Proceedings of the ASME 2011 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/CIE 2011
August 28-31, 2011, Washington, DC, USA

DETC2011-48608

SURROGATE MODELING OF COMPLEX SYSTEMS USING ADAPTIVE HYBRID FUNCTIONS

Jie Zhang[*]
Rensselaer Polytechnic Institute
Troy, New York 12180
Email: zhangj17@rpi.edu

Souma Chowdhury[*]
Rensselaer Polytechnic Institute
Troy, New York 12180
Email: chowds@rpi.edu

Achille Messac[†]
Syracuse University
Syracuse, NY, 13244
Email: messac@syr.edu

Junqiang Zhang[*]
Rensselaer Polytechnic Institute
Troy, New York 12180
Email: zhangj18@rpi.edu

Luciano Castillo[‡]
Rensselaer Polytechnic Institute
Troy, New York 12180
Email: castil2@rpi.edu

ABSTRACT

This paper explores the effectiveness of the recently developed surrogate modeling method, the Adaptive Hybrid Functions (AHF), through its application to complex engineered systems design. The AHF is a hybrid surrogate modeling method that seeks to exploit the advantages of each component surrogate. In this paper, the AHF integrates three component surrogate models: (i) the Radial Basis Functions (RBF), (ii) the Extended Radial Basis Functions (E-RBF), and (iii) the Kriging model, by characterizing and evaluating the local measure of accuracy *of each model. The AHF is applied to model complex engineering systems and an economic system, namely: (i) wind farm design; (ii) product family design (for universal electric motors); (iii) three-pane window design; and (iv) onshore wind farm cost estimation. We use three differing sampling techniques to investigate their influence on the quality of the resulting surrogates. These sampling techniques are (i) Latin Hypercube Sampling (LHS), (ii) Sobol's quasirandom sequence, and (iii) Hammersley Sequence Sampling (HSS). Cross-validation is used to evaluate the accuracy of the resulting surrogate models. As expected, the accuracy of the surrogate model was found to improve with increase in the sample size. We also observed that, the Sobol's and the LHS sampling techniques performed better in the case of high-dimensional problems, whereas the HSS sampling technique performed better in the case of low-dimensional problems. Overall, the AHF method was observed to provide acceptable-to-high accuracy in representing complex design systems.*

KEYWORDS: Complex engineered systems; hybrid surrogate modeling; optimization; product family; response surface; wind farm;

INTRODUCTION

Complex systems such as human bodies, rain forests, aerospace-systems, energy systems and wireless networking generally tend to be highly interdisciplinary. Understanding, designing, building and controlling such complex systems remains a central challenge in the academia and the industry [1].

The determination of complex underlying relationships between system parameters from simulated and/or recorded data

[*]Doctoral Student, Multidisciplinary Design and Optimization Laboratory, Department of Mechanical, Aerospace and Nuclear Engineering, ASME student member.

[†]Distinguished Professor and Department Chair. Department of Mechanical and Aerospace Engineering, ASME Lifetime Fellow. Corresponding author.

[‡]Associate Professor, Department of Mechanical Aerospace and Nuclear Engineering, ASME member

requires advanced interpolating functions, also known as surrogates. The development of surrogates for such complex relationships often requires modeling high dimensional and non-smooth functions with limited information. To this end, the hybrid surrogate modeling paradigm, where characteristically differing surrogate models are intelligently aggregated, offers a robust solution.

Over the past two decades, function estimation methods and approximation-based optimization have progressed remarkably. Surrogate models are being extensively used in the analysis and in the optimization of computationally expensive simulation-based models. Surrogate modeling techniques have been used for a variety of applications in multidisciplinary design optimization to reduce the analysis time and to improve the tractability of complex analysis codes [2, 3].

Experimental Design and Response Surface Method

In the literature, we can find a wide variety of surrogate modeling techniques, including: (i) the Polynomial Response Surface method (PRSM) [4], (ii) the Kriging approach [5,6], (iii) the Radial Basis Functions (RBF) [7], (iv) the Extended Radial Basis Functions (E-RBF) [8, 9], (v) the Artificial Neural Networks (ANN) [10], (vi) the Support Vector Regression (SVR) [11, 12], and (vii) the hybrid surrogate modeling method [13–15]. Table 1 provides a list of standard sampling techniques, surrogate modeling methods, and function-coefficient estimation methods, which extends the preexisting work by Simpson et al. [16].

More recently, researchers have presented a combination of different function-approximation models into a single ensemble model [13, 14, 17–19]. Zepra et al. [13] reported the application of an ensemble of surrogate models to construct a weighted average surrogate for the optimization of alkaline-surfactant-polymer flooding processes. They found that the weighted average surrogate provides better performance than individual surrogates. Goel et al. [14] considered an ensemble of three surrogate models (polynomial response surface, Kriging and radial basis neural network), using the *Generalized Mean Square Cross-validation Error* of the individual surrogate models. Acar and Rais-Rohani [18] treated the selection of weight factors in the general weighted-sum formulation of an ensemble as an optimization problem with the objective to minimize an error metric. The results showed that the optimized ensemble provides more accurate predictions than the stand-alone surrogate model. Acar [20] investigated the efficiency of using various local error measures for constructing an ensemble of surrogate models, and also presented the use of the pointwise cross validation error as a local error measure. Zhou et al. [21] used a recursive process to obtain the values of weights, in which the values of surrogate weights are updated in each iteration until the last ensemble achieves a desirable prediction accuracy.

Motivation and Objectives

The recently developed hybrid surrogate modeling method, the Adaptive Hybrid Functions (AHF), integrates component surrogate models by characterizing and evaluating the local *measure of accuracy* of each model. A novel *crowding distance based trust region* was proposed to capture both the global and the local accuracy of the surrogate model. The weights of the component surrogates are adaptively selected based on the *measure of accuracy* of each surrogate in the trust region. This paper explores the original AHF methodology by

1. applying the AHF to complex engineered systems design, and economic system design problems,
2. implementing three representative sampling techniques (i) Latin Hypercube Sampling (LHS), (ii) Sobol's quasirandom sequence, and (iii) Hammersley Sequence Sampling (HSS), and
3. investigating the effects of the sample size and the problem dimensionality on the performance of the surrogate model.

This paper primarily seeks to validate the wide applicability and the robustness of the AHF methodology.

ADAPTIVE HYBRID FUNCTIONS (AHF)

The AHF methodology was recently developed by Zhang et al. [15, 22]. The AHF formulates a reliable trust region, and adaptively combines characteristically differing surrogate models. The weight of each contributing surrogate model is represented as a function of the input domain, based on a local *measure of accuracy* for that surrogate model. Such an approach exploits the advantages of each component surrogate, thereby capturing both the global and the local trend of complex functional relationships. In this paper, the AHF integrates three component surrogate models: (i) RBF, (ii) E-RBF, and (iii) Kriging, by characterizing and evaluating the local *measure of accuracy* of each model. The hybrid surrogate modeling methodology adopted in this paper introduces a three-step approach:

1. Determination of a trust region: numerical bounds of the estimated parameter (output) as functions of the independent parameters (input vector).
2. Characterization of the local *measure of accuracy* (using probability distribution functions) of the estimated function value, and representation of the corresponding distribution parameters as functions of the input vector.
3. Weighted summation of characteristically different surrogate models (component surrogates) based on the local *measure of accuracy* (modeled in the previous step).

Table 1. TECHNIQUES FOR RESPONSE SURFACE

Sampling/Design of Experiments	Surrogate Modeling	Coefficient Estimation
(Fractional) factorial	Polynomial (linear, quadratic)	Least Squares Regression
Central composite	Splines (linear, cubic)	Best Weighted Least Squares Regression
Latin Hypercube	Kriging	Best Linear Predictor
Hammersley sequence	Radial Basis Functions (RBF)	Log-likelihood
Uniform designs	Extended RBF	Multipoint approximation
Sobol sequence	Support Vector Regression (SVR)	Adaptive response surface
Random selection	Neural Network (NN)	Back propagation
Box-Behnken	Hybrid models	Entropy
Plackett-Burman		Linear Unbiased Predictor
Orthogonal arrays		

We consider a set of training points D, expressed as

$$D = \begin{pmatrix} x_1^1 & x_2^1 & \cdots & x_{n_d}^1 & y^1 \\ x_1^2 & x_2^2 & \cdots & x_{n_d}^2 & y^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_1^{n_p} & x_2^{n_p} & \cdots & x_{n_d}^{n_p} & y^{n_p} \end{pmatrix}$$

The three steps, followed to formulate the AHF surrogate model, is illustrated in Fig.1. In the above expression, x_j^i is the j^{th} dimension of the input vector that represents the i^{th} training point, and y^i is the corresponding output; n_d is the dimension of the input variable, and n_p represents the number of training data points.

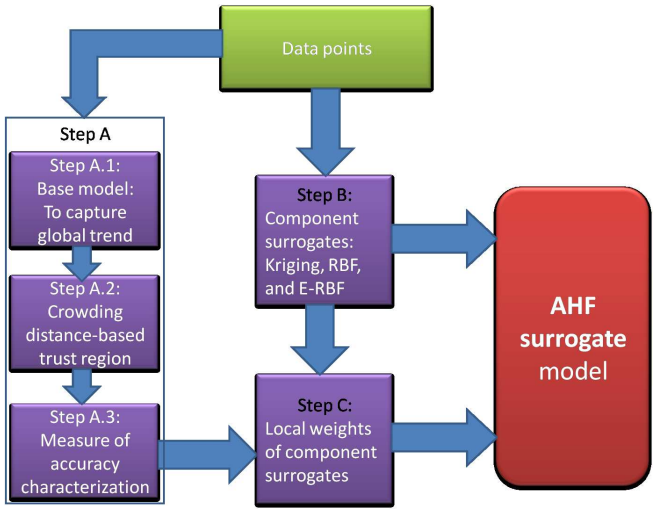


Figure 1. THE FRAMEWORK OF THE AHF SURROGATE MODEL

Step A.1: Determination of the Base Model

The base model is developed using smooth functions to obtain a global approximation for the given set of points, D. This base model could capture the global trend of training points, thereby seeking to improve the global accuracy for the overall surrogate. In the current paper, the base model is constructed using Quadratic Response Surface Model (QRSM). However, the base model also has the flexibility to use other (typically monotonic) smooth functions as well. A typical QRSM can be represented as

$$\tilde{f}_{qrsm}(x) = a_0 + \sum_{i=1}^{n_d} a_i x_i + \sum_{i=1}^{n_d} a_{ii} x_i^2 + \sum_{i=1}^{n_d-1} \sum_{j>i}^{n_d} a_{ij} x_i x_j \quad (1)$$

where the x_i's are the input parameters and the a_{ij}'s are the unknown coefficients determined by the least squares approach.

Step A.2: Formulation of Trust Region Boundaries

In this step, we formulate a trust region for the estimated parameter. The boundaries of the trust region are adaptively constructed using the base model. This process allows the final surrogate model to capture the global accuracy, even though, locally accurate models are used as component surrogates. The trust region boundaries of the surrogate model are constructed according to the base model and the estimated crowding distance of sample points. A set of points are selected on the base model, and the crowding distance is evaluated for each point. Then, the

base model is relaxed along either directions of the output axis to obtain the boundaries of the surrogate.

In the Non-dominated Sorting Genetic Algorithm (NSGA-II), the crowding distance value of a candidate solution provides a local estimate of the density of solutions [23]. In this paper, crowding distance is used to evaluate the density of training points surrounding any point on the base model. Larger crowding distance values at a point reflects lower sample density (fewer points around that point); the *measure of accuracy* of a surrogate is expected to be relatively lower around that point. Therefore, the trust region boundaries should be relaxed at that location in the input domain. Hence, based on the crowding distance value at each point on the base model, we construct adaptive trust region boundaries. This adaptive trust region is called the Crowding Distance-Based Trust Region (CD-TR). In this paper, the crowding distance of the i^{th} point on the base model (CD^i) is evaluated by

$$CD^i = \sum_{j=1}^{n_p} \left| x^j - x^i \right|^2 \tag{2}$$

where n_p is the number of training data points used. A parameter ρ is defined to represent the local density of input data, which is given by

$$\rho^i = \frac{1}{CD^i} \tag{3}$$

The parameter ρ is then normalized to obtain α_i's, as given by

$$\alpha_i = \frac{\max(\rho) - \rho^i}{\max(\rho) - \min(\rho)} \tag{4}$$

The adaptive distance d^i between the i^{th} corresponding point on the boundary and the base model along either directions of the output axis, is expressed as

$$d^i = (1 + \alpha_i) \times \max_{j \in D} \left| \tilde{f}_{qrsm}(x^j) - y^j \right| \tag{5}$$

where D represents the original training data set. It is helpful to note that, in Eq. 5, the index $'j'$ represents training points and the index $'i'$ represents a uniform set of points selected on the base model. In Eq. 5, the adaptive distance is divided into two parts:

1. $\max_{j \in D} \left| \tilde{f}_{qrsm}(x^j) - y^j \right|$, a constant to ensure that all the training points are located between the boundaries; and
2. $\alpha_i \times \max_{j \in D} \left| \tilde{f}_{qrsm}(x^j) - y^j \right|$, an adaptive distance based on the distance coefficients (α_i's).

Crowding distance is evaluated with respect to each of the selected points, based on which $'\alpha'$ is previously formulated. The extent of the boundary region is scaled using the maximum of the training data deviation from the base model. Here, $\tilde{f}_{qrsm}(x^j)$ is the estimated output value of the j^{th} training point using QRSM; $\left| \tilde{f}_{qrsm}(x^j) - y^j \right|$ is the distance from the j^{th} training point to the base model along the direction of the output axis. Subsequently, we could obtain two sets of points, D^U and D^L, for constructing the two boundaries, as expressed by

$$D^U = \begin{pmatrix} x_1^1 & x_2^1 & \cdots & x_{n_d}^1 & y^1 + d^1 \\ x_1^2 & x_2^2 & \cdots & x_{n_d}^2 & y^2 + d^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_1^{n_p} & x_2^{n_p} & \cdots & x_{n_d}^{n_p} & y^{n_p} + d^{n_p} \end{pmatrix}$$

$$D^L = \begin{pmatrix} x_1^1 & x_2^1 & \cdots & x_{n_d}^1 & y^1 - d^1 \\ x_1^2 & x_2^2 & \cdots & x_{n_d}^2 & y^2 - d^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_1^{n_p} & x_2^{n_p} & \cdots & x_{n_d}^{n_p} & y^{n_p} - d^{n_p} \end{pmatrix}$$

Again, QRSM is adopted to estimate the upper and the lower boundary surfaces, using the generated data points D^U and D^L, respectively. The local probability of individual surrogates are determined within these trust region boundaries.

It is important to note that the crowding distance (or sampling density) based trust region estimation is particulary useful for recorded data-based (commercial or experimental data) surrogate modeling. In the case of problems, where the user has control over sampling (simulation-based), the initial sample data is expected to be relatively evenly distributed; significant variation in crowding distance might not be observed.

Step A.3: Estimation of Probability Distributions

With the CD-TR developed above, it is important to determine the *measure of accuracy* of the estimated function value at a given point in the trust region. Based on the *measure of accuracy*, we can adaptively integrate different component surrogate models. In this paper, we develop a credible metric, which we call the *Accuracy Measure of Surrogate Modeling* (AMSM), to represent the uncertainty in the estimated function value.

Function estimation is performed between the two boundary surfaces, using a local *measure of accuracy* technique. The uncertainty in the estimated function value at a location in the input variable domain is modeled using a kernel function. This kernel function is expressed as a function of the output parameter. The corresponding typical coefficients of the kernel function are represented as functions of the input vector, thereby characterizing

the *measure of accuracy* of the estimated function over the entire input domain.

The kernel function used to represent the measure of accuracy must have the following properties.

1. The kernel function value must be a maximum of one at the actual output, $y(x^i)$.
2. The kernel function must be equal to the specified small tolerance value at the upper and the lower boundaries of the trust region.
3. The function must increase monotonically from either boundary to the actual output value.
4. The function must be continuous.

In this paper, the following kernel function is adopted. This kernel function that satisfies the specified requirements is expressed as

$$P(z) = a\exp\left[-\frac{(z-\mu)^2}{2\sigma^2}\right] \tag{6}$$

where the amplitude coefficient a is set to be equal to one; the coefficients μ and σ represent the mean and the standard deviation of the kernel function, respectively. It is helpful to note that other kernel functions that have similar properties can also be used to represent the measure of accuracy.

The distance between the two boundaries is normalized. At each training data point x^i, the output value at the lower and upper boundaries ($f_L^i(x^i)$ and $f_U^i(x^i)$, respectively) are also normalized. We assume that the estimated *measure of accuracy* (kernel function) is a maximum of one at the actual output value $y(x^i)$; and, a minimum of 0.1 at the boundaries (within the trust region). The output value of a training point does not necessarily occur midway between the two boundaries. In order to ensure the continuity of the kernel function, we divide the function into two parts, with distinct standard deviations and the same mean. Then, we can represent the kernel function as

$$P(x^i) = \begin{cases} a\exp\left\{-\frac{\left[y(x^i)-\mu(x^i)\right]^2}{2\sigma_1^2(x^i)}\right\} & \text{if } 0 \le y(x^i) \le \mu(x^i) \\ a\exp\left\{-\frac{\left[y(x^i)-\mu(x^i)\right]^2}{2\sigma_2^2(x^i)}\right\} & \text{if } \mu(x^i) \le y(x^i) \le 1 \end{cases} \tag{7}$$

where the parameters σ_1 and σ_2 are controlled by the *full width at one tenth maximum* (Δx_{10}), given by

$$\begin{aligned} \sigma_1(x^i) &= \frac{\Delta z_{10}(x^i)}{2\sqrt{2\ln 10}} = \frac{2\left[\mu(x^i) - f_L^i(x^i)\right]}{2\sqrt{2\ln 10}} \\ &= \frac{2\mu(x^i)}{2\sqrt{2\ln 10}} = \frac{\mu(x^i)}{\sqrt{2\ln 10}} \qquad ,and \end{aligned} \tag{8}$$

$$\begin{aligned} \sigma_2(x^i) &= \frac{\Delta z_{10}(x^i)}{2\sqrt{2\ln 10}} = \frac{2\left[f_U^i(x^i) - \mu(x^i)\right]}{2\sqrt{2\ln 10}} \\ &= \frac{2\left[1-\mu(x^i)\right]}{2\sqrt{2\ln 10}} = \frac{1-\mu(x^i)}{\sqrt{2\ln 10}} \qquad ,where \end{aligned} \tag{9}$$

$$P(\mu \pm 0.5\Delta z_{10}) = \frac{1}{10} \tag{10}$$

From Eq. 7, we determine the *measure of accuracy* coefficients $\mu(x^i)$ for the i^{th} training point. The coefficient μ is expressed in terms of input variables x_j^i using a Polynomial Response Surface.

Step B: Creation of Surrogate Models Using Existing Methods

In this step, we construct different surrogate models (component surrogates). The selected component surrogate models should be locally accurate - with greater accuracy in the local region close to the training points. Three component surrogates are constructed based on the set of training points D, using Kriging, RBF, and E-RBF. However, we can also integrate other standard surrogates that are locally accurate. For each test point, the estimated function vector can be represented as $\tilde{f} = \left\{\tilde{f}_{kriging}\ \tilde{f}_{rbf}\ \tilde{f}_{erbf}\right\}$. The parameters $\tilde{f}_{kriging}$, $\tilde{f}_{rbf}$ and $\tilde{f}_{erbf}$ represent function values estimated by the Kriging, the RBF and the E-RBF methods, respectively.

Step C: Determining Local Weights of Component Surrogates

Finally, we formulate the Adaptive Hybrid Functions (AHF) surrogate model by adaptive selection of weights for the three component surrogate models (RBF, E-RBF and Kriging). The AHF is a weighted summation of function values estimated by the component surrogates, as given by

$$\tilde{f}_{AHF} = \sum_{i=1}^{n_s} w_i(x)\tilde{f}_i(x) \tag{11}$$

where n_s is the number of component surrogates integrated into the AHF, and $\tilde{f}_i(x)$ represents the estimated value by each component surrogate. The weights w_i's are expressed in terms of the estimated *measure of accuracy*, which is given by

$$w_i(x) = \frac{P_i(x)}{\sum_{i=1}^{n_s} P_i(x)} \tag{12}$$

where $P_i(x)$ is the *measure of accuracy* of the i^{th} surrogate for point x.

PERFORMANCE CRITERIA

The overall performance of the surrogate can be evaluated using three standard performance metrics: (i) Root Mean Squared Error (RMSE) [5, 24], which provides a global error measure over the entire design domain; (ii) Maximum Absolute Error (MAE), which is indicative of local deviations; and (iii) Relative Accuracy Error (RAE). The RMSE is given by

$$RMSE = \sqrt{\frac{1}{n_t}\sum_{k=1}^{n_t}\left(f(x^k)-\tilde{f}(x^k)\right)^2} \tag{13}$$

where $f(x^k)$ represents the actual function value for the test point x^k, $\tilde{f}(x^k)$ is the corresponding estimated function value. The parameter n_t is the number of test points chosen for evaluating the error measure. The MAE is expressed as

$$MAE = \max_k \left|f(x^k)-\tilde{f}(x^k)\right| \tag{14}$$

The RAE is evaluated at each test point, using the formula

$$RAE(x^k) = \frac{|\tilde{f}(x^k)-f(x^k)|}{f(x^k)} \tag{15}$$

Cross-Validation

Cross-validation is a technique that is used to analyze and improve the robustness of a surrogate model. Cross-validation error is the error estimated at a data point, when the response surface is fitted to a subset of the data points not including that point (also called the leave-one-out strategy). A vector of cross-validation errors, $\tilde{e}$, can be obtained, when the response surfaces are fitted to all the other $p-1$ points. This vector is also known as the prediction sum of squares (the PRESS vector).

The leave-one-out strategy is computationally expensive for large number of points, which can be overcome by the q-fold strategy. Q-fold strategy involves (i) splitting the data randomly into q (approximately) equal subsets, (ii) removing each of these subsets in turn, and (iii) fitting the model to the remaining $q-1$ subsets. A loss function L can be computed to measure the error between the predictor and the points in the subset that we set aside at each iteration; the contributions to L are then summed up over the q iterations.

More formally, when the mapping $\zeta : 1,\cdots,n \rightarrow 1,\cdots,q$ describes the allocation of the n training points to one of the q subsets and $\hat{f}^{-\zeta(i)}(x)$ (of the predictor) is obtained by removing the subset $\zeta(i)$, the cross-validation measure is given by

$$PRESS_{SE} = \frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)}-\hat{f}^{-\zeta(i)}(x^{(i)})\right]^2 \tag{16}$$

Hastie et al. [25] recommended compromise values of $q=5$ or $q=10$. Using fewer subsets generally has an additional advantage of reducing the computational cost of the cross-validation process by reducing the number of models that have to be fitted.

COMPLEX ENGINEERED SYSTEMS

The AHF is applied to complex engineering design problems and an economic system, which are: (i) wind farm design; (ii) product family design (for universal electric motors); (iii) three-pane window design; and (iv) onshore wind farm cost estimation.

Wind Farm Power Generation Model

This power generation model is taken from the paper by Chowdhury et al. [26, 27]. The power generated by a wind farm is an intricate function of the configuration and location of the individual wind turbines. The flow pattern inside a wind farm is complex, primarily due to the wake effects and the highly turbulent flow. The power generated by a wind farm (P_{farm}) comprised of N wind turbines is evaluated as a sum of the powers generated by the individual turbines, which is expressed as [27, 28]

$$P_{farm} = \sum_{j=1}^{N} P_j \tag{17}$$

Accordingly, the farm efficiency can be expressed as

$$\eta_{farm} = \frac{P_{farm}}{\sum_{j=1}^{N} P_{0j}} \tag{18}$$

where P_{0j} is the power that $turbine-j$ would generate if operating as a stand-alone entity, for the given incoming wind velocity. Detailed formulation of the power generation model can be found in the paper [26].

The power generated is thus a function of the location co-ordinates of each turbine. Turbine-type and operating conditions remaining fixed, the x-y coordinates are the design variables for wind farm layout optimization. In this paper, we develop a hybrid response surface (using AHF) to represent the net power generation as a function of the turbine location co-ordinates. In the case of a wind farm comprised of N turbines, the power generation model presents a $2N$ dimensional problem.

The power generation model is a complex combination of nonlinear sub-models, including the wake model, the wake overlap model, and the turbine power response model. The overall farm power generation function is thus highly nonlinear. In addition, owing to the inherent nature of the layout design approach, the power generation function is also expected to be multimodal.

Therefore this problem presents significant challenges to accurate representation through surrogate modeling.

We consider four cases: (i) wind farm with 4 turbines (8 variables); (ii) wind farm with 9 turbines (18 variables); (iii) wind farm with 16 turbines (32 variables); and (iv) wind farm with 25 turbines (50 variables). The *GE-1.5MW-xle* [29] turbine is chosen as the specified turbine-type for all cases. The features of this turbine is provided in Table 2.

Table 2. FEATURES OF THE *GE-1.5MW-XLE* TURBINE [29]

Turbine feature	Value
Rated power (P_{r0})	1.5MW
Rated wind speed (U_{r0})	11.5m/s
Cut-in wind speed (U_{in0})	3.5m/s
Cut-out wind speed (U_{out0})	20.0m/s
Rotor-diameter (D_0)	82.5m
Hub-height (H_0)	80.0m

Product Family Design (for Universal Electric Motors)

A product family is a group of related products that are derived from a common product platform to satisfy a variety of market niches [30]. Sharing of a common platform by different products is expected to result in: (i) reduced overhead, (ii) lower per product cost, and (iii) increased profit. The recently developed Comprehensive Product Platform Planning (CP^3) framework [31] formulated a generalized mathematical model to represent the complex platform planning process.

The CP^3 model formulates a generic equality constraint (the *commonality constraint*) to represent the variable based platform formation. The presence of a combination of integer variables (specifically binary variables) and continuous variables can be attributed to the combinatorial process of platform identification. The nonlinearity of this problem can be primarily attributed to the likely nonlinear nature of the system model (typical nonlinear performance functions and nonlinear constraints) for the products.

The general Mixed Integer Non-Liner Programming (MINLP) problem, formulated to represent the design optimization of the product family example, is given by

$$
\begin{aligned}
&\text{Max } f_{perf}(Y)\\
&\text{Min } f_{cost}(Y)\\
&\text{subject to}\\
&\quad \lambda_1^{12}\left(x_1^1 - x_1^2\right)^2 + \lambda_2^{12}\left(x_2^1 - x_2^2\right)^2 + \lambda_3^{12}\left(x_3^1 - x_3^2\right)^2 = 0\\
&\quad g_i(X) \le 0, \quad i = 1,2,....,p\\
&\quad h_i(X) = 0, \quad i = 1,2,....,q
\end{aligned}
\tag{19}
$$

where

$$
\begin{aligned}
&Y = \left\{x_1^1,\, x_2^1,\, x_3^1,\, x_1^2,\, x_2^2,\, x_3^2,\, \lambda_1^{12},\, \lambda_2^{12},\, \lambda_3^{12}\right\}\\
&X = \left\{x_1^1,\, x_2^1,\, x_3^1,\, x_1^2,\, x_2^2,\, x_3^2\right\}\\
&\left(\lambda_1^{12},\, \lambda_2^{12},\, \lambda_3^{12}\right) \in B : B = \{0,\, 1\}
\end{aligned}
$$

and where f_{perf} and f_{cost} are the objective functions that represent the performance and the cost of the product family, respectively. In Eq. 19, g_i and h_i represent the inequality and equality constraints related to the physical design of the product. The first equality constraint in Eq. 19, which involves the parameters λ_j^{ik} is termed the *commonality constraint*.

The authors [32] developed a methodology to reduce the high dimensional binary integer problem to a more tractable integer problem, where the commonality matrix is represented by a set of integer variables. Subsequently, they determined the feasible set of values for the integer variables in the case of families with 2 - 7 kinds of products. The cardinality of the feasible set is found to be orders of magnitude smaller than the total number of unique combinations of the commonality variables. A n-variable product family will produce n additional integer variables that represent individual blocks of the commonality matrix.

In this paper, we develop surrogates to represent the objectives of designing a family of universal electric motors. Universal electric motors are capable of delivering more torque than any other single phase motors, and can operate using both direct current (DC) and alternating current (AC). The design optimization of the family of universal electric motors (in this paper) involves simultaneous (i) maximization of the efficiency of the motors, and (ii) minimization of the cost of the family of motors chiefly attributed to platform planning. The design of each motor involves seven design variables; the corresponding variable limits are given in Table 3.

The complexities of the performance objective and the net system constraint of the product family design problem depend

Table 3. DESIGN VARIABLE LIMITS OF THE ELECTRIC MOTORS

Design Variable	Lower Limit	Upper Limit
Number of turns on the armature (N_a)	100	1500
Number of turns on each field pole (N_f)	1	500
Cross-sectional area of the armature wire (A_{wa})	0.01 mm^2	1.00 mm^2
Cross-sectional area of the field pole wire (A_{wf})	0.01 mm^2	1.00 mm^2
Radius of the motor (r_o)	10.00 mm	100.00 mm
Thickness of the stator (t)	0.50 mm	10.00 mm
Stack length of the motor (L)	1.00 mm	100.00 mm

on the system complexity of the products being designed. The case study in this paper corresponds to the design of a family of universal electric motors, for which the performance function and the system constraint are fairly nonlinear. The commonality constraint is nonlinear and particularly multimodal. Overall, the product family design problem thus presents a set of four complex nonlinear criterion functions.

The AHF method is used to represent the two objectives (performance objective, f_{perf}, and cost objective, f_{cost}) and the two constraints (system constraint and commonality constraint) as functions of design variables. In the case of universal electric motors, the total number of design variables is $(N_{pro}+1)\times 7$, where N_{pro} represents the number of kinds of product in the family. In this paper, we have considered three cases: (i) 2 products (21 variables); (ii) 3 products (28 variables); and (iii) 4 products (35 variables).

Three-Pane Window Design

The performance of the three-pane window varies over one year [33]. Since environmental conditions vary with geographical locations and time, the thermodynamic properties of the windows should adapt accordingly. The heat transfer and power supplies are optimized under a reasonable set of environmental conditions. The indoor temperature is maintained at a comfortable value. The three-pane window design is an improvement upon existing windows. A simplified schematic of the three-pane window is shown in Fig.2.

The heat transfer simulation model of the side channels and the air gap is created using the computational fluid dynamics (CFD) software Fluent. The model simulates the steady-state heat transfer process. In this study, the middle tinted pane is made of a generic bronze glass, and the other two panes are made of clear glass. The CFD model of the three-pane window is a computational expensive and nonlinear model, which presents significant challenges to surrogate modeling.

To reduce the extensive computational expense of the Fluent model, a surrogate model is developed using AHF. The inputs

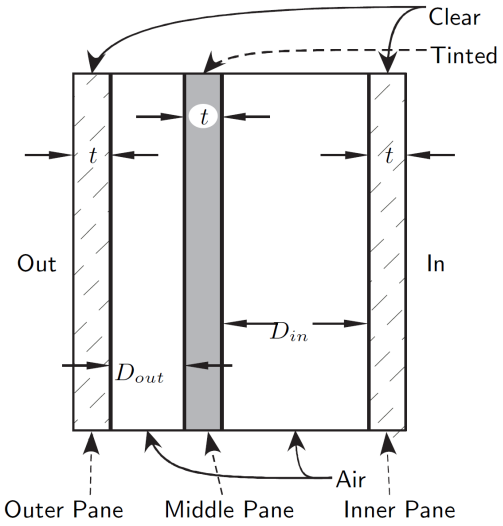


Figure 2. SCHEMATIC OF THE THREE-PANE WINDOW

for the surrogate model are (i) the atmospheric temperature, (ii) the wind speed, and (iii) the solar radiation. The output of the surrogate model is the heat flux through the inner pane, $\dot{Q}_{window}$.

Onshore Wind Farm Cost Model

The Response Surface-Based Wind Farm Cost (RS-WFC) model was introduced by Zhang et al. [34, 35]. In that paper, the RS-WFC model, for onshore wind farms in the U.S, is implemented using the Extended Radial Basis Functions (E-RBF). The RS-WFC model could estimate *the total annual cost of a wind farm per kilowatt installed*, C_t. In the RS-WFC model, C_{LC}, C_{LT} and C_{LM} represent the wage per hour for construction labor, the technician labor and the management labor, respectively; N is the number of turbines in a farm and D is the rotor diameter; the wind turbine lifetime (n), the number of years financed (n_{fi}), the percentage financed (θ), and the interest rate (η) are specified as 20 years, 10 years, 80%, and 10%, respectively. Figure 3 shows

the inputs and output of the RS-WFC model.

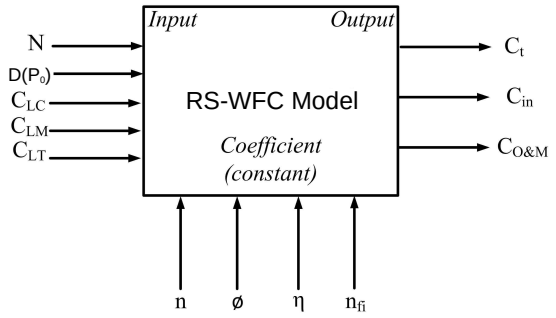


Figure 3. THE INPUT AND OUTPUT OF THE RS-WFC MODEL

In this paper, we estimate the total annual cost of a wind farm using the AHF method. The input parameters to the total annual cost model are (i) the number and (ii) the rated power of wind turbines installed in a wind farm. Data points collected for the state of North Dakota are used to develop the cost model.

EXPERIMENTAL DESIGNS

A sampling technique produces a specified number of data points over the design domain at which the actual function is evaluated. The choice of an appropriate sampling technique is generally considered critical to the performance of any surrogate modeling approach. However, in several practical situations, the designer may not have full control over the choice of design samples (e.g. response surface development from recorded data). In such situations, it is important to select a surrogate modeling approach whose performance is relatively less sensitive to the sampling technique used. In the current paper, we use the following three representative sampling techniques to study their effects on the quality of the resulting surrogates.

Latin Hypercube Sampling

Latin hypercube sampling is a strategy for generating random sample points, which ensures a practically uniform representation of the entire variable domain [36]. A Latin hypercube sample containing n_p sample points (between 0 and 1) over m dimensions is a matrix of n_p rows and m columns. Each row corresponds to a sample point. The n_p values in each column are randomly selected - one each from the intervals, $(0,1/n_p),\ (1/n_p,2/n_p),\ldots,(1-1/n_p,1)$ [36].

Sobol's Quasirandom Sequence Generator

Sobol sequences [37] use a base of two to form successively finer uniform partitions of the unit interval, and then reorder the coordinates in each dimension. The algorithm for generating Sobol sequences is explained in Bratley and Fox, Algorithm 659 [38].

Hammersley Sequence Sampling

The Hammersley Sequence Sampling (HSS) is based on the representation of a decimal number in the inverse radix format, where the radix values are chosen as the first $(m-1)$ prime numbers, m being the number of dimensions. A detailed description of this technique can be found in Kalagnanam and Diwekar [39]. The Hammersley sequence provides a highly uniform set of sample points containing n_p points on an m-dimensional unit hypercube.

Sampling Strategy for Application Problems

For the engineering-design problems in this paper, the training data and test data are obtained through simulation. We consider four sample sizes: (i) 5 times, (ii) 10 times, (iii) 15 times, and (iv) 20 times of the given problem dimension (the number of input variables) for each sampling technique. Table 4 lists the details of the experimental design. The variable limits are given in Table 5. For the onshore wind farm cost problem, we collected data from the *Energy Efficiency and Renewable Energy Program* at the *U.S. Department of Energy* [40].

RESULTS AND DISCUSSION

Selection of Parameters

Through numerical experiments, we found that the following prescribed coefficient values generally produced accurate function estimations. We set $c = 0.9$ for the RBF approach. We use $c = 0.9$ and $\lambda = 4.75$ for the E-RBF approach. The parameter t of the E-RBF approach is fixed at 2 (second degree monomial). For the Kriging method used in this paper, we have used an efficient MATLAB implementation, DACE (design and analysis of computer experiments), developed by Lophaven et al. [41]. The bounds on the correlation parameters in the nonlinear optimization, θ_l and θ_u, are selected to be 0.1 and 20. Under the kriging approach, the order of the global polynomial trend function was specified to be zero. During the cross-validation evaluation process, the number of subsets (q) is specified to be 5. The prescribed values are summarized in Table 6.

Surrogate Modeling Results Using AHF

Complex Engineered Systems Modeling Using AHF The surrogate modeling performance (in terms of the error measures) of the wind farm power generation, the product family design, and the three-pane window design are summarized in Tables 7-11.

Table 4. EXPERIMENTAL DESIGN FOR EACH PROBLEM

Engineering problem	N/N_{pro}	n_{int}	n_{dim}	Sample size (training points)				No. of test points
				$5 \times n_{dim}$	$10 \times n_{dim}$	$15 \times n_{dim}$	$20 \times n_{dim}$	
Wind farm 1	4	0	8	40	80	120	160	100
Wind farm 2	9	0	18	90	180	270	360	200
Wind farm 3	16	0	32	160	320	480	640	320
Wind farm 4	25	0	50	250	500	750	1000	500
Product family 1	2	11	21	105	210	315	420	210
Product family 2	3	13	28	140	280	420	560	280
Product family 3	4	15	35	175	350	525	700	350
Three-pane window		0	3	15	30	45	60	216

N: No. of turbines for wind farm; n_{dim}: No. of variables

N_{pro}: No. of products for product family; n_{int}: No. of integer variables

Table 5. THE LIMITS OF DESIGN VARIABLES

Problem	Variable limit
Wind farm 1	$0 < x_i < 7D_0,\ 0 < y_i < 3D_0,$ where $i = 1,2,3,4$
Wind farm 2	$0 < x_i < 2 \times 7D_0,\ 0 < y_i < 2 \times 3D_0,$ where $i = 1,2,\cdots,9$
Wind farm 3	$0 < x_i < 3 \times 7D_0,\ 0 < y_i < 3 \times 3D_0,$ where $i = 1,2,\cdots,16$
Wind farm 4	$0 < x_i < 4 \times 7D_0,\ 0 < y_i < 4 \times 3D_0,$ where $i = 1,2,\cdots,25$
Product family 1	x_j^1, x_j^2 (see Table 3), $z_j \in \{0,1\},$ where $j = 1,2,\cdots,7$
Product family 2	x_j^1, x_j^2, x_j^3 (see Table 3), $z_j \in \{0,1,2,4,7\},$ where $j = 1,2,\cdots,7$
Product family 3	$x_j^1, x_j^2, x_j^3, x_j^4$ (see Table 3), $z_j \in \{0,1,2,4,7,8,12,16,18,28,32,33,42,52,63\},$ where $j = 1,2,\cdots,7$
Three-pane window	$259K < T_{out} < 309K,\ 0 < U < 21.5m/s,\ 0 < E_{solar} < 1000W/m^2$

Table 6. PARAMETER SELECTION OF THE AHF

Method	Parameter value
E-RBF	$\lambda = 4.75$, c = 0.9, t = 2
RBF	c = 0.9
Kriging	$\theta_l = 0.1$, $\theta_u = 20$
Cross-validation	q = 5

Figures 4, 5, and 6 illustrate the variations in the average RMSE, MAE, and PRESS values (across all functions and sampling techniques), respectively, as a function of the corresponding sample sizes. In the case of wind farm power generation, we averaged all the four wind farms to get the average values of RMSE, MAE, and PRESS for each sample size. For the product family design problem, we averaged the 12 cases (each design with two objectives and two constrains) for this purpose. We observe that the accuracy of the three problems improved with an increase in the sample size.

1. In the case of the three-pane window, it can be seen from Figs. 4(a) and 4(b) that, the HSS technique performs better than the LHS and Sobol techniques.
2. For the wind farm power generation model, the LHS technique performs better than other two sampling methods (Figs. 5(a) and 5(b)).
3. For the product family design problem (with universal motors), a conclusive comparison of the sampling technique performances could not be readily accomplished. In terms of the RMSE values (Fig. 6(a)), both LHS and Sobol seem to perform equally well. However, the HSS yielded better (lower) MAE values (Fig. 6(b)).

Table 7. RESULTS OF THE AHF SURROGATE MODELING ON WIND FARM POWER GENERATION

Problem	No. of	Sobol			HSS			LHS		
	training points	RMSE	MAE	*PRESS*	RMSE	MAE	*PRESS*	RMSE	MAE	*PRESS*
	$5 \times n_{dim}$	0.0587	0.1774	0.0048	0.0356	0.1204	0.0006	0.0355	0.1052	0.0011
Wind farm 1	$10 \times n_{dim}$	0.0403	0.1843	0.0026	0.0360	0.0950	0.0008	0.0314	0.1057	0.0009
4 turbines	$15 \times n_{dim}$	0.0423	0.1797	0.0017	0.0308	0.1159	0.0007	0.0344	0.0960	0.0014
	$20 \times n_{dim}$	0.0426	0.1601	0.0016	0.0310	0.0915	0.0009	0.0318	0.1012	0.0013
	$5 \times n_{dim}$	0.0342	0.1034	0.0017	0.0340	0.1155	0.0004	0.0240	0.0888	0.0005
Wind farm 2	$10 \times n_{dim}$	0.0264	0.0959	0.0009	0.0368	0.1379	0.0005	0.0231	0.0819	0.0005
9 turbines	$15 \times n_{dim}$	0.0257	0.0875	0.0008	0.0278	0.1062	0.0005	0.0231	0.0770	0.0005
	$20 \times n_{dim}$	0.0254	0.0785	0.0007	0.0262	0.1006	0.0004	0.0232	0.0754	0.0005
	$5 \times n_{dim}$	0.0230	0.0871	0.0011	0.0441	0.1174	0.0006	0.0191	0.0652	0.0004
Wind farm 3	$10 \times n_{dim}$	0.0199	0.0831	0.0006	0.0373	0.1387	0.0002	0.0172	0.0584	0.0004
16 turbines	$15 \times n_{dim}$	0.0193	0.0755	0.0005	0.0427	0.2084	0.0002	0.0176	0.0621	0.0003
	$20 \times n_{dim}$	0.0191	0.0772	0.0004	0.0367	0.1229	0.0002	0.0165	0.0661	0.0004
	$5 \times n_{dim}$	0.0231	0.0814	0.0007	0.0610	0.1529	0.0004	0.0171	0.0709	0.0003
Wind farm 4	$10 \times n_{dim}$	0.0178	0.0710	0.0004	0.0565	0.1812	0.0002	0.0162	0.0657	0.0002
25 turbines	$15 \times n_{dim}$	0.0166	0.0612	0.0004	0.0709	0.4725	0.0002	0.0158	0.0663	0.0002
	$20 \times n_{dim}$	0.0166	0.0585	0.0003	0.0541	0.2978	0.0002	0.0160	0.0613	0.0002

Table 8. RESULTS OF THE AHF SURROGATE MODELING ON PRODUCT FAMILY

Problem	No. of	Sobol			HSS			LHS		
	training points	RMSE	MAE	*PRESS*	RMSE	MAE	*PRESS*	RMSE	MAE	*PRESS*
	$5 \times n_{dim}$	0.5814	2.4371	0.4256	0.6249	1.9922	0.2439	0.5715	2.5451	0.8957
2 products	$10 \times n_{dim}$	0.5464	2.4196	0.3220	0.4660	2.2218	0.3432	0.5337	2.3936	0.3301
objective 1	$15 \times n_{dim}$	0.5252	2.4443	0.2950	0.4547	2.1015	0.2761	0.5197	2.3013	0.3128
	$20 \times n_{dim}$	0.5127	2.4201	0.2578	0.4150	2.1498	0.2222	0.4854	2.3846	0.3146
	$5 \times n_{dim}$	0.0700	0.1813	0.0044	0.0489	0.1241	0.0038	0.0721	0.1837	0.0034
2 products	$10 \times n_{dim}$	0.0686	0.1703	0.0039	0.0409	0.1329	0.0011	0.0690	0.1692	0.0037
objective 2	$15 \times n_{dim}$	0.0685	0.1691	0.0038	0.0396	0.1331	0.0007	0.0686	0.1682	0.0041
	$20 \times n_{dim}$	0.0681	0.1653	0.0038	0.0359	0.1309	0.0005	0.0680	0.1655	0.0039
	$5 \times n_{dim}$	1.3669	7.3977	2.7425	1.4026	6.6388	1.2698	1.3336	7.7774	2.1966
2 products	$10 \times n_{dim}$	1.2887	7.7741	2.0228	1.3560	6.0880	1.8955	1.3281	7.4042	1.6712
constraint 1	$15 \times n_{dim}$	1.2572	7.1684	1.7952	1.2995	7.2284	0.9545	1.2937	8.6504	1.4494
	$20 \times n_{dim}$	1.2331	7.2238	1.6565	1.2534	6.7623	0.8641	1.2286	8.1530	1.8657
	$5 \times n_{dim}$	0.0507	0.1586	0.0040	0.0629	0.1762	0.0028	0.0524	0.1840	0.0043
2 products	$10 \times n_{dim}$	0.0492	0.1531	0.0029	0.0516	0.1688	0.0025	0.0479	0.1554	0.0034
constraint 2	$15 \times n_{dim}$	0.0479	0.1450	0.0028	0.0485	0.1849	0.0021	0.0494	0.1502	0.0027
	$20 \times n_{dim}$	0.0463	0.1453	0.0024	0.0460	0.1846	0.0017	0.0486	0.1742	0.0025

Table 9. RESULTS OF THE AHF SURROGATE MODELING ON PRODUCT FAMILY (CONT)

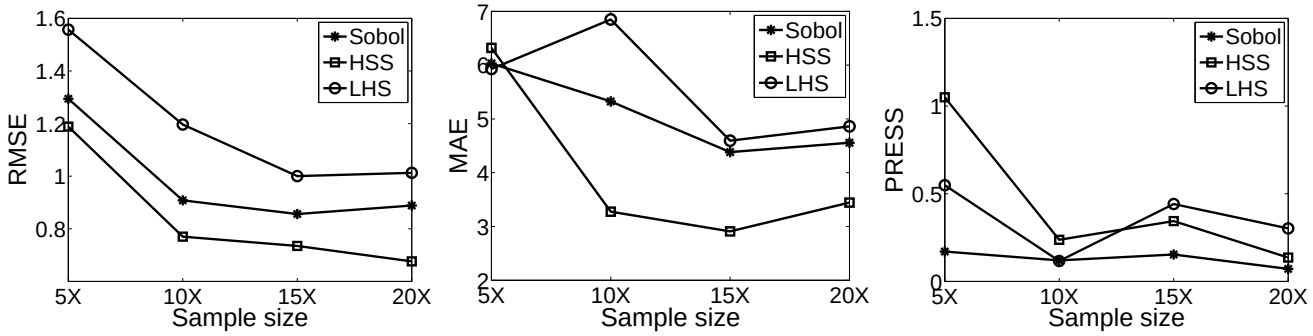
Problem	No. of	Sobol			HSS			LHS		
	training points	RMSE	MAE	*PRESS*	RMSE	MAE	*PRESS*	RMSE	MAE	*PRESS*
	$5 \times n_{dim}$	0.6193	5.3032	0.2488	0.7285	4.5999	0.3069	0.5836	5.1683	0.2558
3 products	$10 \times n_{dim}$	0.5520	4.9811	0.2089	0.5571	4.4519	0.1644	0.5600	5.1275	0.1960
objective 1	$15 \times n_{dim}$	0.5421	4.8942	0.1825	0.5033	3.9131	0.1480	0.5150	4.4526	0.1859
	$20 \times n_{dim}$	0.5088	4.7585	0.1629	0.5197	4.2039	0.1439	0.4947	4.7351	0.2041
	$5 \times n_{dim}$	0.0587	0.1635	0.0028	0.0680	0.2024	0.0016	0.0570	0.1711	0.0029
3 products	$10 \times n_{dim}$	0.0524	0.1589	0.0026	0.0557	0.1548	0.0007	0.0521	0.1675	0.0021
objective 2	$15 \times n_{dim}$	0.0512	0.1532	0.0024	0.0521	0.1535	0.0007	0.0505	0.1495	0.0022
	$20 \times n_{dim}$	0.0506	0.1808	0.0024	0.0503	0.1445	0.0007	0.0498	0.1522	0.0022
	$5 \times n_{dim}$	0.9974	5.6228	1.4164	1.0173	4.7166	0.5065	0.9604	6.5103	0.8506
3 products	$10 \times n_{dim}$	0.9273	5.3516	0.9859	0.9352	4.2291	0.5825	0.8629	5.3556	0.8856
constraint 1	$15 \times n_{dim}$	0.8589	5.3763	0.9515	0.9902	4.9851	0.4871	0.8781	5.5677	0.7500
	$20 \times n_{dim}$	0.8226	4.8800	0.8331	0.9360	4.8892	0.5234	0.8462	5.2641	0.7687
	$5 \times n_{dim}$	0.0281	0.0892	0.0011	0.0333	0.1560	0.0004	0.0288	0.0905	0.0012
3 products	$10 \times n_{dim}$	0.0270	0.1036	0.0009	0.0329	0.1289	0.0005	0.0262	0.0847	0.0008
constraint 2	$15 \times n_{dim}$	0.0269	0.1101	0.0008	0.0308	0.1201	0.0005	0.0261	0.1008	0.0007
	$20 \times n_{dim}$	0.0264	0.1051	0.0008	0.0281	0.1013	0.0004	0.0263	0.1199	0.0007

Table 10. RESULTS OF THE AHF SURROGATE MODELING ON PRODUCT FAMILY (CONT)

Problem	No. of	Sobol			HSS			LHS		
	training points	RMSE	MAE	*PRESS*	RMSE	MAE	*PRESS*	RMSE	MAE	*PRESS*
	$5 \times n_{dim}$	0.4305	3.0440	0.2001	0.5203	2.3004	0.1425	0.3809	3.3622	0.1582
4 products	$10 \times n_{dim}$	0.3922	3.1585	0.1458	0.4824	2.2122	0.0989	0.3764	3.0467	0.1550
objective 1	$15 \times n_{dim}$	0.3730	3.0890	0.1370	0.5018	3.0059	0.0991	0.3651	2.8548	0.1338
	$20 \times n_{dim}$	0.3602	3.0410	0.1259	0.3788	2.7134	0.0832	0.3604	2.9519	0.1325
	$5 \times n_{dim}$	0.0725	0.2511	0.0060	0.0782	0.2156	0.0020	0.0658	0.1913	0.0043
4 products	$10 \times n_{dim}$	0.0610	0.1852	0.0045	0.0847	0.2653	0.0014	0.0624	0.1865	0.0045
objective 2	$15 \times n_{dim}$	0.0603	0.1704	0.0045	0.0859	0.3494	0.0012	0.0645	0.1832	0.0038
	$20 \times n_{dim}$	0.0596	0.1734	0.0042	0.0902	0.2354	0.0014	0.0644	0.1781	0.0036
	$5 \times n_{dim}$	0.7158	3.4054	0.7837	0.9240	4.2043	0.2965	0.7258	3.3029	0.4648
4 products	$10 \times n_{dim}$	0.6885	3.3460	0.6066	0.8930	3.1446	0.3117	0.6951	3.0873	0.4568
constraint 1	$15 \times n_{dim}$	0.6474	3.2213	0.4805	1.0753	5.2102	0.3543	0.6455	3.4550	0.4235
	$20 \times n_{dim}$	0.6317	3.4038	0.4788	0.7328	2.8579	0.3002	0.6524	2.9791	0.4479
	$5 \times n_{dim}$	0.0199	0.0630	0.0004	0.0255	0.1065	0.0002	0.0202	0.0592	0.0005
4 products	$10 \times n_{dim}$	0.0191	0.0562	0.0004	0.0231	0.0946	0.0002	0.0190	0.0571	0.0004
constraint 2	$15 \times n_{dim}$	0.0186	0.0540	0.0004	0.0315	0.1349	0.0002	0.0188	0.0641	0.0004
	$20 \times n_{dim}$	0.0184	0.0532	0.0004	0.0225	0.0845	0.0002	0.0183	0.0638	0.0004

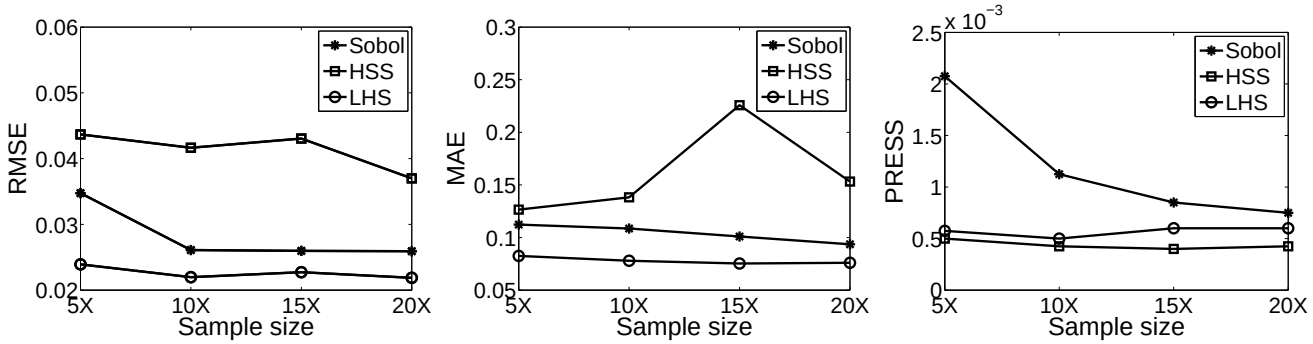
Table 11. RESULTS OF THE AHF SURROGATE MODELING ON THREE-PANE WINDOW

Problem	No. of	Sobol			HSS			LHS		
	training points	RMSE	MAE	*PRESS*	RMSE	MAE	*PRESS*	RMSE	MAE	*PRESS*
	$5 \times n_{dim}$	1.2942	6.0306	0.1703	1.1885	6.3210	1.0498	1.5570	5.9292	0.5488
Three-pane	$10 \times n_{dim}$	0.9081	5.3257	0.1200	0.7702	3.2753	0.2372	1.1964	6.8487	0.1166
window	$15 \times n_{dim}$	0.8562	4.3830	0.1534	0.7348	2.9090	0.3439	1.0002	4.5963	0.4414
	$20 \times n_{dim}$	0.8884	4.5566	0.0713	0.6761	3.4442	0.1359	1.0127	4.8611	0.3021



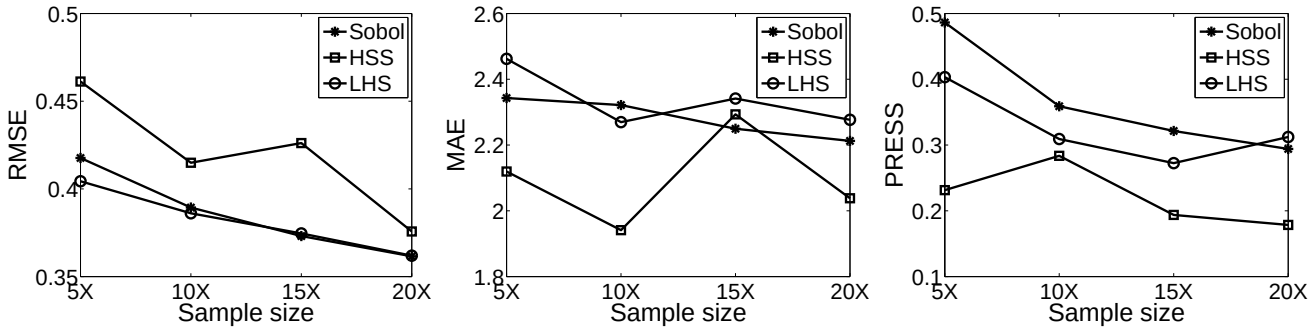
(a) RMSE (b) MAE (c) PRESS

Figure 4. EFFECT OF SAMPLE TECHNIQUE AND SIZE ON SURROGATE MODELING ACCURACY FOR THREE-PANE WINDOW MODEL



(a) RMSE (b) MAE (c) PRESS

Figure 5. EFFECT OF SAMPLE TECHNIQUE AND SIZE ON SURROGATE MODELING ACCURACY FOR WIND POWER GENERATION MODEL



(a) RMSE (b) MAE (c) PRESS

Figure 6. EFFECT OF SAMPLE TECHNIQUE AND SIZE ON SURROGATE MODELING ACCURACY FOR PRODUCT FAMILY MODEL

For each wind farm design scenario (with different numbers of turbines), we estimated the averaged values of RMSE, MAE and PRESS. Figure 7 shows the effect of increasing problem dimensionality on the corresponding surrogate model performance. From Fig. 7 we can observe that: (i) in the case of Sobol and LHS sampling techniques, the values of RMSE, MAE, and PRESS decrease when the dimension increases; (ii) in the case of HSS sampling technique, the AHF method has high accuracy for relatively lower dimensional problems (Figs. 7(a) and 7(b)).

It is helpful to note that, the computational cost for AHF surrogate modeling is slightly higher than that of the component surrogates (QRSM, Kriging, RBF, or E-RBF). Interestingly, the increase in the accuracy of the estimated output accomplished by the AHF approach did not demand an appreciable increase in the computational cost. Such positive attributes further illustrate the applicability of the AHF method to model complex systems.

RS-WFC model Using AHF Method We use 60 training points and 15 test points for the development of the cost model (Table 12). The values of RMSE, MAE, and PRESS are shown in Table 12. From Table 13, it is observed that the largest RAE is 0.48% (estimated at test point 13).

Table 12. EXPERIMENTAL DESIGN FOR THE WIND FARM COST ESTIMATION

Parameter	Value
No. of variables	2
No. of training points	60
No. of test points	15
RMSE	0.2470
MAE	0.5916
PRESS	0.9565

CONCLUSION

This paper presented applications of the Adaptive Hybrid Functions (AHF) to represent complex engineering and economic systems. Three representative sampling techniques (Latin Hypercube Sampling, Sobol's Quasirandom Sequence, and Hammersley Quasirandom Sequence) are selected to study their effects on the quality of the resulting surrogates. We also investigated the influence of the sample size and the problem dimensionality on the performance of resulting surrogate model.

The results show that: (i) the accuracy of the surrogate model improves with an increase in the sample size (which is expected); (ii) the AHF method maintains relatively higher accuracy for high dimensional problems, when the Sobol and the LHS sampling techniques are used; (iii) the AHF method maintains relatively higher accuracy for low dimensional problems, when the HSS sampling technique is used. These case studies successfully establish the wide applicability and the robustness of the AHF method.

ACKNOWLEDGMENT

Support from the National Found from Awards CMMI-0533330, and CMII-0946765 is gratefully acknowledged.

REFERENCES

[1] Braha, D., Minai, A., and Bar-Yam, Y., eds., 2006. *Complex Engineered Systems*. Springer.

[2] Queipo, N., Haftka, R., Shyy, W., Goel, T., Vaidyanathan, R., and Tucker, P., 2005. "Surrogate-based analysis and optimization". *Progress in Aerospace Sciences,* ***41***(1), pp. 1–28.

[3] Wang, G., and Shan, S., 2007. "Review of metamodeling techniques in support of engineering design optimization". *Journal of Mechanical Design,* ***129***(4), pp. 370–380.

[4] Myers, R., and Montgomery, D., 2002. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. Wiley-Interscience; 2 edition.

[5] Forrester, A., and Keane, A., 2009. "Recent advances in surrogate-based optimization". *Progress in Aerospace Sciences,* ***45***(1-3), pp. 50–79.

[6] Wilson, B., Cappelleri, D., Simpson, T., and Frecker, M., 2001. "Efficient pareto frontier exploration using surrogate approximations". *Optimization and Engineering,* ***2***(1), pp. 31–50.

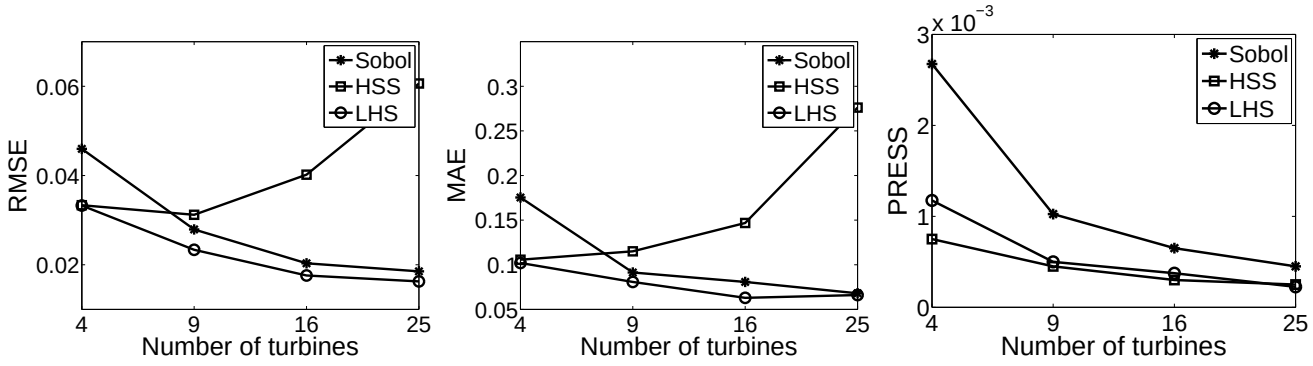
[7] Hardy, R. L., 1971. "Multiquadric equations of topography and other irregular surfaces". *Journal of Geophysical Research,* ***76***, pp. 1905–1915.

[8] Mullur, A., and Messac, A., 2005. "Extended radial basis functions: More flexible and effective metamodeling". *AIAA Journal,* ***43***(6), pp. 1306–1315.

[9] Messac, A., and Mullur, A., 2008. "A computationally efficient metamodeling approach for expensive multiobjective optimization". *Optimization and Engineering,* ***9***(1), pp. 37–67.

[10] Simpson, T., Peplinski, J., Koch, P., and Allen, J., 2001. "Metamodels for computer-based engineering design: Survey and recommendations". *Engineering with Computers,* ***17***(2), pp. 129–150.

[11] Basudhar, A., and Missoum, S., 2008. "Adaptive explicit decision functions for probabilistic design and optimization using support vector machines". *Computers and Structures,* ***86***(19-20), pp. 1904–1917.



(a) RMSE (b) MAE (c) PRESS

Figure 7. EFFECT OF PROBLEM DIMENSIONALITY

Table 13. RESULTS OF ONSHORE WIND FARM COST MODEL

Test point	Rated power $P_0(MW)$	No. of turbines N	Reference annual cost $C_t(\$/KW)$	RS-WFC annual cost $C_t(\$/KW)$	RAE
1	1.00	20	125.07	124.87	0.16%
2	1.00	40	123.71	123.44	0.21%
3	1.00	60	122.43	122.25	0.14%
4	1.25	20	124.72	124.53	0.15%
5	1.25	40	123.06	122.68	0.31%
6	1.25	60	121.51	121.35	0.13%
7	1.50	20	124.41	124.64	0.19%
8	1.50	40	122.65	122.64	0.01%
9	1.50	60	120.89	120.75	0.11%
10	2.00	20	123.93	124.13	0.16%
11	2.00	40	121.45	121.47	0.02%
12	2.00	60	120.32	120.12	0.17%
13	2.50	20	123.40	122.81	0.48%
14	2.50	40	120.29	120.34	0.04%
15	2.50	60	120.33	120.57	0.20%

[12] Yun, Y., Yoon, M., and Nakayama, H., 2009. "Multi-objective optimization based on meta-modeling by using support vector regression". *Optimization and Engineering,* ***10***(2), pp. 167–181.

[13] Zerpa, L., Queipo, N., Pintos, S., and Salager, J., 2005. "An optimization methodology of alkalinesurfactantpolymer flooding processes using field scale numerical simulation and multiple surrogates". *Journal of Petroleum Science and Engineering,* ***47***(3-4), pp. 197–208.

[14] Goel, T., Haftka, R., Shyy, W., and Queipo, N., 2007. "Ensemble of surrogates". *Structural and Multidisciplinary Optimization,* ***33***(3), pp. 199–216.

[15] Zhang, J., Chowdhury, S., and Messac, A., 2011. "Reliability based hybrid functions: Robust surrogate modeling". In AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference.

[16] Simpson, T. W., Peplinski, J. D., Koch, P. N., and Allen, J. K., 2001. "Metamodels for computer-based engineering

design: Survey and recommendations". *Engineering with Computers,* ***17***(2), pp. 129–150.

[17] Sanchez, E., Pintos, S., and Queipo, N., 2008. "Toward an optimal ensemble of kernel-based approximations with engineering applications". *Structural and Multidisciplinary Optimization,* ***36***(3), pp. 247–261.

[18] Acar, E., and Rais-Rohani, M., 2009. "Ensemble of metamodels with optimized weight factors". *Structural and Multidisciplinary Optimization,* ***37***(3), pp. 279–294.

[19] Viana, F., Haftka, R., and Steffen, V., 2009. "Multiple surrogates: How cross-validation errors can help us to obtain the best predictor". *Structural and Multidisciplinary Optimization,* ***39***(4), pp. 439–457.

[20] Acar, E., 2010. "Various approaches for constructing an ensemble of metamodels using local measures". *Structural and Multidisciplinary Optimization,* ***42***(6), pp. 879–896.

[21] Zhou, X., Ma, Y., and Li, X., 2011. "Ensemble of surrogates with recursive arithmetic average". *Structural and Multidisciplinary Optimization*(DOI 10.1007/s00158-011-0655-6).

[22] Zhang, J., Chowdhury, S., and Messac, A., 2011. "An adaptive hybrid surrogate model". *Structural and Multidisciplinary Optimization (accepted)*.

[23] Deb, K., 2001. *Multi-Objective Optimization Using Evolutionary Algorithms*. Wiley.

[24] Jin, R., Chen, W., and Simpson, T., 2001. "Comparative studies of metamodelling techniques under multiple modelling criteria". *Structural and Multidisciplinary Optimization,* ***23***(1), pp. 1–13.

[25] Hastie, T., Tibshirani, R., and Friedman, J., 2001. *The Elements of Statistical Learning*. Springer-Verlag.

[26] Chowdhury, S., Zhang, J., Messac, A., and Castillo, L., 2011. "Unrestricted wind farm layout optimization (uwflo): Investigating key factors influencing the maximum power generation". *Renewable Energy (accepted)*.

[27] Chowdhury, S., Messac, A., Zhang, J., Castillo, L., and Lebron, J., 2010. "Optimizing the unrestricted placement of turbines of differing rotor diameters in a wind farm for maximum power generation". In ASME 2010 International Design Engineering Technical Conferences (IDETC).

[28] Chowdhury, S., Zhang, J., Messac, A., and Castillo, L., 2010. "Exploring key factors influencing optimal farm design using mixed-discrete particle swarm optimization". In 13th AIAA/ISSMO Multidisciplinary Analysis Optimization Conference.

[29] GE, 2010. *GE Energy 1.5MW Wind Turbine Brochure*. General Electric, http://www.gepower.com/.

[30] Simpson, T., Siddique, Z., and Jiao, R., 2006. *Product Platform and Product Family Design : Methods and Applications*. Springer, New York.

[31] Chowdhury, S., Messac, A., and Khire, R., 2010. "Comprehensive product platform planning (cp^3) framework: Presenting a generalized product family model". In AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference.

[32] Chowdhury, S., Messac, A., and Khire, R., 2010. "Developing a non-gradient based mixed-discrete optimization approach for comprehensive product platform planning (cp^3)". In 13th AIAA/ISSMO Multidisciplinary Analysis Optimization Conference.

[33] Zhang, J., Messac, A., Chowdhury, S., and Zhang, J., 2010. "Adaptive optimal design of active thermally insulated windows using surrogate modeling". In AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference.

[34] Zhang, J., Chowdhury, S., Messac, A., and Castillo, L., 2010. "Economic evaluation of wind farms based on cost of energy optimization". In 13th AIAA/ISSMO Multidisciplinary Analysis Optimization Conference.

[35] Zhang, J., Chowdhury, S., Messac, A., Castillo, L., and Lebron, J., 2010. "Response surface based cost model for onshore wind farms using extended radial basis functions". In ASME 2010 International Design Engineering Technical Conferences (IDETC).

[36] McKay, M., Conover, W., and Beckman, R., 1979. "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code". *Technometrics,* ***21***(2), pp. 239–245.

[37] Sobol, I., 1976. "Uniformly distributed sequences with an additional uniform property". *USSR Computational Mathematics and Mathematical Physics,* ***16***(5), pp. 236–242.

[38] Bratley, P., and Fox, B., 1988. "Algorithm 659: Implementing sobol's quasirandom sequence generator". *ACM Transactions on Mathematical Software,* ***14***(1), pp. 88–100.

[39] Kalagnanam, J., and Diwekar, U., 1997. "An efficient sampling technique for off-line quality control". *Technometrics,* ***39***(3), pp. 308–319.

[40] Goldberg, M., 2009. *Jobs and Economic Development Impact (JEDI) Model*. National Renewable Energy Laboratory, Golden, Colorado, US, October.

[41] Lophaven, S., Nielsen, H., and Sondergaard, J., 2002. Dace a matlab kriging toolbox, version 2.0. Tech. Rep. Informatics and mathematical modelling report IMM-REP-2002-12, Technical University of Denmark.