53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference
20th AI
23 - 26 April 2012, Honolulu, Hawaii

AIAA 2012-1429

53^{rd} AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, April 2012, Honolulu, HI

Improving the Accuracy of Surrogate Models Using Inverse Transform Sampling

Junqiang Zhang*
Rensselaer Polytechnic Institute, Troy, New York 12180

Achille Messac†
Syracuse University, Syracuse, New York 13244

Jie Zhang*and Souma Chowdhury*
Rensselaer Polytechnic Institute, Troy, New York 12180

This paper applies inverse transform sampling to sample training points for surrogate models. Inverse transform sampling uniformly generates a sequence of real numbers ranging from 0 to 1 as the probabilities at sample points. The coordinates of the sample points are evaluated using the inverse functions of Cumulative Distribution Functions (CDF). The inputs to surrogate models are assumed to be independent random variables. The sample points obtained by inverse transform sampling can effectively represent the frequency of occurrence of the inputs. The distributions of inputs to the surrogate models are fitted to their observed data. These distributions are used for inverse transform sampling. The sample points have larger densities in the regions where the Probability Density Functions (PDF) are higher. This sampling approach ensures that the regions with higher densities of sample points are more prevalent in the observations of the random variables. Inverse transform sampling is applied to the development of surrogate models for window performance evaluation. The distributions of the following three climatic conditions are fitted: (i) the outside temperature, (ii) the wind speed, and (iii) the solar radiation. The sample climatic conditions obtained by the inverse transform sampling are used as training points to evaluate the heat transfer through a generic triple pane window. Using the simulation results at the sample points, surrogate models are developed to represent the heat transfer through the window as a function of the climatic conditions. It is observed that surrogate models developed using the inverse transform sampling can provide higher accuracy than that developed using the Sobol sequence directly for the window performance evaluation.

I. Introduction

Sampling is an important component of optimization, numerical simulations, design of experiments, motion planning, uncertainty analyses, and risk evaluation. It is the process of selecting a sequence of data points from the sample space. These sample points are used as inputs in regions of interests to generate system outputs. For many modern engineering designs, accurate high fidelity simulations or real-life experiments are used to explore the relationships between inputs and outputs. These simulations or experiments can take excessive computation time for highly complex systems. The results from the expensive simulations or experiments can be used to develop surrogate models of systems, to evaluate system performance or to optimize system designs. Sample points are generated in the design space by appropriate sampling

*Doctoral Student, Department of Mechanical, Aerospace and Nuclear Engineering
†Distinguished Professor and Department Chair, Department of Mechanical and Aerospace Engineering. AIAA Lifetime Fellow. Corresponding author (messac@syr.edu)

algorithms. It is expected that an intelligent selection of sample points can increase the accuracy of surrogate models, expedite optimization, and reduce overall computation time.

Many sampling sequences have been developed to address different design space exploration demands. Latin hypercube sampling,[1] random sampling,[2] pseudorandom sampling,[3] low-dispersion sampling,[4] and low-discrepancy sampling[5] are the major categories of sampling sequences that have been reported in the literature. Since low-discrepancy sampling sequences can distribute points uniformly in a sampling space, they are widely used for optimization, surrogate modeling, and numerical integration. Some well-known and popular low-discrepancy sampling sequences include the van der Corput sequence,[6] the Halton/Hammersley sequence,[7] the Sobol sequence,[8] and the Faure sequence.[9]

Some of the popular approaches to generate sample points from a probability distribution include inverse transform sampling,[10] rejection sampling,[11] importance sampling,[12] Markov Chain Monte Carlo,[13] Metropolis-Hastings Sampling,[14] and Gibbs Sampling.[15] In this research, *inverse transform sampling* is used to generate training points for surrogate modeling.

In order to numerically evaluate systems that operate under varying conditions, sampling is often necessary to specify a discrete set of operating conditions. Certain conditions are expected to occur more frequently, and hence comprise regions of high interest in the condition space. The sample points are expected to have large densities in the regions of high interests. Treating the conditions as random variables, and fitting the appropriated probability distributions, the likelihood of occurrence of different conditions can be effectively captured. This paper applies inverse transform sampling to place more sample points in the regions of higher interests based on estimated probability distribution of the operating conditions. This approach provides efficient exploration of system designs under prevalent conditions, and reduces computational expense in the regions of lower interests.

The procedure of inverse transform sampling is illustrated using an example in Sect. II. Section III shows that this sampling approach can be used with larger numbers of sample points. In Sect. IV, inverse transform sampling is applied to sample variables with multimodal distributions. Section V explains how the inverse transform sampling can represent prevalent conditions. Inverse transform sampling is applied to the window performance evaluation under varying climatic conditions in Sect. VI. Concluding remarks are presented in Sect. VII.

II. Procedure of Inverse Transform Sampling

Sampling methods generate a sequence of points in the sample space. The location of a sample point is determined by its coordinates. The Euclidean distance or other distance metrics are generally used to evaluate the distances between points.[16] Inverse transform sampling can sample more points in the regions where random variables have higher probability densities, and sample fewer points in the regions where random variables have low probability densities. In this regard, the estimated probability of random variables is used as the metric of distances. The sample points do not distribute uniformly in terms of the Euclidean distance. However, they are uniform in terms of the probability differences.

The random variables in different dimensions are assumed to be independent with respect to each other. Using recorded data of the variables, a distribution is fitted for each variable in the sample space. Along each dimension, the value of the Cumulative Distribution Function (CDF) goes from 0 to 1. The proposed approach uses a low-discrepancy sampling sequence of real numbers between 0 and 1 as the values of the CDFs. The corresponding coordinates of sample points are computed using the inverse functions of the CDFs. These points crowd in the regions with higher probability densities.

The procedure of the inverse transform sampling for random variables is implemented by the following four steps.

Step 1: Observe and record the pertinent data for the random variables.

Step 2: Fit a distribution function for each random variable using the recorded data.

Step 3: Generate a low-discrepancy sampling sequence of real numbers between 0 and 1 to serve as the values of the probabilities for the random variables.

Step 4: Evaluate the coordinates of the random variable corresponding to their probabilities using the inverse functions of the CDFs.

The four steps are illustrated using a two-dimensional example. In each dimension, the random variable has a Gaussian distribution. The random variables are preconditioned to be independent with respect to each other.

A. Step 1: Random Variable Observations

Sampling variables are generally system inputs. For example, if air temperature is an input to a simulation, it can be treated as a random variable for sampling purpose. The air temperature should be observed and recorded on a regular basis. In order to fit distributions of random variables, the occurrence of random variables should be sufficiently observed. The observed data will be used to fit an appropriate distribution in Step 2.

In this example, for the purpose of illustration, 100 data points with Gaussian distributions in a two dimensional space are recorded.

B. Step 2: Distribution Function Fitting

Distribution fitting is the procedure of selecting a statistical distribution that best fits the observations of random variables obtained in Step 1. For several design and analysis situations, the occurrence of data is known to follow particular distributions. For example, the heights of people or any species of animals can be assumed to follow a Gaussian distribution because the heights are the results of many independent factors.[17] If the value of a random variable is the result of rare events, such as industrial accidents, the Poisson distribution is appropriate.[18] If the types of distributions are known for particular problems, they can be directly used to fit the observed data.

The least squares method,[19] the least absolute deviations method,[20] the generalized method of moments,[21] and the Maximum Likelihood Estimation (MLE) method[22] are commonly used parameter estimation techniques for probability distributions. The MLE method evaluates model parameters by maximizing the probability of the observed data. There are n independent and identically distributed observations, x_1, x_2, ..., x_n. The joint density function for the n observations is expressed as the product of the PDFs specific to each observation in Eqn. 1. The joint density function can be regarded as a function of the model parameters. The MLE method maximizes the likelihood function with respect to the model parameters.

$$L(x_1, x_2, \cdots, x_n|\theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{1}$$

where x_1, x_2, $\cdots$, x_n are sample variables; and θ represents model parameters.

In the example, the distribution of each random variable follows the Gaussian distribution. The Gaussian PDF function is expressed in Eqn. 2. The parameter, μ, is the mean, and the parameter, σ, is the standard deviation.

$$f_G(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{2}$$

C. Step 3: Generating the Sequence of CDFs

In each dimension, the CDF increases monotonically from 0 to 1. Since sample points are distributed in the region where the CDF is between 0 and 1, the probabilities at these points are also between 0 and 1.

Since low-discrepancy sampling methods can simultaneously generate uniformly distributed sequences in all dimensions of a sample space, they are used to produce sequences within the ranges between 0 and 1. The values of the sequences are used as the values of the CDFs.

In the example, a two dimensional Sobol sequence is generated within the range between 0 and 1.

D. Step 4: Coordinates Evaluation

The sequence generated in Step 3 is used as the values of CDFs. Each point in the sequence is multidimensional. In each dimension, the coordinates of the random variable are evaluated using the inverse function of the CDF. If the inverse function of a CDF can be expressed analytically, the values of x corresponding to different probabilities can be evaluated directly. If it is difficult to express the inverse function of the CDF

analytically, the process to find the values of x can be regarded as solving nonlinear equations. Numerical methods can be applied to solve nonlinear equations. The Newton's method,[23] the Levenberg-Marquardt algorithm,[24] the trust region methods,[25] and other applicable methods are popular in solving nonlinear equations.

In the example, the CDF of the standard Gaussian distribution is expressed as

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-t^2/2)dt = \frac{1}{2}[1 + erf(\frac{x}{\sqrt{2}})],\ x \in R \tag{3}$$

The CDF of the Gaussian distribution can be expressed as

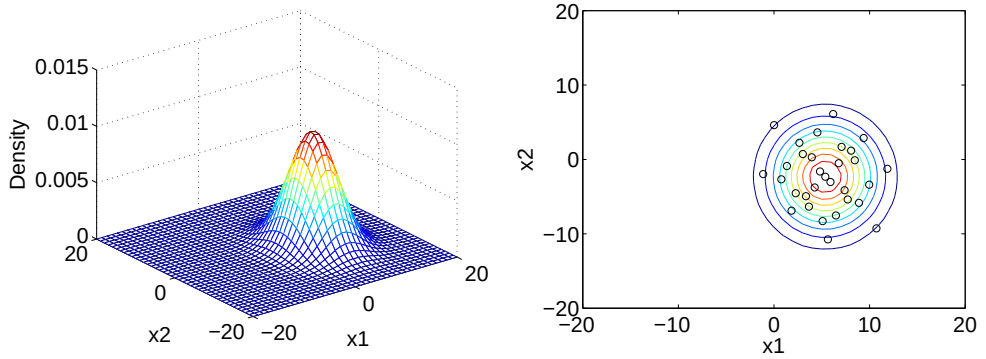
$$F_G(x|\mu,\sigma) = \Phi(\frac{x-\mu}{\sigma}) = \frac{1}{2}[1 + erf(\frac{x-\mu}{\sigma\sqrt{2}})],\ x \in R. \tag{4}$$

where the variable, x, corresponding to the value of the CDF, p, can be expressed as

$$x = F_G^{-1}(p|\mu,\sigma) = \{x : F_G(x|\mu,\sigma) = p\} \tag{5}$$

Equation 5 can be solved using the numerical methods listed above.

The two dimensional joint PDF function is plotted in Fig. 1(a). Figure 1(b) shows the contours of the PDF and the locations of the sample points in the sample space. It can be observed from the figures that the density of points is higher in the region where the PDF is higher. In the region where the PDF is extremely small, there are no sample points.



(a) PDF of unimodal Gaussian distribution (b) Sample points on the contours of the PDF

Figure 1. Locations of sample points

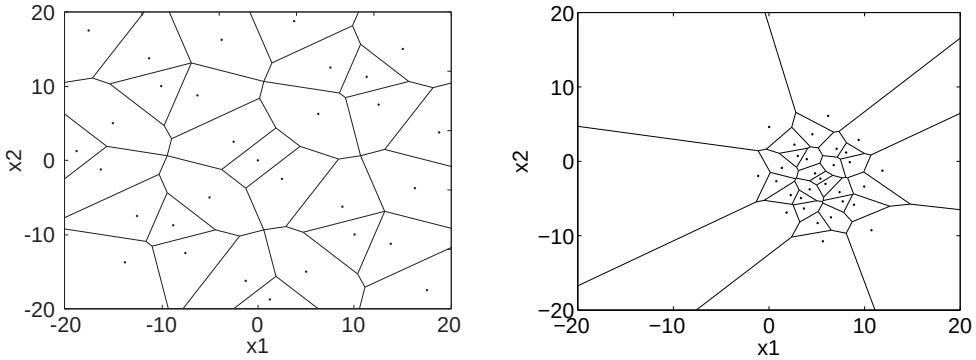
E. Comparisons and Analyses

In the above example, the probability density beyond [-20, 20] in each dimension is extremely low. Using -20 and 20 as the lower and upper bounds of the two dimensions, respectively, the Sobol points generated between 0 and 1 are scaled up to the range [-20, 20] in each dimension.

A Voronoi diagram is a special kind of decomposition of a metric space determined by distances to a specified discrete set of points in the space.[26] The Voronoi diagrams of the scaled Sobol sequence and the inverse transform sample points are generated to analyze their distributions in the sampling space. Figure 2(a) shows the Voronoi diagram of the scaled Sobol sequence, and Fig. 2(b) shows the Voronoi diagram of the points obtained by inverse transform sampling. Each point has a cell that includes the region closer to the point than to any others. The lines are equidistant to the two nearest points. The nodes are the place where their distances to three or more points are equal.

Figure 1(b) shows that the sample points crowd in the region where the PDF is high. As the value of the PDF decreases, the density of the sample points becomes lower. In the region where the probability density is extremely low, there are no points.

The Voronoi diagram of the sample points in Fig. 2(b) shows that the sample points are closer to each other in the region where the density is higher. As the values of the PDF decreases, the distances between the sample points become larger.



(a) Voronoi diagram of the Sobol sequence (b) Voronoi diagram of the sample points

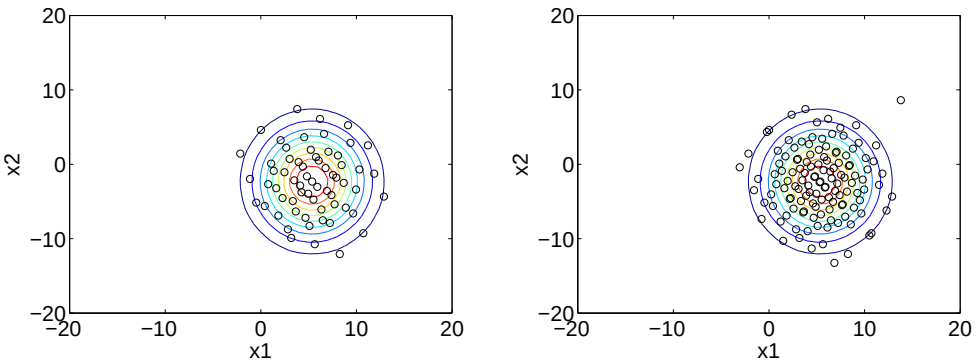
Figure 2. Comparison of the Voronoi diagrams

Inverse transform sampling is also implemented using other low-discrepancy sampling sequences as its values of CDFs. The results show the same effect that there are more sample points in the regions where the PDF is higher.

III. Sampling with Different Numbers of Points

In the example in Sect. II, the number of sample points is 31. Inverse transform sampling is also implemented with 63 and 127 sample points, using the same PDF in the example. Figures 3(a) and 3(b) show the locations of the 63 and 127 sample points in the sampling space, respectively.

If the number of the sample points is increased, the number of sample points in the regions with small PDF can also increase. However, the density of sample points in the regions with high PDF will increase at a higher rate.



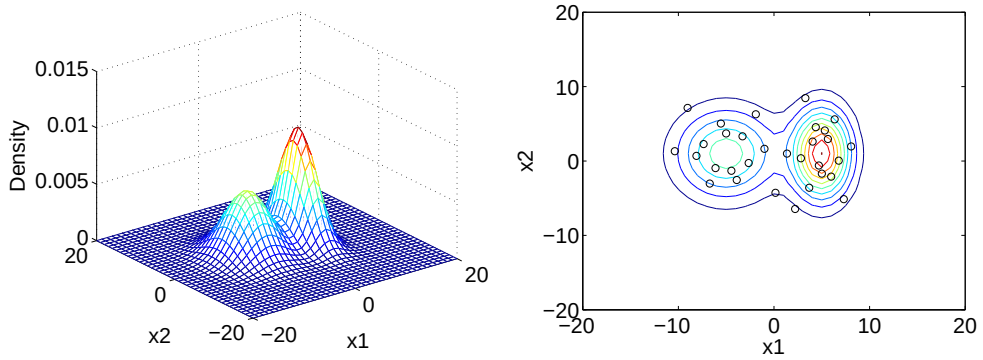
(a) Locations of 63 sample points (b) Locations of 127 sample points

Figure 3. Increasing numbers of sample points

IV. Sampling for Multimodal Distributions

The distribution shown in the example in Sect. II is unimodal. Inverse transform sampling can also be applied to sample multimodal distributions. The sample points of a bimodal distribution and a quad-modal distribution in two dimensional space are generated using their probabilities. The contours of the PDFs and the locations of the sample points are generated for the two cases.

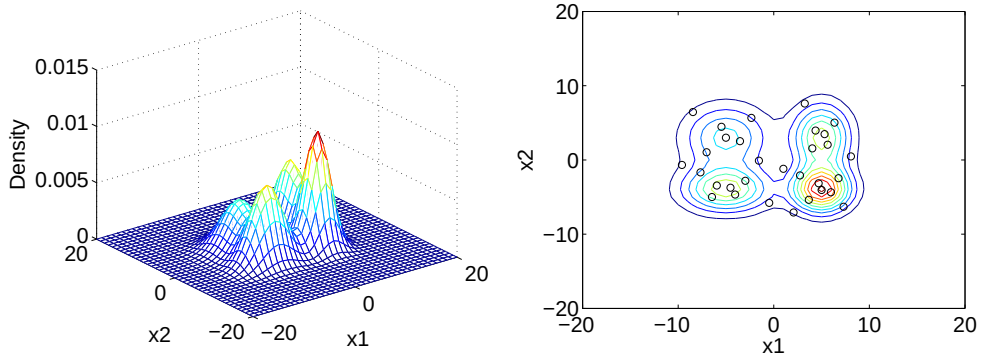
Figure 4(a) shows the PDF of the bimodal distribution, and Fig. 4(b) shows the sample points generated by inverse transform sampling for the bimodal distribution.



(a) PDF of a bimodal distribution (b) Locations of the sample points

Figure 4. Sample points for the bimodal distribution

Figure 5(a) shows the PDF of the quad-modal distribution. The sample points obtained using the inverse transform sampling are shown in Fig. 5(b).



(a) PDF of a quad-modal distribution (b) Locations of the sample points

Figure 5. Sample points for the quad-modal distribution

Observed from Figs. 4(b) and 5(b), the densities of sample points are high in the regions where the PDF is large, which is the same as that of unimodal distributions. Inverse transform sampling can be applied not only to unimodal distributions, but also to multimodal distributions.

V. Explanation for the Change of Point Density

In each dimension, at a sample point, x_i, CDF is the integration of PDF from the negative infinity to the coordinate of this point, which is shown by

$$F(x_i) = \int_{-\infty}^{x_i} f(t)dt \tag{6}$$

The variable of integration, dt, is the infinitesimally small Euclidean distance.

As illustrated in Step 3 in Sect. C, a uniformly-distributed sequence is used to quantify the CDF. In each dimension, the difference between any two consecutive CDF values is the same. The difference between consecutive CDF values between two consecutive points, x_i and x_{i-1} is expressed as

$$F(x_i) - F(x_{i-1}) = \int_{x_{i-1}}^{x_i} f(t)dt \tag{7}$$

The Euclidean distance between x_i and x_{i-1} is expressed as

$$x_i - x_{i-1} = \int_{x_{i-1}}^{x_i} dt \tag{8}$$

Between any two consecutive sampling points, x_i and x_{i-1}, the difference between their CDF values, $F(x_i) - F(x_{i-1})$, is the same. Obtained from Eqn. 7, higher PDF ($f(t)$) requires shorter Euclidean distance between two points, $x_i - x_{i-1}$. Therefore, in the regions where PDF is higher, the densities of sampling points are higher. In these regions, sample points are closer to each other in terms of the Euclidean distance.

VI. Window Performance Evaluation under Varying Climatic Conditions

Inverse transform sampling is applied to generate sample climatic conditions for the performance evaluation of a triple pane window. Three climatic conditions used for simulations are the air temperature, the wind speed, and the solar radiation. The indoor temperature is maintained at a comfortable value. A Computational Fluid Dynamics (CFD) model of the window is created by Fluent.[27] The steady-state heat transfer through the window is simulated by Fluent. The three climatic conditions vary over time.[28] January and August are chosen as the typical months for winter and summer, respectively. The performance of the window is evaluated for these two months.

The three climatic conditions are recorded for a target location. Appropriate types of distributions are fitted to the recorded climate data. Sample sequences are generated using inverse transform sampling. The simulations of the heat transfer through the window are performed using these sample climatic conditions. Using the simulation results, surrogate models are developed to represent the heat transfer through the window in January and August. The procedure is illustrated from Sect. A to Sect. C.

A. Data Collection of Climatic Conditions

The target location for window performance evaluation is Michigan, ND (Latitude: 48.019°, Longitude: -98.172°). The data of climatic conditions used for sampling are obtained from the North Dakota Agricultural Weather Network (NDAWN).[29] Recorded hourly averaged data for the air temperature, the wind speed, and the solar radiation in January and August from 2006 to 2010 are used in this paper. In either January or August from 2006 to 2010, there are 3720 observations of the climatic conditions. The values of correlation coefficients for each pair of climatic conditions are less than 0.003. They indicate the correlation between the three conditions are significantly small.

B. Distribution Fitting of Climatic Conditions

The distribution of the air temperature in a particular month is assumed to be a Gaussian distribution. Since the Weibull distribution generally matches wind speed data, it is used to fit the data of the wind speed.

The distribution of the solar radiation is assumed to be a gamma distribution. At night the solar radiation is zero. The values of many observations of solar radiation are zeros. To fit reasonable gamma distributions, all the zero records are removed. Eighty small numbers, 10^{-3} and 10^{-2}, are added to the January record, and forty small numbers are added to the August record, respectively.

The PDF of the Gaussian distribution is expressed by Eqn. 2.

The PDF of the Weibull distribution can be expressed as

$$f_W(x|\lambda,k) = \begin{cases} \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1} e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

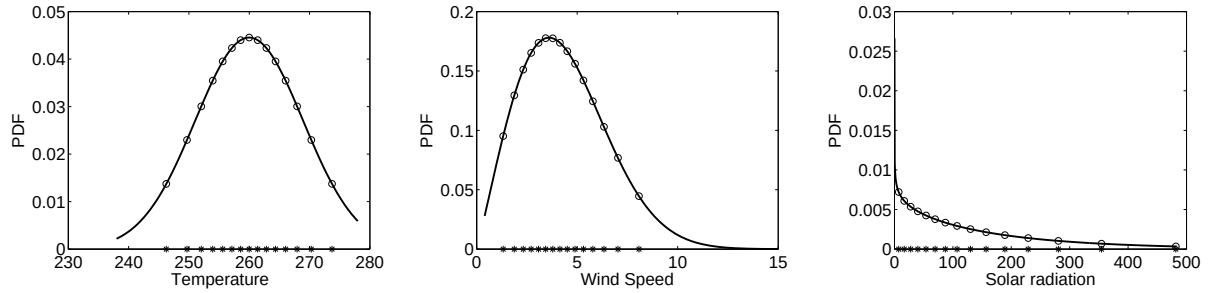
where λ is the scale parameter, and k is the shape parameter.

The equation defining the PDF of the gamma distribution is

$$f_\Gamma(x|k,\theta) = \begin{cases} x^{k-1} \frac{e^{-x/\theta}}{\theta^k \Gamma(k)} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

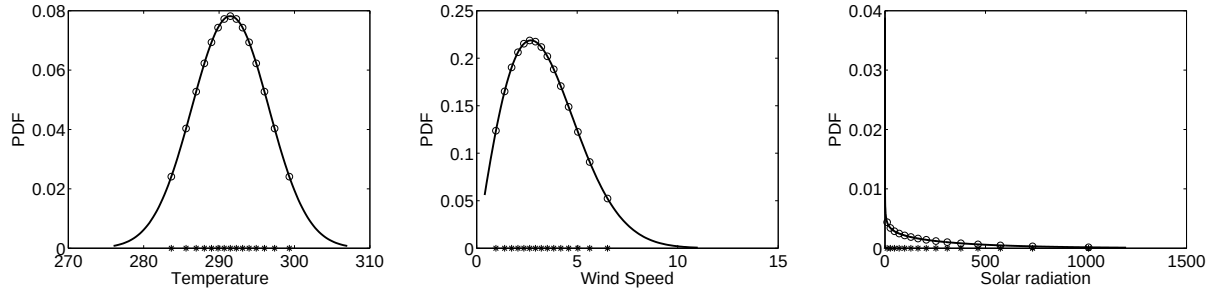
where the shape parameter, k, is a positive integer; the scale parameter, θ, is greater than 0; and $\Gamma(k) = (k-1)!$.

All the three types of distributions are fitted using the Maximum Likelihood Estimation (MLE) method. The curves of the PDFs for the air temperature, the wind speed, and the solar radiation in January and August are shown in Figs. 6 and 7, respectively. The stars and the circles in the figures are the sample points generated in Sect. C.



(a) PDF of air temperature in January (b) PDF of wind speed in January (c) PDF of solar radiation in January

Figure 6. PDFs of three climatic conditions in January



(a) PDF of air temperature in August (b) PDF of wind speed in August (c) PDF of solar radiation in August

Figure 7. PDFs of three climatic conditions in August

C. Climatic Condition Sampling

Using the fitted distributions for the air temperature, the wind speed, and the solar radiation in January and August, the corresponding CDFs are obtained.

The CDF of Gaussian distribution is expressed by Eqns. 3 and 4.

The CDF of the Weibull distribution is given by

$$F_W(x|\lambda,k) = \begin{cases} 1 - e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

The CDF of the gamma distribution is given by

$$F_\Gamma(x|\lambda,k) = \begin{cases} \frac{\gamma(k,x/\theta)}{\Gamma(k)} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

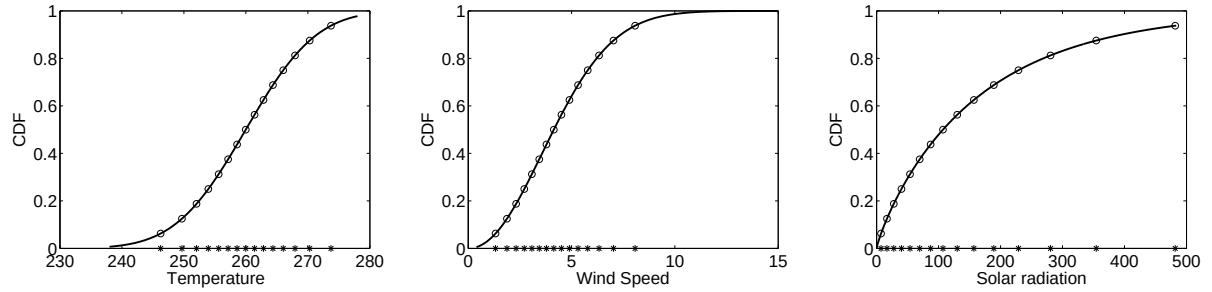
where $\gamma(k, x/\theta)$ is given by

$$\gamma(k, x/\theta) = \int_0^{x/\theta} t^{k-1} e^{-t} dt. \tag{9}$$

The CDFs of the air temperature, the wind speed, and the solar radiation in January and August are plotted in Figs. 8 and 9, respectively.

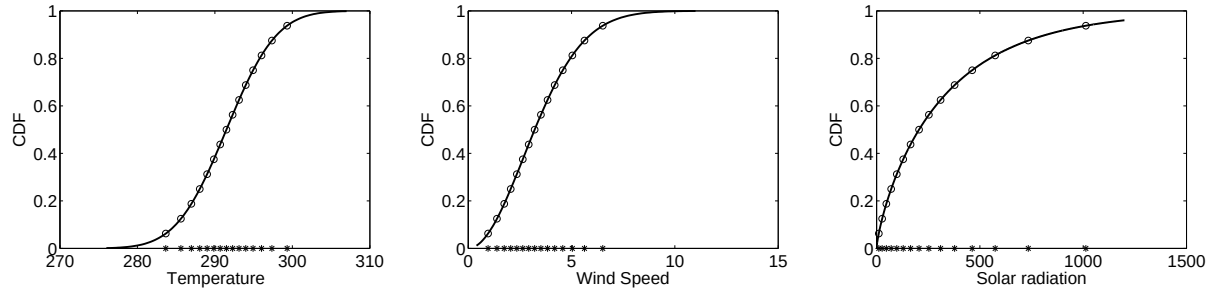
The number of sample climatic conditions is expected to be close to ten times the number of variables.[30] In this research, 31 sample points are generated for each month.

Using the Sobol sequence to quantify the CDFs, the corresponding sample values of the air temperature, the wind speed, and the solar radiation are evaluated. Figures 8 and 9 show the coordinates of the sample points along with the corresponding values of the CDFs. In each figure, the stars on the climatic condition axis are the coordinates of the three climatic conditions, and the circles on the curve show the corresponding values of the CDFs. Each figure only shows the first 15 sample points in each sequence.



(a) CDF of air temperature in January (b) CDF of wind speed in January (c) CDF of solar radiation in January

Figure 8. CDFs of three climatic conditions in January



(a) CDF of air temperature in August (b) CDF of wind speed in August (c) CDF of solar radiation in August

Figure 9. CDFs of three climatic conditions in August

The first 15 sample points are also plotted on the PDF figures, as shown in Figs. 6 and 7. Expectedly, the densities of the sample points are higher in the regions where the values of the PDFs are higher. The higher densities indicate that the corresponding climatic conditions occur more frequently.

Using each of the 31 climatic conditions as boundary conditions, the heat transfer rate through the triple pan window is evaluated. Subsequently, two surrogate models are developed using Kriging[31] to represent the heat transfer rates as functions of the three climatic conditions in January and August, respectively. The development of the Kriging models is performed using DACE Matlab Kriging toolbox.[32]

D. Surrogate Model Performance Criteria

The errors of the surrogate models can be evaluated using standard accuracy measures: (i) the Root Mean Squared Error (RMSE),[31,33] which provides a global error measure over the entire design domain, and (ii) the Maximum Absolute Error (MAE),[34,35] which is indicative of local deviations. It is desirable that both of these error measures are minimized. These metric measures are defined as

$$RMSE = \sqrt{\frac{1}{n_t}\sum_{k=1}^{n_t}\left(h(x^k) - \tilde{h}(x^k)\right)^2} \tag{10}$$

$$MAE = \max_k \left|h(x^k) - \tilde{h}(x^k)\right| \tag{11}$$

where $h(x^k)$ represents the actual function value for the test point x^k; $\tilde{h}(x^k)$ is the corresponding function value estimated by the surrogate model; and n_t is the number of test points chosen for estimating the error measure.

The percentage of errors with respect to the actual values can be measured by the following two metrics: (i) the Root Mean Squared Percentage Error (RMSPE), which provides a measure of the percentage error over the entire design domain, and (ii) the Maximum Percentage Error (MPE), which indicates the maximum percentage error among all the testing points. They are expected to be minimized. These two percentage error measures can be defined as

$$RMSPE = \sqrt{\frac{1}{n_t}\sum_{k=1}^{n_t}\left(\frac{h(x^k)-\tilde{h}(x^k)}{h(x^k)}\right)^2} \tag{12}$$

$$MPE = \max_k \left|\frac{h(x^k)-\tilde{h}(x^k)}{h(x^k)}\right| \tag{13}$$

E. Surrogate Model Performance Comparison

Two other surrogate models are also developed using sample points directly generated by the Sobol sequence. Using the minimum and the maximum values of the climatic parameters as lower and higher bounds, respectively, 31 sets of the climatic conditions are generated directly using the Sobol sequence. These 31 sample points are uniformly distributed between the lower and higher bounds. The heat transfer rate through the window is evaluated for this set of sample climatic conditions. Using these evaluation results, two Kriging models are developed to represent the heat transfer in January and August, respectively.

The performance of the two surrogate models using inverse transform sampling is compared with that using the Sobol sequence.

The 3720 observations of climatic data for January from 2006 to 2010 are used to evaluate the window performance. The evaluation results are used as the actual function values to calculate RMSE, MAE, RMSPE, and MPE for the surrogate models. Table 1 shows the values of RMSE, MAE, RMSPE, and MPE for the surrogate models using different sampling methods.

Table 1. Performance comparison for surrogate models

Month	Sampling Method	RMSE	MAE	RMSPE	MPE
January	Inverse Transform	0.047	0.49	0.64%	7.2%
	Sobol Sequence	0.054	0.30	0.68%	9.3%
August	Inverse Transform	0.079	0.54	11%	318%
	Sobol Sequence	0.094	0.32	85%	4373%

For January, the three error measures, RMSE, RMSPE, and MPE for the surrogate model developed by inverse transform sampling have lower values than those developed using the conventional Sobol sequence. The lower values of RMSE and MPE indicate a better overall performance. The higher value of MAE for inverse transform sampling indicates that it has a larger value of the maximum absolute error. The smaller MPE for inverse transform sampling shows that the maximum percentage error for this sampling approach is lower than that for the Sobol sequence. For August, the comparison result is similar with that of January.

It is observed that the RMSPE and MPE for August are much higher than those for January. For the January climatic conditions, the simulation results of the heat transfer through the triple pane window are between $1.86W$ and $18.73W$, and the average is $10.74W$; for the August climatic conditions, the simulation results are between $-13.93W$ and $5.88W$, and the average is $-1.33W$. Many values of actual heat transfer in August are close to zero. As Eqns. 12 and 13 show, each simulation result, $h(x^k)$, is the denominator used to calculate the percentage error for $\tilde{h}(x^k)$. Although the absolute values of errors for August are of the same order of magnitude as those for January, the percentage error values are higher. The RMSE and RMSPE, which represent the absolute errors, are of the same order of magnitude for January and August. However, for August, the orders of magnitude of RMSPE and MPE are much higher than those for January.

F. Surrogate Model Performance in Increasing Sample Space

All the recorded hourly climatic conditions can be classified into regions with different PDF values in the sample space respectively for January and August. The performance of the surrogate models in increasing sample space is recorded in Tables 2 and 3.

In Tables 2 and 3, the percentages in the first column are the joint probabilities of the random variables. Since there are three climatic condition variables, the probability of each random variable is derived as the cube root of the percentage. For each variable, the probability is the integral of the fitted PDF in an interval $[a, b]$, as shown in Eqn. 14.

$$F(b) - F(a) = \int_a^b f(t)dt \tag{14}$$

The length of the interval, $b - a$, is minimized in order to find the interval with the highest values of $f(x)$. For example, if the percentage for January is 2.7%, the probability of each climatic condition variable is 30%. In Fig. 10, the probability of temperature, $F(b|\mu, \sigma) - F(a|\mu, \sigma)$, is 30%. The length of the interval, $b - a$, is minimized in order to find the interval with the highest values of $f(x|\mu, \sigma)$. After the intervals of the three climatic condition variable are evaluated, all the recorded hourly climatic conditions inside the intervals and their corresponding heat transfer results are used to evaluate the four performance criteria for the surrogate models.

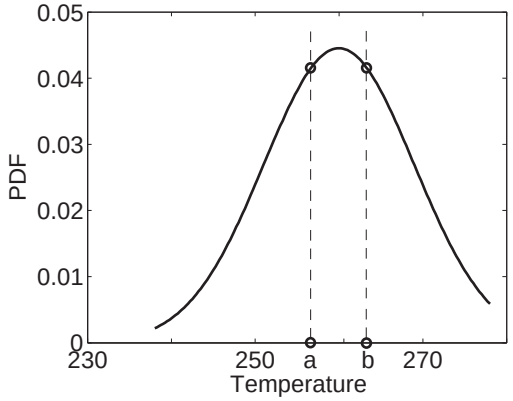


Figure 10. An interval with the highest PDF values

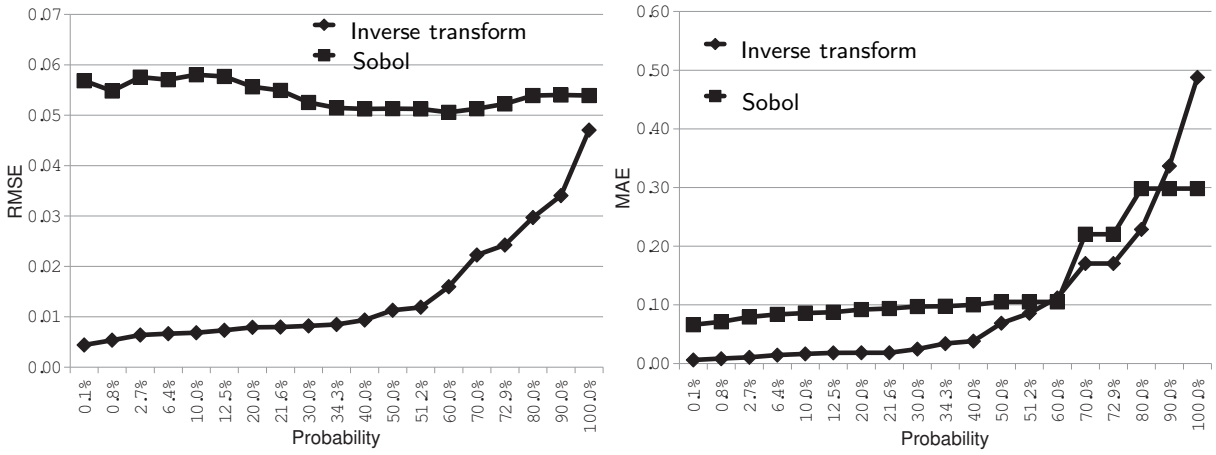
The first column of Tables 2 and 3 shows the probabilities. From the first row to the last row, the probability is increasing. The four surrogate model performance criteria, RMSE, MAE, RMSPE, and MPE, vary with the probability values. The trends of the four criteria are plotted in Figs. 11 and 12 for January and August, respectively.

It is observed from Figs. 11 and 12 that the the surrogate models developed using inverse transform sampling have better overall performance (lower values of RMSE and RMSPE) in all regions listed in Tables 2 and 3 for both January and August. As the percentage in the first column of each table increases, the sample space expands to the regions with low PDF, and the values of RMSE and RMSPE for inverse transform sampling become higher. The RMSE and RMSPE values increase slowly at the beginning. As the percentage approaches 100%, they increase at a higher rate.

In regions with small percentages, the values of MAE for inverse transform sampling are lower than those for the Sobol sequence. As the percentage increases, the MAE for inverse transform sampling increases faster than that for the Sobol sequence. It becomes higher than that for the Sobol sequence as the percentage approaches 100%.

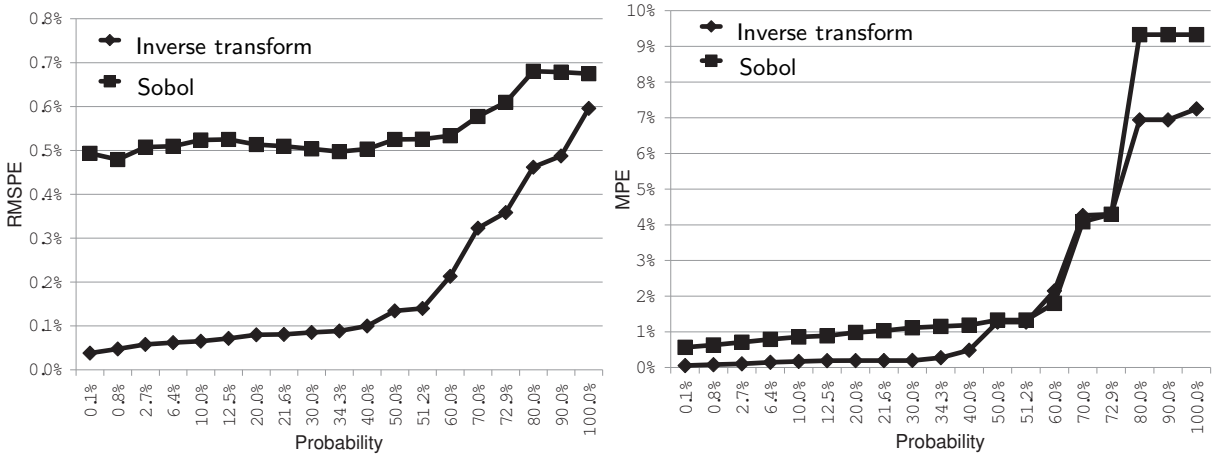
The values of MPE for inverse transform sampling are generally lower than those for the Sobol sequence. The MPE for the Sobol sequence has a sharp increase at 12.5% for August. The climatic condition leads to an actual heat transfer value of 0.0034. This value is very close to zero, and it is used as the denominator to evaluate its percentage error. For the Sobol sequence, its percentage error is 4373%. It causes a sharp increase of the MPE as well as the increase of RMSPE for the Sobol sequence.

As mentioned in Sect. B, during the distribution fitting to the solar radiation, all zeros are ignored, and a number of small values are used instead. The fitted PDFs do not reflect the actual distributions of solar



(a) RMSE for January

(b) MAE for January



(c) RMSPE for January

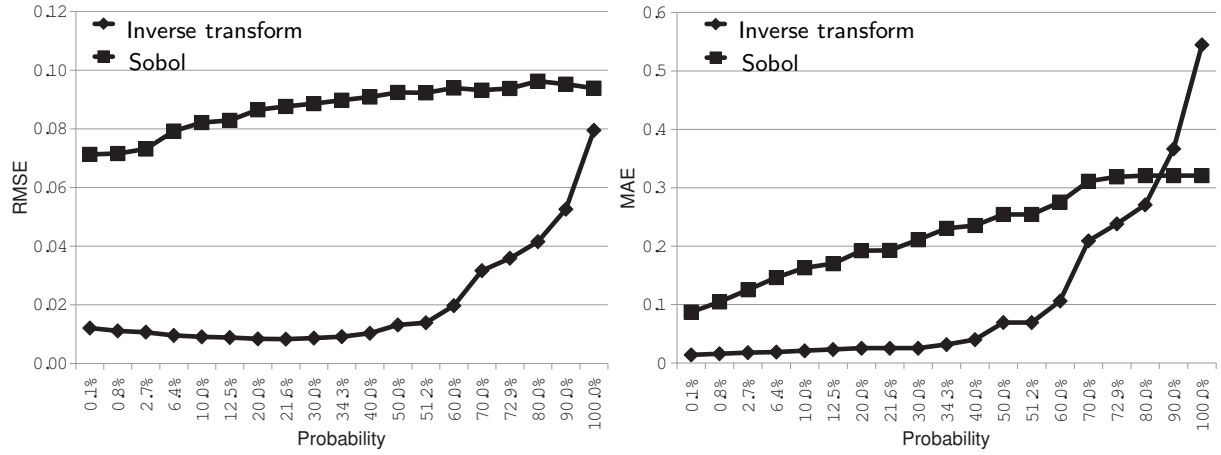
(d) MPE for January

Figure 11. Performance evaluation of surrogate models for January

radiation in January and August. The second column of Tables 2 and 3 shows the numbers of climatic conditions in different percentages of sample space. These numbers are not equal to the products of the corresponding percentages in the first column and the total number of climatic conditions.

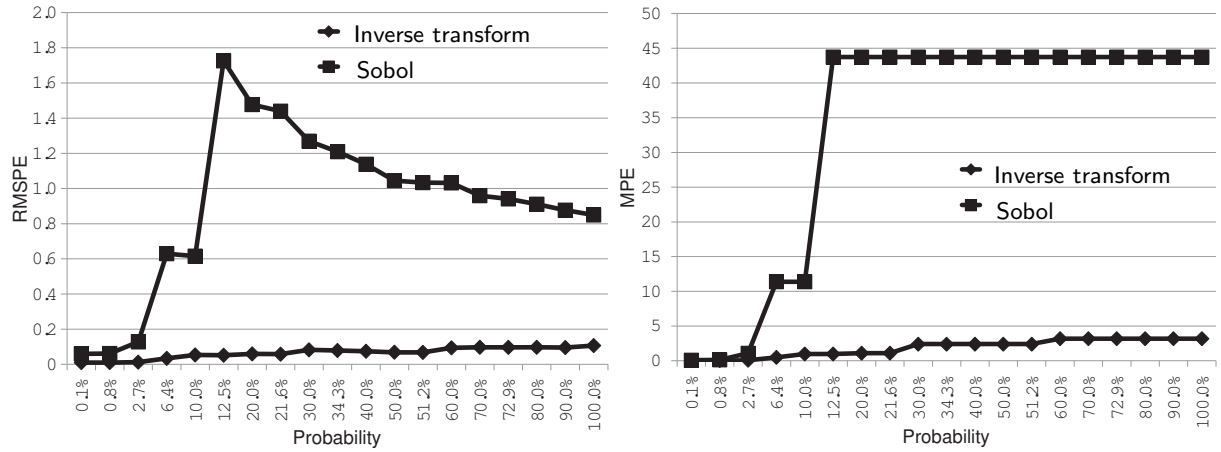
VII. Concluding Remarks

Inverse transform sampling is used in this study to improve the accuracy of surrogate models. Inverse transform sampling can effectively represent the frequency of occurrence of independent random variables. Numerical experiments show that this approach samples more points in the regions with higher probability densities. Inverse transform sampling approach is applied to the development of surrogate models for window performance evaluation. The surrogate models developed using inverse transform sampling have higher accuracy than those developed using the Sobol sequence. The performance of the surrogate models is explored in regions with increasing probabilities. The trends of the performance criteria with the changes of probabilities are studied. It is observed that, in the regions where the probability densities of random variables are higher, inverse transform sampling provides better surrogate model performance than that obtained using conventional Sobol sequence.



(a) RMSE for August

(b) MAE for August



(c) RMSPE for August

(d) MPE for August

Figure 12. Performance evaluation of surrogate models for August

VIII. Acknowledgements

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

References

[1]Husslage, B. G., Rennen, G., van Dam, E. R., and den Hertog, D., "Space-filling Latin Hypercube Designs for Computer Experiments," *Optimization and Engineering*, Vol. 12, 2011, pp. 611–632.

[2]Clarkson, K. L. and Shor, P. W., "Applications of Random Sampling in Computational Geometry, II," *Discrete and Computational Geometry*, Vol. 4, 1989, pp. 387–421.

[3]Goldreich, O., *Computational Complexity: A Conceptual Perspective*, Cambridge University Press, 1st ed., 2008.

[4]LaValle, S. M., *Planning Algorithms*, Cambridge University Press, 2006.

[5]Niederreiter, H., "Point Sets and Sequences with Small Discrepancy," *Monatshefte fr Mathematik*, Vol. 104, December 1987, pp. 273–337.

[6]van der Corput, J. G., "Verteilungsfunktionen," *Nederl. Akad. Wetensch. Proc.*, Vol. 38, 1935, pp. 813–821.

[7]Diaconis, P., "The Distribution of Leading Digits and Uniform Distribution Mod 1," *The Annals of Probability*, Vol. 5, No. 1, Feb 1977, pp. 72–81.

[8]Sobol, I. M., "Uniformly Distributed Sequences with an Additional Uniform Property," *USSR Computational Mathematics and Mathematical Physics*, Vol. 16, 1976, pp. 236–242.

[9]Faure, H., "Discrpances de suites associes un systme de numration en dimension s," *Acta Arithmetica*, Vol. 41, 1982, pp. 337–351.

[10]Miller, F., Vandome, A., and John, M., *Inverse Transform Sampling*, VDM Verlag Dr. Mueller e.K., 2010.

[11]von Neumann, J., "Various Techniques Used in Connection with Random Digits," *Nat. Bureau Stand. Appl. Math. Ser.*, Vol. 12, 1951, pp. 3638.

[12]Marshall, A. W., "The Use of Multi-stage Sampling Schemes in Monte Carlo Computations," *H. A. Meyer (ed.), Symposium on Monte Carlo Methods*, edited by N. Y. John Wiley & Sons, Inc., 1956, p. 123140.

[13]Gilks, W., Gilks, W., Richardson, S., and Spiegelhalter, D., *Markov Chain Monte Carlo in Practice*, Interdisciplinary Statistics, Chapman & Hall, 1996.

[14]Chib, S. and Greenberg, E., "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, Vol. 49, No. 4, November 1995, pp. 327–335.

[15]Kim, C.-J. and Nelson, C. R., *State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications*, Vol. 1 of *MIT Press Books*, The MIT Press, June 1999.

[16]Bryant, V., *Metric Spaces: Iteration and Application*, Cambridge University Press, 1985.

[17]Marusteri, M. and Bacarea, V., "Comparing Groups for Statistical Differences: How to Choose the Right Statistical Test?" *Biochemia Medica*, Vol. 20, No. 1, 2010, pp. 15–32.

[18]Hill, T. and Lewicki, P., "STATISTICS Methods and Applications," 2007.

[19]Rao, C., Toutenburg, H., Fieger, A., Heumann, C., Nittner, T., and Scheid, S., *Linear Models: Least Squares and Alternatives (Springer Series in Statistics)*, Springer, 1999.

[20]Siemsen, E. and Bollen, K. A., "Least Absolute Deviation Estimation in Structural Equation Modeling," *Sociological Methods and Research*, Vol. 36, No. 2, 2007, pp. 227–265.

[21]Hall, A. R., *Generalized Method of Moments (Advanced Texts in Econometrics)*, Oxford University Press, 2005.

[22]Hald, A., "On the History of Maximum Likelihood in Relation to Inverse Probability and Least Squares," *Statistical Science*, Vol. 14, No. 2, 1999, pp. 214–222.

[23]Varona, J. L., "Graphic and Numerical Comparison Between Iterative Methods," *Math. Intell.*, Vol. 24, 2002, pp. 37–46.

[24]Bates, D. M. and Watts, D. G., *Nonlinear Regression Analysis and its Applications*, Wiley, 1988.

[25]Byrd, R. H., Schnabel, R. B., and Shultz, G. A., "A Trust Region Algorithm for Nonlinearly Constrained Optimization," *SIAM J. Numer. Anal.*, Vol. 24, 1987, pp. 1152–1170.

[26]Aurenhammer, F., "Voronoi Diagrams - A Survey of a Fundamental Geometric Data Structure," *ACM Computing Surveys*, Vol. 23, No. 3, 1991, pp. 345–405.

[27]FLUENT, "ANSYS FLUENT 12.0 Getting Started Guide," April 2009.

[28]Zhang, J., Messac, A., Chowdhury, S., and Zhang, J., "Adaptive Optimal Design of Active Thermally Insulated Windows Using Surrogate Modeling," 51^{st} *AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, No. AIAA 2010-2917, Orlando, Florida, April 2010.

[29]NDAWN, "The North Dakota Agricultural Weather Network," 2010.

[30]Colaco, M. J., Dulikravich, G. S., and Sahoo, D., "A Response Surface Method-Based Hybrid Optimizer," *Inverse Probl Sci Eng*, Vol. 6, No. 16, 2008, pp. 717–741.

[31]Forrester, A. and Keane, A., "Recent Advances in Surrogate-based Optimization," *Progress in Aerospace Sciences*, Vol. 45, No. 1-3, 2009, pp. 50–79.

[32]Lophaven, S., Nielsen, H., and Sondergaard, J., "DACE - A Matlab Kriging Toolbox, Version 2.0," 2002.

[33]Jin, R., Chen, W., and Simpson, T., "Comparative Studies of Metamodelling Techniques under Multiple Modelling Criteria," *Structural and Multidisciplinary Optimization*, Vol. 23, No. 1, 2001, pp. 1–13.

[34]Queipo, N., Haftka, R., Shyy, W., Goel, T., Vaidyanathan, R., and Tucker, P., "Surrogate-Based Analysis and Optimization," *Progress in Aerospace Sciences*, Vol. 41, No. 1, 2005, pp. 1–28.

[35]Mullur, A. A. and Messac, A., "Metamodeling Using Extended Radial Basis Functions: a Comparative Approach," *Engineering with Computers*, Vol. 21, No. 3, 2006, pp. 203–217.

Table 2. Performance comparison for surrogate models in increasing space for January

Fraction	No.	Mean	Method	RMSE	MAE	RMSPE	MPE
0.1%	13	11.5	Inv	0.0044	0.0062	0.04%	0.05%
			Sobol	0.0569	0.0662	0.49%	0.56%
0.8%	61	11.4	Inv	0.0054	0.0085	0.05%	0.08%
			Sobol	0.0548	0.0713	0.48%	0.63%
2.7%	161	11.3	Inv	0.0064	0.0107	0.06%	0.10%
			Sobol	0.0576	0.0793	0.51%	0.71%
6.4%	298	11.2	Inv	0.0066	0.0144	0.06%	0.14%
			Sobol	0.0570	0.0838	0.51%	0.79%
10.0%	430	11.2	Inv	0.0068	0.0166	0.07%	0.17%
			Sobol	0.0580	0.0859	0.52%	0.86%
12.5%	529	11.1	Inv	0.0073	0.0182	0.07%	0.19%
			Sobol	0.0577	0.0875	0.53%	0.89%
20.0%	839	10.9	Inv	0.0079	0.0185	0.08%	0.19%
			Sobol	0.0556	0.0921	0.51%	0.98%
21.6%	915	10.8	Inv	0.0080	0.0185	0.08%	0.19%
			Sobol	0.0549	0.0937	0.51%	1.03%
30.0%	1339	10.5	Inv	0.0082	0.0247	0.09%	0.20%
			Sobol	0.0525	0.0971	0.50%	1.11%
34.3%	1537	10.4	Inv	0.0085	0.0341	0.09%	0.27%
			Sobol	0.0515	0.0977	0.50%	1.15%
40.0%	1773	10.3	Inv	0.0094	0.0383	0.10%	0.48%
			Sobol	0.0513	0.1001	0.50%	1.19%
50.0%	2171	10.1	Inv	0.0113	0.0689	0.13%	1.27%
			Sobol	0.0513	0.1052	0.53%	1.33%
51.2%	2225	10.1	Inv	0.0119	0.0858	0.14%	1.27%
			Sobol	0.0513	0.1052	0.53%	1.33%
60.0%	2582	10.0	Inv	0.0160	0.1121	0.21%	2.15%
			Sobol	0.0506	0.1052	0.53%	1.80%
70.0%	3018	10.1	Inv	0.0223	0.1706	0.32%	4.26%
			Sobol	0.0513	0.2204	0.58%	4.08%
72.9%	3124	10.0	Inv	0.0243	0.1706	0.36%	4.29%
			Sobol	0.0523	0.2204	0.61%	4.30%
80.0%	3382	10.1	Inv	0.0297	0.2288	0.46%	6.94%
			Sobol	0.0539	0.2981	0.68%	9.33%
90.0%	3565	10.2	Inv	0.0341	0.3366	0.49%	6.94%
			Sobol	0.0540	0.2981	0.68%	9.33%
100.0%	3720	10.3	Inv	0.0470	0.4878	0.64%	7.25%
			Sobol	0.0539	0.2981	0.68%	9.33%

Table 3. Performance comparison for surrogate models in increasing space for August

Fraction	No.	Mean	Method	RMSE	MAE	RMSPE	MPE
0.1%	25	1.23	Inv	0.0121	0.0140	1.0%	1.2%
			Sobol	0.0713	0.0872	5.9%	7.9%
0.8%	94	1.25	Inv	0.0111	0.0158	1.0%	1.9%
			Sobol	0.0716	0.1046	6.2%	12.9%
2.7%	245	1.19	Inv	0.0107	0.0179	1.2%	6.3%
			Sobol	0.0732	0.1255	13%	107%
6.4%	462	1.13	Inv	0.0096	0.0187	3.4%	49%
			Sobol	0.0792	0.1462	63%	1135%
10.0%	615	1.11	Inv	0.0091	0.0210	5.3%	98%
			Sobol	0.0821	0.1630	61%	1135%
12.5%	722	1.11	Inv	0.0089	0.0232	5.2%	98%
			Sobol	0.0829	0.1700	172%	4373%
20.0%	1016	1.03	Inv	0.0084	0.0254	5.9%	108%
			Sobol	0.0865	0.1924	147%	4373%
21.6%	1072	1.00	Inv	0.0084	0.0254	5.7%	108%
			Sobol	0.0876	0.1928	143%	4373%
30.0%	1386	0.84	Inv	0.0087	0.0254	8.2%	240%
			Sobol	0.0886	0.2109	126%	4373%
34.3%	1528	0.80	Inv	0.0092	0.0317	7.8%	240%
			Sobol	0.0898	0.2304	120%	4373%
40.0%	1727	0.68	Inv	0.0103	0.0400	7.4%	240%
			Sobol	0.0909	0.2354	113%	4373%
50.0%	2056	0.47	Inv	0.0132	0.0692	6.8%	240%
			Sobol	0.0924	0.2543	104%	4373%
51.2%	2106	0.44	Inv	0.0139	0.0692	6.8%	240%
			Sobol	0.0923	0.2543	103%	4373%
60.0%	2456	0.06	Inv	0.0197	0.1060	9.4%	318%
			Sobol	0.0939	0.2755	103%	4373%
70.0%	2871	-0.62	Inv	0.0317	0.2090	9.7%	318%
			Sobol	0.0932	0.3111	95%	4373%
72.9%	2990	-0.81	Inv	0.0359	0.2381	9.6%	318%
			Sobol	0.0938	0.3187	94%	4373%
80.0%	3238	-0.97	Inv	0.0415	0.2709	9.7%	318%
			Sobol	0.0962	0.3208	91%	4373%
90.0%	3495	-1.10	Inv	0.0526	0.3665	9.5%	318%
			Sobol	0.0952	0.3208	87%	4373%
100.0%	3720	-1.33	Inv	0.0795	0.5447	10.6%	318%
			Sobol	0.0939	0.3208	84%	4373%