A COMPARISON OF STATISTICALLY-BASED AND NEURAL NETWORK MODELS OF PLASMA ETCH BEHAVIOR

Christopher D. Himmel, Byungwhan Kim, and Gary S. May
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250

Abstract

Previously, models for semiconductor manufacturing processes have been developed using first principles, statistical experimental design, and a combination of the two. Due to the complexity of modern fabrication processes, these approaches have met with varying degrees of success. Recently, a new adaptive learning approach involving neural networks has been applied to the modeling of polysilicon film growth by low-pressure chemical vapor deposition [1]. In this paper, a neural network modeling methodology is applied to the removal of polysilicon films by plasma etching. The plasma etch process under investigation was previously modeled using the empirical response surface approach. However, in comparing the two neural methods with the statistical technique, it has been determined that the neural network models exhibit superior accuracy and require fewer training experiments. In fact, the results of this study indicate that the predictive capabilities of the neural net models are superior to that of their statistical counterparts for the same experimental data.

1.0 Introduction

Accurate and efficient modeling of semiconductor fabrication processes is necessary for a variety of manufacturing applications. Among these are recipe generation and optimization [2], statistical process control [3], and equipment malfunction diagnosis [4]. In recent years, the technique of ion-assisted plasma etching has become widely used in semiconductor manufacturing. Due to this increased popularity, considerable effort has been expended in developing reliable methods of process characterization for this complex physical and chemical process. Characterization is aimed at modeling the response of process outputs to variations in input parameters with a very high degree of precision. Previously, such models have been developed using a variety of approaches, including first principles [5-6], statistical experimental design and response surface methodology [7-8], and a combination of the two [9].

Plasma modeling from a fundamental standpoint has had limited success. Current physically-based models attempt to derive self-consistent solutions to first-principle equations involving continuity, momentum balance, and energy balance inside a high frequency, high intensity electric field. This is accomplished by means of computationally expensive numerical simulation methods which typically produce distribution profiles of electrons and ions within the plasma sheath. Although simulation is useful for equipment design and optimization, it is subject to many simplifying assumptions, and the true connection between these microscopic models and macroscopic parameters (such as etch rate) has yet to be clearly distinguished.

Since the complexity of practical plasma processes at the equipment level is presently ahead of theoretical comprehension, other efforts have focused on empirical approaches to plasma modeling involving Response Surface Methods (RSM) [10]. Models derived via RSM accurately represent the behavior of a specific piece of equipment under a wide range of etch recipes, thus making them very useful for manufacturing purposes. More recently, however, an adaptive learning technique which utilizes neural networks and influence diagrams in conjunction with statistical methods has been applied to the modeling of polysilicon growth by low-pressure chemical vapor deposition (LPCVD) [1]. In that study, the neural process models were shown to exhibit less experimental error than their statistical counterparts, even when created from less experimental data.

In this paper, two different neural network modeling methodologies are applied to the removal of polysilicon films by plasma etching. In one case, a neural modeling algorithm is implemented via a C program, while in the second case, neural process models are developed using the commercially available *Brainmaker* software package [11]. A recent RSM study focused on modeling the etch characteristics of n$^+$-doped polysilicon in a $CCl_4/He/O_2$ plasma [7]. Etch rate, uniformity, and selectivity to both silicon dioxide and photoresist were modeled as a function of six input parameters: RF power, pressure, electrode spacing, and the three gas flows. Etching took place in a Lam Research Autoetch 490 single-wafer plasma system. In comparing the RSM models derived from these experiments the neural models, it is determined that the neural models exhibit superior accuracy to the RSM models and require fewer training experiments. In fact, the results of this study indicate that the predictive capabilities of the neural net models are superior to that of their statistical counterparts for the same experimental data.

2.0 Experimental Apparatus and Technique

In the experiment described in [7], etching was per-

0-7803-0680-5/92 $3.00 ©1992 IEEE

formed on a test structure designed to facilitate the simultaneous measurement of the etch rates of polysilicon, SiO_2, and photoresist. The patterns were fabricated on 4-in diameter silicon wafers. Approximately 1.2μm of phosphorus-doped polysilicon was deposited over 0.5μm of thermal SiO_2 by LPCVD. The thick layer of oxide was grown to prevent etching through the oxide by the less selective experimental recipes. Oxide was grown in a steam ambient at 1000°C. One micron of Kodak 820 photoresist was spun-on and baked for 60 seconds at 120°C. The etching apparatus consisted of a Lam Research Corporation Autoetch 490 single-wafer parallel-plate system operating at 13.56 MHz.

Film thickness measurements were performed on five points per wafer using a Nanometrics Nanospec AFT system and an Alphastep 200 Automatic Step Profiler. Vertical etch rates were calculated by dividing the difference between the pre- and post-etch thickness by the etch time. Expressions for the selectivity of etching poly with respect to oxide (S_{ox}) and with respect to resist (S_{ph}) and percent nonuniformity (U), respectively, are given below:

$$S_{ox} = \frac{R_p}{R_{ox}} \quad (1)$$

$$S_{ph} = \frac{R_p}{R_{ph}} \quad (2)$$

$$U = \frac{|R_{pc} - R_{pe}|}{R_{pc}} * 100 \quad (3)$$

where R_p is the mean vertical poly etch rate over the five points, R_{ox} is the mean oxide etch rate, R_{ph} is the mean resist etch rate, R_{pc} is the poly etch rate at the center of the wafer, and R_{pe} is the mean poly etch rate of the four points located about one inch from the edge.

3.0 Experimental Design

The experiment was divided into two phases: an initial *variable screening* [12] phase to determine the most significant parameters, followed by a second phase designed to obtain the statistical response models.

3.1 Phase 1 - Variable Screening

Table I: Range of Input Factors

Parameter	Range	Units
RF Power (Rf)	300 - 400	watts
Pressure (P)	200 - 300	mtorr
Electrode Gap (G)	1.2 - 1.8	cm
CCl_4 Flow	100 - 150	sccm
He Flow	50 - 200	sccm
O_2 Flow	10 - 20	sccm

The six input factors and their respective ranges of variation are shown in Table I. A full factorial experiment to determine all effects and interactions for six factors would require 2^6, or 64 experimental runs. To reduce the experimental budget, the effects of higher order interactions were neglected and a fractional factorial design requiring only 32 runs was performed.

Experimental runs were performed in two blocks of 16 trials each in such a way that no main effects or first order interactions were confounded. Three center points were added to the design to provide a check for model nonlinearity. The experimental sequence was randomized in order to avoid biases due to equipment aging during the experiment.

3.2 Phase 2 - RSM Modeling

Analysis of the first stage of the experiment revealed significant nonlinearity in all responses, indicating the necessity of quadratic models. None of the input factors were found to be statistically insignificant. Thus, none were omitted from the second phase. To derive the RSM models, a second experiment which employed a Central Composite Circumscribed (CCC) Box-Wilson design was performed. In this design, the 2-level factorial "box" was enhanced by further replicated experiments at the center as well as symmetrically located "star" points [12]. A complete CCC design for six factors requires a total of 91 runs. To reduce the size of the experiment and combine it with the results from the screening phase, a half replicate design was again employed. The entire second phase required 18 additional runs. A graphic description of central composite designs appears in Figure 1.

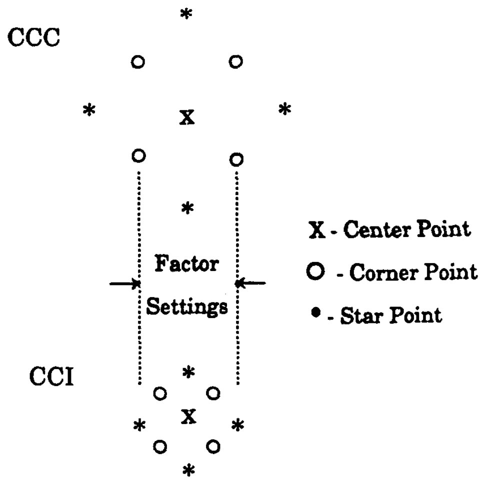

Figure 1 - Central composite experimental designs.

Phase 2 of the experiment yielded quadratic models which describe the interaction between input factors and the four responses. Each model was derived by stepwise regression [13]. *Analysis of variance* (ANOVA) techniques supported both the validity and quality of fit of each model [7]. The models appear below:

$$\begin{gathered} R_p = -245 - 4.24P + 11.0Rf + 0.742CCl_4 \qquad (4) \\ + 11.2He + 523G + 35.9O_2 - 0.034P{*}He + 7.82P{*}G \\ + 0.085Rf{*}CCl_4 - 8.36Rf{*}G - 0.132(CCl_4)^2 \\ + 0.059CCl_4{*}He - 0.059He^2 \quad \text{Å /min} \end{gathered}$$

$$\begin{gathered} U = -11.5 - 0.0385P + 0.0937Rf \qquad (5) \\ + 0.710CCl_4 - 0.415He - 8.90G \\ - (1.77e{-}3)Rf{*}CCl_4 + (1.38e{-}3)Rf{*}He \\ - (1.40e{-}3)CCl_4{*}He + (7.98e{-}4)He^2 \quad [\%] \end{gathered}$$

$$\begin{gathered} S_{ox} = -13.1 + 0.097P + 0.04Rf - 0.06CCl_4 \qquad (6) \\ - 0.059He + 0.079O_2 - (2e{-}4)P{*}Rf \\ + (2.9e{-}4)P{*}CCl_4 - (3e{-}4)P{*}He \end{gathered}$$

$$\begin{gathered} S_{ph} = 7.56 + 0.009P + 0.014Rf - 0.022CCl_4 \qquad (7) \\ + 0.006He - 2.59G - 0.099O_2 \\ - (5e{-}5)P{*}Rf + (1.3e{-}4)P{*}CCl_4 - (7e{-}5)P{*}He \\ + (3.7e{-}4)P{*}O_2 + (2.7e{-}5)Rf^2 + (3.6e{-}5)Rf{*}He \\ - (5e{-}5)CCl_4{*}He + 0.757G^2 \end{gathered}$$

4.0 Process Modeling With Neural Networks

Neural networks have previously demonstrated the capability of learning complex relationships between groups of related parameters [1,14-16]. Such learning abilities are attributed to the fact that neural networks, possessing many simple parallel processing units (called *neurons* or *neurodes*), crudely resemble the architecture of the human brain. These rudimentary processors are interconnected in such a way that knowledge is stored in the weight of the connections between them. Each neuron contains the weighted sum of its inputs filtered by a exponential "squashing" function, thereby endowing neural networks with the ability to generalize with an added degree of freedom not available in statistical regression techniques.

In order to model complex manufacturing processes such as the plasma etching example above, the quantitative relationships which relate the six input parameters are learned via the feed-forward error back-propagation (FFEBP) algorithm. This algorithm has been shown to be very effective in learning arbitrary nonlinear mappings between noisy sets of input and output vectors [16]. Networks consist of one or more layers of neurons which receive, process, and transmit critical information regarding the relationships between the input parameters and corresponding responses. FFEBP networks incorporate "hidden" layers of neurons which do not interact with the outside world, but perform classification and feature extraction tasks on information provided by the input and output layers.

4.1 Plasma Etch Modeling Strategy

The structure of a typical FFEBP network appears in Figure 2. In the plasma etch application, the input layer of neurons receives external information and corresponds to the six adjustable input parameters. The output layer transmits information to the outside world, and thus corresponds to the process responses (etch rate, uniformity, etc.). The number of hidden layer neurons is varied to achieve maximum performance. A neural network with a structure consisting of six input neurons, six hidden neurons and four output neurons is referred to as a "6-6-4" structure.

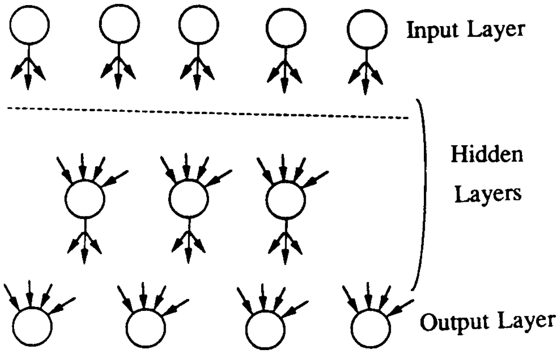

Figure 2 - FFEBP neural network showing input, hidden and output layers.

In the FFEBP modeling algorithm, the network begins with a random set of weights. An input vector is then presented to the network and the output is calculated using this initial weight matrix. Next, the calculated output is compared to the measured output data, and the squared difference between these two vectors determines the system error. The accumulated error for all the input-output pairs is the Euclidean distance in the weight space which the network attempts to minimize. Minimization is accomplished via the *gradient descent* approach, in which the system is adjusted in the direction of decreasing error [15]. Increasing the number of hidden neurons or hidden layers in the network helps this minimization process to converge. It has been shown that FFEBP networks can encode any arbitrary input-output relationship with a maximum of two hidden layers [15-16].

In this study, the above algorithm has been implemented via a C program as well as by using the commercially available *Brainmaker* software [11]. In each case, the network is trained on a subset of the available experimental data which spans the ranges of interest and subsequently tested on other remaining data. Training is assumed to be complete when all outputs are within a previously defined tolerance of the training data.

The C program employs a *madaline* architecture with a sigmoidal activation function and no input bias for forward propagation [17]. The program consists of three major portions: 1) the input module, which reads in the structure of the network and the data to be modeled; 2) a loop which performs the FFEBP algorithm until acceptable error is achieved; and 3) an evaluation module, which calculates the error of the network from the test data. *Brainmaker* also utilizes the FFEBP algorithm. One difference is that in *Brainmaker*, the error in the output of each neuron is allowed to vary within a user-defined range known as the *training tolerance.* The use of a training tolerance speeds network training, but causes a small trade-off in accuracy.

5.0 Results

Both the neural and RSM process models predict the measured responses of the plasma etch equipment with an experimental error σ, which is the square root of the residual sum-of-squares (S_R) for each response. S_R is calculated as follows [12]:

$$S_R = \frac{1}{(n-1)}\sum_{1}^{n}(y_i - \hat{y}_i)^2 \qquad (8)$$

where n is the number of experiments, y_i is the measured value of each response, and $\hat{y}_i$ is the corresponding model prediction. The experimental error for the RSM models (σ_{RSM}), also referred to as *replication error*, is shown in Table II below. This error is due to the inherent variability in the plasma etch equipment [7].

Table II: RSM Model Experimental Error

Model	σ_{RSM}
R_p	309.4 Å /min
U	6.66 [%]
S_{ox}	0.91
S_{ph}	0.26

5.1 Experimental Error (Full Experiment)

The RSM models described in §3.2 were constructed from a total of 53 experimental runs, 35 of which took place in the screening phase of the experiment. In order to compare the error of the neural etch models to their RSM counterparts, neural networks employing a 6-6-1 architecture were trained using the C program for all 53 original data. Table III compares the experimental error of these networks (σ_{NN}) to that of the RSM models (σ_{RSM} above). This comparison is further enhanced by the scatterplots in Figures 3-6, which depict the neural and RSM model predictions versus actual measurements. In these figures, perfect model predictions lie on the diagonal line, while scatter in the data is indicative of experimental error.

For each of the four etch responses, the neural networks exhibited significantly superior accuracy. This can be attributed to the fact that the RSM models were based on the somewhat questionable apriori assumption of quadratic behavior. That is, the RSM procedure began by assuming that quadratic functions adequately described the variation of the responses with the six inputs, and then fit coefficients to the four polynomial response equations (4)-(7) using regression techniques. However, the initial assumption of quadratic behavior is rather arbitrary, and may not be valid. The neural models, on the other hand, are not constrained to any particular form. In fact, due to their use of the previously mentioned exponential squashing function, neural networks are able to interpolate and generalize arbitrary nonlinear relationships.

Table III: Neural versus RSM Experimental Error

Model	σ_{NN}	% Improvement
R_p	98.1Å /min	68.3
U	4.1 [%]	38.2
S_{ox}	0.34	62.8
S_{ph}	0.08	67.8

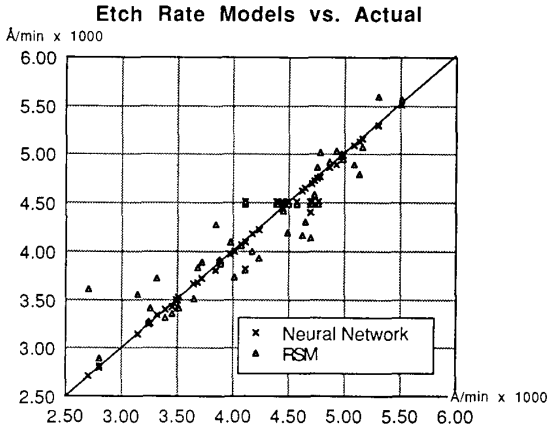

Figure 3 - Predicted vs. measured etch rate.

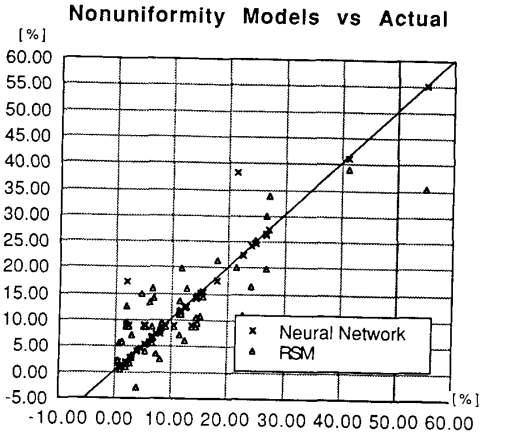

Figure 4 - Predicted vs. measured etch uniformity.

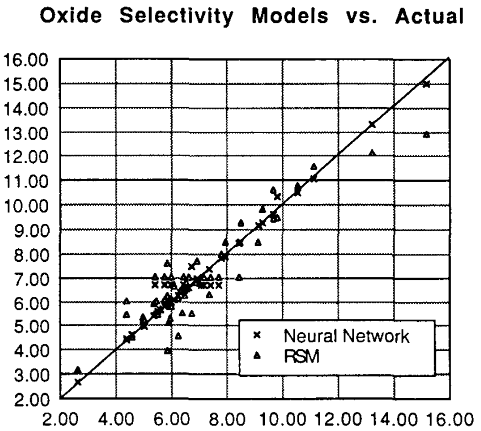

Figure 5 - Predicted vs. measured oxide selectivity.

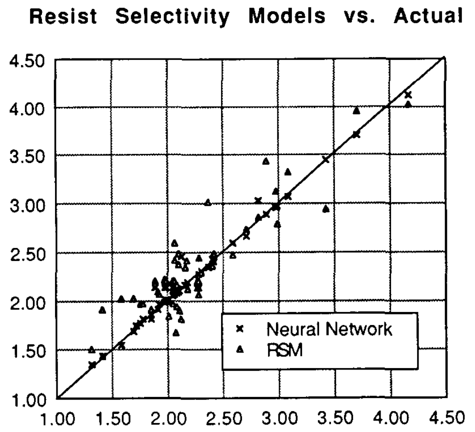

Figure 6 - Predicted vs. measured resist selectivity.

5.2 Predictive Capability (Reduced Experiment)

Due to the enhanced accuracy of the neural models, they can be used to predict etch behavior more precisely. It was also hypothesized that the neural models could exhibit this superior predictive capability with even fewer training experiments than the 53 used to derive the RSM models. In order to verify this hypothesis, the neural network model for the polysilicon etch rate was re-trained on the 35 runs corresponding to the initial screening phase of the RSM experiment only. Once trained, the network was subsequently used to predict the etch rate for 12 additional wafers which were not a part of the original experiment. Nine of these 12 wafers had been etched using the standard recipe of the Berkeley Microfabrication Laboratory, while the remaining wafers were etched using a recipe designed to optimize the four etch responses [7].

Table IV: Etch Rate Error of 12 "Additional" Wafers

σ_{RSM}	σ_{NN}	% Improvement
401.5 Å /min	247.9 Å /min	38.3%

As shown in Table IV, the experimental error of the new neural etch rate model for the 12 extra runs was 38.3% less than the RSM predictions. This result can be visualized in Figure 7. Furthermore, the 247.9 Å /min figure for the reduced neural model is also 19.9% lower than the 309.4 Å /min error derived from the *entire* original RSM experiment (refer to Table II). Thus, with 18 fewer experimental runs, a model of polysilicon etch rate with significantly less error has been obtained using neural networks. Therefore, the neural modeling approach reduces the time necessary to characterize the etch process as well as the alleviating the added cost of more experimental runs.

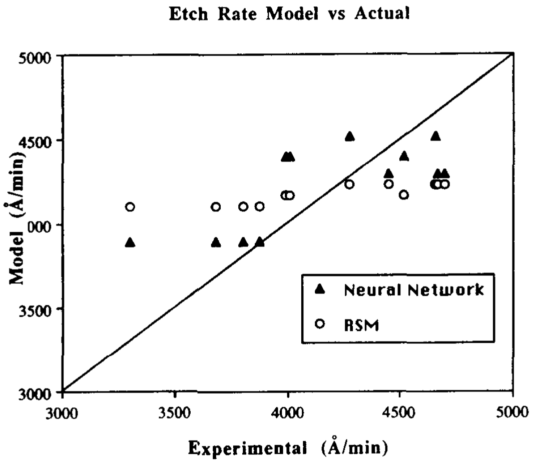

Figure 7 - Predicted vs. measured etch rate for "additional" wafers.

5.3 Brainmaker Results

The above simulations were repeated using the *Brainmaker* package. This software offers the advantage of much faster runtime than the C program. While the C program typically runs for several hours before convergence is achieved, *Brainmaker* training usually lasts on the order of minutes. As a result, many different neural architectures were tested with *Brainmaker* in an effort to determine the optimal architecture. This search yielded models exhibiting the best performance which possessed a variety of structures. The experimental error for these models (σ_B) is compared with the RSM error in Table V.

Table V: Brainmaker vs RSM Experimental Error

Model	Structure	σ_B	% Improvement
R_p	6-8-12-1	267.6 Å /min	13.5
U	5-10-20-1	4.63 [%]	30.5
S_{ox}	6-8-3-1	0.81	11.0
S_{ph}	6-8-12-1	0.18	30.8

For the reduced RSM experiment, the predictive capability of the *Brainmaker* model is not as impressive as the C program, but is still superior to RSM. When the *Brainmaker* etch rate model was re-trained on the 35 screening runs and tested on the same 12 additional wafers as in §5.2, the experimental error of the new *Brainmaker* etch rate model for those 12 runs was 329.2 Å /min, which represents a 18.0% improvement over the RSM result of 401.5 Å /min for these wafers (see Table VI). This implies that although *Brainmaker* training is much faster, better accuracy may be obtained from software customized to a particular application.

Table VI: Etch Rate Error of 12 "Additional" Wafers

σ_{RSM}	σ_B	% Improvement
401.5 Å /min	329.2 Å /min	18.0%

6.0 Conclusion

A neural network modeling technique has been applied to the highly complex process of plasma etching. For a polysilicon etch in a $CCl_4/He/O_2$ plasma, the etch rate, uniformity, and selectivity to both silicon dioxide and photoresist were modeled as a function of RF power, pressure, electrode spacing, and the three gas flows. Neural process models were subsequently compared to models derived by Response Surface Methodology (RSM) for the same data. It was demonstrated that the neural models possess significantly superior performance. Furthermore, the derivation of accurate neural models was shown to require fewer training experiments. As a result, neural network modeling promises to be a faster, more efficient and less expensive method of process characterization and the RSM approach in the future.

Acknowledgement

We are grateful to the staff of the UC-Berkeley Microfabrication Laboratory for aid in performing the original experimentation.

References

[1] F. Nadi, A. Agogino, and D. Hodges, "Use of Influence Diagrams and Neural Networks in Modeling Semiconductor Manufacturing Processes," *IEEE Trans. Semi. Manufac.*, vol. 4, no. 1, Feb., 1991.

[2] K. Lin, "Modeling and Characterization of Semiconductor Manufacturing Equipment: An Application to LPCVD Reactors," *UC-Berkeley ERL Memo. No. UCB/ERL M90/44*, May, 1990.

[3] H. Guo, C. Spanos, and A. Miller, "Real Time Statistical Process Control for Plasma Etching," *ISMSS*, May, 1991.

[4] G. May and C. Spanos, "Automated Malfunction Diagnosis of Semiconductor Fabrication Equipment: A Plasma Etch Application," submitted to *IEEE Trans. Semi. Manufac.*, June, 1991.

[5] A. Gerodolle and J. Pelletier, "Two-Dimensional Implications of a Purely Reactive Model for Plasma Etching," *IEEE Trans. Elec. Dev.*, vol. 38, no. 9, Sept., 1991.

[6] A. Paranjpe, J. McVittie, and S. A. Self, "Numerical Simulation of 13.56 MHz Symmetric Parallel Plate RF Glow Discharges in Argon," *Proc. 41st Gas. Elec. Conf.*, Oct., 1988.

[7] G. May, J. Huang, and C. Spanos, "Statistical Experimental Design in Plasma Etch Modeling," *IEEE Trans. Semi. Manufac.*, vol. 4, no. 2, May, 1991.

[8] M. Jenkins, M. Mocella, K. Allen, and H. Sawin, "The Modeling of Plasma Etching Processes Using Response Surface Methodology," *Sol. St. Tech.*, Apr., 1986.

[9] K. Lin and C. Spanos, "Statistical Equipment Modeling for VLSI Manufacturing: An Application for LPCVD," *IEEE Trans. Semi. Manufac.*, vol. 3, no. 4, Nov., 1990.

[10] G. Box and N. Draper, *Empirical Model-Building and Response Surfaces*, New York: Wiley, 1987.

[11] *Brainmaker Neural Network Simulation Software User's Guide and Reference Manual*, California Scientific Software, 1990.

[12] G. Box, W. Hunter, and J. Hunter, *Statistics for Experimenters*, New York: Wiley, 1978.

[13] R. Hocking, "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, vol. 32, March, 1976.

[14] S. Rangwala and D. Dornfeld, "Learning and Optimization of Machine Operations Using Computing Abilities of Neural Networks," *IEEE Trans. Sys. Man Cybern.*, Mar., 1989.

[15] R. Lippman, "An Introduction to Computing with Neural Nets," *IEEE ASSP Mag.*, Apr., 1987.

[16] F. Nadi, "Modeling Complex Manufacturing Processes Via Integration of Influence Diagrams and Neural Networks," *UC-Berkeley ERL Memo. No. UCB/ERL M89/123*, Nov., 1989.

[17] J. Freeman and D. Skapura, *Neural Networks*, New York: Addison-Wesley, 1991.