*Proceedings of the 2006 Winter Simulation Conference*
*L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds.*

# A NEW METRIC FOR MEASURING METAMODELS QUALITY-OF-FIT FOR DETERMINISTIC SIMULATIONS

Husam Hamad

Electronic Engineering Department
Hijjawi College of Engineering Technology
Yarmouk University
Irbid, JORDAN

## ABSTRACT

Metamodels are used to provide simpler prediction means than the complex simulation models they approximate. Accuracy of a metamodel is one fundamental criterion that is used as the basis for accepting or rejecting a metamodel. Average-based metrics such as root-mean-square error RMSE and R-square are often used. Like all other average-based statistics, these measures are sensitive to sample sizes unless the number of test points in these samples is adequate. We introduce in this paper a new metric that can be used to measure metamodels fit quality, called metamodel acceptability score MAS. The proposed metric gives readily interpretable meaning to metamodels acceptability. Furthermore, initial studies show that MAS is less sensitive to test sample sizes compared to average-based validation measures.

## 1 INTRODUCTION

Metamodels are approximations to simulation models. They are built and validated using simulation results for samples of data points in the input space. Two fundamental criteria are used as the basis for accepting or rejecting a metamodel: efficiency and accuracy. Efficiency is indicative of how expeditiously predictions can be obtained; accuracy is indicative of how good these predictions are.

Efficiency of a metamodel can be determined prior to metamodel construction, and without any computational cost in terms of the simulation runs needed, e.g., the time taken to evaluate a second-order polynomial metamodel in a given number of dimensions is the same regardless of the underlying simulation model. On the other hand, determining the accuracy of a metamodel is closely linked to the number of data points used in error calculations.

The accuracy of a metamodel is determined using objective methods or subjective methods (Balci 1989, Sargent 2004, and Hamad 2005). Objective methods are based on statistical tests that make certain assumptions about the data's correlation, distribution, etc. A central assumption for the application of these statistics is related to the number of data points used. Even when other assumptions are satisfied, statistical tests are meaningful only if the data used is sufficient in number. It is often the case that, even if the system is completely observable, obtaining a sufficient number of observations is impractically expensive. For such cases, average-based metrics such as RMSE and R-square -two of the most popular statistics used for metamodel accuracy assessment-may be 'sensitive' to the number of observations used.

The objective of this paper is to introduce a new measure to assess the acceptability of metamodels quality of fit based on their accuracy of predictions. The term given to this metric is MAS: metamodel acceptability score. MAS can be used particularly for situations where there is not enough validation test data or the validation data is too expensive to generate. The proposed metric gives readily interpretable meaning to metamodels acceptability. Furthermore, and by comparison to average-based measures such as RMSE and R-square, initial studies show that MAS is less sensitive to test sample sizes.

The remainder of this paper is organized as follows. In section 2, the MAS metric is introduced and contrasted with the RMSE and R-square metrics. Test results using these three metrics are compared by examples in section 3. The paper is then concluded by section 4.

## 2 MAS: METAMODEL ACCEPTABILITY SCORE

The proposed metric MAS is defined in this section as a measure of metamodel acceptability with regard to prediction accuracy. The discussion of this new metric is presented in the context of the two average-based statistics of RMSE and R-square. In the following discussion, $y_i$ denotes the response modeled by $\hat{y}_i$ for the $i^{th}$ data point in a validation test sample having $n$ observations.

1-4244-0501-7/06/$20.00 ©2006 IEEE

### 2.1 Average-Based Statistics

Two of the more important measures used for model accuracy assessment including deterministic simulation models are RMSE and R-square. They are defined by

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \qquad (1)$$

$$R^2 = 1 - \frac{\mathrm{MSE}}{\sigma^2} \qquad (2)$$

where MSE is the mean square error and $\sigma^2$ is the variance.

Note that results returned by RMSE cannot be interpreted without referring to the context of the problem. For example, if in a test problem RMSE is found to be 0.05 and 500 in another, then it could be that the metamodel is more accurate for the latter problem; e.g., if $y_i$ values are within the range 0.04 to 0.06 for the first problem and 10000 to 20000 in the second one. Furthermore, the size of RMSE is strongly influenced by the number of observations n in the test sample if the sample size is inadequate.

Variations with sample size n in the R-square statistic can be in some cases less pronounced, since there is a chance that the influence of such variations cancels out in the numerator and dominator of Equation 2 above. However, care must be taken when interpreting results returned by R-square for response values $y_i$ which are nearly constant. In such cases, the dominator of the second term in Equation 2 is small, and even if the metamodel returns nearly zero MSE, the value of the second term in Equation 2 resulting from dividing two small numbers is not without numerical problems. A similar situation leading to questioning the validity of R-square results that is dealt with in the literature rises for the case when n is close to the number of coefficients q in the metamodel. For such cases, R-square is 'adjusted' to accommodate the relative size of n to q; see (Kliejnin 2007).

### 2.2 Metamodel Acceptability Score

Metamodel acceptability score MAS is defined in this subsection. It will be shown how intuitive thinking lead naturally to MAS development and, hopefully, its subsequent adoption for deterministic simulation metamodel validation.

The starting step in metamodel validation activities is a list of n responses $y_i$ and a corresponding list of n metamodel $\hat{y}_i$ values. To determine how close the constituents of each of these n response/metamodel pairs are to each other either the difference $(\hat{y}_i - y_i)$ or the metamodel-to-response ratio $(\hat{y}_i/y_i)$ may be used, broadly speaking. A difference $(\hat{y}_i - y_i)$ of zero or a ratio $(\hat{y}_i/y_i)$ of one indicate a perfect match for the $i^{th}$ pair. Leave out the case $y_i$ equals zero for now; this issue will be addressed in a short while.

Like the case mentioned above for RMSE, a given difference $(\hat{y}_i - y_i)$ value is not readily interpretable without context. On the contrary, and again exempting the case of zero response $y_i$ for the moment, results for the ratio $(\hat{y}_i/y_i)$ require no context for their interpretation; a ratio of 0.999 is always taken positively while a ratio of 999 is on the other hand always taken negatively, regardless of the response modeled.

Before going down the list of n response/metamodel pairs, a criterion is set for acceptability of a given pair. Options for such a criterion are numerous and application dependant. For example, a pair may be accepted if its response and metamodel values are within 20% of each other, i.e., if $(0.8 \le \hat{y}_i/y_i \le 1.2)$. In another application, a conservative metamodel may be favored, thus the acceptability criterion may change to $(0.8 \le \hat{y}_i/y_i \le 1.0)$ so that it is guaranteed that for worst case a higher value for the response is obtained, e.g., to optimize an engineering system design.

Now, given a required acceptability criterion, the list of n response/metamodel pairs is traversed and the number of accepted pairs is counted. Let this number be m; MAS is then defined as

$$\mathrm{MAS} = \frac{m}{n} \times 100 \qquad (3)$$

Given a certain MAS level along with the acceptability criterion used, there is no ambiguity in interpreting the results leading to acceptance or rejection of the metamodel. For instance, a MAS of 90% may be set as the minimum acceptable MAS level for a given metamodel with the acceptability criterion of $(0.8 \le \hat{y}_i/y_i \le 1.2)$. For this situation, a metamodel with MAS = 65%, say, is rejected because 35% of its tested space fails vis-à-vis the given acceptability criterion.

In addition to interpretability of results, MAS sensitivity to variations in sample-to-sample size should be to a lesser degree in comparison to average-based metrics. MAS represents a relative count of test points throughout the input space. Changing the test sample size n leads to a corresponding change in the number of acceptable points m, provided that the distribution of test points over the space in the new sample is not changed; e.g., test points are drawn from the entire space in all cases without putting more emphasis for one sub region over another. .

To clarify this point, suppose that a given test sample has n =100 observations giving a certain MAS level. Now, if only 10 points are used instead, but with a distribution in the input space similar to the distribution of the 100-points sample, e.g., covering the entire input space not only the first half, then for most cases the difference in MAS levels should not be detrimental. Of course, such an argument still needs further scientific justifications, may be benefiting from the well-established sampling theorems such as those used for digital signal processing methods in electronic engineering. Nonetheless, initial investigations show that MAS is less sensitive to test sample sizes compared to average-based validation measures such as RMSE, as expected for the reasons just mentioned, and as will be shown in the examples of the next section.

We now return to the situation where for one or more of the n observations the response $y_i$ is zero. This issue is treated further in (Hamad 2006). Let the number of such points be $\nu$. Then as far as MAS calculation is concerned, two approaches may be taken depending on the size of $\nu$ relative to n:

- If $\nu$ is only a small fraction of n, as may be set by the analyst, then these $\nu$ points are simply removed. This should have almost no effect on MAS levels provided that $n >> 1$.
- If on the other hand $\nu$ is a sizable fraction of n, then, if validation using MAS levels is planned at the outset, an approach may be based on modeling using a 'shift-transformation' of the response $y_i$ to $y_{si} = y_i + \delta$, by adding a constant $\delta$ to all of the n data points $y_i$. Here, $\delta$ is chosen to make the shifted response $y_{si}$ greater than zero. This way, validation using MAS results may be carried out for the complete list of n data points after transformation.

The sample-to-sample variability of MAS results are contrasted to those of RMSE and R-square for the examples in the next section.

## 3 EXAMPLES

We compare in this section variations with sample size for MAS on one hand, and RMSE and R-square on the other hand, via two examples. The first example uses a one-dimensional analytic function for the response. For this example, two metamodels having different number of coefficients q are studied. The second example involves simulation results for an electronic circuit with three design variables (inputs). For these examples:

- Polynomials metamodels are used. The number of coefficients q for a polynomial in k dimensions with a degree d is $q = (k + d)!/k!d!$
- Latin hypercube validation test samples are used. Sample sizes of $\omega q$ are used, where the multiplier $\omega$ is varied in steps of 1 starting at $\omega = 1$. Latin hypercube sampling is used to provide flexibility with sample sizes and good uniformity over the input space.
- The sample-to-sample variation is measured by the difference-mode to common-mode ratio DMCMR. The DMCMR for the two quantities $\varsigma_1$ and $\varsigma_2$ is defined by

$$\mathrm{DMCMR} = \frac{\varsigma_1 - \varsigma_2}{\frac{1}{2}(\varsigma_1 + \varsigma_2)} \times 100 \qquad (4)$$

Absolute values are taken to calculate the common-mode component in the dominator of Equation 4.

### 3.1 Example 1

The following response is defined for the space $x \in [-1,1]$:

$$y = 1 \times 10^{-5}\left(e^{20x} - 1\right) + 1000 \qquad (5)$$

Two metamodels are derived for this response: the first one is a second-order polynomial built using a minimum bias design having four points, and the second metamodel is a fifth-order polynomial derived using another minimum bias design with 10 points.

Accuracy tests are carried out using fifty samples for each metamodel. The number of observations for these samples are $\omega q$ ; $\omega = 1, 2, \ldots, 50$. The number of coefficients q for the second-order and fifth-order polynomials is three and six, respectively.

Calculations of RMSE, R-square, and MAS are carried out for the second-order polynomial metamodel using the fifty test samples in turn. Acceptability criterion is set at $(0.8 \le \hat{y}_i/y_i \le 1.2)$ for MAS calculations. Results are shown in Figure 1(a)-(c) for one trial, while part (d) of the figure shows the results using three other trials for each of the fifty Latin hypercube test samples for MAS calculations.

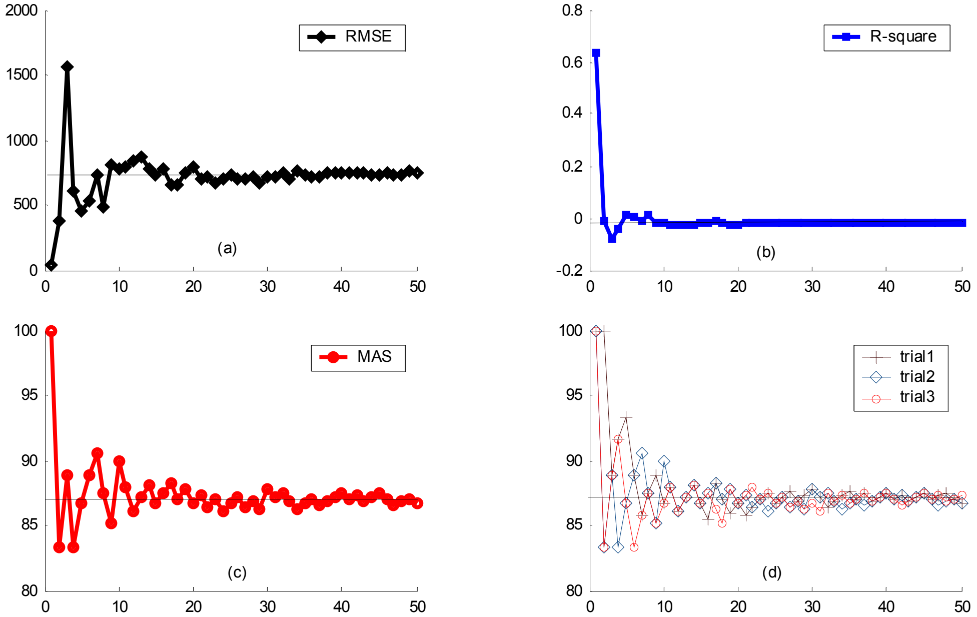

Figure 1: Validation Results for the Second-Order Polynomial Metamodel vs. the Number of Coefficients Multiplier $\omega$ for (a) RMSE, (b) R-Square, and (c) MAS; (d) Three Trials for Each of the 50 Latin Hypercube Test Samples for MAS

Note from the figure that each metric settles at a 'final' value shown by solid lines crossing the plots in the middle. However, variation around the final value is smallest for MAS, as can be seen from Figure 1c which shows that MAS levels are at nearly $(87 \pm 6)$ after $\omega = 1$.

MAS supercedes not only in terms of sensitivity to sample size variation, but also in terms of interpretability of results. To clarify, it can be seen by reference to Figure 1a that RMSE is at nearly 700 for adequate sample sizes. Now, with this RMSE size in mind, is the second-order polynomial a good metamodel? Moving on to part (b) of the figure, R-square values reveal that the metamodel is inadequate.

For MAS results in Figure 1c on the other hand, information about the metamodel quality-of-fit is not ambiguous. The figure reveals that for around 87% of the tested sample observations the metamodel is within $\pm 20\%$ of the response-the criterion used in the example for MAS calculations as mentioned. With such a readily interpretable result, the decision of whether to accept or reject the metamodel becomes far more easier. Note that R-square results point out to a completely contradictory conclusion. This is because y in Equation 5 is almost constant for more than 85% of the input space, a condition which is not favorable for R-square application as mentioned in the previous section.

In order to better compare sample-to-sample variations for the three metrics, DMCMR's corresponding to the data of Figure 1 are calculated and plotted in Figure 2. The same scales can now be used for all three metrics as shown.

Figure 2a shows that RMSE can vary by as much as 170% for small sample sizes. The sample-to-sample variation for R-square can be as high as 200% for the smaller sample sizes, as revealed by Figure 2b. On the other hand, sample-to-sample variation for MAS is less than 10% for worst cases, as depicted in Figure 2c, with only 2-3% changes for sample sizes as small as three to four times the number of polynomial coefficients q.

The order of the metamodel polynomial is changed to five, and the corresponding variations of RMSE, R-square and MAS are calculated. Results for DMCMRs are shown in Figure 3 superimposed for the three metrics. The same scales as in Figure 2 are used for easy comparison. As seen in Figure 3, although sample-to-sample variations has improved for RMSE, however, DMCMR for RMSE is still the worst, settling down to small percentages after nearly $\omega > 15$; i.e., for sample sizes with $\omega q > 15 \times 6 = 90$ observations.

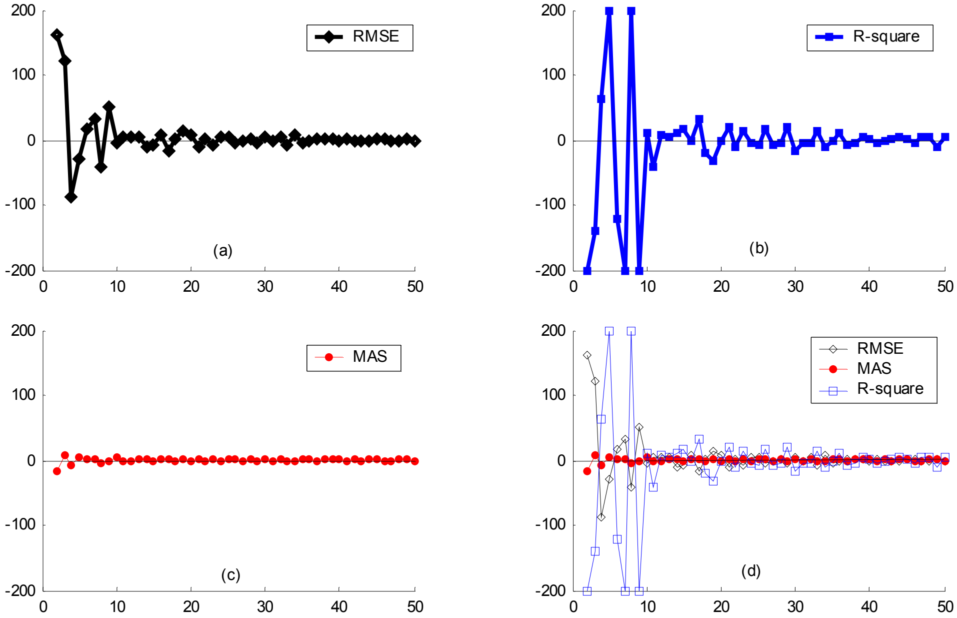

Figure 2: Sample-to-Sample Variations Expressed by DMCMR for the Second-Order Polynomial Metamodel vs. the Number of Coefficients Multiplier $\omega$ for (a) RMSE, (b) R-square, and (c) MAS; (d) Superimposition of the Three Plots

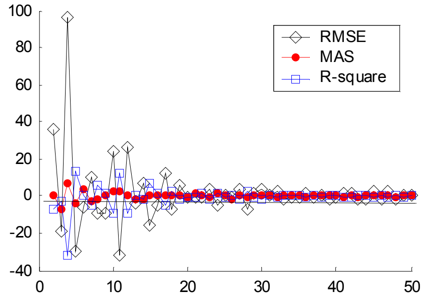

Figure 3: DMCMRs for the Fifth-Order Polynomial Metamodel vs. $\omega$

### 3.2 Example 2

The three-dimensional problem in this subsection is an engineering problem that relates the portion H of the input signal that appears as an output in the circuit of Figure 4.

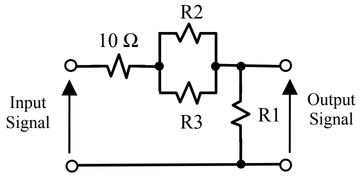

Figure 4: Electric Circuit for Example 2

The portion H is dependant upon the three design variables $R_1$, $R_2$, and $R_3$ connected as shown in the figure. Using a circuit simulator gives results which are identical to those given by the following equation obtained from elementary circuit analysis techniques:

$$H = \frac{R_1R_2 + R_1R_3}{R_1R_2 + R_1R_3 + R_2R_3 + 10R_2 + 10R_3} \tag{6}$$

A second-order polynomial is constructed from a minimum bias experimental design having seventeen points in the space $[1,100]^3$. Accuracy tests are then carried out using fifty samples. The number of observations for these samples are $\omega q$ ; $\omega = 1, 2, \ldots, 50$. The number of coefficients q for this case is ten. Calculations of RMSE, R-square, and MAS are carried out for the metamodel using the fifty test samples in turn. Acceptability criterion is set again at $(0.8 \le \hat{y}_i/y_i \le 1.2)$ for MAS calculations. Results are shown for RMSE and MAS in Figure 5 for one trial for each of the fifty Latin hypercube test samples used; R-square results are omitted to save space.

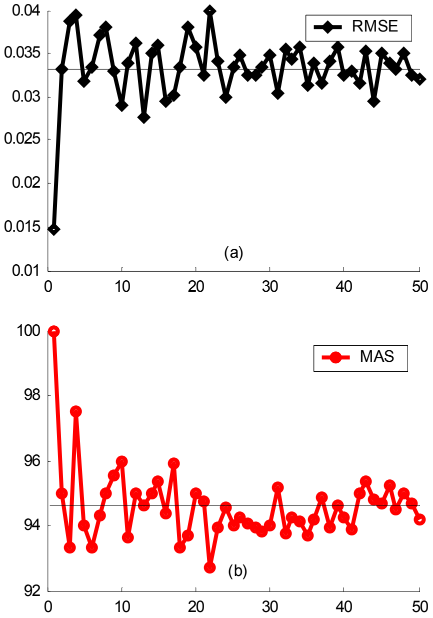

Figure 5: (a) RMSE vs. $\omega$ (b) MAS vs. $\omega$

Sample-to-sample variations are determined for the three metrics via DMCMR calculations. Results are shown in Figure 6a for the three metrics superimposed for comparison, while Figure 6b shows DMCMR results for MAS.

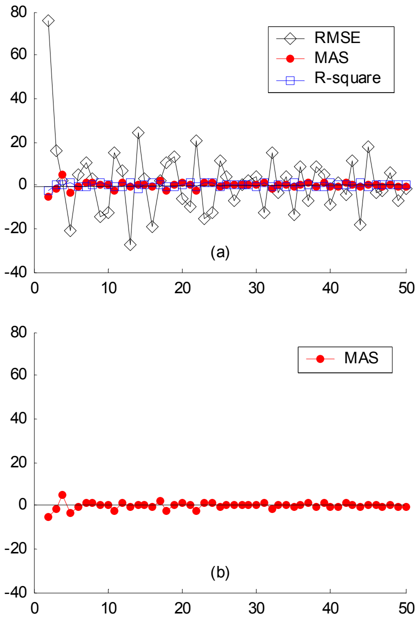

Figure 6: DMCMR (a) for Superimposition for the Three Metrics (b) for MAS Only

Figure 5a shows that RMSE varies from a little over 0.025 to 0.04 for sample sizes up to about 20; is this metamodel accurate based on these RMSE results? Note that the answer to this question is readily obtained by reference to MAS results in Figure 5b. The figure shows that the metamodel is accepted for $(95\pm3)\%$ of the observations for any test sample size greater than $\omega=1$ times the number of coefficients q; i.e., for sample sizes greater than ten in this case. Sample-to-sample variations for MAS are depicted by Figure 6a, showing that DMCMR for any change in sample size is less than 5%. Note from Figure 6b that sample-to-sample variation is worse for RMSE, with the R-square measure performing nearly as good as MAS.

## 4 CONCLUSION

This paper presented a new metric as a measure of metamodel acceptability with regard to prediction accuracy. The term used for this metric is 'metamodel acceptability score' or MAS. The discussion of this new metric is presented in the context of the two well-known average-based statistics of RMSE and R-square. MAS differs from such existing statistics in nature; a MAS level for a test sample is established by counting points in the sample which satisfy a required metamodel performance rather than taking average performance across the sample points. This aspect of MAS makes it less sensitive to sample size variations. Furthermore, MAS results are extremely easy to interpret, leading to a decision for accepting or rejecting a metamodel based on solid grounds.

## REFERENCES

Balci, O. 1989. How to assess the acceptability and credibility of simulation results. In *Proc. 1989 Winter Simulation Conf.,* ed. E. A. MacNair, K. J. Musselman, and P. Heidelberger, 62-71.

Hamad, H, and S. Al-Hamdan. 2005. Two new subjective validation methods using data displays. In *Proc. 2005 Winter Simulation Conf.,* ed. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 2542-2545.

Hamad, H., and S. Al-Hamdan. 2006. Discovering metamodels' quality-of-fit via graphical techniques, *European Journal of Operational Research* (In Press). [online]. Available via <www.sciencedirect.com> (doi: 10.1016/j.ejor.2006.01.026) [accessed April, 2006].

Kleijnen, J. P. C., and D. Deflandre. 2006. Validation of regression metamodels in simulation: bootstrap approach, *European Journal of Operational Research,* 170(1), pp. 120-131.

Sargent, R. G. 2004. Validation and verification of simulation models. In *Proc. 2004 Winter Simulation Conf.,* ed. R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 13-24.

## AUTHOR BIOGRAPHY

**HUSAM HAMAD** is an assistant professor in the Electronic Engineering Department at Yarmouk University in Jordan. He received his B.S. in Electrical Engineering from Oklahoma State University in 1984, M.S. in Device Electronics from Louisiana State University in 1985, and PhD in Electronic Systems Engineering from the University of Essex, England, in 1995. He was a member of PHI KAPPA PHI Honor Society during his study in the U.S. His research interests include analog integrated circuits analysis and design, electronic design automation, CAD, signal processing, and metamodel fitting and validation techniques. His e-mail address is <husam@yu.edu.jo>.