10th World Congress on Structural and Multidisciplinary Optimization
May 19 -24, 2013, Orlando, Florida, USA

New Model Validation Metrics for Correlated Multiple Responses

Wei Li[1], Zhenzhou Lu[2], Wei Chen[3], Zhen Jiang[4], Yu Liu[5]

[1]Northwestern Polytechnical University, Xi'an, PR China, Predoctoral visiting fellow, Northwestern University, Evanston, IL, USA, wei.li@northwestern.edu
[2]Northwestern Polytechnical University, Xi'an, PR China, zhenzhoulu@nwpu.edu.cn
[3]Northwestern University, Evanston, IL, USA, corresponding author: weichen@northwestern.edu
[4]Northwestern University, Evanston, IL, USA, ZhenJiang2015@u.northwestern.edu
[5]University of Electronic Science and Technology of China, Chengdu, PR China, YuLiu@uestc.edu.cn

1. Abstract
Validating models with correlated multivariate outputs involves the comparison of multiple quantities. Considering both uncertainty and correlations among multiple responses from model and physical observations impose challenges. Existing marginal comparison methods and the hypothesis testing based methods either ignore the correlations among responses or are only suitable for reaching Boolean conclusions (yes or no) without accounting for the amount of discrepancy between model and the underlying reality. A new validation metric is needed to quantitatively characterize the overall agreement of multiple responses considering the correlations among responses and the uncertainty in both model predictions and physical observations. In this paper, by extending the concept of "area metric" and the "u-pooling method" developed for validating a single response, we propose two new model validation metrics for validating correlated multiple responses using multivariate probability integral transformation (PIT). One new metric is the PIT area metric for model validation at a single validation site, which measures the distance between the PIT distribution of the joint cumulative distribution function (CDF) of model predictions and the empirical CDF of transformed observations; the other is the t-pooling metric, similar to the idea of u-pooling for a single response, that allows for pooling observations of multiple responses observed at different validation sites by comparing the empirical CDF of the twice transformed observations with the standard uniform distribution. The proposed metrics provide objective measures of the accuracy of multi-response prediction either at a specified site or at multiple sites for assessing the global predictive capability. The proposed metrics have many favorable properties that are well suited for the validation assessment of models with correlated responses. The two metrics are examined and compared with the direct area metric and the marginal u-pooling method respectively through numerical case studies to illustrate their validity and potential benefits.
2. Keywords: Validation, Uncertainty, Correlation, Area Metric, Multiple Responses, Multivariate PIT

3. Introduction
Due to the expensive cost for conducting full-scale physical experiments, the prediction of performance of complex engineering systems has increasingly relied on the use of computational models. Validation of these models is becoming a major issue as a model needs to be either accepted or rejected; sometimes a choice has to be made among alternative models. Model validation is defined as the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended use of the model [1, 2]. In our earlier work [3], four categories of existing model validation methods and the associated metrics are classified, examined, and compared along with the desired features of validation metrics [4, 5]. A better understanding of the pros and cons of each type of metrics has been provided. Among these, the area metric based methods proposed by Ferson et al. [5] are shown to be promising due to the many favorable features they have [6]. By direct measuring the area difference between the cumulative distribution function (CDF) of the simulation and the empirical CDF of the experimental data, the metric provides an objective comparison of the whole distributions between model predictions and experimental measurements [7]. Most valuably, by applying the so-called 'u-pooling' technique, physical observations collected at different validation sites can be incorporated into a single metric to assess the global predictive capability of a model for its intended use [3, 8].

Despite their advantages, the existing area metric (for single site) and u-pooling metric (for multiple sites) only make comparisons between marginal distributions, which are better suited for validating a single response model or multiple uncorrelated responses rather than correlated multi-responses. There are mainly two situations resulting in multivariate responses: (1) A physical experiment and the computational model generate multiple responses or measurements [9]. These responses usually have distinct magnitudes and scales, e.g., acceleration versus displacement; (2) The responses of interest measured from the same experiment is a function of spatial [10] and temporal [4] variables. In both cases, there is a strong correlation between any pair of response quantities from the

same experiment. Though several methods for validation of multivariate models have been proposed based on the classical or Bayesian hypothesis testing [11, 12], these methods are under normality assumption. The Boolean results from the hypothesis test based methods do not quantitatively measure the discrepancy between predictions and observations and is not applicable for the case where the asymptotic limit of uncertainty goes to zero.

Considering uncertainty in prediction and physical observations is important in model validation. Several different sources of uncertainty can be identified in engineering computer models and experiments based on the work of Kennedy and O'Hagan [13]. For instance, the lack of knowledge uncertainty resulting from model parameter uncertainty and model inadequacy; the algorithmic uncertainty introduced from numerical implementations of the computer model such as numerical integration; experimental uncertainty in the form of measurement error, systematic error, and random errors; and interpolation uncertainty due to lack of samples. To account for both uncertainty and multi-response correlation, one intuitive extension of the area metric would be comparing the joint cumulative distribution function (CDF) of the model responses and the multivariate empirical cumulative distribution function (ECDF) of the observed data [3]. This method, which we term as the "direct area metric", though plausible, is not suitable for pooling experimental data of multiple quantities measured at different spatial or temporal points. Besides, the metric would suffer severely from the "curse of dimensionality" for computing the high dimensional integration the multivariate ECDF. In lack of proper methods for incorporating multiple response quantities observed at multiple validation sites, a practical treatment is to conduct u-pooling separately for each response based on the marginal distributions, and then take the average or a weighted sum of the u-pooling metrics for multiple responses without considering the correlation among them.

In this work, two new metrics are proposed for the validation assessment of models with correlated multiple responses by extending the idea of "area metric" and the "u-pooling method" [5] through introducing the multivariate probability integral transformation (PIT) theorem [14]. The first one is the PIT area metric for observations collected at single validation site. The joint CDF of the model responses and the multivariate experimental observations are transformed into a univariate CDF and random data sequence following the multivariate PIT theorem. The PIT area metric provides a comparison between the empirical CDF of the data sequence and the PIT distribution of the joint CDF of model responses. The second metric is the t-pooling metric for observations from multiple validation sites, in which case the PIT distributions of the joint CDFs of model responses at multiple sites are transformed into a same standard uniform distribution - the observations are simultaneously transformed twice based on the relevant joint CDFs and PIT distributions into a univariate data sequence and compared with the uniform distribution. With the uncertainty and correlation information captured respectively by the transformed data sequences and the multivariate PITs, the issues of both uncertainty and correlation can be addressed. Also owing to the univariate nature of the multivariate PIT, the proposed metrics are all univariate integrations regardless of the number of response quantities, which significantly cuts down the computational costs compared to the direct area metric.

In the remainder of this paper, a brief introduction of the probability integral transformation (PIT) theorem and the area metric/u-pooling technique is provided in Section 4. The proposed PIT area metric and t-pooling metric are presented in Section 5. In Section 6, the proposed metrics are tested and compared with the direct area metric and the marginal u-pooling method through illustrative numerical examples to show their advantages in assessing both the correlation and uncertainty of predictive models. The closure of the paper is provided in Section 7.

4. Background information

4.1. Probability integral transformation

The probability integral transformation (PIT) for a single random variable is well established in the literature. Given any random variable Y with a continuous cumulative distribution function $F_Y(y)$, the PIT of Y is a standard uniform random variable V that transformed via the relation $V = F_Y(y)$, i.e. $V \sim U(0,1)$ [15]. The general proof of this theorem is given in an advanced undergraduate textbook [16 Page: 52-54] by Casella and Berger. However, the PIT for higher dimensions is far less understood [3].

Assume a random vector $\mathbf{Y} = (Y_1, \cdots, Y_d)$ of multi-response is jointly distributed as $F_{\mathbf{Y}}(y_1, \cdots, y_d)$ $(d > 1)$, with random variable $V = F_{\mathbf{Y}}(Y_1, \cdots, Y_d)$ as the analogue d-dimensional PIT of $\mathbf{Y}$, the CDF of V is denoted by $K_V(v)$ on [0,1], by definition, $K_V(v) = \mathrm{P}\big(V \le F_{\mathbf{Y}}(y_1, \cdots, y_d)\big)$. However, $K_V(v)$ is not generally a uniform distribution as the PIT of a random variable. It has been recognized that $K_V(v)$ contains valuable information of the correlation structure underlying the joint CDF of $\mathbf{Y}$, but it does not depend on the marginal distributions of the random variables in $\mathbf{Y}$. The computation of the PIT distribution $K_V(v)$ is quite tractable: the bivariate PIT distributions can be derived analytically based on the underlying copulas of the joint CDFs [14]. While in general cases, due to the fact that the

PIT random variable V is a function of the random vector $\mathbf{Y}$, V can be sampled from the function values of the random numbers that are generated by the joint CDF of $\mathbf{Y}$, and subsequently an empirical $K_V(v)$ can be simulated based on the random numbers of V. Figure 1 illustrates the PIT of a bivariate CDF. In Figure 1 (a), a random sample $(y_1, y_2)_j$, $j = 1, \cdots, n$ generated by the bivariate CDF is put into the distribution to obtain the corresponding sample v_j of the PIT random variable V. As sufficient samples of V are collected in $\{v_j\}_{j=1}^n$ with the increase of samples, a smooth empirical CDF of V can be simulated, as shown in Figure 1 (b). The multivariate PIT has been successfully applied in literature to obtaining the maximum likelihood estimation of dependence parameters [17], testing the copula goodness of fit of the dependence structure [18], and evaluating the conditional density forecast in the econometric mainstream [19], etc.

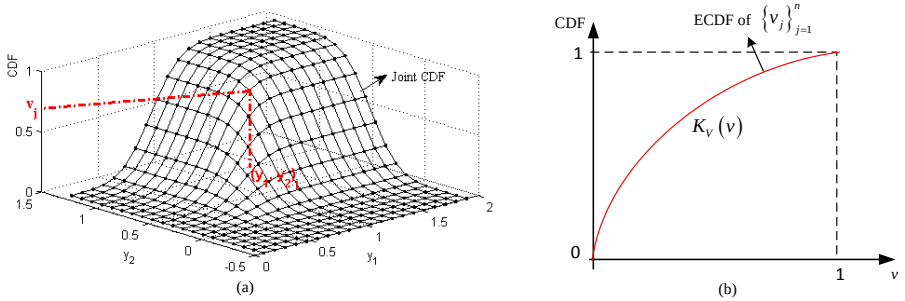


Figure 1: Empirical PIT distribution of a bivariate CDF

4.2. Area Metric and U-pooling Metric

The area metric proposed by Ferson et al. [5, 7] aims at measuring the disagreement of the entire distributions of predictions and observations, of which the metric operator is shown in Eq. (1). For a model that predicts a response y at a single validation site, by taking integral over the area differences between the CDF of the model response $F^m(y)$ and the empirical CDF of the observed data $S_n^e(y)$, the overall disagreement between the predictive model and physical experiment can be measured.

$$d\left(F^m, S_n^e\right) = \int_{-\infty}^{+\infty} \left|F^m(y) - S_n^e(y)\right| dy \tag{1}$$

A larger area difference would indicate larger disagreement at the specified validation site, and vice versa.

The u-pooling metric was proposed upon the idea of the area metric by Ferson et al. [5] to measure the overall disagreement between the predictive model and the physical experiment at multiple validation sites. By applying the probability integral transformation for in univariate case, different prediction distributions can all be transformed into a standard uniform distribution, i.e. $U(0,1)$. Figure 2 provides an illustration of the u-pooling method for three experimental data y_k^e $(k = 1,2,3)$ observed at multiple validation sites which corresponding to different prediction distributions $F_k^m(\cdot)$. In Figure 2 (b), for each observed data y_k^e, a corresponding u-value is calculated according to $F_k^m(\cdot)$, i.e. $u_k = F_k^m(y_k^e)$ [3]. Then the area differences (shaded areas in Figure 2 (a)) between empirical CDF of the u-values and the standard uniform distribution are added together to provide a single metric that accounts for the overall accuracy of a model at multiple validation sites. The value of the metric is between 0 and 1/2, with 0 indicating a perfect match between the model and experiments, and 1/2 indicating a worst match.

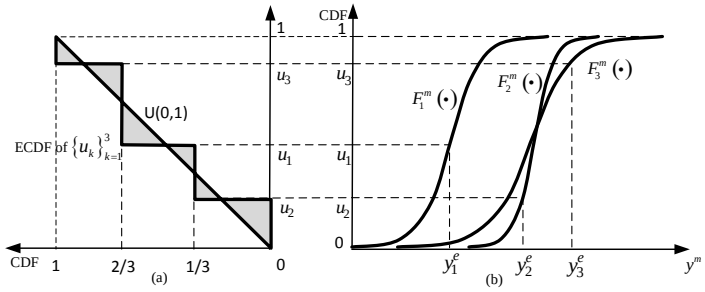


Figure 2: Illustration of the U-pooling metric

For the measurements of multiple responses at different validation sites, however, the joint CDFs of model responses will be transformed into different PIT distributions, and this u-pooling is not possible when data are compared against different distributions. As mentioned in Section 3, an average or weighted-sum approach is often used in practice without considering the correlation among the responses. The limitations of the marginal u-pooling approach will be further examined by comparing it to the proposed t-pooling metric through numerical studies in Section 6.1.2.

5. Proposed validation metrics for multiple responses

In this section, we extend the idea of area metric/u-pooling method and propose two metrics for assessing the predictive capability of models with correlated multiple responses considering uncertainty in both physical experiments and predictive models. The first one is the PIT area metric for observations collected at a single validation site, and the second is the t-pooling metric for observations at multiple validation settings of interest. Multivariate PIT is conducted in the validation process of both metrics to incorporate the correlation among the responses.

5.1. PIT area metric

The PIT area metric defined in this subsection provides a comparison between the probabilistic predictions and empirical observations of multiple response quantities at a single spatial or temporal location. The physical experiment considered has multiple responses, $y_i^e(\mathbf{x}), i = 1, \cdots, d,$ where $\mathbf{x}$ is a vector of the controllable inputs which determines the intended validation sites and d is the number of responses. The candidate computer model for predicting these responses are denoted by $y_i^m(\mathbf{x}, \boldsymbol{\theta})$, with $\boldsymbol{\theta}$ a vector of model parameters. To address the issues of both uncertainty and correlation in the validation assessment, the probabilistic predictions of the computer model can be characterized by the joint cumulative distribution function $F^m(y_1, \cdots, y_i, \cdots, y_d)$ of the model responses, which contains the information of both marginal distributions and correlations.

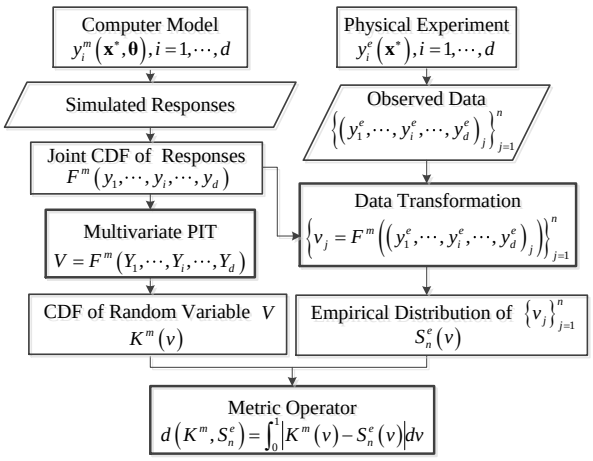


Figure 3: Flowchart of PIT area metric for single input site

A flowchart of the validation process by the proposed PIT area metric is shown in Figure 3. Based on the flowchart, the predictive capability of a computer model is assessed according to the following steps:

Step 1: On the right side of the flowchart, a number of data sets $\{(y_1^e, \cdots, y_i^e, \cdots, y_d^e)_j\}_{j=1}^n$ are collected from the physical experiments at a specified validation site $\mathbf{x}^*$, where $(y_1^e, \cdots, y_i^e, \cdots, y_d^e)_j$ is the j-th observed data set and n is the number of the data sets .

Step 2: On the left side, the candidate computer model is run at the same validation site $\mathbf{x}^*$ to generate simulated responses for constructing the joint CDF of the model responses, i.e. $F^m(y_1, \cdots, y_i, \cdots, y_d)$.

Step 3: The multivariate joint CDF of the responses is transformed into a univariate CDF $K^m(v)$ according to the multivariate probability integral transformation theorem, where V is the multivariate PIT random variable of the model responses.

Step 4: Correspondingly, the experimental data sets are transformed into a one-dimensional data sequence

$\{v_j\}_{j=1}^{n}$, in which every *v*-value is the joint CDF value of the relevant set of data, i.e. $\{v_j = F^m((y_1^e, \cdots, y_i^e, \cdots, y_d^e)_j)\}_{j=1}^{n}$. Then the empirical distribution $S_n^e(v)$ is computed based on these *v*-values. If we assume that the probabilistic predictions are exactly the same as the responses of the physical experiment, these v-values would be the samples of random variable V, and therefore $S_n^e(v)$ will have the same distribution as $K^m(v)$. If, however, there is a constant difference between the two distributions that cannot be eliminated by adding more observations, we can infer that there is some disagreement between the predictions and the physical experiment.

Step 5: As a resulting, the two distributions, $K^m(v)$and$S_n^e(v)$, are compared according to the metric operator $d(K^m, S_n^e) = \int_{-\infty}^{+\infty} |K^m(v) - S_n^e(v)|\, dv$, which is the area difference between the two CDF curves. The integral is taken over a unit interval [0, 1] because the PIT random variable V is sampled from CDF values.

It should be noticed that both the information of correlation and uncertainty has been incorporated in the proposed PIT area metric, either by the transformed PIT distribution or the transformed data sequence. Also, different from the hypothesis testing based methods, there is neither assumption of normality regarding to the distribution of the model responses nor the experimental observations. The metric is applicable for general multi-response problems.

5.2. t-pooling metric

The u-pooling metric is feasible for pooling incomparable data of single response problems due to the fact that the probability integral transformation for any one dimensional CDF is a standard uniform distribution. For multiple responses, however, different joint CDFs will be transformed into different PIT distributions. Therefore, the original u-pooling metric is not applicable for pooling data of multiple responses observed at different validation sites.

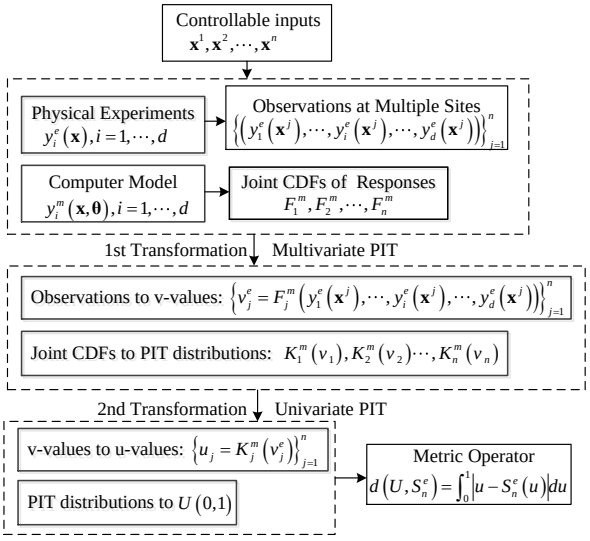


Figure 4: Flowchart of the t-pooling metric for multiple input sites

In this subsection, another transformation based area metric, namely, the t-pooling metric, is proposed for integrating the evidence from all relevant data of multi-response quantities over the intended validation domain into a single measure of the overall disagreement. A flowchart of the t-pooling metric for data observed at different controllable input locations is provided in Figure 4. The CDFs of the model responses and observations are respectively transformed twice into a standard uniform distribution and comparable data set. Given a series of input sites $\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^n$, which are often selected by the design of experiments methods [20, 21], the experimental data $\left\{\left(y_1^e(\mathbf{x}^j), \cdots, y_i^e(\mathbf{x}^j), \cdots, y_d^e(\mathbf{x}^j)\right)\right\}_{j=1}^{n}$ can be observed by measuring the physical experiment responses $y_1^e(\mathbf{x}), \cdots, y_i^e(\mathbf{x}), \cdots, y_d^e(\mathbf{x})$ at these input settings. The computer model is also simulated at theses input sites in the same sense to construct the relevant joint CDFs $F_1^m, \cdots, F_j^m, \cdots, F_n^m$ of the predictive responses.

Following the illustration in Figure 4, in the 1st transformation, all joint CDFs $F_1^m, \cdots, F_j^m, \cdots, F_n^m$ of the model responses are transformed to their corresponding PIT distributions $K_1^m, \cdots, K_j^m, \cdots, K_n^m$. Simultaneously, the observations are transformed by their relevant joint CDFs into a sequence of v-values with

$v_j^e = F_j^m\left(y_1^e(\mathbf{x}^j), \cdots, y_i^e(\mathbf{x}^j), \cdots, y_d^e(\mathbf{x}^j)\right)$. Each of the v-values can be compared with the PIT distribution transformed by its relevant joint CDF. This step is the same as the PIT area metric, however, v-value is only comparable with a specified PIT distribution, for example, v_1^e is comparable with K_1^m, v_2^e is comparable with K_2^m, etc. The pooling is not possible when data are compared against different distributions.

Therefore, a 2nd transformation needs to be conducted for the v-values and the PIT distributions. Due to the univariate nature of the PIT distributions, $K_1^m, \cdots, K_j^m, \cdots, K_n^m$ can all be transformed into a standard uniform distribution $U(0,1)$. Meanwhile the v-values are transformed into a set of u-values $\left\{u_j = K_j^m\left(v_j^e\right)\right\}_{j=1}^n$, which are all comparable with the CDF of $U(0,1)$. This step is similar to the idea of the u-pooling metric, except that the distributions and the data are transformed. Furthermore an empirical CDF $S_n^e(u)$ is obtained based on these u-values to compare against the CDF of the uniform distribution. As a result, all evidence of the mismatch between the observations and the predictions at different validation sites can be summarized in the metric operator $d(U, S_n^e) = \int_0^1 |u - S_n^e(u)| du$, and therefore, provides a global assessment for multi-response models.

For illustration purpose we assume only one observation for each input site, hence the number of observations is the same as that of the validation sites. If multiple observations exist at each validation site, the proposed t-pooling metric can still be used, and the only difference is that every observation needs to be transformed by the distribution relevant to the specified validation site for the observation.

The metrics proposed in this section have inherited many good features of area metric/u-pooling method for single response [5]. For example, the proposed metrics do not include of any criterion or belief of accepting a model [3]. On the other hand, the value of the original area metric could be anything larger than 0, thus a distinctive disadvantage of the area metric is the difficulty in determining the model acceptance threshold. For the proposed PIT area metric and t-pooling metric, however, given the fact that the PIT distribution and the empirical distribution of the transformed data are both distributed over [0, 1], the range of the proposed metrics are both ideally normalized, so the accuracy requirement of the model acceptance can be easily determined without considering the magnitude of individual responses.

6. Numerical case studies

In this section, a series of numerical studies are designed to compare the proposed metrics with the existing area metrics based methods for testing their validity. The PIT area metric is compared with the direct area metric method defined in Eq. (2). The PIT area metric is an immediate extension of the original single response area metric in Eq. (1) with higher dimensional integrals:

$$d\left(F^m, S_n^e\right) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \left| F^m\left(y_1, \cdots, y_i, \cdots, y_d\right) - S_n^e\left(y_1, \cdots, y_i, \cdots, y_d\right) \right| dy_1 dy_2 \cdots dy_d \ , \tag{2}$$

where $F^m(y_1, \cdots, y_i, \cdots, y_d)$ is the joint CDF of model responses and $S_n^e(y_1, \cdots, y_i, \cdots, y_d)$ is the multivariate ECDF of the experimental data. The t-pooling metric is compared with the marginal u-pooling method.

For demonstration purpose, the experimental observations in this section are generated using the following two responses:

$$\begin{aligned} y_1^e(x,\theta) &= \sin\left(2x - 0.5\pi\right)\theta + \varepsilon_1 \\ y_2^e(x,\theta) &= \cos\left(0.25\pi x\right)\theta + 0.2x + \varepsilon_2 \end{aligned} \ , \tag{3}$$

where y_1^e and y_2^e stand for experimental responses, x $(0 \le x \le 6)$ is a deterministic control variable, and θ is a model parameter that equals to 1.5. The measurement errors of the two responses, ε_1 and ε_2, both follow a zero mean Gaussian distribution $N(0, \sigma^2)$ with standard deviation $\sigma = 0.2$ and a correlation coefficient $\rho_{\varepsilon_1,\varepsilon_2} = 0.5$ between them. The responses generated from the above two functions are treated as experimental observations in the validation process. It can be noted that since both the function forms of y_1^e and y_2^e and the measurement errors ε_1 and ε_2 are correlated, the experimental responses are set to be correlated random variables.

Two test settings are created by using different predictive models. The predictive computer models of the two tests are summarized in Table 1. Test 1 aims at examining whether the proposed two metrics can differentiate correct and incorrect models. Test 2 focuses on studying whether the metric can differentiate between models with larger and lesser uncertainty when the correlation between the model responses varies from site to site. The measurement errors are included as a part of the predictive models but with different correlation coefficients in different cases.

Table 1: Formulas of the predictive (computer) models in two test cases

Test	Model ID	Formulas	Model description
Test 1	1	$y_1^{m_1}(x) = y_1^e(x, \theta = 1.5)$ $y_2^{m_1}(x) = y_2^e(x, \theta = 1.5), \rho_{\varepsilon_1,\varepsilon_2} = 0.5$	Exactly as the experimental data source.
	2	$y_1^{m_2}(x) = y_1^e(x, \theta = 1.2)$ $y_2^{m_2}(x) = y_2^e(x, \theta = 1.2), \rho_{\varepsilon_1,\varepsilon_2} = 0.5$	Model parameter is incorrect.
	3	$y_1^{m_3}(x) = y_1^e(x, \theta = 1.2)$ $y_2^{m_3}(x) = y_2^e(x, \theta = 1.2), \rho_{\varepsilon_1,\varepsilon_2} = -0.6$	Both model parameter and correlation coefficient are incorrect.
Test 2	4	$y_1^{m_4}(x) = y_1^e\left(x, \theta \sim N(1.5, 0.2^2)\right)$ $y_2^{m_4}(x) = y_2^e\left(x, \theta \sim N(1.5, 0.2^2)\right), \rho_{\varepsilon_1,\varepsilon_2} = 0.5$	Mean of the model parameter is exact, uncertainty is smaller.
	5	$y_1^{m_5}(x) = y_1^e\left(x, \theta \sim N(1.5, 0.4^2)\right)$ $y_2^{m_5}(x) = y_2^e\left(x, \theta \sim N(1.5, 0.4^2)\right), \rho_{\varepsilon_1,\varepsilon_2} = 0.5$	Mean of the model parameter is exact, uncertainty is larger.

6.1. Test 1: differentiating correct and incorrect Models

In this section, three predictive models are validated against the physical observations generated by Eq. (3). Model 1 is considered to be a correct predictive model with model parameter θ and the measurement correlation coefficient $\rho_{\varepsilon_1,\varepsilon_2}$ matches exactly as that of the experimental data source; model 2 is set to have an incorrect model parameter with $\theta = 1.2$; model 3 is assumed to not only have an incorrect model parameter $\theta = 1.2$, but also have a wrong measurement correlation coefficient $\rho_{\varepsilon_1,\varepsilon_2} = -0.6$.

6.1.1. Single validation site: $\boldsymbol{x = 2.0}$

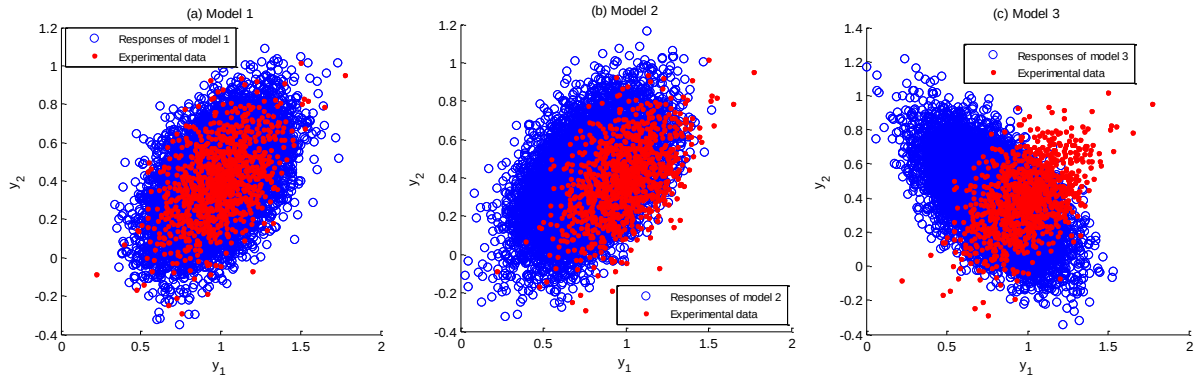


Figure 5: Graphical comparisons between observations and model responses at input site $x = 2.0$

1000 observations are generated from Eq. (3) at a representative input site $x = 2.0$ for validating each of the three models. The scatter plots in Figure 5 provide graphical comparisons between the data from physical experiments and the three models, 1000 sets of observations (marked by red dots in Figure 5) are respectively compared with 10000 sets of simulated responses (the blue circles) generated by each predictive model for visualizing the differences between the predictions and the physical experiments. As shown in Figure 5 (a), because model 1 is set the same as the experimental data source, the data cloud of the predictions of model 1 overlaps extremely well with the cloud of the experimental observations. For model 2 in plot (b), the shapes and orientations of the two data clouds are almost the same, but there is a noticeable distance between their centroids, which makes sense because model 2 is set to have a discrepancy in the model parameter θ. The scatter plot (c) of model 3 demonstrates the worst agreement between the predictions and observations: since neither the correlation information nor the marginal distributions of the two responses is correctly modeled - not only the centroids of the two clouds are far from each other, but their orientations behave quite differently too. Ideally, a good metric should indicate that model 1 is better than model 2, and model 3 is the worst among the three models.

Figure 6 illustrates when applying the conventional "direct area metric", the metric measures the differences between the joint CDF of the model responses (transparent surfaces) and the multivariate ECDF (colored surface) of

the observed data. Figure 7 shows when applying the proposed PIT area metric, the metric provides is a comparison between the PIT distribution of the joint CDF (blue curve) and the ECDF of the transformed observations (red curve). All multivariate ECDF surfaces and the ECDF curves in Figure 6 and 7 are smooth, indicating that the amount of observations are sufficient to draw trust-worthy metric measures. Results of the two metrics for these three models are summarized in Table 2.

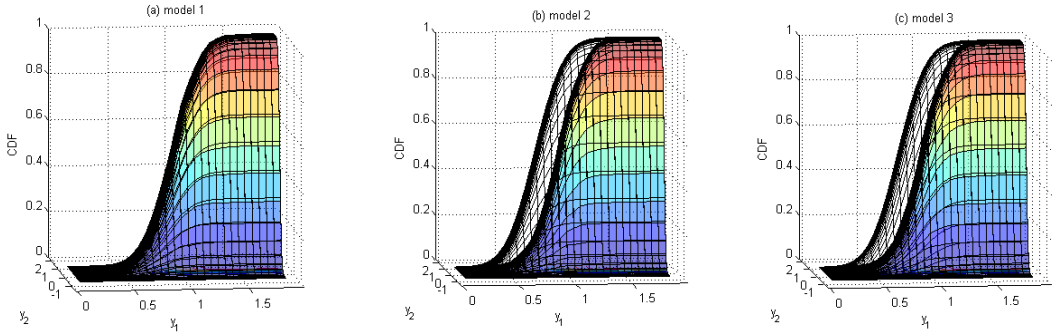


Figure 6: Direct area metric for model 1, model 2 and model 3

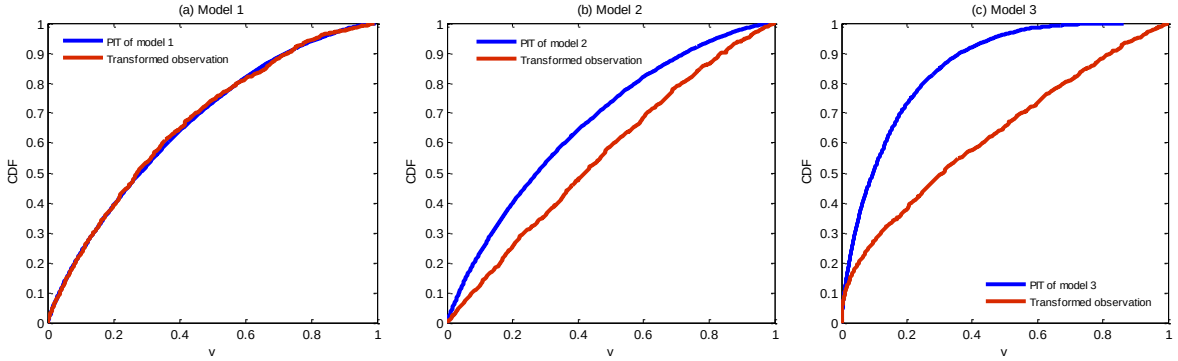


Figure 7: PIT area metric for model 1, model 2 and model 3

Table 2: Comparison of Metric results of the three models for both single and multiple validation sites

Methods/Models	Model 1	Model 2	Model 3
Direct Area Metric	0.041	**0.471**	**0.442**
PIT Area Metric	0.009	0.105	0.184
Separate U-pooling	0.003	**0.083**	**0.083**
T-pooling Metric	0.012	0.103	0.144

Since model 1 is an accurate model, the metric value of model 1 is expected to be 0. As shown in Table 2, the proposed PIT area metric provides a more accurate assessment (0.009) versus the result from direct area metric (0.041); the inaccuracy of the latter approach is due to multi-dimensional computations. The results of both metrics suggest that model 2 is less accurate than model 1. However, for model 3, the result of the PIT area metric shows correctly that model 3 is less accurate than model 2, but the direct area metric result suggests that model 3 is almost as good as or even slightly better than model 2. Our comparison shows that the direct area metric is incapable of differentiating models with right or wrong correlation coefficients.

6.1.2. Multiple validation sites

The proposed t-pooling metric is tested against the marginal u-pooling method in this study to assess the global predictive capability of the three models. 1000 sets of observations are collected at multiple validation sites with only one observation at each site. These validation sites are uniformly distributed on the interval [0, 6] of *x*. The

graphical comparisons of observations and the responses of the three models are provided in Figure 8; 10 sets of simulated responses are generated from the models at multiple validation sites for shaping the data clouds of the models. Thus the predictive capability of the three models can be easily judged based on the similarity in the shape and density of the data clouds between the experiments and the models. Clearly, the data cloud of model 1 shows the best match with that of the observations, and model 2 has a better match than model 3. Yet we need to quantify this degree of similarity by metrics to provide a quantitative measure of the disagreement between the computational model and the experiments.

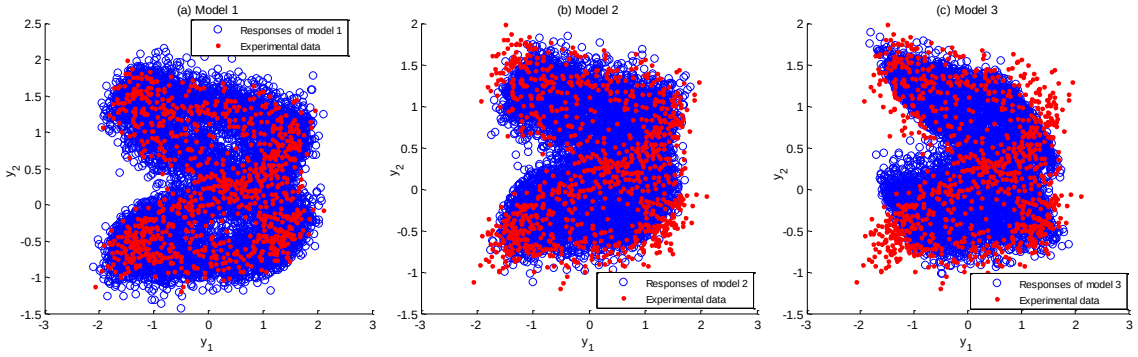
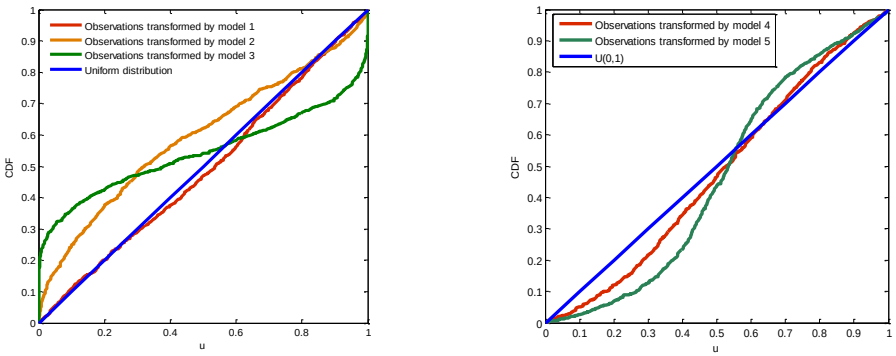


Figure 8: Graphical comparisons between observations and model responses over the entire interval of x

For pooling these observations of multiple quantities at different sites, the t-pooling metric provides a comparison between the ECDF of the transformed data and the standard uniform distribution. The ECDFs of the observations transformed by the three models and the CDF of $U(0,1)$ are compared in Figure 9 (a) to show the accuracy of the three models. Since the marginal u-pooling approach provides the average of the u-pooling metrics of the two responses, the correlation information neither in the observations nor in the model responses is considered. The metric results of the marginal u-pooling metric and the t-pooling metric for the three models are listed in Table 2. It is noted that the marginal u-pooling result for model 2 and model 3 are exactly the same. This is due to the reason that the metric only measures the difference of marginal distribution in each response, and totally ignores the correlation between them. Since the data used in the proposed t-pooling metric is transformed by relevant joint CDFs and the PIT distributions, both correlations and uncertainty among the responses are captured in the metric. This explains why the t-pooling metric provides a result consistent with the real accuracy of the three models.



(a) Model 1, model 2 and model 3 (b) Model 4 and model 5

Figure 9: T-pooling metric for the competing models

6.2. Test 2: differentiating models with smaller or larger uncertainty

In Test 2, we consider two candidate models for testing whether the proposed two metrics can differentiate between models of greater and lesser uncertainty. Both predictive models in this test set have an uncertain model parameter θ due to the lack of knowledge. As shown in Table 1, the uncertainty parameter θ in model 4 follows a

Gaussian distribution with a smaller standard deviation, while in model 5, θ follows a Gaussian distribution with a larger standard deviation. For testing the performance of the proposed PIT area metric, 1000 sets of observations at validation site $x = 3.0$ are compared with the model responses. The observations are plotted together with 10000 simulated responses of the two models in Figure 10. Although the data clouds of model 4 covers fewer observations than model 5, a large amount of predictions generated by model 5 are quite far away from the boundary of the observations. Ideally, the metric should indicate that model 4 is more accurate than model 5.

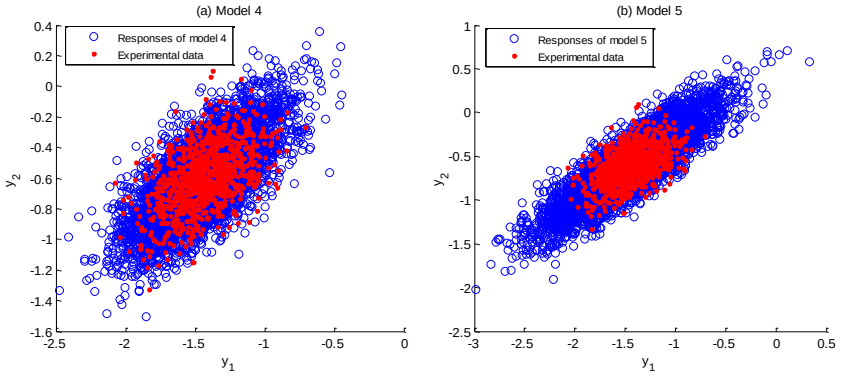


Figure 10: Graphical comparison between the observations and responses of model 4 and model 5 at site $x = 3.0$

The PIT distributions of the two models and the ECDFs of the transformed observations are plotted in Figure 11 to show the disagreement between the predictive models and the physical observations. The area difference for model 5 is 0.117, which is relatively larger than 0.48 for model 4, thus shows that the proposed PIT area metric is capable of differentiating between models with more or less uncertainties in their predictions. For testing the t-pooling metric, we use the same sets of observations collected in Section 6.1.2 to validate the two models in a global sense. The empirical CDFs of data transformed by the joint CDFs and PIT distributions of the two models are compared with standard uniform distribution in Figure 9 (b), while the corresponding metric results are summarized in Table 3 together with the results of the PIT are metric for the two models. The correlation coefficients between the responses of the two models are different from site to site due to the change of model input variable *x*; the correlations are captured by the PIT distributions at each site during the t-pooling process. The results (0.039 for model 4 and 0.080 for model 5) show that the t-pooling metric is capable of differentiating between models with lesser or greater even though the correlation information of the models varies from site to site.

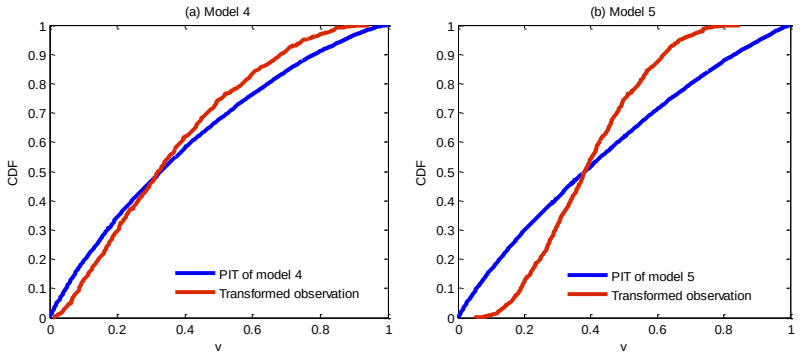


Figure 11: PIT area metric for model 4 and model 5 at input site $x = 3.0$

Table 3: Metric results for model 4 and model 5

Methods/Models	Model 4	Model 5
PIT Area Metric	0.048	0.117
T-pooling Metric	0.039	0.080

6.3. Additional comments on the merits of proposed metrics

Because the direct area metric depends on the physical units in which the prediction distributions and data are expressed, the units of multiple responses are merged in the metric and cannot represent the scale of any of the response. It also becomes difficult to set up an appropriate threshold for rejecting or accepting a model. As the number of responses increases, the computational cost of the direct area metric would be huge due to the high dimensional integrals and the multivariate ECDFs. On the other hand, the proposed PIT area metric is intelligently normalized by the transformation which makes it independent from the scales of the responses, thus the model acceptance threshold can be easily determined. Besides, the comparisons of multivariate data are transformed into one dimensional integral during the validation process. Hence the computational cost of the proposed metric is much lower than the direct area metric. The test in subsection 6.1.2 shows both the direct area metric and the marginal u-pooling metric lack the ability to incorporate the correlation information among the responses, while the proposed two metrics could capture the disagreement of uncertainty information as well as the correlation owing to the multivariate PITs. In addition, the proposed metrics have many desired features that are inherited from the area metric/u-pooling for single response, such as the capability of being used when the amount of physical experiments is small and the capability of providing measures of the global accuracy of a model. However, it should be noted that when the predictions are sparse, for example, when a computer model is extremely expensive to run to extract the uncertainty and correlation information, the risk of underestimation or overestimation the metrics should be seriously considered. This risk can be reduced when more information is collected.

7. Conclusion

For the validation assessment of models with correlated multiple responses, it is important to address both the issues of uncertainty and correlations among the responses. In this paper, two metrics are developed for extending the area metric/u-pooling based methods into multivariate cases by using multivariate probability integral transformations. With the PIT area metric, the experimental data sets observed at a specified validation site are transformed into a univariate data sequence based on the relevant joint CDF of the model responses, and then an empirical expression of the data sequence is compared with the PIT distribution of the joint CDF to show the disagreement between the predictions and observations. For observations of multiple quantities that are collected at different sites, the t-pooling metric is developed for integrating all the evidence from these sites together to assess the global predictive capability of the multivariate predictive models. The pooling is made possible because the observations are transformed according to the corresponding model CDFs and PIT distributions into a data sequence that is comparable with the standard uniform distribution. The differences in uncertainty and correlations between the predictions and observations are addressed through the joint CDFs of the model responses and PIT distributions in the proposed metrics without normality assumptions. After respectively compared with the direct area metric and the marginal u-pooling method through numerical test studies, we found that in addition to the metrics of the area metric, the proposed approaches (1) could sufficiently capture the correlation information among the responses, (2) are convenient for setting model acceptance threshold, and (3) have lower computational cost in face of the multivariate models. These features allow the proposed metrics to be well suited for the validation assessment of multi-response models, especially in handling the correlation among responses.

Further research is needed on several issues, including: (1) How the confidence bounds of the metrics can be quantified based on the sufficiency of data to reduce the risk of underestimate or overestimate the real discrepancy of a model, and (2) How the predictive capability of individual response in a model can be judged based on the proposed metrics considering the physical units of the response. It would also be interesting to investigate the possibility of transforming observations once instead of twice before comparison in the t-pooling metric.

Acknowledgements

The grant support from the China Scholarship Council and US National Science Foundation (CMMI-1233403) are greatly acknowledged. The views expressed are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] W. L. Oberkampf, T. G. Trucano and C. Hirsch, Verification, Validation, and Predictive Capability in Computational Engineering and Physics, *Applied Mechanics Reviews*, 57(3), 345–384, 2004.

[2] D. Sornette, A. B. Davis, K. Ide, K. R. Vixie, V. Pisarenko and J. R. Kamm, Algorithm for Model Validation: Theory and Applications, *Proceedings of the National Academy of Sciences of the United States of America*, 104 (16), 6562–6567, 2007.

[3] Y. Liu, W. Chen, P. Adrent and H. Huang, Toward a Better Understanding of Model Validation Metrics,

ASME Journal of Mechanical Design, 133 (7), 071005, 2011.
[4] W. L. Oberkampf, and M. F. Barone, Measures of Agreement between Computation and Experiment: Validation Metrics, *Journal of Computational Physics*, 217(1), 5–36, 2006.
[5] S. Ferson, W. L. Oberkampf and L. Ginzburg, Model Validation and Predictive Capability for the Thermal Challenge Problem, *Computer Methods in Applied Mechanics and Engineering*, 197 (29-32), 2408-2430, 2008.
[6] C. J. Roy, W. L. Oberkampf, A Comprehensive Framework for Verification, Validation, and Uncertainty Quantification in Scientific Computing, *Computer Methods in Applied Mechanics and Engineering*, 200 (25-28), 2131–2144, 2011.
[7] S. Ferson, W. L. Oberkampf, Validation of Imprecise Probability Models, *International Journal of Reliability and Safety*, 3 (1–3), 3–22, 2009.
[8] Y. Ling and S. Mahadevan, Quantitative model validation techniques: new insights, *Reliability Engineering & System Safety*, 111, 217–231, 2013.
[9] R. G. Hills, Model Validation: Model Parameter and Measurement Uncertainty, *ASME Journal of Heat Transfer*, 128 (4), 339–351, 2006.
[10] K. J. Dowding, M. Pilch and R. G. Hills, Formulation of the Thermal Problem, *Computer Methods in Applied Mechanics and Engineering*, 197 (29–32), 2385–2389, 2008.
[11] R. Rebba and S. Mahadevan, Validation of Models with Multivariate Output, *Reliability Engineering & System Safety*, 91(8), 861–871, 2006.
[12] X. Jiang and S. Mahadevan, Bayesian Validation Assessment of Multivariate Computational Models, *Journal of Applied Statistics*, 35 (1), 49–65, 2008.
[13] M. C. Kennedy and A. O'Hagan, Bayesian Calibration of Computer Models, *Journal of the Royal Statistical Society*, 63 (3), 425–464, 2001.
[14] C. Genest and L. P. Rivest, On the Multivariate Probability Integral Transformation, *Statistics & Probability Letters*, 53, 391–399, 2001.
[15] J. E. Angus, The Probability Integral Transformation and Related Results, *SIAM Review*, 36 (4), 652-654, 1994.
[16] G. Casella and R. L. Berger, *Statistical inference*, Duxbury Press, 2002.
[17] A. Chakak and L. Imlahi, Multivariate Probability Integral Transformation: Application to Maximum Likelihood Estimation, *RACSAM*, 95 (2), 201-212, 2001.
[18] C. Genest, J. C. Quessy and B. Remillard, Goodness-of-fit Procedures for Copula Models Based on the Probability Integral Transformation, *Board of the Foundation of the Scandinavian Journal of Statistics*, 33, 337–366, 2006.
[19] I. Ishida, Scanning multivariate conditional densities with probability integral transforms, *CIRJE F-Series*, Faculty of Economics, University of Tokyo, CIRJE-F-369, 2005.
[20] G. E. Box, W. G. Hunter and J. S. Hunter, *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, Wiley, New York, 1978.
[21] J. Sacks, W. J. Welch, T. J. Mitchell and H. P. Wynn, Design and Analysis of Computer Experiments, *Statistical Science*, 4(4), 409-423, 1989.