for Shape Optimization for Bluff-Body Facilitated Mixing, 43rd AIAA Aerospace Sciences Meeting and Exhibit, Jan 10-13 2005 (Reno, Nevada) AIAA 2005-0333.

[12.] Stander, N., Roux, W., Giger, M., Redhe, M., Fedorova, N., Haarhoff, J., A Comparison of Meta-modeling Techniques for Crashworthiness Optimization, In proceedings of 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Albany, NY, Aug 2004; AIAA-2004-4489.

[13.] Fang, H., Rais-Rohani, M., Liu, Z., Horstemeyer, M.F., A Comparative Study of Metamodeling Methods for Multi-objective Crashworthiness Optimization, Computers and Structures, Vol. 83, 2005, pp. 2121-2136.

[14.] Zerpa, L., Queipo, N.V., Pintos, S., Salager, J., An Optimization Methodology of Alkaline-Surfactant-Polymer Flooding Processes using Field Scale Numerical Simulation and Multiple Surrogates, Journal of Petroleum Science and Engineering, Vol. 47, 2005, pp. 197-208 (also presented at SPE/DOE 14th symposium on improved oil recovery, 2004, April 17-21, Tulsa).

[15.] Orr, M.J.L., Introduction to Radial Basis Neural Networks, Center for Cognitive Science, Edinburgh University, EH9LW, Scotland, UK. http://anc.ed.ac.uk/rbf/, 1996.

[16.] Sanchez, E., Pintos, S., Queipo, N.V., Toward and Optimal Ensemble of Kernel-based Approximations with Engineering Applications, In proceedings of IEEE World Conference on Computational Intelligence, Vancouver B.C., Canada, July 16-21, 2006, Paper ID. 1265.

[17.] Dixon, L.C.W., Szegö, G.P., Towards Global Optimization 2, North-Holland, Amsterdam, 1978.

[18.] Mack, Y., Shyy, W., Haftka, R.T., Griffin, L., Snellgrove, L. Huber, F., Radial Turbine Preliminary Aerodynamic Design Optimization for Expander Cycle Liquid Rocket Engine, In: 42nd AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit, Sacramento CA, 9-12 July, 2006, AIAA-2006-5046.

[19] Huber, F., Turbine Aerodynamic Design Tool Development, In: Space Transportation Fluids Workshop, Marshall Space Flight Center AL, 2001.

[20] MATLAB®, The Language of Technical Computing, Version 6.5 Release 13. © 1984-2002, The MathWorks Inc.

[21] Hesterberg, T., Moore, D.S., Monaghan, S., Clipson, A., Epsterin, R., Bootstrap Methods and Permutation Tests, W H Freeman, New York, Chapter 14, 2005.

[22] Ueberhuber, C.W., Numerical Computation 2: Methods, Software and Analysis, Springer-Verlag: Berlin, p.71, 1997.

[23] Myers, R.H., Montgomery, D.C., Response Surface Methodology-Process and Product Optimization using Designed Experiments, John Wiley & Sons, Inc: New York, 1995.

[24] Martin, J.D., Simpson, T.W., Use of Kriging Models to Approximate Deterministic Computer Models, AIAA Journal Vol. 43(4), 2005, pp. 853-863.

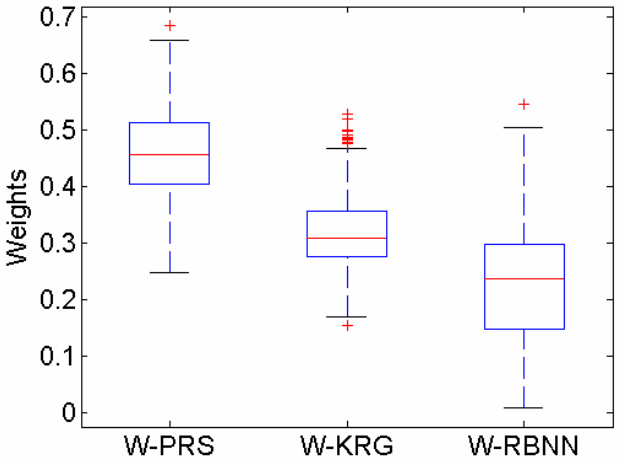
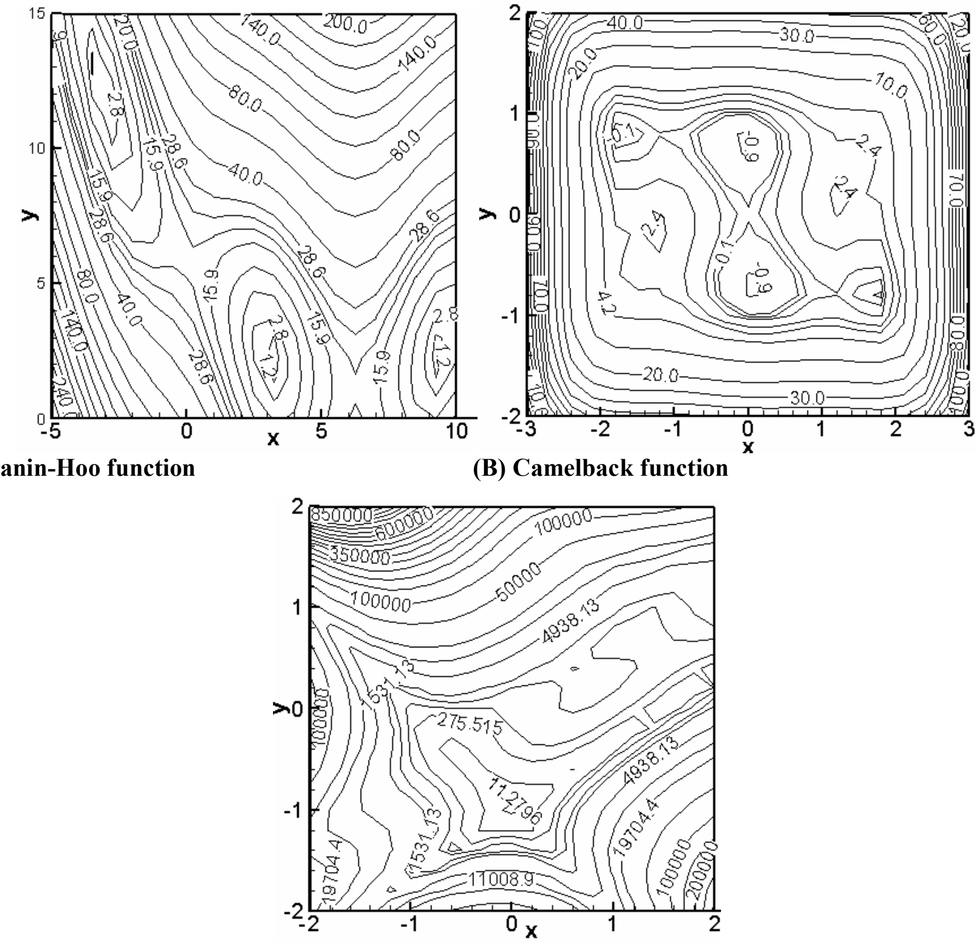


Figure 1 Boxplots of weights for 1000 DOE instances (Camelback function) W-PRS, W-KRG and W-RBNN are weights associated with polynomial response surface approximation, Kriging and radial basis neural network models respectively.



(A) Branin-Hoo function **(B) Camelback function**

(C) Goldstein-Price

Figure 2 Contour plots of two variable test functions.

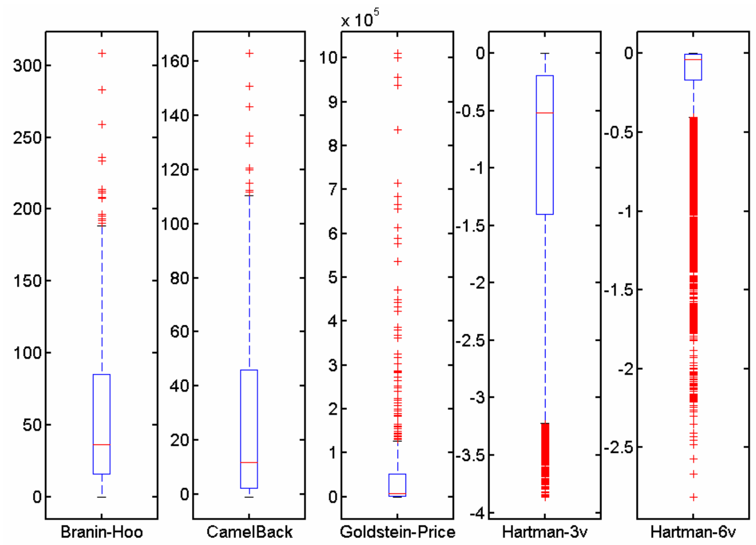
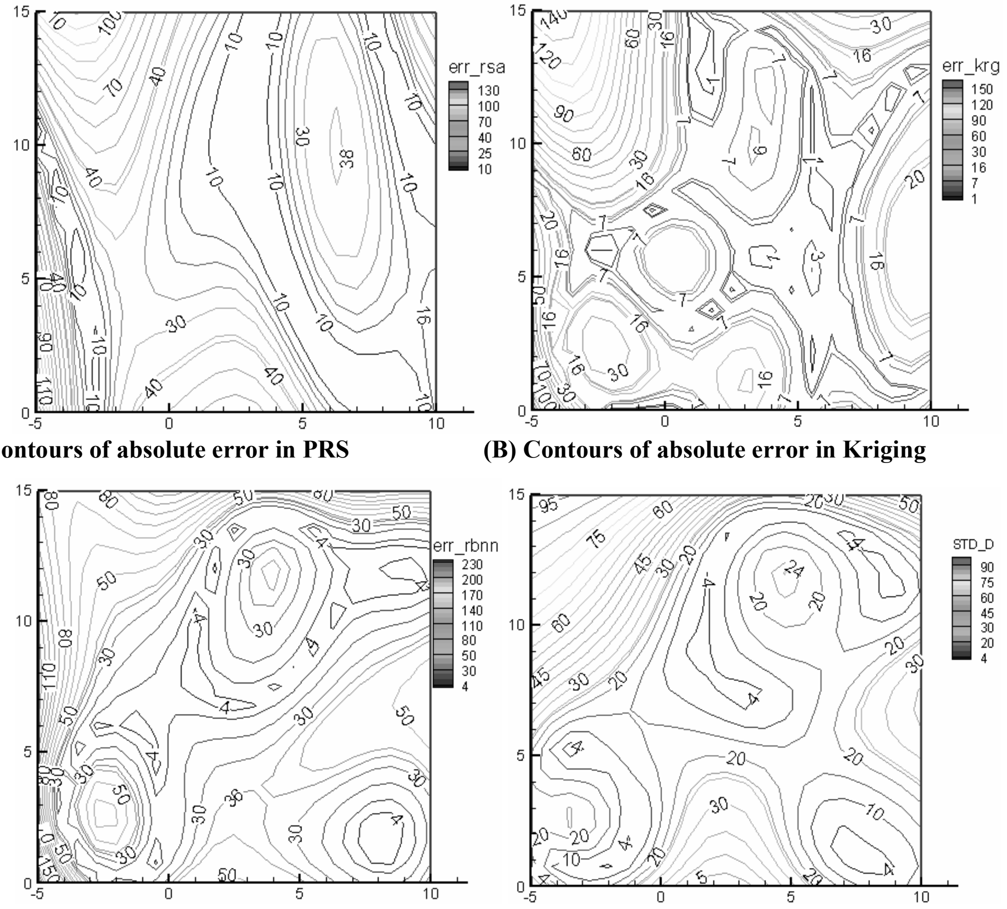


Figure 3 Boxplots of function values of different analytical functions.



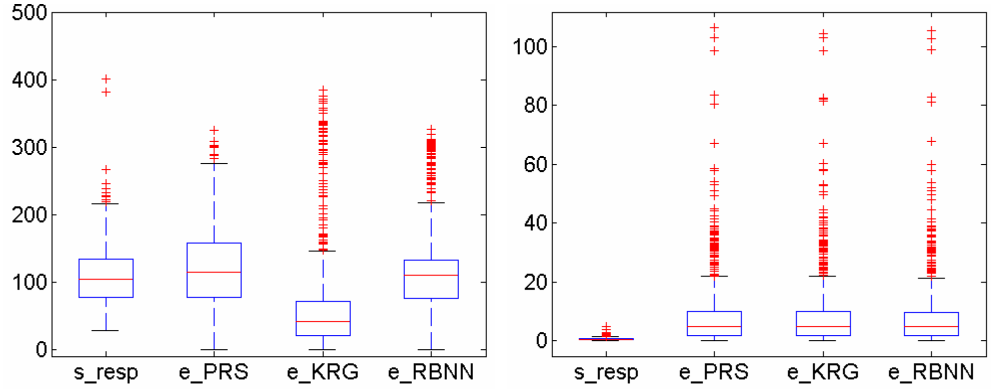
(A) Contours of absolute error in PRS

(B) Contours of absolute error in Kriging

(C) Contours of absolute error in RBNN

(D) Standard deviation of predictions

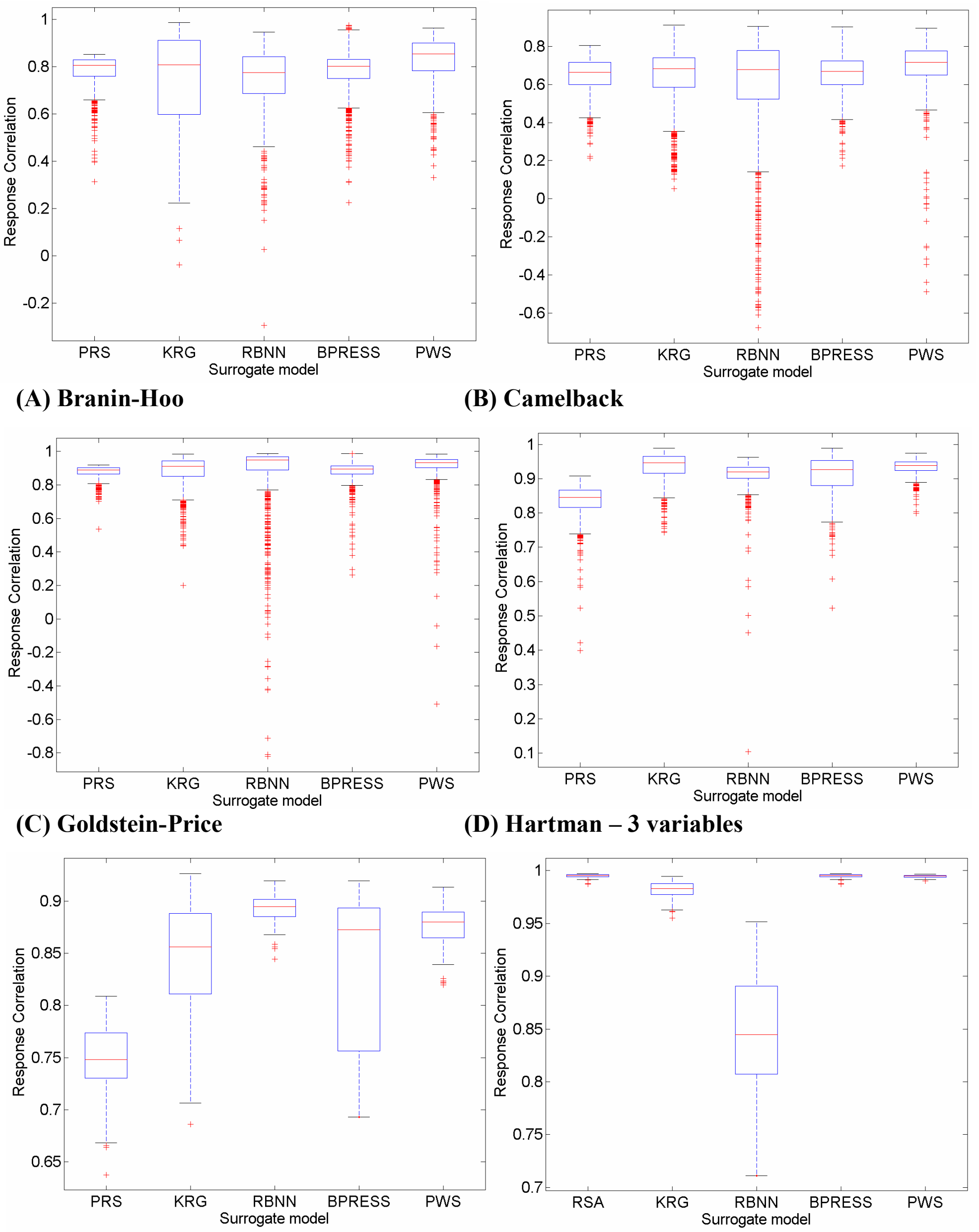
Figure 4 Contour plots of errors and standard deviation of predictions considering PRS, KRG, and RBNN surrogate models for Branin-Hoo function.



(a) Maximum standard deviation and corresponding actual errors

(b) Minimum standard deviation and corresponding actual errors

Figure 5 Maximum/Minimum standard deviation of responses and actual errors in prediction of different surrogates at corresponding locations (boxplots of 1000 DOEs using Branin-Hoo function) (s_resp is standard deviation of responses, e_PRS, e_KRG, e_RBNN are actual errors in PRS, KRG and RBNN).



(A) Branin-Hoo **(B) Camelback**

(C) Goldstein-Price **(D) Hartman – 3 variables**

(E) Hartman – 6 variables **(F) Radial turbine design**

Figure 6 Correlations between actual and predicted response for different test problems. 1000 instances of DOEs were considered for all test problems except Hartman-6 and radial turbine design problem for which we show results based on 100 samples. The center line of each boxplot shows the median value and the box encompasses the 25th- and 75th-percentile of the data. The leader lines (horizontal lines) are plotted at a distance of 1.5 times the inter-quartile range in each direction or the limit of the data (if the limit of the data falls within 1.5 times inter-quartile range).

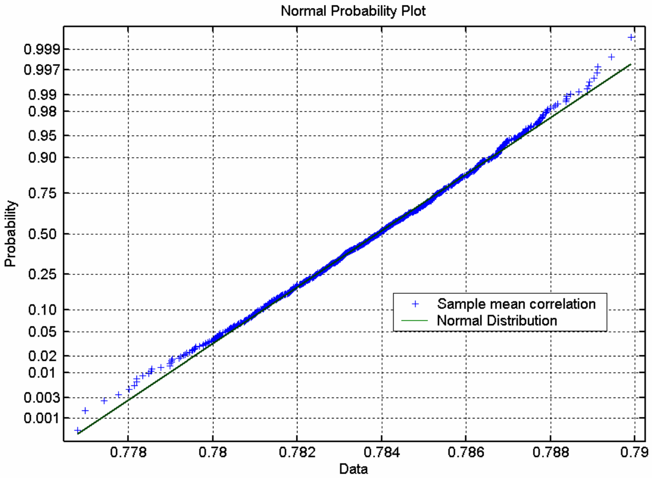
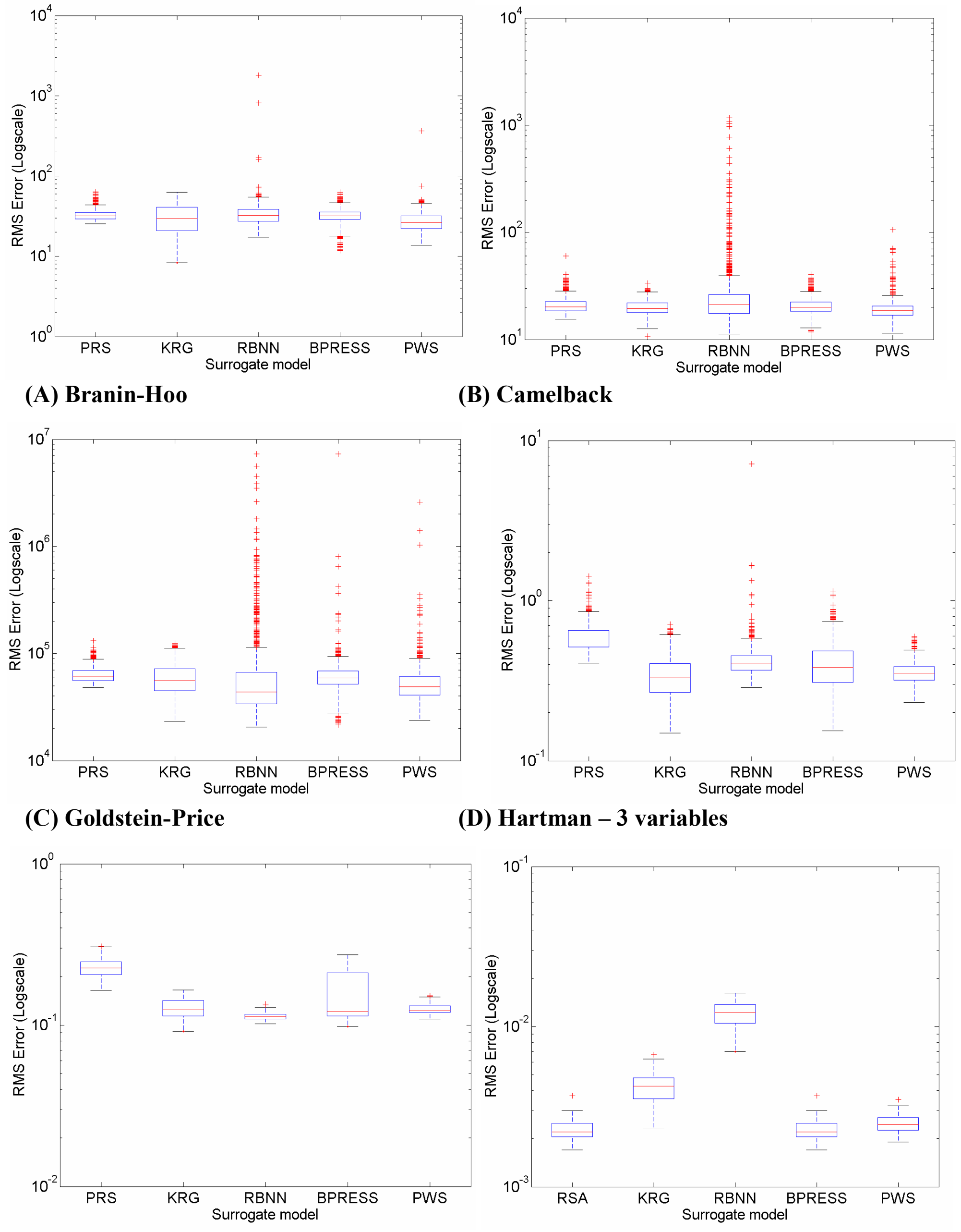


Figure 7 Normal distribution approximation of the sample mean correlation coefficient data obtained using 1000 bootstrap samples (KRG, Branin-Hoo function).



(A) Branin-Hoo

(B) Camelback

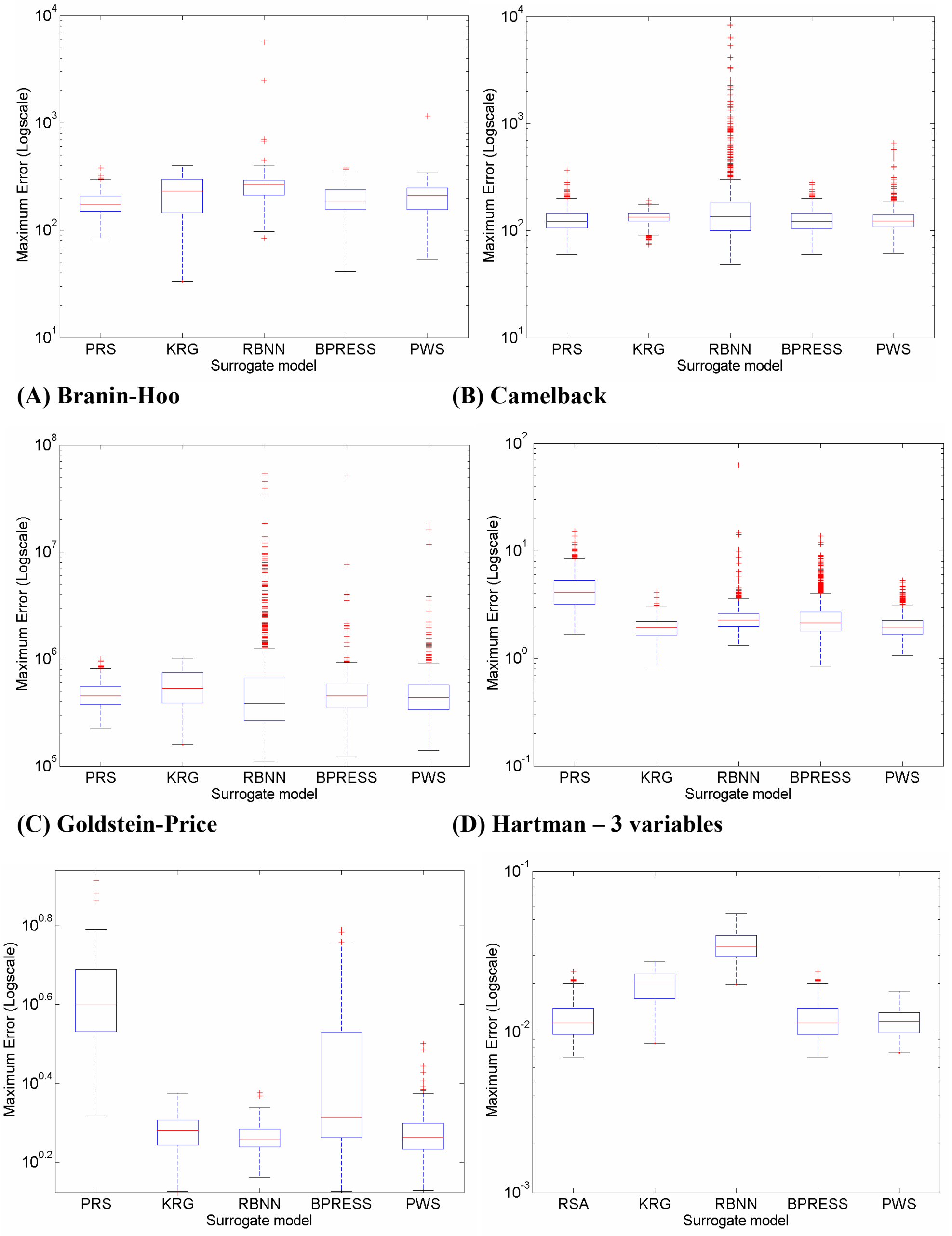
(C) Goldstein-Price

(D) Hartman – 3 variables

(E) Hartman – 6 variables

(F) Radial turbine design problem

Figure 8 RMS errors in design space for different surrogate models. 1000 instances of DOEs were considered for all test problems except Hartman-6 and radial turbine design problem for which we show results based on 100 samples.



(A) Branin-Hoo

(B) Camelback

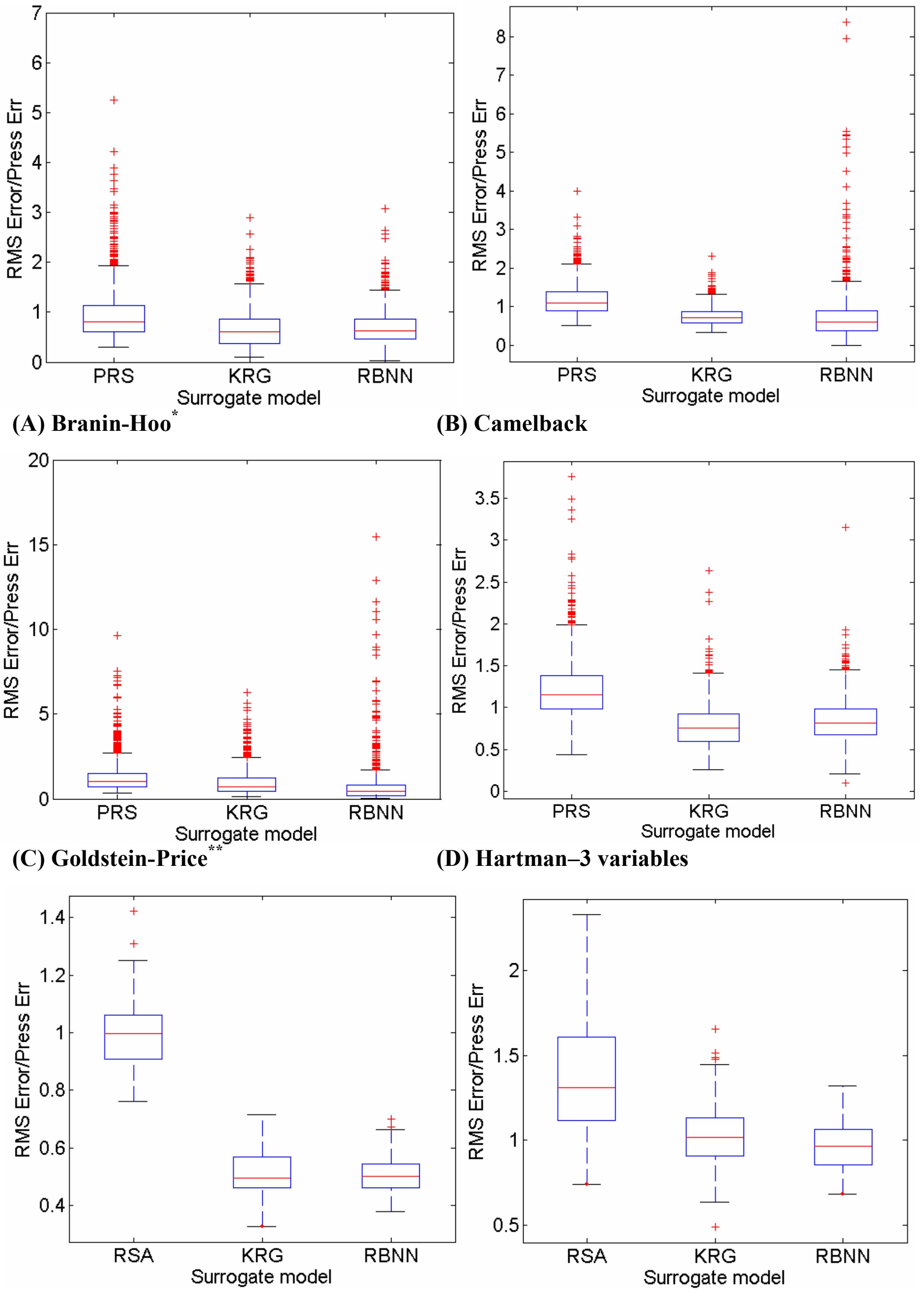
(C) Goldstein-Price

(D) Hartman – 3 variables

(E) Hartman – 6 variables

(F) Radial turbine design problem

Figure 9 Maximum absolute error in design space for different surrogate models. 1000 instances of DOEs were considered for all test problem except Hartman-6 and radial turbine design problem for which we show results based on 100 samples.



(A) Branin-Hoo*

(B) Camelback

(C) Goldstein-Price**

(D) Hartman–3 variables

(E) Hartman–6 variables

(F) Radial turbine design problem

Figure 10 Boxplots of ratio of RMS error and *PRESS* over 1000 DOEs for different problems (* For Branin-Hoo function, one simulation yielded RMSE/PRESS ratio ~O(20) for PRS, ** For Goldstein-Price problem, three simulations yielded high ratio of RMS error and PRESS error (20-80) for RBNN).

Table 1 Parameters used in Hartman function with three variables.

i	a_{ij}			c_i	p_{ij}		
1	3.0	10.0	30.0	1.0	0.3689	0.1170	0.2673
2	0.1	10.0	35.0	1.2	0.4699	0.4387	0.7470
3	3.0	10.0	30.0	3.0	0.1091	0.8732	0.5547
4	0.1	10.0	35.0	3.2	0.03815	0.5743	0.8828

Table 2 Parameters used in Hartman function with six variables.

i	a_{ij}						c_i
1	10.0	3.0	17.0	3.5	1.7	8.0	1.0
2	0.05	10.0	17.0	0.1	8.0	14.0	1.2
3	3.0	3.5	1.7	10.0	17.0	8.0	3.0
4	17.0	8.0	0.05	10.0	0.1	14.0	3.2
i	p_{ij}						
1	0.1312	0.1696	0.5569	0.0124	0.8283	0.5886	
2	0.2329	0.4135	0.8307	0.3736	0.1004	0.9991	
3	0.2348	0.1451	0.3522	0.2883	0.3047	0.6650	
4	0.4047	0.8828	0.8732	0.5743	0.1091	0.0381	

Table 3 Mean, coefficient of variation (COV) and median of different analytical functions.

	Branin-Hoo	Camelback	Goldstein-Price	Hartman-3v	Hartman-6v
Mean	49.5	19.1	49179	-0.8	-0.06
COV	1.0	1.8	3.9	-1.2	-5.1
Median	36.7	11.8	8114	-0.5	-0.04

Table 4 Range of design variables for radial turbine design problem.

Variable	Description	Minimum	Maximum
RPM	Rotational speed	100000	150000
Reaction	Percentage of stage pressure drop across rotor	0.40	0.57
U/C_{isen}	Isentropic velocity ratio	0.56	0.63
Tip Flow	Ratio of flow parameter to a choked flow parameter	0.30	0.53
$D_{hex\%}$	Exit hub diameter as a % of inlet diameter	0.1	0.4
AN^2Frac	Used to calculate annulus area (stress indicator)	0.68	0.85

Table 5 Numerical setup for the test problems.

	Branin-Hoo	Camelback	GoldStein-Price	Hartman3	Hartman6	Radial Turbine
# of variables	2	2	2	3	6	6
# of design points	12	20	25	40	150	56
# of test pts[*]	21	21	21	21	5	254
Order of polynomial	2	3	3	3	3	2
Spread	0.2	0.3	0.5	0.4	0.5	1
Goal	10	10	2500	0.05	0.05	0.01

[*] Total number of points is number of points along a direction raised to the power of the number of variables (e.g. 21^3 for Hartman problem with three variables). For the radial turbine problem, 254 indicate total number of test points. Spread controls the decay rate of radial basis function and Goal is the desired level of accuracy of the RBNN model on training points.

Table 6 Median, 1st and 3rd quartile of the maximum standard deviation and actual errors in predictions of different surrogates at the location corresponding to maximum standard deviation over 1000 DOEs for different test problems.

	Branin-Hoo	Camelback	Goldstein-Price	Hartman-3	Hartman-6	Radial turbine
Median (Max std dev. of response)	105	53	2.7e5	2.5	2.2	0.020
Median (Actual error in PRS)	114	61	2.9e5	3.9	3.9	0.0016
Median (Actual error in KRG)	42	111	3.6e5	0.7	0.2	0.004
Median (Actual error in RBNN)	110	95	2.5e5	0.6	0.1	0.033
1st/3rd Quartile (Max std dev. of response)	77/ 134	38/ 85	1.0e5/ 4.2e5	2.0/ 3.2	1.9/ 2.7	0.017/ 0.022
1st/3rd Quartile (Actual error in PRS)	78/ 158	32/ 92	1.0e5/ 4.7e5	2.8/ 5.2	3.3/ 4.9	0.0008/ 0.0027
1st/3rd Quartile (Actual error in KRG)	21/ 71	66/ 131	1.4e5/ 6.5e5	0.3/ 1.4	0.1/ 0.4	0.002/ 0.006
1st/3rd Quartile (Actual error in RBNN)	76/ 132	42/ 161	1.9e5/ 5.7e5	0.3/ 1.1	0.1/ 0.3	0.028/ 0.038

Table 7 Median, 1st and 3rd quartile of the minimum standard deviation and actual errors in predictions of different surrogates at the location corresponding to minimum standard deviation over 1000 DOEs for different test problems.

	Branin-Hoo	Camelback	Goldstein-Price	Hartman-3	Hartman-6	Radial turbine
Median (Min std dev. of response)	0.41	0.26	492	0.0019	0.0011	2.1e-4
Median (Actual error in PRS)	4.7	1.7	1630	0.063	0.06	1.0e-3
Median (Actual error in KRG)	4.6	1.7	1513	0.062	0.07	1.1e-3
Median (Actual error in RBNN)	4.7	1.7	1510	0.064	0.07	1.0e-3
1st/3rd Quartile (Min std dev. of response)	0.25/ 0.67	0.15/ 0.40	280/ 770	0.0012/ 0.0029	0.0007/ 0.0017	1.5e-4/ 3.2e-4
1st/3rd Quartile (Actual error in PRS)	1.7/ 9.8	0.7/ 4.4	697/ 3854	0.025/ 0.143	0.03/ 0.11	5.0e-4/ 1.9e-3
1st/3rd Quartile (Actual error in KRG)	1.8/ 9.9	0.6/ 4.2	525/ 3842	0.025/ 0.143	0.03/ 0.11	5.0e-4/ 1.9e-3
1st/3rd Quartile (Actual error in RBNN)	1.8/ 9.7	0.6/ 4.2	535/ 3871	0.024/ 0.142	0.03/ 0.11	5.0e-4/ 2.1e-3

Table 8 Median, 1st and 3rd quartile of the maximum standard deviation and maximum actual errors in predictions of different surrogates over 1000 DOEs for different test problems (Number after Branin-Hoo and Camelback functions indicates the number of data points used to model the function).

	Branin-Hoo12	Branin-Hoo31	Camelback-20	Camelback-40	Goldstein-Price	Hartman-3	Hartman-6	Radial turbine
Median (Max std dev. of response)	105	88	53	42	2.7E+05	2.5	2.2	0.020
Median (Max actual error in PRS)	175	32	122	37	4.5E+05	4.1	4.0	0.087
Median (Max actual error in KRG)	232	25	135	37	5.3E+05	1.9	1.9	0.087
Median (Max actual error in RBNN)	268	173	135	80	3.9E+05	2.3	1.8	0.082
1st/3rd Quartile (Max std dev. of response)	77/ 134	61/ 116	38/ 85	31/ 58	1.0e5/ 4.2e5	2.0/ 3.2	1.9/ 2.7	0.017/ 0.022
1st/3rd Quartile (Max actual error in PRS)	150/ 209	27/ 39	106/ 127	31/ 44	3.7e5/ 5.5e5	3.2/ 5.3	3.4/ 4.9	0.082/ 0.093
1st/3rd Quartile (Max actual error in KRG)	146/ 298	16/ 38	123/ 145	26/ 59	3.9e5/ 7.5e5	1.7/ 2.2	1.7/ 2.0	0.082/ 0.093
1st/3rd Quartile (Max actual error in RBNN)	214/ 294	119/ 233	100/ 181	61/ 107	2.7e5/ 6.7e5	2.0/ 2.6	1.7/ 1.9	0.077/ 0.087

Table 9 Effect of design of experiment: Number of cases when an individual surrogate model yielded the least PRESS error (based on 1000 DOEs).

	PRS	KRG	RBNN
Branin-Hoo	715	131	154
Camelback	880	59	61
GoldStein-Price	659	143	198
Hartman3	229	511	260
Hartman6	400	119	481
Radial Turbine	1000	0	0

Table 10 Mean and coefficient of variation (in parenthesis) of correlation coefficient between actual and predicted response (based on 1000 DOEs) for different surrogate models.

	PRS	KRG	RBNN	Best *PRESS*	PWS
Branin-Hoo	0.79	0.76	0.75	0.79	0.84
	(0.08)	(0.24)	(0.18)	(0.12)	(0.11)
Camelback	0.69	0.69	0.62	0.69	0.73
	(0.13)	(0.19)	(0.50)	(0.14)	(0.20)
GoldStein-Price	0.88	0.87	0.86	0.88	0.91
	(0.041)	(0.11)	(0.28)	(0.083)	(0.12)
Hartman3	0.80	0.92	0.90	0.89	0.92
	(0.073)	(0.052)	(0.059)	(0.074)	(0.028)
Hartman6	0.61	0.79	0.85	0.75	0.81
	(0.079)	(0.082)	(0.018)	(0.15)	(0.032)
Radial Turbine	0.9951	0.9814	0.8495	0.9951	0.9946
	(0.0015)	(0.0088)	(0.062)	(0.0015)	(0.0013)

Table 11 The mean and the coefficient of variation (in parenthesis) of RMS errors in design space (based on 1000 instances of DOEs) for different surrogate models.

	PRS	KRG	RBNN	Best PRESS	PWS
Branin-Hoo	32.8	30.7	36.1	32.5	27.7
	(0.15)	(0.38)	(1.70)	(0.20)	(0.46)
Camelback	21.0	20.0	36.1	20.7	19.4
	(0.17)	(0.16)	(2.27)	(0.17)	(0.30)
GoldStein-Price	6.40e4	6.00e4	1.12e5	6.97e4	5.98e4
	(0.17)	(0.33)	(3.52)	(3.32)	(1.66)
Hartman3	0.60	0.34	0.43	0.41	0.36
	(0.20)	(0.28)	(0.55)	(0.34)	(0.16)
Hartman6	0.23	0.13	0.11	0.15	0.13
	(0.14)	(0.12)	(0.051)	(0.34)	(0.074)
Radial Turbine	0.0023	0.0043	0.0120	0.0023	0.0025
	(0.15)	(0.23)	(0.18)	(0.15)	(0.13)

Table 12 The mean and the coefficient of variation (in parenthesis) of maximum absolute error in design space (based on 1000 instances of DOEs).

	PRS	KRG	RBNN	Best PRESS	PWS
Branin-Hoo	182	222	258	198	202
	(0.25)	(0.41)	(0.75)	(0.29)	(0.35)
Camelback	127	133	236	126	128
	(0.24)	(0.12)	(2.41)	(0.23)	(0.33)
GoldStein-Price	4.74e5	5.63e5	1.08e6	5.56e5	5.31e5
	(0.28)	(0.37)	(3.55)	(2.96)	(1.64)
Hartman3	4.40	1.94	2.47	2.59	2.05
	(0.38)	(0.21)	(0.86)	(0.54)	(0.28)
Hartman6	4.24	1.89	1.84	2.62	1.90
	(0.29)	(0.11)	(0.092)	(0.43)	(0.17)
Radial Turbine	0.0120	0.0196	0.0346	0.0120	0.0118
	(0.28)	(0.21)	(0.22)	(0.28)	(0.20)

Table 13 The mean and the coefficient of variation of the ratio of RMS error and *PRESS* over 1000 DOEs. *Branin-Hoo and Goldstein-Price functions had significant difference in the mean and median values of RBNN.

	PRS	KRG	RBNN
Branin-Hoo*	0.97 (0.57)	0.67 (0.60)	0.72 (1.07)
Camelback	1.20 (0.34)	0.76 (0.31)	0.73 (0.98)
GoldStein-Price*	1.32 (0.75)	0.99 (0.83)	0.89 (3.33)
Hartman3	1.22 (0.31)	0.78 (0.33)	0.85 (0.32)
Hartman6	0.99 (0.12)	0.50 (0.17)	0.50 (0.14)
Radial Turbine	1.34 (0.24)	1.02 (0.21)	0.97 (0.14)

Table 14 Studying the impact of modeling high gradients using Branin-Hoo function (Branin-Hoo 12 is the case when we used 12 points for modeling and Branin-Hoo31 is the case when 31 points were used to model the response) We used 1000 DOEs samples to get mean and COV.

		PRS	KRG	RBNN	Best PRESS	PWS
Correlations	Branin-Hoo12	0.79 (0.08)	0.76 (0.24)	0.75 (0.18)	0.79 (0.12)	0.84 (0.11)
	Branin-Hoo31	0.989 (0.003)	0.999 (0.001)	0.93 (0.076)	0.998 (0.003)	0.997 (0.014)
RMS Error	Branin-Hoo12	33 (0.15)	31 (0.38)	36 (1.70)	33 (0.20)	28 (0.46)
	Branin-Hoo31	7.6 (0.11)	2.3 (0.53)	19.3 (1.27)	2.6 (0.64)	4.1 (0.54)
Max Error	Branin-Hoo12	182 (0.25)	222 (0.41)	258 (0.75)	198 (0.29)	202 (0.35)
	BraninHoo31	34 (0.31)	30 (0.63)	183 (0.80)	31 (0.60)	41 (0.53)

Table 15 Studying the impact of modeling high gradients using Camelback function (Camelback20 is the case when we used 20 points for modeling and Camelback40 is the case when we used 40 points to model the response). We used 1000 DOEs to get mean and COV.

		PRS	KRG	RBNN	Best PRESS	PWS
Correlations	Camelback20	0.69 (0.13)	0.69 (0.19)	0.62 (0.50)	0.69 (0.14)	0.73 (0.20)
	Camelback40	0.97 (0.010)	0.98 (0.039)	0.92 (0.080)	0.98 (0.015)	0.98 (0.010)
RMS Error	Camelback20	21 (0.17)	20 (0.16)	36 (2.27)	21 (0.17)	19 (0.30)
	Camelback40	6.9 (0.15)	4.7 (0.74)	11 (0.35)	5.0 (0.42)	5.4 (0.27)
Max Error	Camelback20	127 (0.24)	133 (0.12)	236 (2.41)	126 (0.23)	128 (0.33)
	Camelback40	39 (0.34)	48 (0.64)	90 (0.52)	40 (0.47)	43 (0.39)

Table 16 Effect of parameters in parametric surrogate filter used for PWS. Three settings of parameters α and β were selected. We show median, 1st and 3rd quartile data based on 1000 DOEs for Goldstein-Price problem.

		PRS	KRG	RBNN	Best PRESS	PWS α=0.05, β=-1	PWS α=0.5, β=-1	PWS α=0.05, β=-5
Correlations	Median	0.89	0.90	0.94	0.89	0.93	0.93	0.91
	1^{st}-quartile	0.86	0.84	0.88	0.86	0.90	0.90	0.88
	3^{rd}-quartile	0.90	0.94	0.97	0.91	0.95	0.95	0.94
RMS Error	Median	6.14e4	5.56e4	4.35e4	5.92e4	4.91e4	4.84e4	5.41e4
	1^{st}-quartile	5.58e4	4.51e4	3.39e4	5.18e4	4.10e4	4.07e4	4.47e4
	3^{rd}-quartile	6.92e4	7.21e4	6.65e4	6.86e4	6.07e4	6.05e4	6.46e4
Max Error	Median	4.52e5	5.32e5	3.88e5	4.54e5	4.35e5	4.32e5	4.52e5
	1^{st}-quartile	3.74e5	3.91e5	2.65e5	3.56e5	3.38e5	3.33e5	3.47e5
	3^{rd}-quartile	5.52e5	7.49e5	6.68e5	5.87e5	5.75e5	5.73e5	5.82e5

Appendix: Generalized Mean Square Cross-validation Error

In general, the data is divided into k subsets (k-fold cross-validation) of approximately equal size. A surrogate model is constructed k times, each time leaving out one of the subsets from training, and using the omitted subset to compute the error measure of interest. The generalization error estimate is computed using the k error measures obtained (e.g., average). If k equals the sample size, this approach is called leave-one-out cross-validation (also known as PRESS in the polynomial response surface approximation terminology). Equation (A1) represents a leave-one-out calculation when the generalization error is described by the mean square error (GMSE).

$$GMSE = \frac{1}{k}\sum_{i=1}^{k}(f_i - \hat{f}_i^{(-i)})^2 \qquad \text{(A1)}$$

where $\hat{f}_i^{(-i)}$ represents the prediction at $\mathbf{x}^{(i)}$ using the surrogate constructed using all sample points except ($\mathbf{x}^{(i)}$, f_i). Analytical expressions are available for that case for the GMSE without actually performing the repeated construction of the surrogates for both polynomial response surface approximation (Myers and Montgomery[23], 1995, Section 2.7) and Kriging (Martin and Simpson[24], 2005) however here we used brute-force. The advantage of cross-validation is that it provides nearly unbiased estimate of the generalization error and the corresponding variance is reduced (when compared to split-sample) considering that every point gets to be in a test set once, and in a training set k-1 times (regardless of how the data is divided).