Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

ACCURACY VS. ROBUSTNESS: BI-CRITERIA OPTIMIZED ENSEMBLE OF METAMODELS

Can Cui
Teresa Wu

School of Computing, Informatics, Decision
Systems Engineering
Arizona State University
699 S. Mill Ave.
Tempe, AZ 85281, USA

Mengqi Hu

Department of Industrial and Systems Engineering
Mississippi State University
260M McCain Hall
Starkville, MS 39762, USA

Jeffery D. Weir

Department of Operational Sciences
Air Force Institute of Technology
2950 Hobson Way
Wright-Patterson Afb, Ohio 45433, USA

Xianghua Chu

Shenzhen Graduate School,
Harbin Institute of Technology
Xili University Town
Guangdong 518055, CHINA

ABSTRACT

Simulation has been widely used in modeling engineering systems. A metamodel is a surrogate model used to approximate a computationally expensive simulation model. Extensive research has investigated the performance of different metamodeling techniques in terms of accuracy and/or robustness and concluded no model outperforms others across diverse problem structures. Motivated by this finding, this research proposes a bi-criteria (accuracy and robustness) optimized ensemble framework to optimally identify the contributions from each metamodel (Kriging, Support Vector Regression and Radial Basis Function), where uncertainties are modeled for evaluating robustness. Twenty-eight functions from the literature are tested. It is observed for most problems, a Pareto Frontier is obtained, while for some problems only a single point is obtained. Seven geometrical and statistical metrics are introduced to explore the relationships between the function properties and the ensemble models. It is concluded that the bi-criteria optimized ensembles render not only accurate but also robust metamodels.

1 INTRODUCTION

The growing complexity of real-world systems drives research in developing simulation models to imitate the underlying functionality of the actual system (Banks et al. 2001). However, development of a simulation model requires expertise and is sometimes time consuming. A Metamodel, or a surrogate model, also known as a “model of the model” (Kleijnen 1995), is often built when the simulation is not computationally easily implemented. Metamodels can be used to approximate and thus replace the complex simulation model for engineering decisions. A comprehensive review of metamodeling applications in computer-based engineering design and design optimization can be found in previous references (Simpson et al. 1997; Wang and Shan 2007).

Researchers have conducted comprehensive comparisons of different metamodeling techniques. With respect to accuracy (see Section 3.1 for definition), they have concluded that none of the metamodels can perform uniformly well on diverse problems. Yet, to the best of our knowledge, the comparison studies

978-1-4799-7486-3/14/$31.00 ©2014 IEEE

were conducted without considering various uncertainties (see Section 4.2 for definition and classification) on design variables and parameters. It is expected that the accuracy of metamodels may deteriorate with uncertainties existing in the system. In addition, adding uncertainties into the testing problems will introduce another evaluation metric: robustness, the quality of being robust (see Section 3.1 for definition). Thus, in this research, twenty eight test functions from the literature are used for comparison study. For each function, uncertainties (from a Gaussian distribution) are purposely added, in order to exploit the robustness performance of metamodels. A robust model endeavors to uniformly behave well on each task it performs. In the case of metamodeling, the model is supposed to give uniformly accurate predictions across all sample points, rather than asymmetrically highly accurate predictions in some areas (or points) while poorly accurate predictions in some other areas (or points), which results in a high variability of the local prediction accuracy. Therefore, to develop a metamodel, global accuracy is no longer the only objective. In this context, we must incorporate another evaluation metric which entitles the metamodel insensitivity to uncertainties. Thus, we are looking for a model that is both accurate and robust, in the sense that it maintains a harmonically balanced control on boosting accurate global predictions while preventing local exaggerators. This has practical meaning. Imagine we are developing a model for forecasting the hourly energy consumption for a building. Without consideration of a robustness control, we might build a prediction model with good global accuracy, but a number of local outliers which will deteriorate the decision making and make the system unreliable. Driven by this motivation, a bi-criteria (accuracy and robustness) ensemble optimization framework of three well-known metamodel techniques, namely Kriging (Matheron 1960), Support Vector Regression (SVR) (Drucker et al. 1996; Clarke, Griebsch, and Simpson 2005) and Radial Basis Function (RBF) (Dyn, Levin, and Rippa 1986) is developed. We observe that Pareto frontiers are obtained for most problems, while for a small subset of problems, the Pareto front converges to a single point, which is due to the dominant advantage of one metamodel over others. Moreover, to gain insight on single-point Pareto frontier, we introduce seven geometric and statistical properties describing the function. The relationships between the function properties and metamodel performance are then summarized.

This paper is organized as follows. In Section 2, background knowledge on metamodeling techniques, and ensemble are introduced. The Bi-objective optimization framework for ensemble weight selection is proposed in Section 3, in which accuracy and robustness are elaborated upon. In Section 4, comparison experiments are detailed with insights on the results provided. Conclusions and future research are discussed in Section 5.

2 REVIEW ON METAMODELING TECHNIQUES

Three metamodeling techniques are of interests in this research: Kriging, SVR and RBF, due to their extensive use in surrogate modeling. Each is reviewed in this section. In addition, research in developing an ensemble model (single criteria) is reviewed.

2.1 Kriging

Kriging (also known as Gaussian process regression) is an interpolation method that assumes that the simulation output may be modeled by a Gaussian process; it gives the best linear unbiased prediction of simulation output not yet observed. It generates the prediction in the form of a combination of a global model with local random noise:

$$y(x) = f(x)\gamma + Z(x), \tag{1}$$

where x is the input vector, γ is the weight vector, and $Z(x)$, is a stochastic process with zero mean and stationary covariance of

$$COV\left[Z(x_i), Z\left(x_j\right)\right] = \sigma^2 R(x_i, x_j), \tag{2}$$

where σ^2 is the process variance, $R(x_i, x_j)$ is an n by n correlation matrix where n is the sample size of the training data. R is usually depicted by a Gaussian correlation function, $exp(-\theta(x_i - x_j)^2)$ with parameter θ. Kriging is one of the most intensively studied metamodels among the others because it is flexible with a number of correlation functions and regression functions (with polynomial degree of 0, 1 or 2) to choose from. It is generally acknowledged that the Kriging model outperforms other metamodels on nonlinear problems but is not easy to develop due to the time consumed by Maximum Likelihood Estimation of the correlation parameters in R which is a multi-dimensional optimization problem (Simpson et al. 2001). In this study, we use deterministic Kriging.

2.2 Support Vector Regression (SVR)

Support vector regression is analogous to support vector classification, which tries to maximize the distance between two classes of data by selecting two hyperplanes in a way that they separate the training data with no points between them, the mathematical form of SVR is:

$$f(x) = \langle \omega \cdot x \rangle + b, \tag{3}$$

where ω is the norm vector to the hyperplane and $\frac{b}{\|\omega\|}$ determines the offset of the hyperplane from the origin. The goal is to find a hyperplane that separates the data points optimally without error and separates the closest points with the hyperplane as far as possible. Thus, it can be constructed as an optimization problem:

$$\text{Minimize: } \frac{1}{2}|\omega|^2$$
$$\text{subject to } \begin{cases} y_i - \langle \omega \cdot x_i \rangle - b \leq \varepsilon \\ \langle \omega \cdot x_i \rangle + b - y_i \leq \varepsilon \end{cases}. \tag{4}$$

According to the duality principle, the nonlinear regression problem is given by:

$$f(x) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i)\, k\langle x_i \cdot x_j \rangle + b, \tag{5}$$

where α_i^* and α_i are two introduced dual variables, and $k\langle x_i \cdot x_j \rangle$ is a kernel function, e.g. Gaussian kernel. It is interesting to note that research has demonstrated SVR performs well on a number of problems, and even outperforms Kriging for some cases (Wang and Shan 2007). However, most studies have been empirical.

2.3 Radial Basis Function (RBF)

RBF is used to develop interpolation on scattered multivariate data. A RBF is a linear combination of a real-valued radially symmetric function, $\emptyset(x)$, based on distance from the origin,

$$f(x) = \sum_{i=1}^{n} \theta_i\, \emptyset(\|x - x_i\|), \tag{6}$$

where θ_i is the unknown interpolation coefficient determined by the least-squares method, n is the number of sampling points and $\|x - x_i\|$ is the Euclidean norm of the radial distance from design point x to the sampling point x_i. Fang et al. (2005) found RBF performs well on highly nonlinear problems.

2.4 Single-Criteria Ensembles

Ensemble is a technique of combining multiple models in order to create a stronger overall representation of the system studied. Acar and Rais-Rohani (2009) proposed an ensemble of surrogate models and demonstrate that the resulting hybrid model improves the prediction accuracy.

In general, a weighted average surrogate model is (Bishop 1995):

$$\hat{y}_{avg}(x) = \sum_{j=1}^{M} w_j(x) \hat{y}_j(x), \tag{7}$$

where $\hat{y}_{avg}(x)$ is the ensemble response prediction by the weighted sum of each surrogate model response prediction $\hat{y}_j(x)$, $w_j(x)$ is the weight corresponding to the j^{th} surrogate, and M is the number of surrogate models. In addition, to achieve an unbiased estimator of the ensemble, the weights $\omega_j(x)$ must sum to one.

It is expected that an outperforming surrogate is deemed to be assigned larger weight while an underperforming surrogate deserves smaller weight in an ensemble. The evaluation metrics on "goodnessof-fit" are considered to be a confident measurement of the model accuracy. Acar and Rais-Rohani (2009) proposed a weight selection approach by solving an optimization problem to identify the weight for each surrogate that would minimize a selected error metric (e.g., root mean square error). That is,

$$\text{Minimize:} \quad \varepsilon_e = Err\{\hat{y}_e\left(w_j(x), \hat{y}_j(x), y_j(x)\right), j = 1, \dots M\}$$
$$\text{subject to} \quad \sum_{j=1}^{M} w_j(x) = 1,$$
$$w_j(x) \geq 0, j = 1, \dots, M, \tag{8}$$

where *Err{}* is the error metric of the built ensemble predicted response. We want to note that this framework used a single criteria (in this case, accuracy) to develop the ensemble. In the next section, we propose a bi-objective optimization problem that not only considers accuracy but also robustness in a weight selection mechanism.

3 FRAMEWORK OF BI-CRITERIA OPTIMIZATION ON ENSEMBLE WEIGHT FACTORS SELECTION

3.1 Metamodel Performance Criteria

In a real-world engineering system, the impacts of uncertainties and noises may not be negligible. This requires a model (surrogate, ensemble) to perform not only accurately but also robustly. Accuracy is defined as how well the metamodel predicts the unknown data, which is usually evaluated by an error measurement. Robustness measures how consistently a model performs over different problems (Li et al. 2010), or how consistently a model performs over different design regions on one problem. The accuracy metrics reflect the degree of closeness of the metamodel measurement outputs $\hat{y}$ to true output *y* which is obtained from the deterministic input. One global measurement for accuracy is Normalized Root Mean Square Error (*NRMSE*)

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} / (y_{max} - y_{min}). \tag{9}$$

Robustness considers the metamodel's ability to consistently achieve similar accuracy over the whole design space, thus it could be evaluated by the standard deviation of local prediction error (*SDPE*):

$$SDPE = std. |y_i - \hat{y}_i| \; \forall i \in n, \tag{10}$$

where $|y_i - \hat{y}_i|$ is the absolute difference between the true output and prediction output at data point *i*, and *n* is the number of data points. Formulation of Normalized SDPE is given in the next section.

3.2 Bi-Criteria Optimization Framework

Given *NRMSE* and *NSDPE*, a bi-objective optimization problem is introduced:

$$\text{Minimize:} \quad (NRMSE_{en}, NSDPE_{en}),$$

where

$$NRMSE_{en} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[y_i - \sum_{j=1}^{M}(w_j * \hat{y}_{ij})]^2} / (y_{max} - y_{min}),$$
$$NSDPE_{en} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left\{\left|y_i - \sum_{j=1}^{M}(w_j * \hat{y}_{ij})\right| - \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \sum_{j=1}^{M}(w_j * \hat{y}_{ij})\right|\right\}^2} / (y_{max} - y_{min}),$$
$$\text{subject to} \quad \sum_{j=1}^{M} w_j = 1,$$
$$w_j \geq 0, j = 1, \dots, M, \tag{11}$$

where M is the number of metamodel techniques in the ensemble (in this study, M=3, $j \leq M$, j=1 is Kriging; j=2 is SVR; j=3 is RBF), y_i is the true fitness value at point i; y_{max} and y_{min} are the maximum and minimum value of y_i over n sample points; $\hat{y}_{ij}$ is the prediction at point i by metamodel j, w_j is the ensemble weight of metamodel j. Note both objectives are normalized for comparison convenience among functions of various fitness magnitudes.

We use a weighted sum to transform the multi-objective problem into a single objective function as in (Hwang and Masud 2012):

$$\begin{aligned} &\text{Minimize: } (\beta_1 * NRMSE_{en} + \beta_2 * NSDPE_{en}), \\ &\text{subject to } \beta_1 + \beta_2 = 1, \\ &\beta_1,\ \beta_2 \geq 0, \end{aligned} \tag{12}$$

where β_1 and β_2 are importance weights for the two objectives. In this research, a sequential quadratic programming algorithm (SQP) (Nocedal and Wright 2006), an iterative method for nonlinear optimization, is used to solve this bi-criteria optimization problem.

4 METAMODELING EXPERIMENTS

4.1 Test Functions

The test functions used in this research are composed of 28 black-box benchmark functions that are originally proposed for competition on real parameter single objective optimization (Liang and Suganthan 2013). There are three main categories of black-box functions: 5 unimodal functions, 15 multimodal functions and 8 composition functions, each of which has the range $[-100,100]^D$, where D denotes the number of dimensions of the function. These three function categories have distinguishing characteristics, for example, a unimodal function has only one global optima (valley/peak), while a multimodal can have many local optima (valleys/peaks), and a composition function is composed from unimodal and multimodal functions.

4.2 Design of Experiment

4.2.1 Training, Validation and Test Data Generation

Latin hypercube sampling (LHS) is a statistical sampling method used in construction of computer experiments for good uniformity and coverage from a multidimensional distribution (Eglajs and Audze 1977). It is widely used because the sample size is not strictly determined by the number of dimensions of the underlying simulation model (Zhang et al. 2013). For fitting a quadratic polynomial model, the minimum number of sample points is $(d + 1) * (d + 2)/2$, where d denotes the number of variables (Jin, Du, and Chen 2003; Shan and Wang 2005). We generate training data of two distinct dimensions of design variables, 5 and 10, with the sampling size varying from 21 to 66 respectively, using LHS. With comparative settings of different dimensions of the design variables, we are able to see if the dimensionality impacts the performance of the metamodels. In order to avoid over fitting, we implement a k-fold cross-validation process to the training data (Kohavi 1995). Here k=2, because we have a relatively small number of sampling data points. Test data is randomly generated with 1000 data points over the design space, which is treated as a testing data set for the metamodel performance evaluation.

Given the test problems, we purposely add uncertainties to the inputs and the outputs of the functions. Specifically, parametric uncertainty (variability on each input variable) and residual uncertainty (variability on the outputs) are added in the black-box functions. For the parametric uncertainty, a random number within 10% of each design variable range of [-100,100], [-10, 10] is generated and added. For the residual uncertainty, we add a random number from a Normal distribution ~N (0, σ^2), where σ^2=10% times the logarithm of the difference between the maximum and minimum of the training output for each black-box problem. It is easily understood that with the existence of uncertainties, the same input may not generate the same output, thus 25 simulation replicates are conducted to mitigate the randomness. An average value of the 25 output replicates ($\bar{y}$) is taken as the output for training data while the input for training data takes its nominal value (x), which is the exact value without noise contamination. And the same operation is applied to test data.

4.2.2 Parameter Tuning

The grid search method (Chang and Lin 2011) is applied in this study to select the optimal parameters that give the minimum validation error, then the test data is applied to the optimally trained model to obtain its generalization error. For example, based on experiments, the degree of polynomial regression function for Kriging is selected as 0 for a Sphere function, while for Composition Function 1, it is 2. Table 1 summarizes the parameter search space for each model.

Table 1: Parameter Search Space for Each Metamodel.

Technique	Parameters
Kriging	Regression function: polynomial (degree=0, 1, 2); Correlation function: cubic ($1-3\varepsilon^2+2\varepsilon^2, \varepsilon = min\{1, \theta\lvert x_i - x_j\rvert\}$), exponential ($exp(-\theta\lvert x_i - x_j\rvert)$) and Gaussian.

SVR	SVR type: nu-SVR (nu=0,0.25,0.5,0.75,1);
	Kernel function: RBF;

RBF	RBF center: set as equal to input; Basis function: Gaussian or polyharmonicspline.

4.3 Result Analysis

In this section, the performance of the three metamodels simulated on 28 black-box problems under parametric and residual uncertainties is analyzed with statistical and graphical methods. An ensemble solely based on error minimization is first built and compared to the three metamodels in terms of accuracy. Then a bi-criteria ensemble is built for analysis on the relationship between accuracy and robustness, and between function properties and metamodel performance.

4.3.1 Metamodels' Performance Comparison using Accuracy

In response to different distributions among the metamodels, a nonparametric statistics test is preferable as it has fewer restrictive assumptions about data levels and underlying probability distributions. The Friedman test (Friedman 1937), a non-parametric statistical test, is applied to detect differences in error measurements across the three metamodels and the ensemble.

It is observed that there is no significant difference on the model performance by statistically comparing the mean of each model's NRMSE between the two different settings of dimensions, 5 and 10, thus we only report the results of testing problems from the 10 dimension problems. The average NRMSE and standard deviation of NRMSE across 28 functions, which provides a rough indication of accuracy and robustness, for Kriging, SVR, RBF and their ensemble are 0.106, 0.131, 2.726, and 0.086, and 0.465, 0.055, 10.034 and 0.050, respectively. It is noticeable that RBF gives two outliers, function 8 and function 20. From a geometric point of view, these two functions share one common characteristic: an almost flat response surface for most of the space with an abrupt altitude change within a small area. Moreover, the change in altitude is not uniformly spreading, but is an asymmetrical pattern centered from a point. By examining the fundamental modeling mechanism of RBF, which makes inferences based on radial symmetrical distance, the outliers can be explained. Specifically, RBF might be inaccurate for a response surface with an asymmetrical shape. Thus, to effectively compare the three metamodels performance, we remove the outliers, which results in an average NRMSE of 0.110 that is comparable to Kriging and a standard deviation of NRMSE of 0.118 that is still inferior to the others. As a result, we need to keep in mind that RBF might not be a good choice for a robust model, especially when the response surface of the system is unknown. The Friedman's test results indicate that the four groups of metamodels have significant differences due to significant P-values given in Table 2. The Nemenyi's test (Nemenyi 1963) further divides the four metamodels types into three groups labeled A, B, and C, in which Kriging and RBF have similar rankings, whereas ensemble is ranked first and SVR is ranked last, given by Table 3. These statistical tests to some extent demonstrate that Kriging and RBF generally perform better than SVR when all of the 28 functions are aggregately considered, while even so, the ensemble of the three metamodels performs even better than any stand-alone metamodel, which is indicated by the sum of ranks and mean of ranks values in the two-tailed test. Therefore, building an ensemble to take advantage of each metamodel's strength and mitigate their weakness is an effective way to avoid biased and sub-optimal solutions. In a word, the comprehensive performance among all models could be summarized as: ensemble outperforms all, followed by Kriging, and RBF with two outliers, and the last is SVR.

However, the ensemble we obtained to this extent is built solely based on NRMSE minimization, which is not sufficient when uncertainties exist, so in the next section, a bi-criteria ensemble is built to further incorporate robustness.

Table 2: Friedman's Test on NRMSE.

Q (Observed value)	Q (Critical value)	DF	p-value (Two-tailed)	alpha
52.325	7.815	3	< 0.0001	0.05

Table 3: Multiple Pairwise Comparisons Using Nemenyi's Procedure/ Two-tailed Test on NRMSE.

Sample	Sum of ranks	Mean of ranks	Groups		
Ensemble	34.000	1.214	A		
RBF	71.500	2.554		B	
Kriging	71.500	2.554		B	
SVR	103.000	3.679			C

4.3.2 Bi-Criteria Ensemble

In this section, the case of dimension=10 is selected as an illustrative example. 28 bi-criteria ensembles are built from the optimization framework defined in Section 3. One hundred and one single objective optimization problems are solved by changing β_1 (see (12)) from 0 to 1 with 0.01 step size. The Pareto optimality plots of most of black-box problems show that the two objectives are conflictive.

For clarification, we take one un-normalized unimodal black-box problem, Rotated High Conditioned Elliptic Function's Pareto Frontier as an example, which is shown in Figure 1. The two end points in the Pareto frontier indicate when RMSE fully controls the optimization problem, it reaches its minimum, while SDPE reaches its maximum, and vice versa. The middle point reflects when RMSE and SDPE are of equal importance. To illustrate, see Figure 2, the sum of the two optimal solutions for each normalized objectives is plotted over all iterations, as the point is located in the lower middle of the curve, we can intuitively reconsider the two objectives as two costs, then to reach the minimum cost, we can plot the sum of the two objectives w.r.t. the solutions, then in this case, we find that the 51st solution gives the minimum cost.

As we stated, most of the problems have a typical Pareto Frontier as Figure 1 shows, however, some of them do not, instead, their Pareto Frontier is a single point. The reason for this phenomenon is because for some of the black-box functions, one metamodel performs significantly better than others. In each iteration of the optimization problem, it is assigned the full weight of 1 while the others are evicted from the ensemble. It is necessary to study these functions' properties so we can gain some useful insights on the relationship between the response surface properties and metamodel performance. In the next Section, we employ some geometrical and statistical metrics to evaluate the properties of the black-box functions with a single-point Pareto frontier.

4.3.3 Properties of Black-box Functions with Single-point on Pareto Frontier

In this section, 7 properties are introduced. The definitions are listed in Table 4. The response surface of each function is meshed by star-studded sampling points. Each dimension of sampling points are uniformly distributed within the range of [-100,100], with the interval of 1. “Gradient of Response Surface Point” denotes the first derivative of a function evaluated at a sampling point on the surface. This measures the change rate of bumpiness. A single-point Pareto frontier is detected in 7 out of the 28 functions and its corresponding predictor metamodel is given in Table 5.

To intuitively understand the data given in Table 5, plots of the seven metrics for the 28 functions are provided in Figures 3-9.

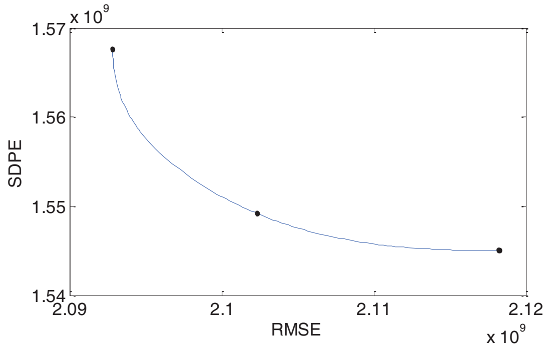

Figure 1: Pareto Frontier of Rotated High Conditioned Elliptic Function.

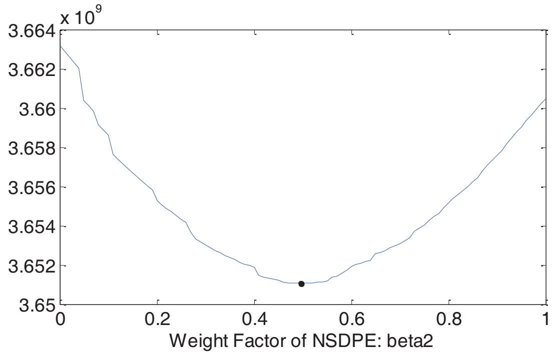

Figure 2: Sum of RMSE & SDPE on Rotated High Conditioned Elliptic Function.

Table 4: Definition and Calculation Methods of 7 Geometric and Statistical Properties.

Properties		Description
Gradient of Response Surface Point	Mean	Mean of absolute values of gradients on all response surface points generally evaluates how steep and rugged the surface is by looking into its rate of change on each point
	Median	Median of absolute values of gradients on all response surface points finds the median value of all rates of change on each point
	Std.	Standard deviation of absolute values of gradients on all response surface points evaluates the rate of the rate of change on each point
	Min	Minimum of absolute values of gradients on all response surface points gives a lower bound of rate of change on each point
	Max	Maximum of absolute values of gradients on all response surface points gives an upper bound of rate of change on each point
Std. of Function values		Standard deviation of function values on all design points can evaluate how bumpy the surface is by looking into its each value’s deviation from the mean
Biggest difference		Averaged local biggest difference of function values can evaluates the average bumpiness by looking into the difference between “valley” and “peak” on each local area

It is noticeable that the single metamodel solution for the Pareto point is either Kriging (functions 3, 7, 8, 20) or RBF (functions 4, 12, 19). This indicates that Kriging and RBF perform better than SVR in terms of both accuracy and robustness on these functions. In addition, except for max of gradient of response surface point for Kriging, and std. of function values for RBF, all the rest of the plots show that the functions with a single Pareto point have relatively smaller values of the seven evaluation metrics. In fact, functions with higher values of evaluation metrics tend to be more complicated and changeful, and vice versa. Therefore, we can intuitively interpret the values of the evaluation metrics as a measure of a function's complexity, i.e. the lower value indicates a simpler form of the response surface. Therefore, we can conclude that for a black-box problem with a relatively simple form of a response surface, there is a higher chance that one metamodel will overwhelmingly perform better than others and take full charge of the ensemble development. In another words, because the function is simple to predict, any metamodel can easily produce a good enough approximation, so the performance difference among all the metamodels is marginal. When dealing with a complicated response surface, it is difficult with only one metamodel to achieve a good approximation, and the strength of all metamodels must be aggregated in order to derive a good approximation of the surface, so an ensemble built from all metamodels is naturally required.

Table 5: Evaluation Metrics Statistics Summary for Black-box Functions with Single Pareto Point.

Properties	Gradient of Response Surface Point					Std. of Function values	Biggest difference	Single Pareto
Function Number	Mean	Median	Std.	Min	Max			
3	0.0001	0.0000	0.0030	0.0000	0.1841	0.0148	0.0032	Kriging
4	0.0030	0.0027	0.0024	0.0000	0.0109	0.1842	0.0596	RBF
7	0.0004	0.0000	0.0056	0.0000	0.5441	0.0163	0.0049	Kriging
8	0.0158	0.0142	0.0111	0.0000	0.2804	0.0399	0.1256	Kriging
12	0.0018	0.0008	0.0027	0.0000	0.0158	0.1319	0.0308	RBF
19	0.0021	0.0003	0.0043	0.0000	0.0272	0.1389	0.0415	RBF
20	0.0134	0.0000	0.0395	0.0000	0.4541	0.0554	0.0804	Kriging

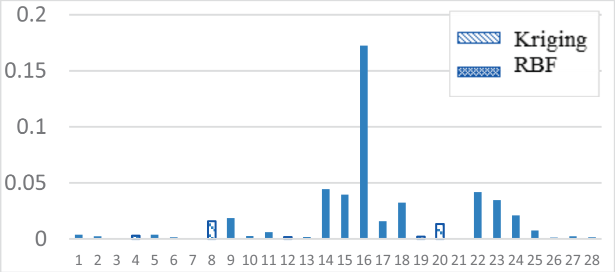

Figure 3: Mean of Gradient of Response Surface.

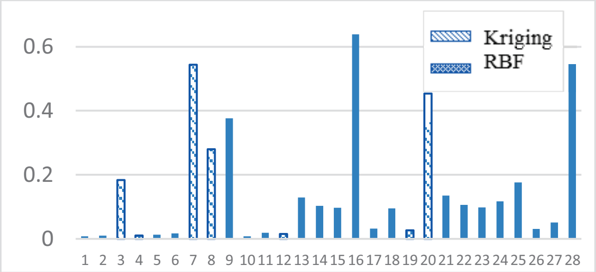

Figure 4: Max of Gradient of Response Surface.

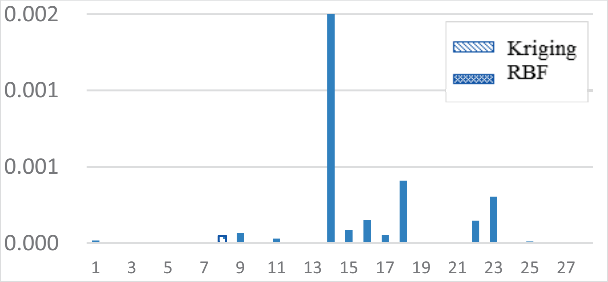

Figure 5: Min of Gradient of Response Surface.

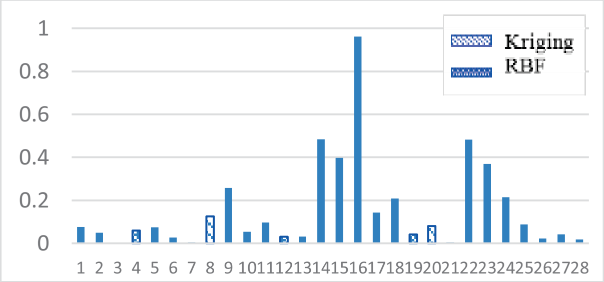

Figure 6: Biggest Difference.

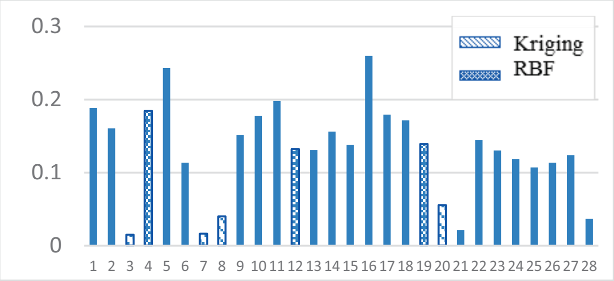

Figure 7: Std. of Function Values.

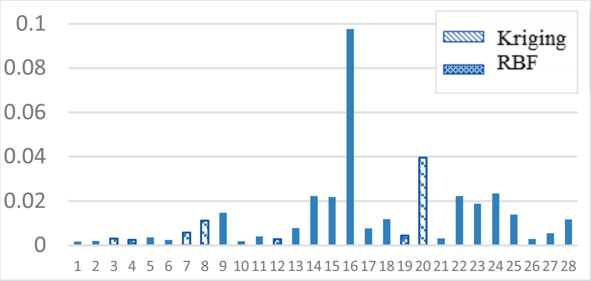

Figure 8: Std. of Gradient of Response Surface.

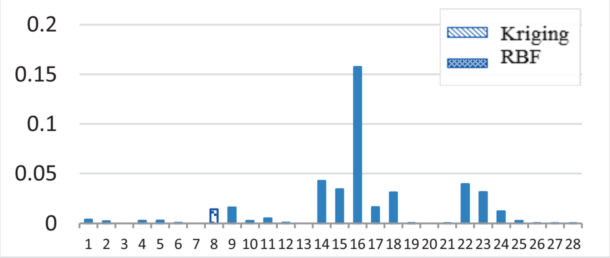

Figure 9: Median of Gradient of Response Surface.

5 CONCLUSION

A new approach on selecting weight factors for developing an ensemble of metamodels is proposed in this research. Based on earlier efforts, this approach treats the development of ensembles by extending a single objective optimization to a bi-criteria optimization problem, which is achieved by simultaneously optimizing on both accuracy and robustness. The addition of uncertainties is of practical rationale as in real world design engineering, noises and errors are inevitable. For a complete study, 28 black-box problems of distinct characteristics are set as test problems to evaluate the performance of each metamodel in terms of accuracy and robustness and to test our proposed approach.

When both accuracy and robustness are considered in selecting weight factors of the ensemble, different Pareto Frontiers are achieved. 21 out of the 28 functions have typical Pareto Frontiers where the two objectives present a conflictive relationship. However, 7 out of the 28 functions display a single Pareto point, which is because one metamodel overwhelmingly performs better than the others. Seven evaluation metrics are introduced to empirically analyze the single Pareto point phenomenon. The results show that functions with a single Pareto point are more likely to have a simpler form of response surface than others.

In summary, some of the key findings in this study are:

- It’s not conclusive that any metamodel predicts better than others across diverse problems;
- RBF performs poorly on problems with an asymmetrically distributed response surface, which indicates it is less robust than the other two models;
- Kriging and RBF statistically perform better than SVR in terms of accuracy. An ensemble of three metamodels statistically performs better than any stand-alone metamodel. Therefore, it is recommended that if computationally permissible, an ensemble is built instead of a stand-alone metamodel, given that sufficient prior knowledge regarding each member of ensemble is fully understood by users;
- A novel ensemble weight selection mechanism is successfully implemented, which renders not only accurate but also robust surrogate models, and provides us with new insights on the relationship between accuracy and robustness, given uncertain conditions;

- In most cases, accuracy and robustness have a conflictive relationship, of which the Pareto Frontier looks like the left half of a convex parabola. We suggest equal weight on both objectives for a balanced solution.

In future research, we plan to implement bi-criteria ensemble metamodel to replace an expensive simulation model of a real world problem, e.g., building energy consumption, in order to achieve an accurate and robust metamodel under uncertain conditions with unequal variance. Some algorithms that are more adaptive to a stochastic environment will be considered instead, e.g., stochastic Kriging. What's more, relationship between the problem properties and metamodeling performance will also be further explored. According to the meta-learning concept (Vilalta and Drissi 2002), which studies how to properly select a model for a specific problem based on experience, it would be more efficient and effective if connections between the training data characteristics and metamodel performance can be established.

REFERENCES

Acar, E., M. Rais-Rohani. 2009. "Ensemble of Meta-models with Optimized Weight Factors." *Structural and Multidisciplinary Optimization* 37: 279–294.

Banks, J., J. S. Carson, B. L. Nelson, and D. M. Nicol. 2000. *Discrete-Event System Simulation* 3rd ed. Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*. New York: Oxford University Press.

Chang, C., C. Lin. 2011. "LIBSVM: A Library for Support Vector Machines." *ACM Transactions on Intelligent Systems and Technology (TIST)* 2: Article No. 27.

Clarke, S. M., J. H. Griebsch, T. W. Simpson. 2005. "Analysis of Support Vector Regression for Approximation of Complex Engineering Analyses." *Journal of Mechanical Design* 127(11):1077–1087.

Drucker, H., C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. 1996. "Support Vector Regression Machines," *Advances in Neural Information Processing Systems* 9, NIPS, 155–161, MIT Press.

Dyn, N., D. Levin, S. Rippa. 1986. "Numerical Procedures for Surface Fitting of Scattered Data by Radial Basis Functions." *SIAM Journal on Scientific and Statistical Computing* 7(2): 639–659.

Eglajs, V., P. Audze. 1977. "New Approach to the Design of Multifactor Experiments," *Problems of Dynamics and Strengths* 35: 104–107.

Fang, H., M. Rais-Rohani, Z. Liu, M. F. Horstemeyer. 2005. "A Comparative Study of Metamodeling Methods for Multiobjective Crashworthiness Optimization." *Computers & Structures* 83: 2121–2136.

Friedman, M. 1937. "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *Journal of the American Statistical Association* 32 (200): 675–701.

Hwang, C. L., A. S. M. Masud. 1979. "Multiple Objective Decision Making, Methods and Applications: A State-of-the-art Survey." *Economics and Mathematical Systems* 164.

Jin, R., X. Du, and W. Chen. 2003. "The Use of Metamodeling Techniques for Optimization under Uncertainty." *Structural and Multidisciplinary Optimization* 25: 99-116.

Kleijnen, J. 1995. "Theory and Methodology Verification and Validation of Simulation Models." *European Journal of Operational Research* 82:145-162.

Kohavi, R. 1995. "A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection." *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* 2 (12): 1137–1143.

Li, Y. F., S. H. Ng, M. Xie, and T. N. Goh. 2010. "A Systematic Comparison of Metamodeling Techniques for Simulation Optimization in Decision Support Systems." *Applied Soft Computing* 10: 1257–1273.

Liang, J. J., B. Y. Qu, and P. N. Suganthan. 2013. "Problem Definitions and Evaluation Criteria for the CEC 2013 Special Session on Real-Parameter Optimization." Technical Report Computational

Intelligence Laboratory, Zhengzhou University, Zhengzhou China and Technical Report, Nanyang Technological University, Singapore.

Matheron, G. 1960. "Krigeage d'un Panneau Rectangulaire Par sa Périphérie," *Note géostatistique* no.28, CG, Ecole des Mines de Paris.

Nemenyi, P. B. 1963. "Distribution-free Multiple Comparisons," *PhD thesis*, Princeton University.

Nocedal, J., and S. J. Wright. 2006. *Numerical Optimization*. Springer.

Shan, S., and G. G. Wang. 2005. "An Efficient Pareto Set Identification Approach for Multiobjective Optimization on Black-Box Functions." *Transactions of the ASME* 127: 866-874.

Simpson, T. W., J. Peplinski, P. N. Koch, and J. K. Allen. 1997. "On the Use of Statistics in Design and the Implications for De terministic Computer Experiments." *ASME Design Engineering Technical Conferences* DTM-3881.

Simpson, T. W., J. Peplinski, P. N., Koch, and J. K. Allen. 2001. "Metamodels for Computer-based Engineering Design: Survey and Recommendations." Engineering with Computers 17(2): 129-150.

Vilalta, R., and Y. Drissi. 2002. "A Perspective View and Survey of Meta-Learning." *Artificial Intelligence Review* 18: 77–95.

Wang, G. G., and S. Shan. 2007. "Review of Metamodeling Techniques in Support of Engineering Design Optimization." *Journal of Mechanical Design* 370.

Zhang, S., P. Zhu, W. Chen, and P. Arendt. 2013. "Concurrent Treatment of Parametric Uncertainty and Metamodeling Uncertainty in Robust Design," *Structural and Multidisciplinary Optimization* 47: 63-76.

AUTHOR BIOGRAPHIES

Can Cui is a research assistant of Industrial Engineering program in Arizona State University. Her research mainly deals with dynamic system modeling, simulation and optimization. She is a Ph.D. student. Her email address is ccan1@asu.edu.

Mengqi Hu received the Ph.D. degree in industrial engineering from Arizona State University, in 2012. He is currently an Assistant Professor with the Department of Industrial and Systems Engineering, Mississippi State University. His current research interests include complex system modeling, simulation and optimization, swarm intelligence and evolutionary computation, with applications in energy systems, healthcare systems. His email address is mhu@ise.msstate.edu.

Jeffery D. Weir received the Ph.D. degree in industrial and systems engineering from the Georgia Institute of Technology. He is currently an Associate Professor with the Department of Operational Sciences, Air Force Institute of Technology. His current research interests include decision analysis and transportation modeling. His email address is Jeffery.Weir@afit.edu.

Xianghua Chu is currently working toward his Ph.D. in Management Science, Harbin Institute of Technology, China. His current research interests include swarm intelligence, evolutionary algorithms and intelligent decision support. His email address is xianghua.chu@gmail.com.

Teresa Wu received the Ph.D. degree in industrial engineering from the University of Iowa, in 2001. She is currently a Professor with the School of Computing, Informatics, Decision Systems Engineering, Arizona State University. Her current research interests include distributed decision support, distributed information systems, and health informatics. Her email address is Teresa.Wu@asu.edu.