Statistically Based Parametric Yield Prediction for Integrated Circuits

David S. Gibson, Ravi Poddar, Gary S. May, *Senior Member, IEEE,* and Martin A. Brooke, *Member, IEEE*

***Abstract*—This paper presents a novel procedure for predicting integrated circuit parametric performance and yield when provided with sample transistor test results and a circuit schematic. Two enhancements to the existing Monte Carlo simulation procedures are described: 1) a multivariate nested model is used to reproduce random process-induced device variations, rather than the multivariate multinormal model typically used, and 2) the stochastic Monte Carlo method for mapping process variability into a performance distribution is replaced with a deterministic mapping technique. The use of multivariate nested distributions allows estimation not only of correlation between various model parameters, but also allows each of those variations to be apportioned among the various stages of the process (i.e., wafer to wafer, lot to lot, etc.). This allows matched devices to be more accurately simulated, without having to develop customized models for each configuration of matching, and provides focus for process improvement efforts into those areas with the maximum potential reward. The use of deterministic mapping provides simulation results which are repeatable and do not rely on chance to insure that the process parameter space has been evenly explored. A software package which implements the entire procedure has been written in C++.**

***Index Terms*—Monte Carlo simulation, multivariate statistics, parametric yield.**

I. Introduction

THE parametric performance of integrated circuits depends on both the circuit design and the fabrication process used to build the design. The ability to predict this performance is essential to those attempting to design integrated circuits, modify fabrication processes, plan production schedules, or specify product operating characteristics. Typically, this prediction is accomplished using a three-step Monte Carlo method [1]: 1) a statistical model is built to characterize the fabrication process to be used; 2) a circuit design is created using a circuit simulator (such as SPICE) and nominal device values for the target process; and 3) randomly generated instances of the process model are simulated in a "Monte Carlo" fashion to produce a representative set of output performance characteristics. The impact of random process variations can be inferred from these simulations, and yield can then be estimated using the percentage of that sample which meets the performance requirements.

Several variations on this technique have been presented in an effort to either improve accuracy or reduce computational complexity. These efforts have occurred primarily in three areas: selection of the variables in which the process model is to be built, development of methods for predicting the performance of a particular circuit instance, and determination of techniques for guiding the selection of Monte Carlo instances to improve the accuracy of yield estimates for a fixed sample size.

Input variables are often examined using analysis of variance (ANOVA) techniques [2] to eliminate statistically insignificant parameters [3], [4]. Uncorrelated multinormal distributions are often used, with many authors choosing the same four supposedly independent parameters for MOSFET devices: length, width, gate oxide thickness, and flatband voltage [5]–[10]. Others have preferred to use manufacturing parameters such as diffusion times and temperatures [3], [11]. Often, correlation is accounted for through the use of principal component analysis [12], [13].

Given a generated circuit instance, many techniques have been used to predict the resulting performance parameters. The most common of these, and the standard by which the accuracy of most techniques is measured, is the use of a traditional circuit simulator such as SPICE. Since pure Monte Carlo precision can be directly tied to the number of simulations performed, it is desirable to minimize the calculation time for each simulation. The data resulting from a handful of well chosen SPICE simulations can be used to develop a polynomial or neural network model [14], [15], which can then be used in place of SPICE for future, faster calculations. Polynomial circuit models have been successfully generated using linear [7], piecewise linear [16], quadratic [17], [18], and cubic [19] models. Transformations are often made to improve the accuracy of these models. Analytic expressions can also be used [20], as can hierarchical model generation [4], [21], [22], [23]. Simplicial approximation attempts to reduce the need for simulation by identifying the acceptance region in input parameter space [24].

Several techniques are available for improving the accuracy of a Monte Carlo yield estimate without increasing sample size. Each is dependent on a controlled alteration of the random generation of the input instances. Importance sampling [25], [26] forces more samples in regions where yield is close to 50%. Parametric sampling [27] allows samples from

Manuscript received October 28, 1996; revised February 4, 1997. This work was supported by the National Science Foundation, Grant DDM-9358163, and Analog Devices, Inc.

D. S. Gibson was with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332–0250 USA. He is now with the Rose-Hulman Institute of Technology, Terre Haute, IN 47803 USA (e-mail: David.Gibson@Rose-Hulman.edu).

R. Poddar, G. S. May and M. A. Brooke are with School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332–0250 USA (e-mail: poddar@azalea.mirc.gatech.edu; gsm@kuumba.mirc.gatech.edu; martin.brooke@ece.gatech.edu).

Publisher Item Identifier S 0894-6507(97)07555-6.

TABLE I
EXAMPLE OF A NESTED DISTRIBUTION

Device	Group Level	Device Level	Parameter	Mean Value	Group Value	Device Value
First matched device	1	1	t_{ox}	τ_A	μ_{A_1}	$\varepsilon_{A_{11}}$
			V_t	τ_B	μ_{B_1}	$\varepsilon_{B_{11}}$
Second matched device	1	2	t_{ox}	τ_A	μ_{A_1}	$\varepsilon_{A_{12}}$
			V_t	τ_B	μ_{B_1}	$\varepsilon_{B_{12}}$
First unmatched device	2	1	t_{ox}	τ_A	μ_{A_2}	$\varepsilon_{A_{21}}$
			V_t	τ_B	μ_{B_2}	$\varepsilon_{B_{21}}$

previous circuit designs to be added to samples for more recent versions. Stratified sampling [26] divides the process region into a number of nonoverlapping regions, and then calculates the yield in each region. The control variate technique [26], [28] can be used to overcome inaccuracies in the mapping function (typically employed only when non-SPICE maps are used). Acceptance sampling [29] performs a similar function, but includes elements of stratified sampling as well. Latin hypercube sampling insures that a wide range of values are used for each variable in a multivariate sample [30].

Given this history, there is room for improvement in at least three areas: the simulation of analog circuits which use matched devices, extracting information from the process model to maximize the effect of process improvement efforts, and converting the probability density function (pdf) for the process model to a pdf for circuit output characteristics.

Designers often rely on device matching to improve the performance of circuits which incorporate differential amplifiers (op amps), DAC's, ADC's, and other analog components. If two matched devices are simulated using identical device values, then unrealistically good performance is predicted. Two approaches have attempted to rectify this problem. The first simulates within-die variations by adding a random component to each device [31]. This falls short in that it attempts to use the same variability for all devices in a design, regardless of the degree of matching they exhibit. The second method develops an empirical model for each specific matched device configuration, and uses that model to generate specific parameter information for each matched device [32], [33]. This technique requires explicit model development for each new matching configuration (for example, if three matching devices were used rather than two), and assumes that the only within-die variations which matter are those on matched devices (other devices are modeled with no variation).

Another shortcoming of current techniques is that they do not facilitate process improvement. Manufacturing yield is maximized when a process is centered (i.e., mean value is in the middle of the desired range) and the random component of process variation is minimized. Since process variation comes from many sources, each of which contributes a different amount to the total variation, it is important to attempt reductions of those sources which have the most potential for reducing the variation in output parameters. For example, variation from run to run might not affect the variation in output parameters nearly so much as variation within a single circuit, especially when matched devices are used. In such a case, a dramatic reduction in run-to-run variance might not affect output performance as much as a more modest improvement in within-circuit variance. In the course of developing a process model, much of this information is available. However, models created for most simulation needs are not in a form which allows ready conclusions to be drawn about which process steps should be targeted for improvement. An improved technique would make this information directly available.

While Monte Carlo simulation has served well for a number of years, it has an inherent weakness in that it is a stochastic process. Also, the most straightforward method for improving accuracy is to increase the number of instances generated. This can result in prohibitively large needs for computing time, since each additional instance requires an additional simulation. Furthermore, some instances will be statistically "close" together, others will be "far" from their nearest neighbor. Consequently, it cannot be assured that all regions of the model parameter space will be explored equally. Finally, since Monte Carlo simulation is a stochastic process, it will yield slightly different results each time it is performed.

In this paper, two techniques are proposed to address the three issues identified above. The first models a process using a multivariate nested probability density function rather than the typical multinormal pdf. This results in a model which can more accurately reflect the impact of within-circuit variation, while providing insight into profitable courses of action for variability reduction. The second technique is a deterministic probability mapping method offered as an alternative to Monte Carlo analysis, and addresses those difficulties associated with its stochastic nature.

II. PROCESS MODEL DEVELOPMENT

The modeling of semiconductor manufacturing processes can be improved by using a multivariate, nested, normal distribution. Nested distributions are appropriate when the phenomena being modeled can be divided into "stages" such that the variability within one stage is independent of the variability within another stage [34]. In semiconductor manufacturing, the potential stages are processes, lots, wafers, circuits, groups (of matched devices), and devices. These are generally mutually independent in their variability. For

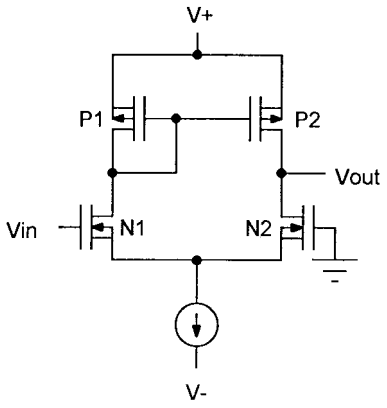

Fig. 1. Circuit used for simulations.

TABLE II
VARIANCES USED IN GENERATING FIG. 2 AND CALCULATED RESULTS (ALL VALUES $x10^{-13}$ m^2)

Variable	Target	Multinormal	Nested
L1	1.00	0.91	1.04
W1	1.00	0.98	0.99
L2	1.00	1.04	0.99
W2	1.00	1.08	0.97
L1 vs. W1	-0.50	-0.46	-0.47
L1 vs. L2	0.50	-0.05	0.57

example, large variations in a device parameter from one lot to the next is no indication that large variations will occur in that same device parameter on two adjacent devices in the same circuit. Lot variations might be due to the concentration of a particular batch of solution used in a wet processing step, while device level variations might be more dependent on how evenly the solution is applied. This independence of cause (and consequential independence of variance) makes semiconductor manufacturing an excellent candidate for nested modeling.

In addition to being nested, this distribution must also be multivariate. That is, the modeling of the various different parameters from one device are modeled using a multinormal distribution, including the relevant variances and correlations. The nested aspect of the distribution will reproduce similarities in different devices from the same stage. The multivariate aspect is needed to insure that when more than one parameter from a single device is modeled, the appropriate correlations are considered. For example, there is a well-documented correlation between MOSFET threshold voltage and gate oxide thickness. Using a multivariate model assures that these two parameters will have the appropriate correlation. Thus, a multivariate nested distribution not only reproduces the correlation of different parameters within a device, but also reproduces the correlation resulting from within-stage similarities.

A. Description of Nested Modeling

In a nested distribution, each sample can be represented as a global mean plus a variance contribution from each stage. Consider two correlated variables A and B with two stages of nesting. A sample from this design can be referenced using (A_{ij}, B_{ij}) where i is the level of the first stage (the highest stage—stage 1) and j is the level of the second stage for that particular sample. No two samples can have the same levels for both stages. A_{ij} can be represented as

$$A_{ij} = \tau_A + \mu_{A_i} + \varepsilon_{A_{ij}} \tag{1}$$

where τ_A is the grand average for A (the same for all samples), μ_{A_i} is the stage 1 component for A (the same for all samples with level i of the first stage), and $\varepsilon_{A_{ij}}$ is the stage 2 component of A (unique for each level of the second stage within the first). B_{ij} can be represented similarly, with all A subscripts replaced with a similar B subscript.

Defining a nested probability distribution requires a knowledge of the global means for each parameter, and a covariance matrix for each stage of the model to characterize the distributions of μ_i and ε_{ij}. Let the covariance matrices be given by $\Sigma_1 = \begin{vmatrix} \sigma^2_{A_1} & \sigma_{AB_1} \\ \sigma_{AB_1} & \sigma^2_{B_1} \end{vmatrix}$, and $\Sigma_2 = \begin{vmatrix} \sigma^2_{A_2} & \sigma_{AB_2} \\ \sigma_{AB_2} & \sigma^2_{B_2} \end{vmatrix}$, where Σ_1 contains the stage 1 covariances (the distributions of μ_{A_i} and μ_{B_i}) and Σ_2 contains the stage 2 covariances (the distributions of $\varepsilon_{A_{ij}}$ and $\varepsilon_{B_{ij}}$). The nested model for these two variables has been defined when all eight of these variables have been identified (two means, four variances, and two covariances). This distribution considers correlation not only among the various parameters for a single device (as modeled by the covariances in the Σ matrices), but also considers correlation among samples of the same parameter from different devices in the same stage (as modeled by the common variance components of common higher stages).

B. Multivariate Nested Modeling Example

Consider a practical application of a nested model in simulating a simple circuit with three similar MOSFET devices, two of which have been carefully matched. Assume each transistor has two model parameters which vary statistically: oxide thickness (t_{ox}) and threshold voltage (V_t). If t_{ox} is considered variable A and V_t is variable B, then the parameter values for these three devices can be expressed as in Table I by summing the mean, group, and device values.

The parameters of the nested pdf determine the distributions of the values in Table I. The two means, τ_A and τ_B, are constants, and consequently have no distribution. The group values μ_{A_1} and μ_{B_1} are distributed multinormally with a 0 mean vector and a covariance matrix of Σ_1, as are μ_{A_2} and μ_{B_2}. Each of the three device value pairs is also distributed multinormally with a 0 mean vector and a covariance matrix of Σ_2. The first two devices will probably be closer matched than the last device since they contain the same values for τ and μ, while the third device only shares the same value of τ. Also, the correlation between t_{ox} and V_t is preserved at each stage via the Σ covariance matrices.

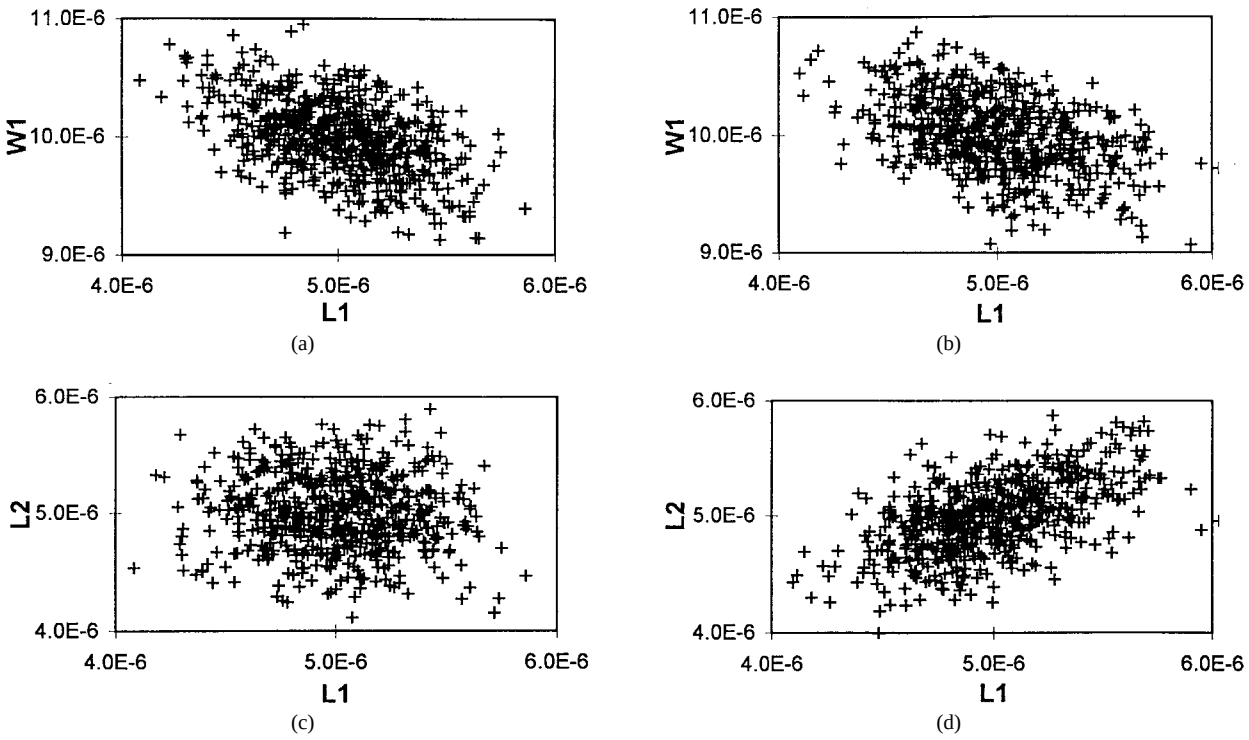

Fig. 2. Input generation differences between multinormal and nested distributions. The four random variables, $L1$, $W1$, $L2$, and $W2$, are projected into two dimensions to facilitate visualization.

C. Multivariate Nested Distributions as Process Improvement Tools

One additional benefit of using a nested distribution to model a process is that sources of variability are more easily identified. In the above example, if it was determined that the variability of t_{ox} was too high, then an examination of $\sigma^2_{A_1}$ and $\sigma^2_{A_2}$ would reveal how much of that variability was present at the group level and how much was at the device level. Larger variation at the group level might suggest that the variability cause is due to a phenomenon which varies across a circuit, such as etch uniformity. Larger variation at the device level might suggest that the variability cause is due to more fundamental limits in the ability to repeatably build devices. These two conditions suggest different plans of attack for process improvement. Furthermore, simulation of the circuit using this distribution with reduced values of device and group level variation will reveal which has a greater effect on variations in output performance.

III. Comparison of Multinormal versus Nested Models

To demonstrate the impact which a process distribution model can have on performance prediction, Monte Carlo simulations were performed using both multinormal and nested distributions. All simulations were performed using the simple differential amplifier circuit shown in Fig. 1. This circuit was chosen for two reasons: it is a common subcircuit used in IC design, and its performance is strongly dependent on the matching of devices $N1$ and $N2$. In an effort to keep these examples simple, all parameters except the length (L) and width (W) of devices $N1$ and $N2$ were held constant for all simulations.

A. Using Nested Models to Reveal the Impact of Device Matching

The advantage gained by using nested distributions as opposed to multinormal distributions arises from differences in the input instances they generate. With a multivariate multinormal distribution, correlations among parameters on a single device can be reproduced, but there is no mechanism for preserving correlations between similar parameters on matched devices. Since designs which incorporate matched devices often have their performance driven largely by how closely those similar parameters match, multivariate multinormal distributions can result in misleading performance predictions.

To illustrate this phenomenon, the differential amplifier shown in Fig. 1 was modeled twice using a Monte Carlo analysis procedure. The first time, input instances were generated using a multinormal distribution, while a multivariate nested distribution was used for comparison. Table II shows the variances used to build each of the models, with similar parameters used for both. For the nested model, the total variance of each device parameter was split evenly between the two levels of variance in evidence here: circuit to circuit and device to device. Fig. 2 shows the inputs generated by performing these two simulations.

Fig. 2(a) and (c) each show the instances generated using the multinormal distribution, while Fig. 2(b) and (d) show the instances generated using the nested distribution. Fig. 2(a) and (b) show the two sets of input instances plotted versus $L1$ and $W1$. As expected, both models reproduce the targeted negative correlation between length and width, as reflected in the tendency of the instances to cluster along a diagonal line sloping from upper left to lower right. Plotting the same

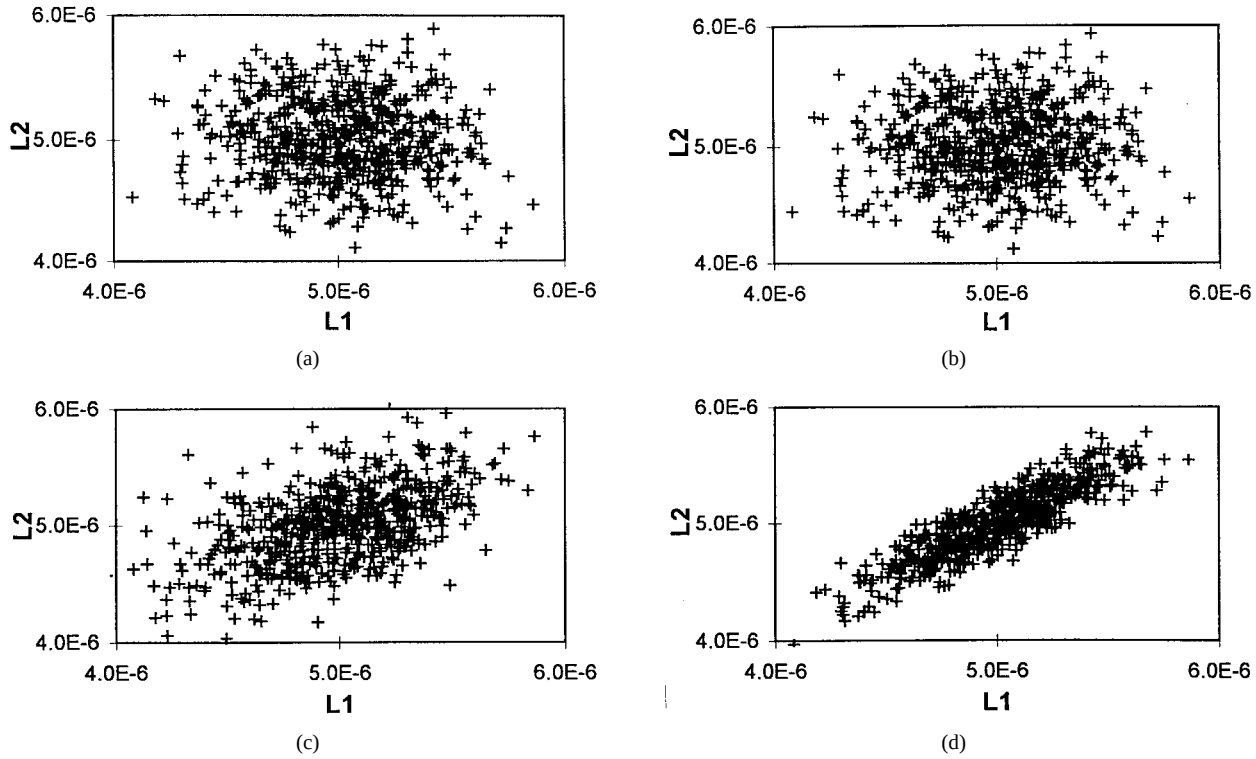

Fig. 3. L values of generated input instances showing varying degrees of variance distribution. Moving from (a) to (d) reflects an increasing portion of the distribution allocated to within-circuit, rather than between-circuit sources.

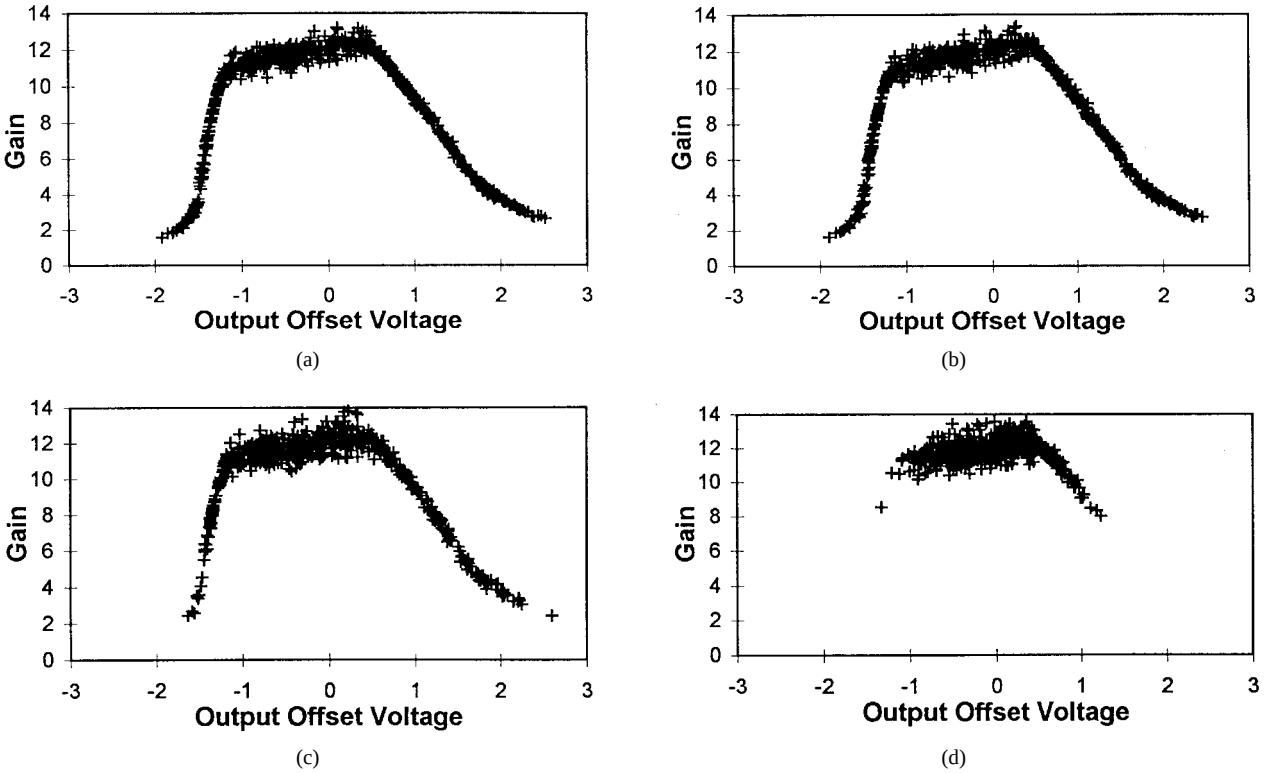

Fig. 4. Output performance for the circuit instances shown in Fig. 3.

data using $L1$ and $L2$ as axes produces Fig. 2(c) and (d). Notice that while no correlation is seen between these two parameters in Fig. 2(c) (as evidenced by the apparently random distribution of instances), Fig. 2(d) exhibits a clear positive correlation. This observation is verified by the statistical analysis of the generated instances shown in Table II. Note that all target parameters are reasonably reproduced in both models (within statistical margins) with the exception of $L1$ versus

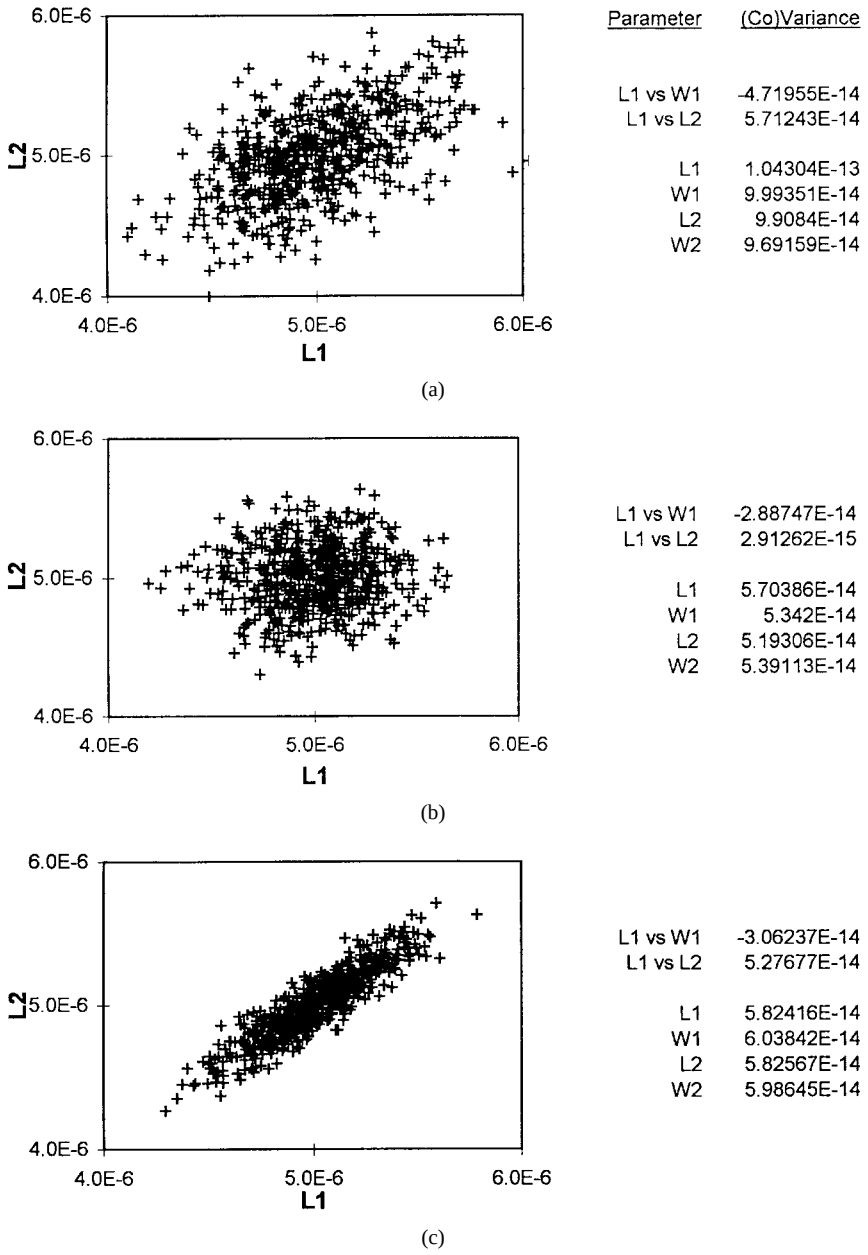

Fig. 5. The impact of stage choice on inputs for variance reduction efforts. (b) reflects a 90% reduction in the between-circuit variance from that shown in (a), while (c) reflects a 90% reduction in the between-device variance.

$L2$ correlation—the multinormal model cannot reproduce this characteristic. This correlation will be referred to hereafter as "device" correlation, since it measures the correlation between similar parameters in different devices, and plays a major factor in the performance of circuits with matched devices.

Fig. 3 shows the instances generated using four different distributions of variance between the circuit and device levels: (a) 100% device (traditional multinormal modeling), (b) 10% circuit, 90% device, (c) 50% circuit 50% device, and (d) 90% circuit, 10% device. As expected, the instances exhibit a stronger relationship between $L1$ and $L2$ as more variability is apportioned to the circuit level (leaving less variability between devices). Notice that the total variance for $L1$ and $L2$ has not been changed in any of these plots, only the manner in which it is distributed.

Fig. 4 shows the corresponding results of simulations on the instances depicted in Fig. 3, with the letters (a)–(d) taking on the same significance. Note that as more of the variance is apportioned to circuit level, the variation in the output performance is reduced (as evidenced by a tightening of the output distribution), even though the variance of the input parameters has remained the same. This is a direct consequence of two facts: 1) better matching of transistors $N1$ and $N2$ is a key factor in reducing output variation (more important than their absolute values) and 2) with constant total variance, more circuit level variation means less device level variation, leading to better matching of $N1$ and $N2$ in any particular circuit. Clearly, by assuming that all variation is device level (as is done with a multinormal model), gross mistakes can be made in estimating circuit performance—including the generation

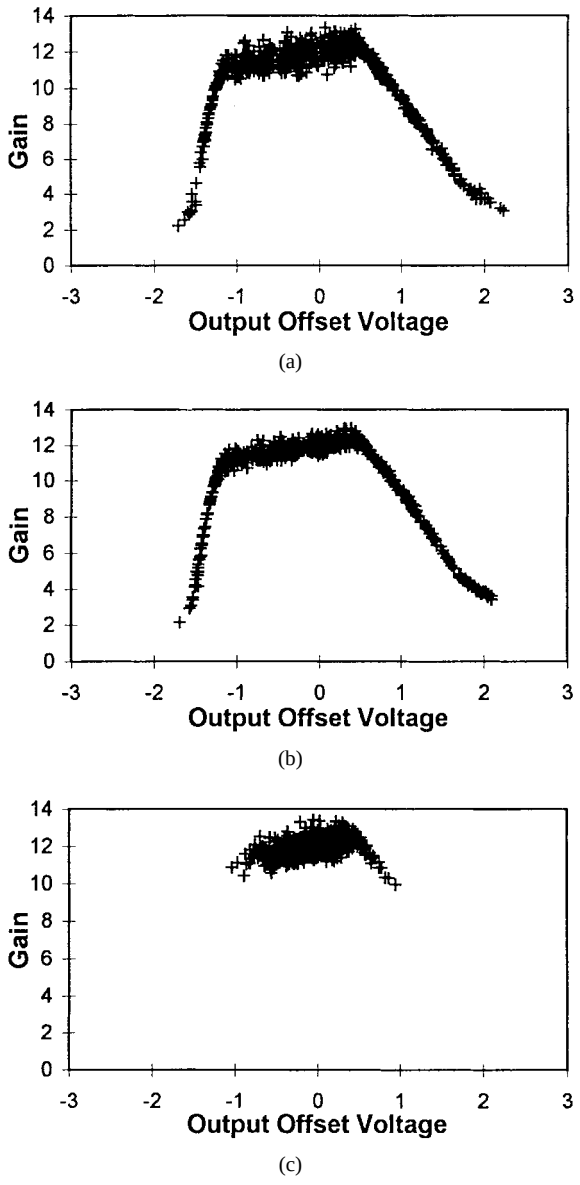

Fig. 6. Output performance for the circuit instances shown in Fig. 5.

of inaccurate yield estimates.

B. Using Nested Models to Target Process Improvement

It is not unusual for either designed or fabricated circuits to exhibit output performance with a greater variability than desired. Several approaches can be taken to reduce this variability, including process and circuit design modification. Two questions must be answered to optimize work in this area: 1) to what types of variability are the output parameters most sensitive; and 2) to what degree are each of those variability sources currently present. An answer to the first question is necessary to avoid working to reduce factors where, even if successful, that work will not significantly improve output performance. The second piece of information can be useful when several sources of variation are found to be significant—typically it is easiest to obtain incremental improvement in the parameter which varies the most.

Consider a hypothetical example using the circuit in Fig. 1, once again only considering the four parameters for $N1$ and $N2$ previously described. The first step in improving variability is to construct a process model. If the multinormal model from above were used, it would reveal a process in which all four parameter variances were essentially equal. The impact of improving device matching could not be explored on paper, since the multinormal model cannot reproduce device correlation. If a decision was made to reduce process variation for one of the four parameters, it would still be unclear how to proceed. For example, circuit to circuit variation might be improved by reducing the time between processing of successive wafers, while device to device variation might be improved by modifying control of gas flow over each individual wafer.

Suppose, on the other hand, that a nested model were used, and that it revealed a distribution similar to that shown in Fig. 3(c). In other words, all four parameter variances were equal, and those variances were split 50/50 between circuit level and device level. A typical set of instances generated from such a distribution is shown in Fig. 5(a), with the corresponding sample variances and covariances shown to the immediate right. Fig. 5(b) and (c) show the effects of reducing the circuit and device level components, respectively, by 90%. Once again, sample statistics for each of the plots are shown to the right of the figures. Note that Fig. 5(b) and (c) show a similar reduction in the four parameter variances, but in Fig. 5(b) this was accomplished at the expense of a dramatic reduction in covariance between $L1$ and $L2$.

The consequences of these two paths are shown in Fig. 6. The conditions in (b) do provide some reduction in variance, but (c) provides a clearly superior route. This suggests that, for this example, the best course of action is to pursue those options which would reduce device to device variance within a circuit, such as improving device matching through revised layout or reducing process variability across a circuit.

IV. Mapping Output Performance

Monte Carlo simulation is nothing more than a method of applying a deterministic mapping function to a joint pdf to generate a resultant joint pdf. When simulating circuits, the initial pdf is the distribution of device parameters, the mapping function is the SPICE simulation program, and the resultant pdf is the distribution of performance parameters. The resultant pdf can also be obtained using a technique known as Jacobian transformation [35].

Suppose A and B are correlated device parameters with a joint probability distribution pdf(A,B), and X and Y are two output parameters with a joint pdf of pdf(X,Y). To use the Jacobian technique, two inverse functions must be calculated such that for all instances of outputs, (x,y), the input instance, (a,b), which generated that output can be calculated as $a = g_A(x,y)$ and $b = g_B(x,y)$. This requires that each output map into one and only one input. The Jacobian matrix, $J(x,y)$, can then be calculated as follows:

$$J(x,y) = \begin{vmatrix} \frac{\partial a}{\partial x} & \frac{\partial a}{\partial y} \\ \frac{\partial b}{\partial x} & \frac{\partial b}{\partial y} \end{vmatrix}. \tag{2}$$

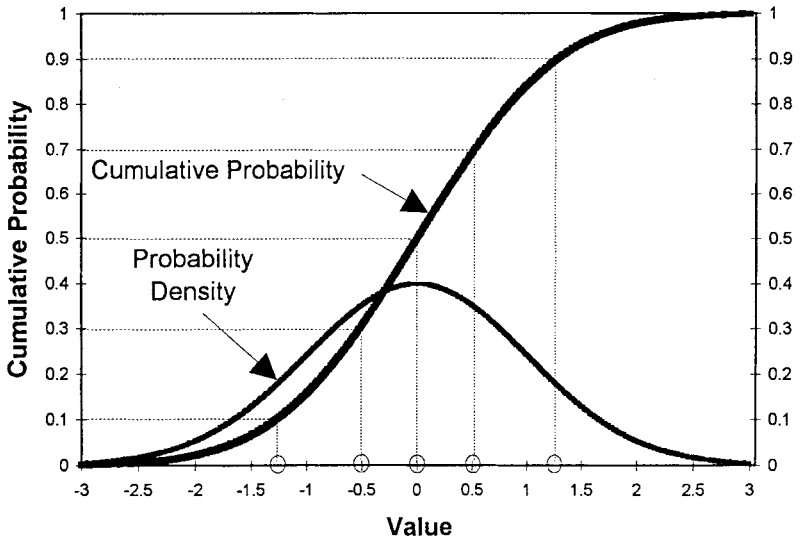

Fig. 7. Use of the pdf discretization procedure on the standard normal distribution. A discretization to five points yields the circled values on the ordinate.

The joint density of x and y is then given by

$$\mathrm{pdf}(x,y) = \mathrm{pdf}(g_A(x,y), g_B(x,y))J(x,y). \tag{3}$$

Unfortunately, Jacobian transformation requires knowledge of an inverse map and partial derivatives of the mapping function—neither of which is available with a SPICE mapping routine. These problems can be overcome by discretizing the input and output probability density functions using the novel methodology described below.

A. Use of Jacobian Concepts for Discrete Functions

Many of the restrictions imposed on using the Jacobian transformation for mapping continuous probability density functions disappear when evaluating discrete pdfs. Let X be a single realization of an input pdf. If there are n possible discrete values for X given by $I_1, I_2, \cdots, I_n$, with the probability for I_b given by $P(I_b)$, and m finite values for the output Y given by $O_1, O_2, \cdots, O_m$, then the probability for each $O_a, P(O_a)$, can be given as:

$$P(O_a) = \sum_b P(I_b)J_b, \quad J_b = \begin{cases} 1 & f(I_b) \in O_a \\ 0 & f(I_b) \notin O_a \end{cases} \tag{4}$$

where $f(X) = Y$ is a deterministic mapping function which maps each X into a single value of Y. The only requirements for this procedure are that: 1) the I_b are defined and their probabilities, $P(I_b)$, are known; 2) a mapping function, $f(I_b)$, is defined for each I_b; and 3) each I_b maps to exactly one O_a. While the J_b are not a Jacobian matrix, they serve a similar purpose in allowing the creation of a new distribution via deterministic mapping of an original distribution. Once the O_a and their probabilities have been determined, a discrete version of the output pdf can be constructed if a "volume" in the output parameter space, V_a, can be associated with each O_a. The resulting the pdf at O_a, pdf_a, is given by

$$\mathrm{pdf}_a = \frac{P(O_a)}{V_a}. \tag{5}$$

By selecting an appropriately fine resolution for the O_a, the output pdf can be reproduced to the desired precision.

B. Discretization of Multinormal Model Density Functions

To use this technique, a finite number of points in the input pdf space must be identified, with a probability assigned to each point. To best represent the original function, each point selected should represent a region of reasonably homogeneous behavior. Unfortunately, there is no *a priori* method for determining where the output function will change rapidly with respect to the process model variables. Consequently, the method used here is to distribute the sample points "evenly" throughout the input probability space. This results in a probability of $1/n$ associated with each point if n points are selected. This is true since "even" spacing implies that all points will represent the same probability volume, and since the sum of all probability densities must equal 1 for the entire volume to be mapped. By increasing n, the volume represented by each point can be reduced as much as desired until it is small enough that no significant variations occur within the region represented by a single point.

This concept is illustrated by the one-dimensional example in Fig. 7, in which a standard normal pdf is discretized into five points. The lower curve shows a standard normal probability density function. The upper curve is the standard normal cumulative probability function, obtained by integrating the pdf. The first step is to divide the distribution into five regions, each representing 1/5, or 20%, of the total distribution. For this example, these regions are 0 to 20%, 20% to 40%, 40% to 60%, 60% to 80%, and 80% to 100%. The percentages representing the centers of each of those regions are 10%, 30%, 50%, 70%, and 90%, respectively. The values associated with each of those percentages are −1.3, −0.5, 0.0, 0.5, and 1.3, respectively.

For multinormal distributions of n variables, the pdf will contain n dimensions. If each of these dimensions is made an independent normal distribution (for example,

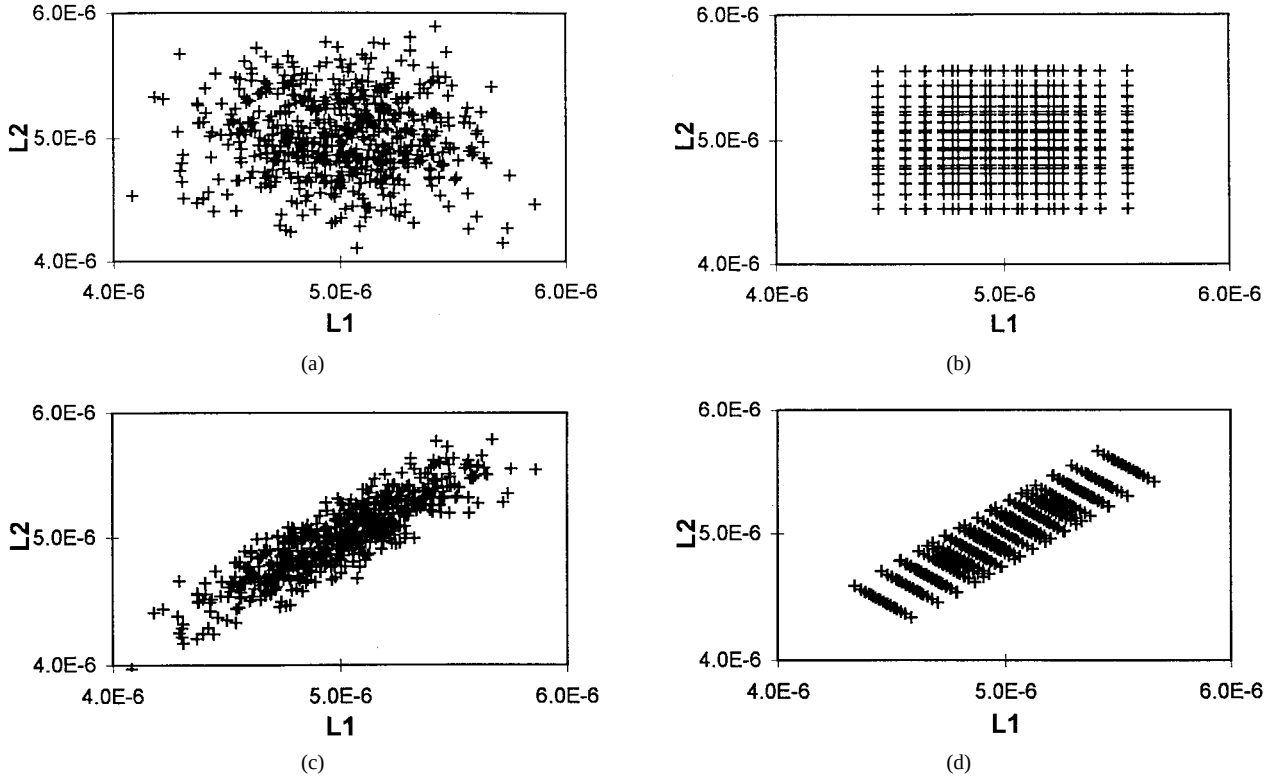

Fig. 8. Comparison of inputs generated using various combinations of distribution model and instance generation technique.

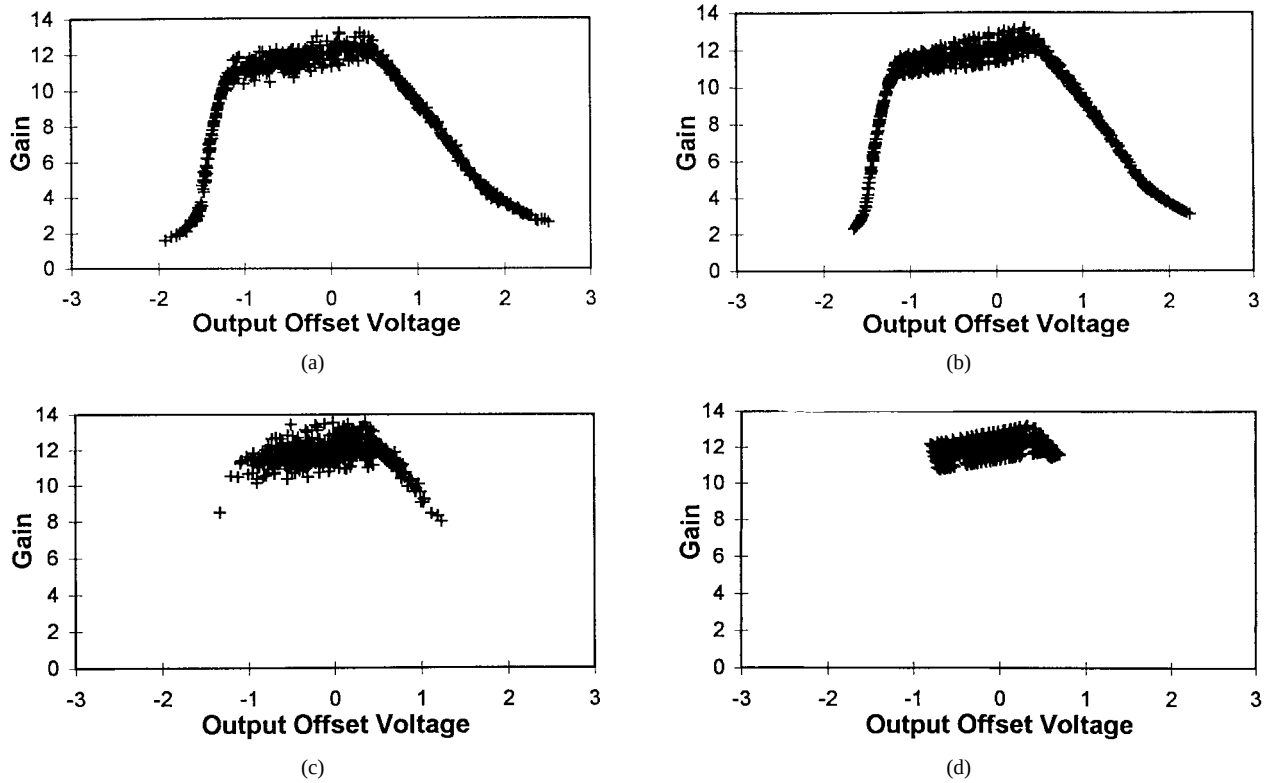

Fig. 9. Output performance for the circuit instances shown in Fig. 8.

through the use of principal component analysis [36]), then this procedure can simply be repeated once for each dimension. Creating a two-dimensional pdf using the distribution in Fig. 7 would result in the same five points for both axes, with the resulting 25 center points being $(-1.3,-1.3)$, $(-1.3,-0.5)$, $(-1.3,0.0)$, $(-1.3,0.5)$, $(-1.3,1.3)$, $(-0.5,-1.3),\cdots,(1.3,1.3)$, each of which has a probability of 1/25.

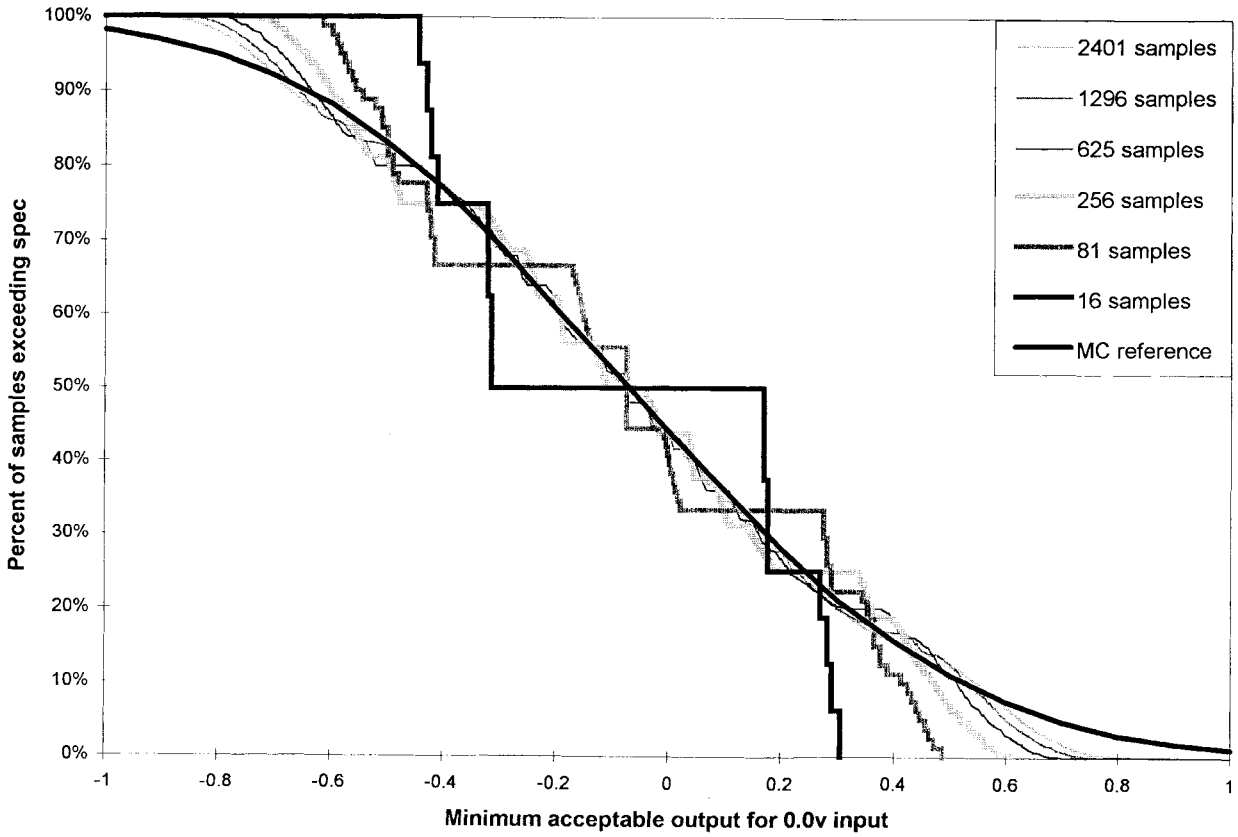

Fig. 10. The impact of grid resolution on yield prediction capability. The dark solid line represents a large sample Monte Carlo prediction (taken as the correct value), while other lines represent increasingly fine pdf grid discretizations.

C. Conversion of Multivariate Nested Distributions to Multinormal Distributions

To discretize a pdf as described above, a multinormal representation of the distribution must be available. Fortunately, any nested distribution with a particular sample plan can be converted to a multinormal distribution. Consider two correlated variables A and B with two stages of nesting. Let A_{ij} and B_{ij} be a pair of instances of A and B which have a stage 1 value of i and a stage 2 value of j. Also, define $\sigma^2_{A_{12}} = \sigma^2_{A_1} + \sigma^2_{A_2}$, $\sigma^2_{B_{12}} = \sigma^2_{B_1} + \sigma^2_{B_2}$, and $\sigma_{AB_{12}} = \sigma_{AB_1} + \sigma_{AB_2}$. For each circuit instance to be generated, many (A, B) pairs must be determined. For example, if A and B are transistor parameters, a pair of (A, B) values have to be generated for each transistor in the circuit design in order to create one instance of the circuit. If N transistors were present in the circuit design, then $2 * N$ random numbers would have to be generated, requiring the creation of a multinormal distribution in $2 * N$ variables.

Covariance values between each of the $2 * N$ variables depend on which stages have identical levels, so that the covariance between the A values for two different devices is given by

$$\mathrm{COV}(A_{ij}, A_{ab}) = \begin{cases} \sigma^2_{A_1} + \sigma^2_{A_2} & i = a \quad \text{and} \quad j = b \\ \sigma^2_{A_1} & i = a \quad \text{and} \quad j \neq b \\ 0 & i \neq a \end{cases} \tag{6}$$

the covariance between the B values for two different devices is given by

$$\mathrm{COV}(B_{ij}, B_{ab}) = \begin{cases} \sigma^2_{B_1} + \sigma^2_{B_2}, & i = a \quad \text{and} \quad j = b \\ \sigma^2_{B_1}, & i = a \quad \text{and} \quad j \neq b \\ 0, & i \neq a \end{cases} \tag{7}$$

and the covariance between any A and B values is given by

$$\mathrm{COV}(A_{ij}, B_{ab}) = \begin{cases} \sigma_{AB_1} + \sigma_{AB_2}, & i = a \quad \text{and} \quad j = b \\ \sigma_{AB_1}, & i = a \quad \text{and} \quad j \neq b \\ 0, & i \neq a. \end{cases} \tag{8}$$

As an example, consider the data used in preparing Table I. A six variable multinormal distribution can be constructed to represent the three devices defined, each of which is modeled with two parameters. The first two devices are matched, but the third is not. Allowing A to represent t_{ox} and B to represent V_t, the means vector for the new multinormal pdf is given by

$$\tau = |\tau_A \quad \tau_B \quad \tau_A \quad \tau_B \quad \tau_A \quad \tau_B| \tag{9}$$

and the covariance matrix is given by

$$\Sigma = \begin{vmatrix} \sigma^2_{AB_{12}} & \sigma_{AB_{12}} & \sigma^2_{A_1} & \sigma_{AB_1} & 0 & 0 \\ \sigma_{AB_{12}} & \sigma^2_{B_{12}} & \sigma_{AB_1} & \sigma^2_{B_1} & 0 & 0 \\ \sigma^2_{A_1} & \sigma_{AB_1} & \sigma^2_{A_{12}} & \sigma_{AB_{12}} & 0 & 0 \\ \sigma_{AB_1} & \sigma^2_{B_1} & \sigma_{AB_{12}} & \sigma^2_{B_{12}} & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{A_{12}} & \sigma_{AB_{12}} \\ 0 & 0 & 0 & 0 & \sigma_{AB_{12}} & \sigma^2_{B_{12}} \end{vmatrix}. \tag{10}$$

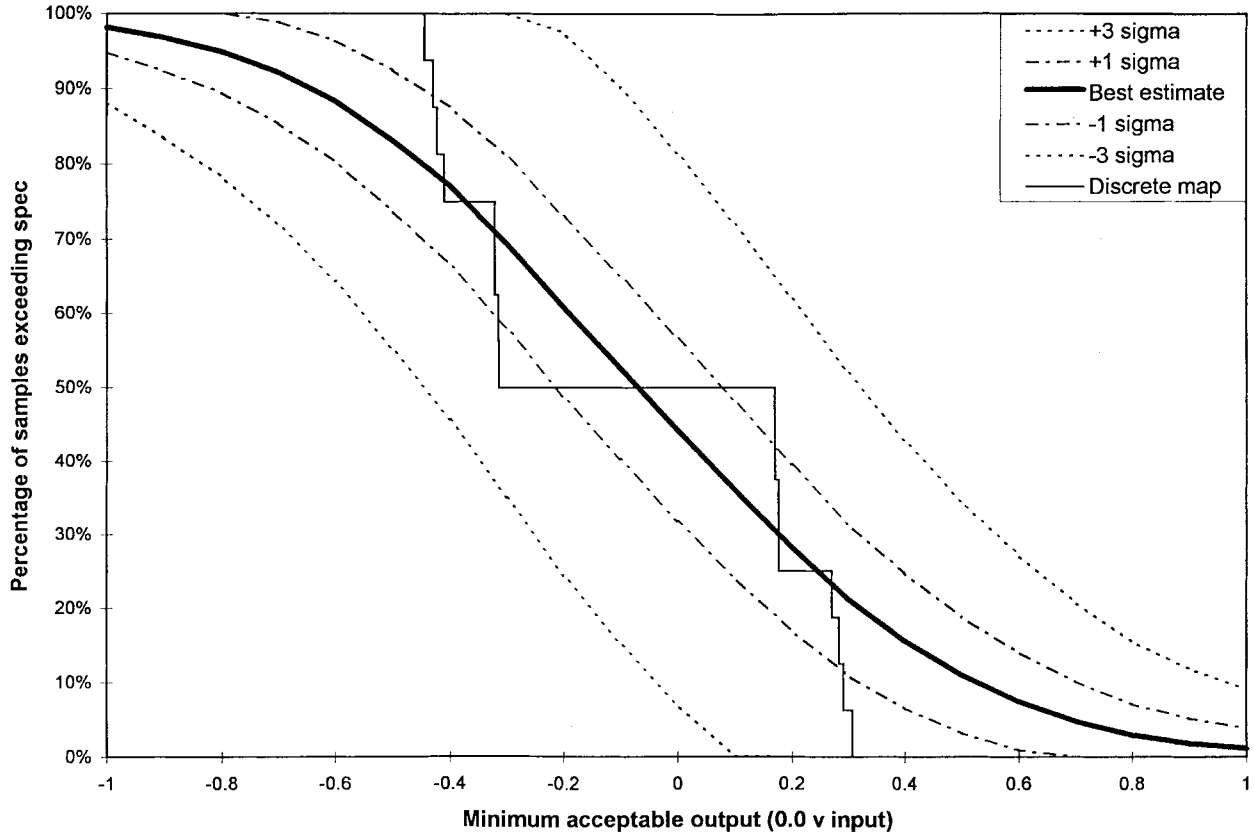

Fig. 11. Yield prediction capability comparison using two points in each dimension (16 samples). Upper and lower confidence intervals for the discrete prediction are shown in broken lines.

Note that each sample of A has the same mean value, τ_A, and the same variance, $\sigma^2_{A_{12}}$. As intuition dictates, only the covariances change from one sample of A (or B) to the next.

V. Comparison of Monte Carlo versus Discrete PDF Simulations

The circuit in Fig. 1 was used once again to demonstrate the characteristics of the discrete pdf mapping and how they differ from traditional Monte Carlo simulation. Unless otherwise noted, all discrete pdf simulations were performed using five divisions in each of the pdf's four dimensions for a total of 625 samples. To allow meaningful comparison, all Monte Carlo simulations were also performed using 625 samples. Furthermore, nested simulations were performed using the nested pdf model introduced with Fig. 2. Based on consistent results obtained with other simulations, the results documented below are believed to have general applicability to all circuit analysis performed using this technique.

A. Interaction between Multinormal and Nested Distributions, Monte Carlo and Discrete Simulations

The circuit in Fig. 1 was simulated using each of the four permutations of multinormal and nested distributions and Monte Carlo and discrete transforms. The impact of using each of these four combinations is most apparent in Fig. 8, where the input parameters used for each of the four conditions are displayed. Note that although there are four inputs, only $L1$ and $L2$ are plotted, so that two additional dimensions representing $W1$ and $W2$ have been projected onto the $L1/L2$ plane.

Fig. 8(a) shows a multinormal distribution with inputs generated for Monte Carlo analysis. Using the same multinormal distribution but generating inputs using the discrete technique yields the very regular input distribution in Fig. 8(b). Note that (b) would be even more obviously regular if it could be viewed in the four dimensions of the full input parameter space. Fig. 8(c) and (d) show inputs generated using a nested distribution. Again the inputs are regular in (d), but they are skewed to reflect the correlation which is now present between $L1$ and $L2$.

Fig. 9 shows the outputs generated from Fig. 8(a)–(d). Three significant features can be observed. First, as noted before, the nested distribution reflects the improved performance expected when matching effects are considered, as reflected by the tighter grouping in (c) and (d) than is found in (a) and (b). Second, the outputs from the discrete technique also display a degree of regularity [as seen in (b) and (d)], although not so pronounced as was evident in the inputs. Finally, the tails of the distributions are not as heavily populated in the outputs using the discrete transform, which is especially obvious in Fig. 9(d). These last two points will be further explored in the next example.

B. Impact of Sample Size on Yield Prediction Capability

Parametric performance for the circuit in Fig. 1 was estimated using a Monte Carlo technique with 10 000 samples

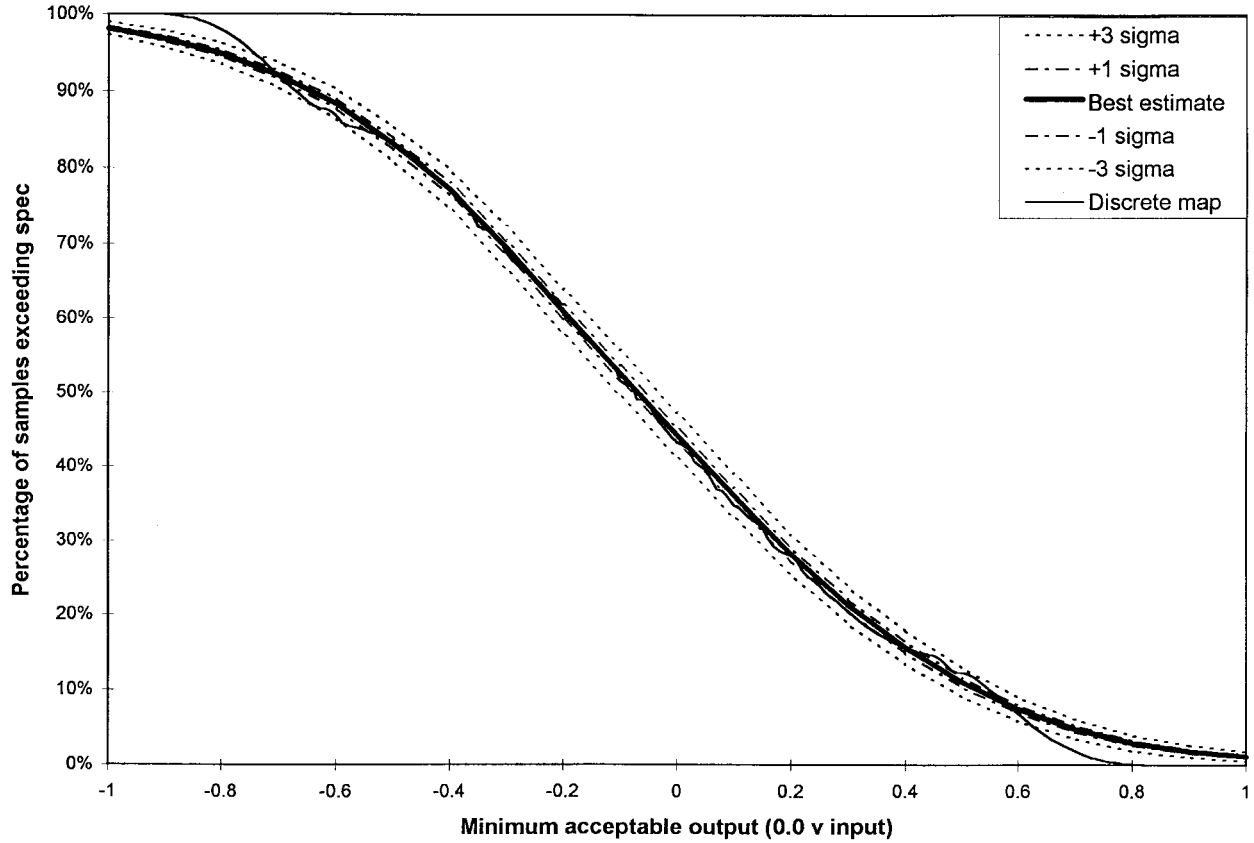

Fig. 12. Yield prediction capability comparison using seven points in each dimension (2401 samples). Upper and lower confidence intervals for the discrete prediction are shown in broken lines.

and a nested input distribution. An estimate for yield, Y^*, could then be calculated by computing the percentage of circuit instances passing a given output criterion (since each instance has identical probability of occurring). The yield has a standard deviation, σ_{Y^*} given by

$$\sigma_{Y^*} = \sqrt{\frac{Y^*(1-Y^*)}{10\,000}}. \tag{11}$$

This results in a prediction of yield which has a worst case standard deviation of only 0.5%. Consequently, the 10 000-sample Monte Carlo prediction is used as a reference to evaluate the discrete transform method. Fig. 10 shows this reference yield (labeled "MC reference") as a function of a single output parameter—output voltage for a zero volt input. Instances displaying an output voltage above the minimum value are categorized as passing for the purpose of calculating yield. As the minimum acceptable output voltage is increased, the percentage of passing instances is reduced until the yield falls to zero at a sufficiently high output requirement. (Although this is not typically the way in which an output offset voltage would be specified, it serves well the purpose at hand.)

Also shown in Fig. 10 are the yields predicted using the discrete transform with various sample sizes. As with the Monte Carlo technique, yield is estimated by calculating the percentage of passing instances. The different sample sizes are the result of varying degrees of coarseness in the input pdf grid. With only two divisions in each of four dimensions, 2^4 points are generated, while seven divisions in each dimension leads to 7^4 (or 2401) points. Three observations can be made from this plot. First, it is apparent that increasing the number of samples improves accuracy, since the curves for successively larger samples sizes come closer and closer to the reference line. Second, errors in the center of the plot (near 50% yield) seem centered about the reference, while errors near the extremes ($>$95% and $<$10%) tend to consistently underestimate the population of the distribution tails. Third, changes in the yield estimated seem to occur in dramatic steps for smaller sample sizes. The first observation is intuitively obvious, but the remaining points bear further study.

To understand why the discrete technique seems to underestimate the population of the distribution tails, consider Fig. 7. Note that the lower 20% of the distribution will always be represented by the 10% point, so the lowest value selected will always be 1.3. If a Monte Carlo simulation were performed using the same distribution and 5 points, then there would be 41% chance (given by $1–0.9^5$) that at least one point would be selected with a value less than 1.3. If a second independent dimension is added to Fig. 7 so that 5^2 points are used, then the Monte Carlo probability of getting a value less than 1.3 increases to 93% ($1–0.9^{25}$), while the discrete technique will never generate a value less than 1.3. If performance varies significantly with nominal decreases in this value, then yield will be misrepresented. A similar argument can be made for the upper 20% of the distribution. Clearly, the discrete technique is most accurate when circuit performance is relatively constant within each individual cell of the input grid.

The dramatic changes in yield seen in Fig. 10 are explained with a similar simple analysis. Note that for the 16-sample case, there are four distinct points where the yield drops dramatically, the first of which occurs at a minimum acceptable output voltage of about -0.45, when the yield drops from 100% to 75%. This is a consequence of the regularity of the sampled points associated with the discrete transform. Since these points tend to be aligned in rows, any output specification which defines a constraint parallel to a row of points will see significant changes in yield as that specification passes the row.

A quantitative comparison of the accuracy of the discrete and Monte Carlo techniques is demonstrated in Figs. 11 and 12. Fig. 11 shows the discrete map, 16-sample yield prediction once again, but it is contrasted with the $\pm 1\sigma$ and 3σ confidence intervals for a 16-sample Monte Carlo prediction [where σ is calculated using (11)], while Fig. 12 shows the same for 2401 samples. Note that for both maps, the discrete prediction is rarely outside of the $\pm 1\sigma$ confidence interval, especially when estimating yields between 85% and 15%. Conversely, the Monte Carlo prediction can be expected to be outside this interval roughly 30% of the time.

This discussion leads to certain conclusions about the utility of this technique. It provides an excellent tool for obtaining quick estimates of yield during the early stages of design, when large numbers of simulations for each design iteration can be excessively time consuming. With small sample sizes, the Monte Carlo technique is risky, since an "unfortunate" sample could lead to erroneous conclusions about the yield space and send the designer in an incorrect direction. The discrete mapping technique, being deterministic, would be immune to such chance occurrences. Furthermore, the regularity of successive discrete output plots would allow a better qualitative understanding of the impact of design changes, and the inherent inaccuracy at extreme values of yield would not be a liability until final design modifications were being incorporated to "squeeze out" the last percentage points of yield.

This suggests a modeling procedure for development work which would incorporate simulations with a small number of points using the discrete technique until the design is nearly optimized. Then, if precise understanding of the expected yield is required, and if the design is expected to be high yielding, a thorough Monte Carlo simulation can be performed with numerous instances to determine exact yield. The only situation in which Monte Carlo analysis offers a clear advantage is in this latter situation: a high (or low) yielding design with need for a precise estimate of yield in a situation where running numerous circuit simulations is not a serious liability.

VI. Conclusion

Two techniques have been described which enhance the capability to predict circuit parametric performance. The use of multivariate nested distributions has been shown to offer distinct advantages over the multivariate normal distributions most commonly used, particularly when device matching has an impact on circuit performance. These multivariate nested distributions can be easily extracted from test structure electrical measurements, and provide improved modeling of output parameter distributions while offering guidance for process improvement efforts. Also, a novel technique has been presented which offers a deterministic method of predicting performance, rather than the traditional stochastic Monte Carlo procedure. In certain cases, this technique can reduce the number of simulations required for accurate yield prediction, and is guaranteed to provide repeatable results with uniform sampling of the process space.

Acknowledgment

The authors would like to thank the National Science Foundation and Analog Devices, Inc. for their support of this research.

References

[1] J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods*, London, U.K.: Methuen, 1964.
[2] D. C. Montgomery, *Design and Analysis of Experiments*. New York: Wiley, 1991.
[3] M. Rencher, "Analog statistical simulation," in *IEEE Custom Integrated Circuits Conf.*, 1991, pp. 29.2.1–29.2.4.
[4] C. M. Kurker, J. J. Paulos, R. S. Gyurcsik, and J. C. Lu, "Hierarchical yield estimation of large analog integrated circuits," *IEEE J. Solid-State Circuits*, vol. 28, pp. 203–209, Mar. 1993.
[5] P. Cox, P. Yang, S. S. Mahant-Shetti, and P. Chatterjee, "Statistical modeling for efficient parametric yield estimation of MOS VLSI circuits," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 391–398, Feb. 1985.
[6] J. P. Brockman and S. W. Director, "Predictive subset testing: Optimizing IC parametric performance testing for quality, cost, and yield," *IEEE Trans. Semiconduct. Manufact.*, vol. 2, pp. 104–113, Aug. 1989.
[7] T. K. Yu, S. M. Kang, L. N. Hajj, and T. N. Trick, "Statistical modeling of VLSI circuit performances," *IEEE Int. Conf. Computer-Aided Design*, 1986, pp. 224–227.
[8] D. E. Hocevar, P. F. Cox, and P. Yang, "Parametric yield optimization for MOS circuit blocks," *IEEE Trans. Computer-Aided Design*, vol. 7, no. 6, pp. 645–658, June 1988.
[9] P. Cox, P. Yang, S. S. Mahant-Shetti, and P. Chatterjee, "Statistical device characterization and parametric yield estimation," *Solid State Technol.*, pp. 154–161, Aug. 1985.
[10] P. Feldman and S. Director, "A macromodeling based approach for efficient IC yield optimization," *IEEE Symp. Circuits and Systems*, 1991, pp. 2260–2263.
[11] I. C. Kizilyalli, T. E. Ham, K. Singhal, J. W. Kearney, W. Lin, and M. J. Thoma, "Predictive worst case statistical modeling of 0.8-μm BICMOS bipolar transistors: A methodology based on process and mixed device/circuit level simulators," *IEEE Trans. Electrom Devices*, vol. 40, pp. 966–972, May 1993.
[12] C. K. Chow, "Projection of circuit performance distributions by multivariate statistics," *IEEE Trans. Semiconduct. Manufact.*, vol. 2, pp. 60–65, May 1989.
[13] E. D. Boskin, C. J. Spanos, and G. J. Korsh, "A method for modeling the manufacturability of IC designs," *IEEE Trans. Semiconduct. Manufact.*, vol. 7, pp. 298–305, Aug. 1994.
[14] Q. J. Zhang and M. Nakhla, "Signal integrity analysis and optimization of VLSI interconnects using neural network models," in *IEEE Symp. Circuits and Systems*, 1994, pp. 459–462.
[15] R. M. M. Chen and W. W. Chan, "An efficient tolerance design procedure for yield maximization using optimization techniques and neural network," in *Proc. Int. Symp. Circuits and Systems*, 1991, vol. 10, pp. 895–903.
[16] C. M. Berrah, "Parametric yield estimation for a MOSFET integrated circuit," *IEEE Symp. Circuits and Systems*, 1990, pp. 2260–2263.
[17] T. K. Yu, S. M. Kang, J. Sacks, and W. J. Welch, "Parametric yield optimization of CMOS analogue circuits by quadratic statistical circuit performance models," *Int. J. Circuit Theory Applicat.*, vol. 19, pp. 579–592, 1991.
[18] T. K. Yu, S. M. Kang, J. Sacks, and W. J. Welch, "An efficient method for parametric yield optimization of MOS integrated circuits," *IEEE Int. Conf. Computer-Aided Design*, 1989, pp. 190–193.
[19] D. S. Boning and P. K. Mozumder, "DOE/Opt: A system for design of experiments, response surface modeling, and optimization using process

and device simulation," *IEEE Trans. Semiconduct. Manufact.*, vol. 7, pp. 233–244, May 1994.
[20] G. G. E. Gielen, H. C. C. Walscharts, and W. M. C. Sansen, "ISAAC: A symbolic simulator for analog integrated circuits," *IEEE J. Solid-State Circuits*, vol. 24, pp. 1587–1597, Dec. 1989.
[21] T. Koskinen and P. Y. K. Cheung, "Statistical and behavioral modeling of analogue integrated circuits," *Proc. Inst. Elect. Eng., Part G: Circuits, Devices and Systems*, vol. 140, no. 3, pp. 171–176, June 1993.
[22] T. Koskinen and P. Y. K. Cheung, "Hierarchical tolerance using behavioral models," in *IEEE Custom Integrated Circuits Conf.*, 1992, pp. 3.4.1–3.4.4.
[23] T. Koskinen and P. Y. K. Cheung, "Modeling behavior and tolerances in analogue cells," in *IEEE Custom Integrated Circuits Conf.*, 1991, pp. 8.7.1–8.7.4.
[24] S. W. Director, G. D. Hatchel, and L. M. Vidigal, "Computationally efficient yield estimation procedures based on simplicial approximation," *IEEE Trans. Circuits Syst.*, vol. CAS-25, pp. 121–130, Mar. 1978.
[25] J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods*, London, U.K: Methuen, 1964.
[26] D. E. Hocevar, M. R. Lightner, and T. N. Trick, "A study of variance reduction techniques for estimating circuit yields," *IEEE Trans. Computer-Aided Design*, vol. CAD-2, no. 3, pp. 180–192, July 1983.
[27] K. Singhal and J. F. Pinel, "Statistical design centering and tolerance using parametric sampling," *IEEE Trans. Circuits Syst.*, vol. CAS-28, no. 7, pp. 692–701, July 1981.
[28] R. S. Soin and P. J. Rankin, "Efficient tolerance analysis using control variates," *Proc. Inst. Elect. Eng., Part G: Circuits, Devices and Systems*, vol. 132, no. 4, pp. 131–142, Aug. 1985.
[29] N. J. Elias, "Acceptance sampling: An efficient, accurate method for estimating and optimizing parametric yield," *IEEE J. Solid-State Circuits*, vol. 29, pp. 323–327, Mar. 1994.
[30] M. D. McKay, R. J. Beckman, and W. J. Conover, "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 21, no. 2, pp. 239–245, May, 1979.
[31] S. W. Pan and Y. H. Hu, "PYFS—A statistical optimization method for integrated circuit yield enhancement," *IEEE Trans. Computer-Aided Design*, vol. 12, pp. 296–309, Feb. 1993.
[32] T. Mukherjee and L. R. Carley, "Rapid yield estimation as a computer aid for analog circuit design," *IEEE J. Solid-State Circuits*, vol. 26, pp. 291–299, Mar. 1991.
[33] J. Oehm and K. Schumacher, "Quality assurance and upgrade of analog characteristics by fast mismatch analysis option in network analysis environment," *IEEE J.Solid-State Circuits*, vol. 28, pp. 865–871, July 1993.
[34] J. Neter, W. Wasserman, and M. H. Kutner, *Applied Linear Statistical Models*. Homewood IL: Irwin, 1990.
[35] W. W. Hines and D. C. Montgomery, *Probability and Statistics in Engineering and Management Science*. New York: Wiley, 1990.
[36] Manly, F. J. Bryan, *Multivariate Statistical Methods*. London, U.K.: Chapman & Hall, 1994, pp. 76–91.

David S. Gibson received the B.S.E.E. and M.S. degrees from Auburn University, Auburn, AL, in 1978 and 1983, respectively. In 1992, he entered the Georgia Institute of Technology, Atlanta, pursuing of the Ph.D. degree in electrical engineering.

From 1983 to 1992, he was employed by Harris Semiconductor, Melbourne, FL. From 1983 to 1990, he performed a variety of tasks in failure analysis and failure mechanism modeling/prediction, including studies of electromigration, dielectric breakdown, and hot carrier injection. From 1990 to 1992, he was a Program Manager supervising the fabrication of semiconductors for major strategic weapons systems. While pursuing the Ph.D. degree, he has consulted in the field of semiconductor reliability and statistical analysis as Gibson Engineering. He is now with the faculty of the Rose-Hulman Institute of Technology, Terre Haute, IN.

Ravi Poddar received the B.S. degree in electrical engineering with Highest Honors in 1991 and the M.S. degree in 1995 from the Georgia Institute of Technology, Atlanta, and is currently enrolled in the Ph.D. program at the same university.

His research interests include statistcal modeling and yield prediction, computer-aided design and simulation of integrated circuits, transistor modeling, predictive modeling and characterization of integrated passive devices and interconnect, and high-speed automated testing.

Gary S. May (S'85–M'91–SM'97) received the B.S. degree in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1985 and the M.S. and Ph.D. degrees in electrical engineering and computer science from the University of California, Berkeley, in 1987 and 1991, respectively.

He is currently an Associate Professor in the School of Electrical and Computer Engineering and Microelectronics Research Center at Georgia Tech. He was previously a Member of the Technical Staff at AT&T Bell Laboratories, Murray Hill, NJ. His research is in the field of computer-aided manufacturing of integrated circuits, and his interests include semiconductor process and equipment modeling, process simulation and control, automated process and equipment diagnosis, and yield modeling.

Dr. May is a National Science Foundation "National Young Investigator," and is Editor of the IEEE TRANSACTIONS ON SEMICONDUCTOR MANUFACTURING. He was an NSF and an AT&T Bell Laboratories graduate fellow. He is Chairperson of the National Advisory Board of the National Society of Black Engineers (NSBE).

Martin A. Brooke (M'90) received the B.E. (elect.) degree (1st Class Honors) from Auckland University, New Zealand, in 1981, and the M.S. and Ph.D. in electrical engineering from the University of Southern California, Losa Angeles, in 1984 and 1988, respectively. His doctoral research focused on reconfigurable analog and digital integrated circuit design.

He is currently the Analog Device Career Development Professor of Electrical Engineering at the Georgia Institute of Technology, Atlanta, and is developing electronically adjustable parallel analog circuit building blocks that achieve high levels of performance and fault tolerance. His current research interests are in high-speed, high-precision signal processing. Current projects include development of adaptive neural network and analog multipliers and dividers, precision analog amplifiers, communications circuits, and sensor signal processing circuitry. To support this large analog and digital systems research, he actively pursues systems level circuit modeling research. He has developed software that reduce the complexity of large analog electronic system models to a user specified tolerance.

Prof. Brooke won the only 1990 NSF Research Initiation Award given in the analog signal processing area.