Variability Modeling and Statistical Parameter Extraction for CMOS Devices

by

Kun Qian

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Costas J. Spanos, Chair
Professor Chenming Hu
Professor Philip B. Stark

Spring 2015

Variability Modeling and Statistical Parameter Extraction for CMOS Devices

Abstract

Variability Modeling and Statistical Parameter Extraction for CMOS Devices

by

Kun Qian

Doctor of Philosophy in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Costas J. Spanos, Chair

Semiconductor technology has been scaling down at an exponential rate for many decades, yielding dramatic improvements in power, performance and cost, year after year. Today's advanced CMOS transistors have critical dimensions well below 24nm. This means that controlling the manufacturing process is increasingly difficult. Process and material fluctuations cause device and circuit characteristics to deviate from design goals, and introduce significant device-to-device variability due to spatial variations across silicon wafers. Accurate modeling of these spatial process variations has become critical to both foundries and circuit designers that seek optimal power/speed/area balance.

To understand the nature of spatial process variations, we first carried out a comprehensive variability analysis of data measured from thousands of variability-sensitized test structures, including ring oscillators, SRAM bit cells and their internal transistors. We manufactured these test chips using early stage 90nm and 45nm commercial semiconductor processes. We proposed a hierarchical variability model to capture the systematic and random components of device parameter variations across silicon wafers, and across chips. The detailed decomposition of the process variation profile reveals significant across-wafer systematic component for the delay and leakage of ring oscillators, and across-chip systematic component for the read/write margins of SRAM bit cells, as well as their internal transistors. The proper modeling of each hierarchical component proved to be crucial for the accurate estimation of the statistics of device performance distribution and its parametric yield.

The knowledge gained about process variation from carefully designed test structures was leveraged into estimating the variation and parametric yield of new devices and circuits. This was accomplished by improved the statistical compact model parameter extraction

methodology, and by proposing a stepwise parameter selection method. We used a normalized notional confidence interval and, and the sum of squares of fitting residuals as extraction and fitting quality criteria. This allowed us to determine the essential model parameters for accurate fitting over a large number of transistors. We applied this methodology to EKV and PSP with both simulated and experimental data, demonstrating its effectiveness. Finally, we combined the results from statistical parameter extraction with the hierarchical spatial variability model. This, compared to traditional methods, produced much-improved estimates of device performance and manufacturing yield.

To my mom

Contents

Acknowledgments

Chapter 1

Introduction

1.1 Motivation: Process Variations

For almost five decades, the semiconductor industry has, phenomenally, kept pace with Moore's Law [1]: Every 18 months, transistor density has doubled, as a result of reducing key device dimensions such as channel length and oxide thickness. However, decreasing dimensions further is increasingly difficult as CMOS technology scaling continues into sub-100nm feature size. Among the many emerging challenges, the increased importance and complexity of process variations is one of the most prominent.

Many variations during manufacturing process impact physical properties of devices and circuits. Lithographic variations [2], line-edge roughness [3][4], random dopant fluctuations [5], layout-dependent stress variations, rapid thermal annealing (RTA) temperature induced variations [6][7], well-proximity effects (WPE) [8], deposition and growth processes, and chemical mechanical polishing (CMP), all cause variations in device parameters such as dimensions, oxide thickness, doping concentrations, diffusion depth, and mobility.

The non-uniformity of transistor characteristics produces timing variations of circuit critical paths [9], smaller read/write noise margins for SRAM memory cells [10], and higher off-state leakage currents, which culminate in yield losses. In general, circuits need to be designed conservatively to cope with performance losses introduced by process variations, which requires devices to have larger area and higher power consumption.

It is critical to understand and quantify process-induced variability to avoid unnecessarily pessimistic designs. Improving the characterization and modeling of variability can help designers optimize performance, power, area, and yield.

1.2 Variability Models

Currently, foundries track on-wafer monitoring structures, including all sorts of active and passive devices, to estimate the performance distribution of devices and circuits. *I-V* data collected from test structures are later used to calibrate compact device models, such as BSIM [11] or PSP [12], and statistical models of device characteristics, which are used in circuit simulations.

Two types of statistical device models are conventionally used by modelers and designers to account for device parameter variations resulting from manufacturing process fluctuations: corner models and Monte Carlo models.

Corner models, often referred to as "worst-case design," seek to characterize worst-case and best-case device parameters. There are typically five worst-case corners, each identified by a two-letter acronym that indicates the relative performance of the n-channel and p-channel devices. Each letter summarizes the device performance of one channel type as typical (T), fast (F), or slow (S). The first letter indicates the performance of the n-channel device and the second letter indicates the p-channel device. Combinations of the performance levels for the n-channel and p-channel devices form the following list of corner cases:

- TT (typical n-channel, typical p-channel): the nominal or typical device performance the manufacturing process targets.
- FF (fast n-channel, fast p-channel): model parameters that reflect a process shift that yields fast operation for both the n- and p-channel devices.
- SS (slow n-channel, slow p-channel): model parameters that yield slow operation for both the n- and p-channel devices.
- FS (fast n-channel, slow p-channel): the n-channel device is fast, and the p-channel device is slow, which could represent the asymmetry of the rising and falling edge of signals in a critical path.
- SF (slow n-channel, fast p-channel): the n-channel device is slow and the p-channel device is fast.

Monte Carlo device models, on the other hand, attempt to represent the unpredictable characteristics of devices, rather than extreme behavior. Monte Carlo methods model device parameters as stochastic, typically assuming that each parameter is a realization of a Normal or uniform distribution. Monte Carlo inputs to SPICE simulations generally also assume that device model parameters are independent across instances of each transistor. Device and circuit performance distributions are derived from the assumed stochastic distributions of model parameters using Monte Carlo simulation. Often, Monte Carlo simulations are used to calibrate worst-case corner models for device performance parameters, such as *Ion* and *Ioff*.

Figure 1.1 depicts the typical relationship between CMOS transistor device parameter space, performance space, and the corresponding worst-case corners. Parameters include device characteristics such as the threshold voltage parameter *Vt* for NMOS and PMOS. Performance space refers to the distribution of device performance metrics, such as *Ion* of NMOS and PMOS. For any assumed distribution of device parameters, Monte Carlo device models can be created to simulate the performance of any circuit of interest; on the other

hand, device parameters are usually extracted from the device performance space, using techniques such as corner lots, statistical process control (SPC) [13], or process and device simulations. Worst-case corner models are commonly defined as the most "probable" combination of device parameters (in the parameter space) that would produce 3σ departures of combination of the performance parameters of the *n*-channel and *p*-channel devices (in the performance space) from their nominal values. For example, the FF corner is the combination of NMOS and PMOS when both *Ion,n* and *Ion,p* are at their high 3σ point, and the SS corner is when both are at their lower 3σ point. At the FS corner, the difference between *Ion,n* and *Ion,p* reaches its higher 3σ point, while at the SF corner it reaches its lower 3σ point. Modelers can then search the parameter space for the most probable (under the assumed model) combination of *Vth,n* and *Vth,p* that attains the worst-case corner values. In more complicated cases, the performance parameter can be a critical characteristic of a specific circuit, such as the delay of a ring oscillator, which is useful to characterize the worst-case operation of large-scale circuits.
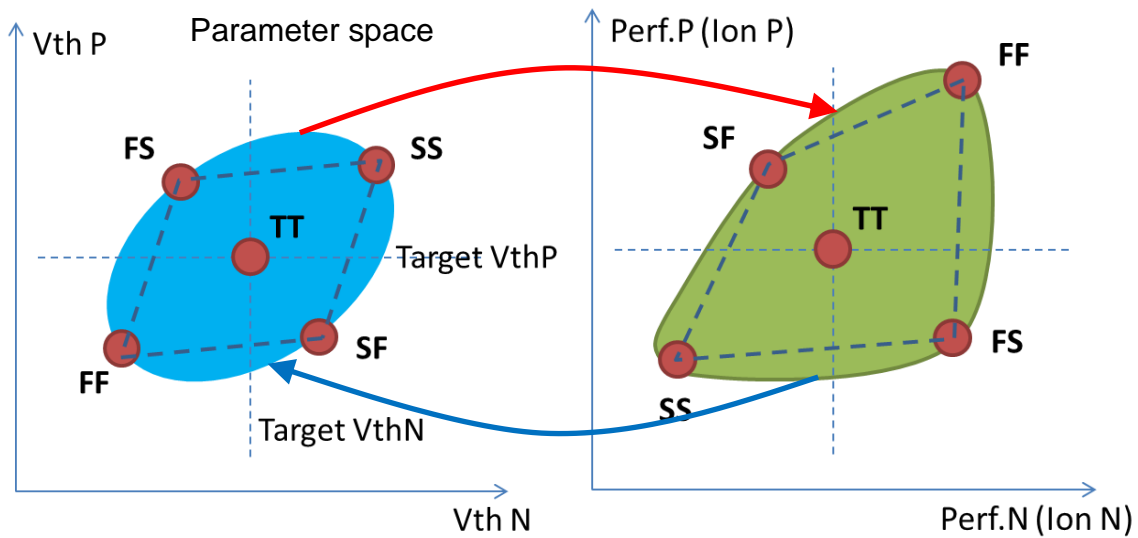
Figure 1.1: Typical CMOS transistor device parameter space and performance space with corresponding worst-case corners.

1.3 Research Goal

The goal of this research is to develop a method that can accurately model the stochastic transistor behavior induced by process variations, so that circuit designers can accurately estimate the parametric yield for a given design.

This goal can be achieved through the following steps:

❖ **Characterize and analyze the composition and structure of process-induced variation in CMOS devices and circuits.**

A key factor in calibrating statistical device models is to have an accurate representation of both the systematic (deterministic) and random (stochastic) variations. In the context of chip fabrication, the concept of systematic and random variations is inevitably entangled with the spatial hierarchy. Traditionally, lot-to-lot, wafer-to-wafer, and chip-to-chip variations are treated as a single pseudo-random component "global variation" from the perspective of individual chips. Global variation is supposed to be similar for all transistors on the same chip. Within-chip variations, on the other hand, are further decomposed into across-chip systematic and local random variations. These variations are modeled by independent Gaussian distributions. A number of previous studies have analyzed process-induced variability and its impact on circuit power performance. Asenov's team built an atomic-level simulation framework for predicting and modeling the intrinsic random variations of transistor parameter variations [14]–[19]. Boning and his students designed ring oscillator arrays for fast delay characterization and a within-chip spatial variability study [20], [21]. Wafer-level and die-level spatial variability of for inter layer dielectric (ILD) thickness variations were studied and modeled with ANOVA [22], [23].

In this thesis, electrical measurements are collected using arrays of standard-variability monitoring structures, such as transistors, ring oscillators, and SRAM bit cells. Several tens of chips are measured for each of the three test wafers fabricated in early commercial 90nm and 45nm low-power CMOS processes, providing full wafer-scale spatial coverage. A hierarchical model of variability is used to analyze the measured device characteristics, decomposing the total variability into random and systematic components at the wafer level and die level.

❖ **Develop a methodology that accurately and robustly translates the variability characteristics from the electrical measurements to the industry standard statistical device models.**

4

Circuit designers rely on statistical compact-device models to estimate performance variations of devices and circuits. The two most commonly used statistical models are the worst-case corner models and Monte Carlo models. Corner models use a finite set of compact model parameters to represent the typical and worst-case conditions of transistors, while the Monte Carlo model involves inventing a joint probability distribution for compact model parameters, then simulating realizations of those distributions. The key task of variability modeling in both cases is to accurately translate the measured *I-V* characteristics into the distribution of compact model parameters.

Traditionally, a few key compact model parameters with clear physical meanings are used to capture the variability in the *I-V* characteristics [24]–[26]. The extracted populations of these parameters are correlated, due to their physical relationship and due to the numerical procedure for estimating them. For this reason, some studies use principal component analysis (PCA) to extract statistically independent components of device variability, which are later used to simulate the compact model parameter variations [27], [28].

This dissertation shows that existing methodologies can be improved in two ways. First, the selection of the model parameters for direct extraction can be tailored more precisely to the silicon data. Only parameters found to be statistically significant will be retained in the fitted stochastic model of device variation; the remaining parameters will be fixed to their nominal values from the typical corner extraction. This is expected to reduce covariance among parameter estimates without relying on combinations of parameters (such as those PCA produces) that are incompatible with SPICE simulations. Second, we show that a full hierarchical model of spatial variability for the extracted compact model parameters allow a more faithful reproduction of device performance variations in Monte Carlo simulations.

❖ **Study the impact of spatial-process variations on the performance and parametric yield of real silicon devices using the improved statistical transistor modeling methodology.**
We use an improved statistical parameter-extraction procedure to identify parameters that can be extracted reliably, and analyze their spatial variability in detail. A statistical device model is extracted from the transistor *I-V* measurements of the SRAM bit cells from the 45nm test chips using this improved methodology. It illustrates good accuracy in predicting the read/write margins collected from the same set of SRAM cells. The accuracy of the spatially hierarchical statistical

transistor model is compared to the conventional method; and its advantage in yield estimation accuracy is evaluated.

1.4 Dissertation Outline

Chapter 2 reviews the sources of variability in the modern semiconductor fabrication process. Based on the stochastic or deterministic nature of the variations and their respective spatial scope of effect, a hierarchical model is proposed to describe the combined effect of the process variations.

Chapter 3 presents the measurement results from two sets of variability characterization test chips, the first from a commercial, general purpose 90nm CMOS process and the other from a commercial 45nm strained-Si CMOS process. Ring oscillator (RO) frequency and leakage data from both sets of test chips are evaluated. The hierarchical model of variability is proved to be very effective in fitting to the RO data. The same hierarchical variability analysis is applied to the read/write margin and the transistors *I-V* measurements collected from the SRAM bit cells from the 45nm test chips.

Chapter 4 details our improved method for modeling the spatial variability of transistors with compact models. A parameter extraction procedure is developed and tested with simulated data for two popular compact models: EKV [29] and PSP [12].

Chapter 5 applies the statistical compact-model extraction methodology to the actual silicon data collected from the SRAM bit cells from the 45nm test chip. For each of the EKV and PSP models, a set of model cards is extracted, to which the hierarchical model of variability is then applied to create a custom statistical compact model. These statistical compact models illustrate better accuracy in predicting device performance variations than conventional method.

Chapter 6 summarizes the highlights of this dissertation and discusses future research directions.

1.5 Statistical Notes

This dissertation uses a variety of statistical methods and concepts, including linear and nonlinear regression, hierarchical models, expectations, variances, hypothesis tests including t-tests, significance levels, p-values, and confidence intervals. However, the generative stochastic model for the data that would be required for those methods to apply

as statistical methods does not hold; moreover, if such a model did hold, in general, there would be more efficient methods than those employed here. Rather, all uses of statistical concepts and methods in this dissertation are to be considered *algorithmic*, rather than *statistical*. The justification for using the methods is not any underlying statistical theory, but instead the empirical performance of the resulting model for the task at hand: understanding process variability and predicting device performance and yield.

Chapter 2

Hierarchical Model for Spatial Variations

2.1 Classification of Process Variations

2.1.1 Environmental, Temporal, and Spatial Variations

Process variability can be environmental, temporal, or spatial [30]. Environmental variations consist of variability in the surrounding temperature, power supply voltage, and even cosmic radiation. Temporal variations, as the name suggests, refer to device-performance change over periods of time ranging from nanoseconds, for the SOI history effects [14] and self-heating effect, to seconds or hours, for the negative bias temperature instability (NBTI) [31], to years, for dielectric material deterioration after repeated programming and erasing operations in flash memories. Spatial variations,are performance differences among devices that depend on the distances between the devices or the locations of the devices on a chip. Typical spatial variations, such as line width or film thickness non-uniformity, universally exist across lots, across wafers, across chips and dies, and between circuit blocks and devices (Figure 2.1). As a result, the circuit performance of chips from wafers produced with the same design and process over a period of manufacturing time will never be the same.

Of the three types, environmental and temporal variability are often accounted for using reliability models, while spatial variability is commonly part of statistical device models. This dissertation studies the spatial variations.
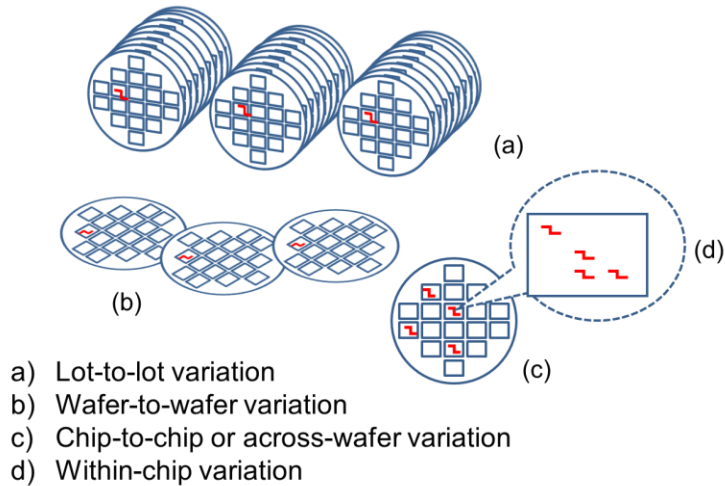
Figure 2.1: Illustration of the spatial process-variation hierarchy: (a) lot-to-lot, (b) wafer-to-wafer (c) chip-to-chip, and (d) within-die variations

2.1.2 Systematic and Random Variations

Systematic variations, also called deterministic variations, are repeatable deviations from nominal device characteristics depending on the device's spatial position on the die and on the wafer and/or the layout context surrounding the device being tested. Common sources of systematic variability include the non-ideality of the lithographic system, such as defocus, misalignment, and line-width roughness [32]; chamber effects that contribute to across-wafer patterns [33]; and various layout-dependent effects, such as WPE [34], optical proximity effects [2], strained silicon effects [35], and CMP [22].

Random variations, or stochastic variations, are unpredictable components of device variability, such as non-uniformities resulting from random fluctuations in the fabrication process, microscopic fluctuations of the number and location of dopant atoms in the transistor channel [17], [36], LER due to photoresist granularity [4], and atomic-scale oxide-thickness variation [16].

Systematic and random variations differ in how they impact device and circuit performance. Systematic spatial variation can cause large differences in performance among devices that are far apart on the die. From a modeling point of view, such an effect in the chips may directly contribute to the spatial correlation among transistors [37]. Random variations, however, are usually treated as independent fluctuations at their corresponding spatial hierarchy level (lot level, wafer level, chip level, etc.).

The classification of systematic and random variations is not absolute. In practice, the running status of equipment or the exact location of the device and circuit on the wafer and chip are often unavailable to circuit designers, rendering it impossible to predict the exact amount of systematic variation. In such cases, systematic variations are often treated as random. Such an approximation is an important source of error in estimating the actual devices' variability and yield.

2.1.3 Global and Local Variations

Another commonly used classification divides device variability into global variation and local variations [38], [39]. As illustrated in Figure 2.1, there are multiple hierarchies above the actual chips in the manufacturing process. From the point of view of an individual chip, variability from the higher hierarchies, such as lot-to-lot, wafer-to-wafer, and chip-to-chip, will be almost equally applied to every transistor on the chip. These variations, whether systematic or random, are lumped together and called "global variations". Correspondingly, the remaining within-chip variations are referred to as "local variations." In SPICE Monte Carlo simulations, the same global-variation component is generated for all devices of the same model, while each device will have its own unique local-variation component. The accurate modeling of the global and local variations plays an important role in estimating the power and performance scaling with circuit complexity, as the local variations will get averaged out among large number of transistors or long critical paths, while the global variations will add up and shift the average power/performance of the entire chip.

2.2 Common Sources of Process Variations

❖ **Lithographic variations**

The uniformity of the printed feature sizes depends heavily on the control of the lithographic imaging system. It affects the two key requirements in integrated circuit manufacturing: the critical dimension (CD) and the overlay control. In a typical step-and-scan lithography stepper (Figure 2.6), the mask reticle and the wafer are simultaneously moving in opposite directions while a slit of light scans the whole mask and projects the image onto the wafer [40]. Even tiny vibrations in the scanner system and variations of the movement speed of the wafer and reticle stage may lead to significant non-uniformities in the depth of focus (DOF) and the light-exposure dose. This can lead to non-uniformity of the critical dimension (CD) of printed lines and may vastly change the speed and leakage of CMOS transistors. Meanwhile, errors in aligning the reticle to the features on the wafer will create variations in misalignment

[41], which can be a crucial problem in achieving the intended line width and good electrical contact between the existing patterns and the new layers of the circuit.
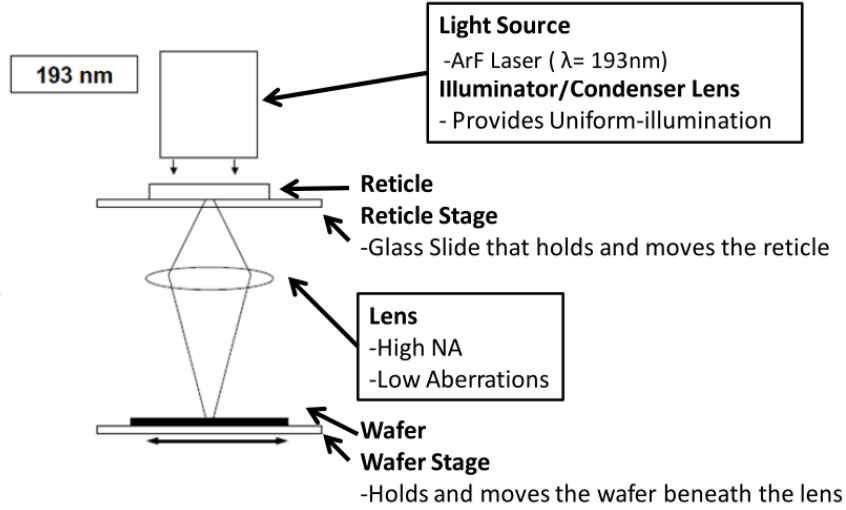
Figure 2.2: A typical lithography imaging system [41]

Another key source of variation in the lithographic patterning process is the post-exposure bake (PEB). The PEB step involves rapidly heating up and cooling down the entire wafer to activate additional chemical reactions and the diffusion of the chemicals within the photoresist. All these phenomena are very sensitive to the PEB temperature trajectory; thus, the uneven temperature in the plate may cause significant CD variations afterward.
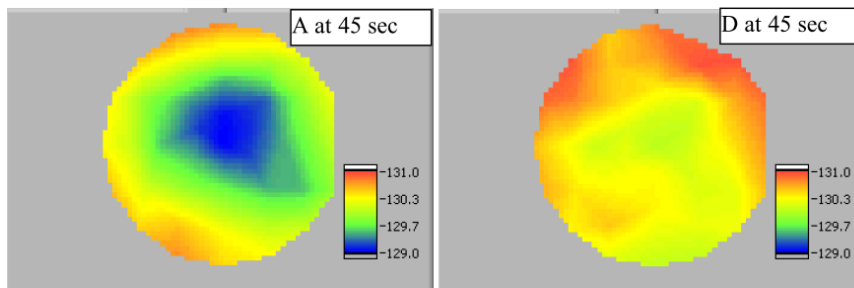
Figure 2.3: Plate temperature non-uniformity near the end of the PEB step [42]

❖ **Line-edge roughness**

As the gate critical dimension shrink continues into the sub-100nm scale, the tolerance of gate line-width control becomes comparable to the size of a resist polymer unit [4], [43]. The granularity of the photoresist creates a non-uniform channel length along the poly gate. This leads to an increased overall leakage current as the off-state current increases exponentially with the reduction in effective channel length. This phenomenon is called line-edge roughness (LER). It contributes to additional threshold voltage variations and degrades the short channel characteristics of transistors. LER is generally considered an intrinsic, random variation.
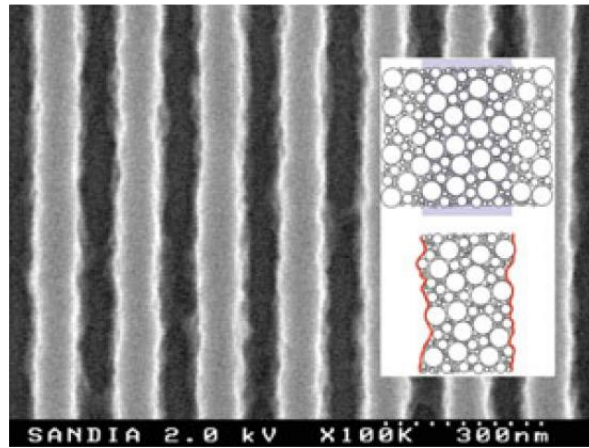
Figure 2.4: Typical LER in a photoresist (Sandia Labs) [44]

❖ **Random dopant fluctuation**

Random dopant fluctuation (RDF) refers to the random microscopic fluctuation of the number and location of dopant atoms in the MOSFET channel region. It causes fluctuations of the transistor electric parameters, such as the threshold voltage (Vt), short channel effect, and drain-induced barrier lowering (*DIBL*). With the gate CD scaling down to sub-100nm, the total number of dopant atoms under the gate is reduced to thousands or even hundreds (Figure 2.5), leading to significant variations in the threshold voltage and drive current [45]. RDF is the single most important source of random variations in the modern CMOS process.
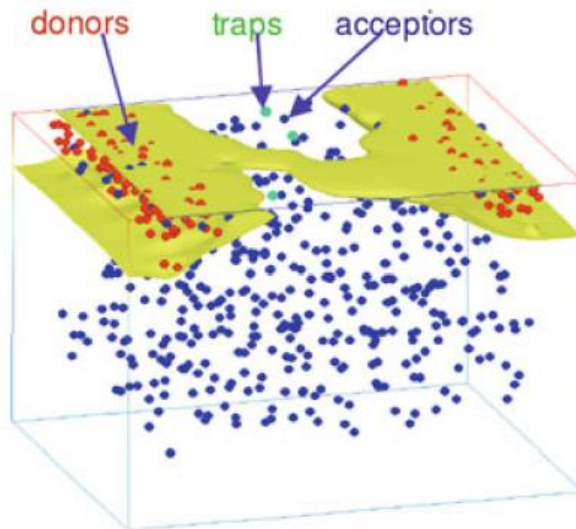
Figure 2.5: Electron distribution in a 30nm "atomistic" MOSFET at threshold [44]

❖ **Well-proximity effect**

The well-proximity effect is an important layout-dependent effect in the deep submicron manufacturing process. It originates from the lateral scattering of implantation ions during the well-implantation step. The incoming high-energy ions collide with the edge of the photoresist on top of the shallow trench isolation (STI), and they get reflected into the channel area before the poly-silicon gate is actually formed. The closer the transistor gate is to the edge of the well, the higher the dopant concentration inside the channel. As a result, transistors with a smaller gate-to-STI distance will have higher threshold voltages.
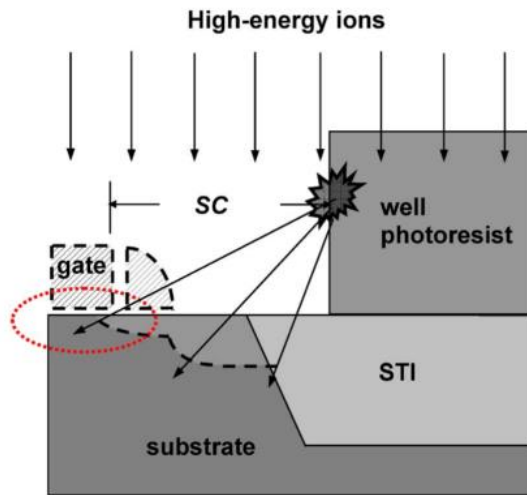
Figure 2.6: Origin of well-proximity effect. High-energy dopant ions scatter at the well photoresist edge during well ion implantation and are reflected into the channel before the gate is formed [34]

❖ **Strained-silicon effects**

The strained-silicon effect is another important source of layout-dependent variation. Currently, advanced CMOS processes intentionally introduce mechanical stress over the channel to enhance the carrier mobility of transistors [46]–[48]. Experiments have shown an electron mobility increase of more than 20% for NMOS with a tensile silicon nitride capping layer and a hole mobility enhancement of more than 50% for PMOS [46] using selective epitaxial $Si_{1-x}Ge_x$ in source and drain. STI stress can also be modulated with gap-fill material to increase the transistor performance by up to 12% with proper layout design and wafer/channel orientation [48]. Studies have shown that the stress profile in the channel can be very sensitive to the length of diffusion (LOD) [49]. Consequently, transistors with the same gate size but a different LOD may have very different speeds.

Figure 2.7: An example of the uniaxial strained-silicon process [46]

❖ **Other variability sources**

Other sources of spatial process variation include the pattern density dependency of the CMP process [22], oxide thickness non-unifo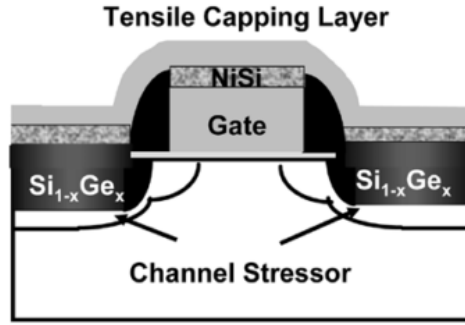rmity [17], non-uniformity in reactive ion etching (RIE), traps and defects in material, etc. These variations will have their own unique impact on transistor characteristics and require extra margin in the design.

2.3 Variation Modeling with Hierarchical Model

2.3.1 Variability Decomposition

From the modeling perspective, process variations can be decomposed in several different ways. Circuit designers commonly treat process variation as a combination of global variation and local variation. This method lumps all the chip-to-chip and wafer-to-wafer variations into one global variation component, and the remaining variations as one local variation component. These are also referred to as "inter-chip" and "intra-chip" variations. With the assumption that variations at different hierarchy levels have very little cross-interaction, the variation in a given device parameter P can be simply decomposed as:

$$\Delta P = \Delta P_{Global} + \Delta P_{Local}$$
$$= \Delta P_{inter-chip} + \Delta P_{intra-chip}$$

(2.1)

Here the inter-chip variation component is the lumped sum of the lot-to-lot, wafer-to-wafer, and chip-to-chip components:

$$\Delta P_{inter-chip} = \Delta P_{lot-to-lot} + \Delta P_{wafer-to-wafer} + \Delta P_{chip-to-chip} \qquad (2.2)$$

Conventionally, the global and local variations are modeled as two independent, normally distributed random variables.

$$\Delta P_{inter-chip} \sim N(\mu_{inter-chip}, \sigma^2_{inter-chip}) \qquad (2.3)$$

$$\Delta P_{intra-chip} \sim N(\mu_{intra-chip}, \sigma^2_{intra-chip}) \qquad (2.4)$$

With the increasingly significant systematic variability, such as layout-dependent effects in the process, some variability models added an additional across-chip systematic component of parameter P to the equation, which is modeled as a normally distributed random variable independent of the other components:

$$\Delta P = \Delta P_{inter-chip} + \Delta P_{intra-chip\ random} + \Delta P_{across-chip\ systematic} \qquad (2.5)$$

$$\Delta P_{across-chip\ systematic} \sim N(\mu_{across-chip\ systematic}, \sigma^2_{across-chip\ systematic}) \quad (2.6)$$

A prior variability study, however, shows that the systematic variations, particularly at the wafer and chip level, will cause the device parameter distribution to deviate from normal distributions at extreme quantiles [50]. To improve the accuracy of the variability model, the systematic and random components should be individually characterized for each level of the fabrication hierarchy, mainly at the wafer level (chip-to-chip) and chip level (device-to-device).

❖ **Variation at lot level and above**

State-of-the-art semiconductor manufacturing involves various batch processes that apply to multiple wafers at the same time for a high wafer throughput. For example, the chemical vapor deposition (CVD) heats up multiple wafers in the furnace, where the reactive gas forms a thin film on the surface of the wafers [13]. The batches are usually referred to as lots, which conventionally contain 25 wafers each. As a result, some process conditions are applied to all the wafers in the same lot but there are changes from lot to lot, leading to *lot-to-lot variation*. Meanwhile, wafers within the same lot are also subject to non-uniformity in the chamber environment, such as the temperature and the speed of the gas flow, which results in *within-lot variations*.

During single-workpiece processes, such as lithographic imaging and reactive ion etching (RIE), each wafer is processed individually. Naturally, this will lead to variability between different wafers, which is called *wafer-to-wafer variation.*

In theory, one can model the lot-to-lot and wafer-to-wafer variations using time-series models [13] and fit the systematic signatures of within-lot variations. In practice, however, such a practice requires long-term monitoring over a significant number of lots and wafers. It is often more convenient to lump them together as a single variation component that varies from wafer to wafer, denoted as ΔP_{W2W}.

In this thesis, without loss of generality, we assume that the wafer-to-wafer variation ΔP_{W2W} can be sufficiently modeled as a normally distributed variable independent to the other variation components. This assumption typically holds well in a reasonably mature semiconductor process without process splits. Thus, the *random wafer-to-wafer variation* is described by

$$\Delta P_{W2W} \sim N(0, \sigma_{W2W}^2) \tag{2.7}$$

❖ **Variation at the wafer level**

Wafer level non-uniformity can come from deposition, photoresist spinning effects, temperature non-uniformity in post-exposure baking or plasma etching, and other equipment non-uniformities that result in a smooth, low-frequency across-wafer variation pattern. In particular, wafer-level variation often exhibits symmetric radial ("dome" or "bull's eye") patterns [51]. We call such repeatable wafer-level variability *systematic across-wafer variation*. Since the chip size is usually much smaller than the wafer diameter, we can assume that the across-wafer pattern is approximately constant within a chip's scale. Therefore, for a device from chip location (x_W, y_W) on the wafer, the systematic across-wafer variability component can usually be sufficiently represented by an elliptic paraboloid function, denoted as $\Delta P_{AW}(x_W, y_W)$:

$$\Delta P_{AW}(x_W, y_W) = a_W \cdot x_W^2 + b_W \cdot x_W + c_W \cdot y_W^2 + d_W \cdot y_W$$
$$+ e_W \cdot x_W y_W + f_W \tag{2.8}$$

In addition, process variations, such as the focus and exposure fluctuation in lithographic imaging, may introduce additional variability from chip to chip. Lumped together with the fitting residual of the systematic across-wafer variation, we call it the *random chip-to-chip variation* (or across-wafer random variation), denoted as ΔP_{AWR}. It may be modeled by a Gaussian variable, as described in Equation 2.9.

$$\Delta P_{AWR} \sim N(0, \sigma_{AWR}^2) \tag{2.9}$$

❖ **Variation at the die/chip level**

Intra-die variation or within-die variation refers to the fluctuation of device properties on the same chip/die. Similar to wafer-level variations, chip-level variations also consist of systematic and random components. Typical sources of systematic spatial variations include stepper-induced variations (illumination, lens aberrations) [52], reticle imperfections, and CMP [50], [53]. The *systematic across-chip variation* of a device with location (x_C, y_C) on the die/chip can often be approximated by an elliptic paraboloid as well, as described by Equation 2.10 [50], [52], [53]:

$$\Delta P_{AC}(x_C, y_C) = a_C \cdot x_C^2 + b_C \cdot x_C + c_C \cdot y_C^2 + d_C \cdot y_C$$
$$+e_C \cdot x_C y_C + f_C \tag{2.10}$$

The *random across-chip variations* (device-to-device or local mismatches), on the other hand, include intrinsic variability, such as RDF, interface-trapped charge fluctuations, atomic oxide-thickness fluctuations, and LER. These intrinsic random variations are dominant at the deep-submicron device scale. They are modeled as a normally distributed random variable independent to the other variation components, which is denoted as ΔP_{ACR}:

$$\Delta P_{ACR} \sim N(0, \sigma_{ACR}^2) \tag{2.11}$$

Last, *the layout-dependent variations,* such as those due to optical-proximity effects, strained-silicon effects, and plasma micro loading, will cause devices with similar design parameters but different layout designs and/or sounding layout contexts to differ significantly in device characteristics, such as gate CD and mobility. In this work, we adopt the assumption that the layout-dependent effects do not interact with the rest of the spatial variations in the system; thus, the layout-dependent variation is a simple additive term described by Equation 2.12.

$$\Delta P_{layout} = F(layout\ pattern) \tag{2.12}$$
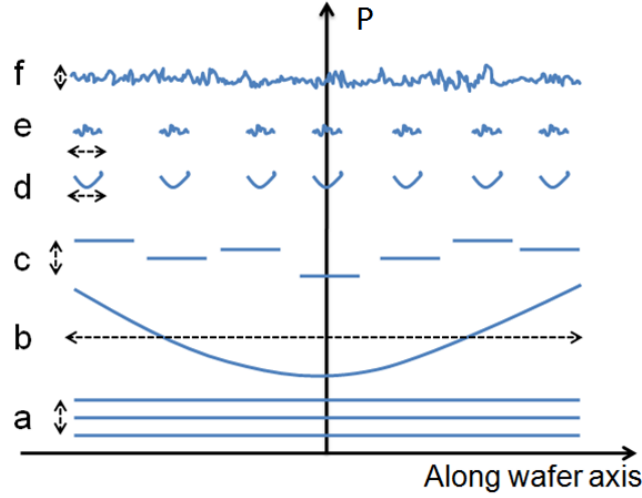
2.3.2 Hierarchical Variability Model

Figure 2.8: One-dimensional illustration of the hierarchical variability components of parameter *P*: a) lot-to-lot and wafer-to-wafer random, b) across-wafer systematic, c) chip-to-chip random, d) across-chip systematic, e) layout-dependent, and f) device-to-device random.

Figure 2.8 is an illustration of the hierarchical variability model we proposed for capturing the systematic and random components in the integrated circuit manufacturing process. Assume device parameter *P* is a process-related physical quantity and that its variations from different sources or hierarchy levels have relatively small interactions. In this case, the total variation of parameter *P* can be simply modeled as the sum of the different variability components: $\Delta P = \Delta P_{source-1} + \Delta P_{source-2} + \cdots \Delta P_{source-N}$.

Given the variability decomposition scheme previously described, the total spatial variation of parameter *P* can be decomposed as the sum of the wafer-to-wafer random variations, across-wafer systematic variations, chip-to-chip (across-wafer) random variations, across-chip systematic variations, device-to-device (across-chip) random variations, and layout-dependent variations.

$$\Delta P = \Delta P_{W2W} + \Delta P_{AW} + \Delta P_{AWR} + \Delta P_{AC} + \Delta P_{ACR} + \Delta P_{Layout} \qquad (2.13)$$

Such an additive hierarchical model can also be applied to the estimates of those device or circuit parameters that are linearly proportional to additive physical quantities. For example, the effective gate length (*Leff*) of transistors is often such a physical parameter that satisfies the additive requirement. Consequently, the ring oscillator stage delay, which is proportional to *Leff* to the first order, can also be modeled in this additive fashion

(Equation 2.14). The additive model cannot be applied to the ring oscillator frequency, however, as it follows 1/*Leff* (Equation 2.15).

$$\Delta Delay = k \Delta L_{eff} = k\left(\Delta L_{eff,source-1} + \Delta L_{eff,source-2}\right)$$

$$= \Delta Delay_{source-1} + \Delta Delay_{source-2}$$

(2.14)

$$\Delta freq = \frac{1}{k \Delta L_{eff}} = \frac{1}{k\left(\Delta L_{eff,source-1} + \Delta L_{eff,source-2}\right)}$$

$$= (\Delta freq_{source-1}^{-1} + \Delta freq_{source-2}^{-1})^{-1}$$

(2.15)

$$\neq \Delta freq_{source-1} + \Delta freq_{source-2}$$

2.4 Summary

In this chapter, we first reviewed the various classifications of process variability. By nature, process variations can be environmental, temporal, or spatial. With regard to repeatability, process variations can be systematic or random, and with regard to their scope of impact, process variations are divided into global variations and local variations.

The common sources of the systematic and random components of spatial process variation are then discussed in detail. The most prominent effects include variations in lithographical imaging and post-exposure baking (PEB), random dopant fluctuations (RDF), line-edge roughness (LER), well-proximity effects (WPE), strained silicon effects, chemical mechanical polishing (CMP), thin film-thickness fluctuation, etc.

Lastly, an additive hierarchical variability model was proposed to capture the various components of spatial process variations. The total variability of device parameter *P* was modeled as the sum of the random wafer-to-wafer, chip-to-chip, and device-to-device variations; the systematic across-wafer and across-chip variations; and the layout-dependent variations. This simple but effective spatial variability model will be used in characterizing the variability profile in the near-mature commercial-quality silicon data in Chapter 3 and Chapter 5.

Chapter 3

Test Chip Design, Characterization, and Variability Analysis

3.1 Introduction

We experimented with multiple test wafers with custom test structures to investigate the influence of process variability in modern semiconductor manufacturing and to understand the underlying mechanism. The test circuits were designed by BWRC students and faculty [54]–[57], and fabricated by our foundries partners using the 90nm and 45nm bulk process. We characterized key variability test structures, including ring oscillator (RO) arrays for delay and leakage current measurement, SRAM arrays, and individually measurable padded-out transistors of the SRAM cells.

3.2 The 90nm Ring Oscillator Test Chip

3.2.1 Chip Design Overview

A test chip is designed and implemented in a general-purpose 90nm CMOS technology process from STMicroelectronics to characterize the process-induced circuit variations [54]. The approach we use is to measure the oscillating delay and transistor source-drain leakage currents of an array of ring-oscillator test structures.

The test chip is made up of 10 rows $\times$ 16 columns of tiles of test structures. Each tile contains twelve 13-stage RO and 12 off-state NMOS transistors, one for each of the 12 different layout styles (Figure 3.2). The tiles are separated by 62.5um horizontally and 100um vertically. The total array area is 1mm $\times$ 1mm, and the overall die size, including the peripherals, is about 1.8mm $\times$ 1.4mm. Layout pattern styles include gate stacks that consist of 1 to 3 Poly-Si fingers with varied length of diffusion (LOD). The Poly-Si pitch of neighboring dummy features is varied, and one layout has a Poly-Si orientation rotated by 90 degrees. Asymmetric masks are used to test the coma effect. The first metal layer coverage over gates is varied as well. The test chip also includes a leakage current measurement circuit, which sits right beside the ROs with the same layout.
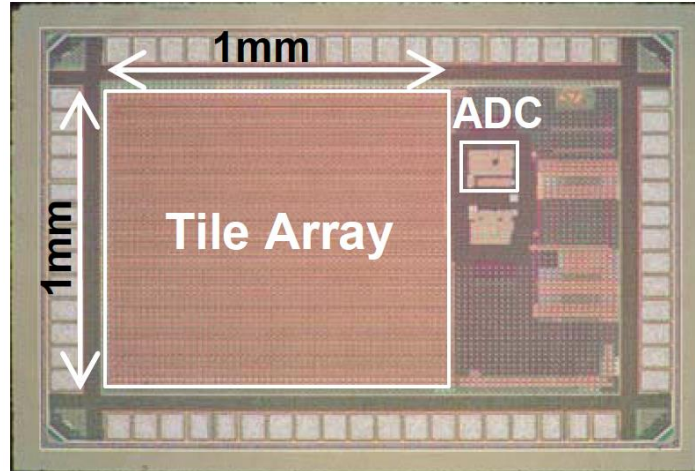
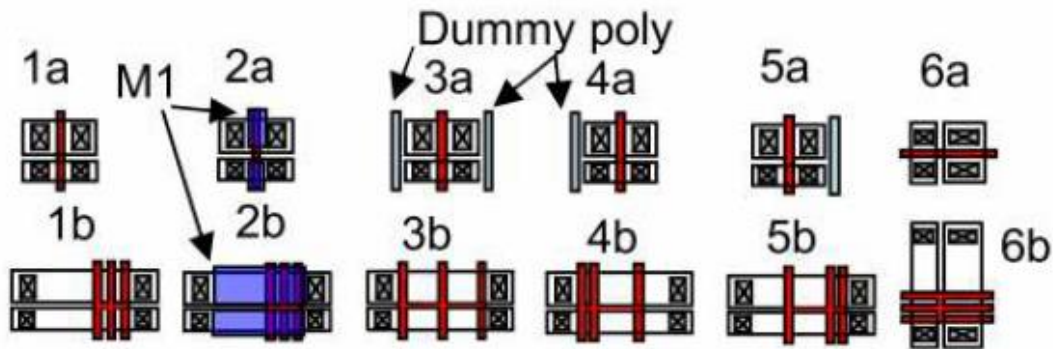Figure 3.1: Die photo of 90nm test chip [54]

Figure 3.2: Layout configuration in the 90nm test chip [54]

3.2.2 Sampling and Measurement Scheme

There are two requirements on the sampling scheme based on our variability model. First, there must be enough measuring points inside each chip to capture the systematic and random components at chip level. Second, these points should be spread out across the wafer to capture the wafer level variations.

For this 90nm test wafer, we examined the delay and static leakage (IDDQ) data collected from 36 chips distributed mostly across the right half of the wafer. Each chip was measured exhaustively to get a complete and statistically significant spatial coverage over the 1mm $\times$ 1mm RO array. The ring oscillator delays were measured off-chip with a

22

20GSPS oscilloscope and averaged over about 100 periods. The transistor off-state currents were measured using an on-chip single-slope analog-to-digital converter (ADC) [58]. The wafer-level measurement plan and the collected RO data are shown in Figure 3.4.
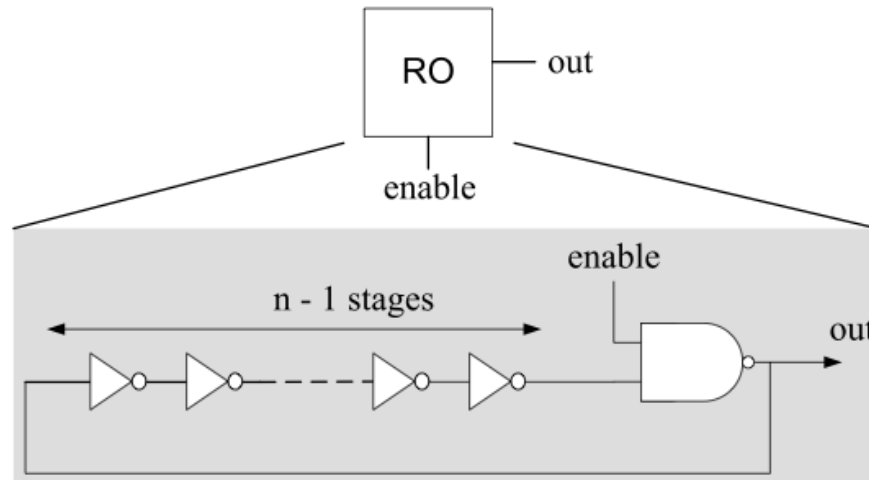

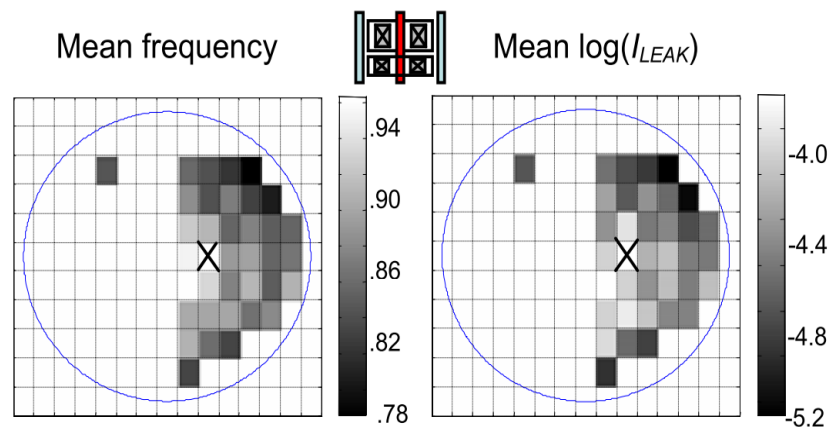
Figure 3.3: Ring oscillator with n stages [58]

Figure 3.4: Wafer maps of mean RO frequency and mean Log(I_{LEAK}) for layout 3A [54]

3.2.3 Variability Observation

Devices of three representative layouts—3A, 4A, and 5A—are selected for studying the wafer-level, chip-level, and layout-to-layout variations. Wafer-level RO delay and leakage (IDDQ) variation of these layouts are shown in Figure 3.5 and Figure 3.6. Each

data block stands for the average value over all the 160 tiles, which is noted by symbol $D\langle-DWP\rangle$ and $I\langle-DWP\rangle$. Similarly, chip-level RO delay and leakage variations are shown in Figure 3.7 and Figure 3.8. Each data block stands for the average value of all 36 dies, which is noted by symbol $D\langle T-WP\rangle$ and $I\langle T-WP\rangle$.

If we assume that the majority of systematic variations are from the effective gate length (L_{eff}) variability, then the following simple model describes the RO delay and leakage (SPICE simulations confirm that this is a good approximation when the gate length variation is small):

$$delay: D = D_0 \left(\frac{L_{eff}}{L_0}\right)$$

$$leakage: \log IDDQ = \log IDDQ_0 \left(\frac{L_0}{L_{eff}}\right)$$

$$(3.1)$$

As a simple function of the physical quantity L_{eff}, RO delay is a good candidate for the application of the additive hierarchical variability model. According to the hierarchical model, the total variation of devices of a given layout pattern on a single wafer can be decomposed into across-wafer systematic (AW), across-wafer random (AWR), across-chip systematic (AC), and across-chip random (ACR). Statistical analysis shows that the across-wafer gate RO delay variation can be approximated adequately by a second-order polynomial, of the form in Equation 3.2. Note that due to the lack of the left half of the wafer, the quadratic term in the X-direction is statistically insignificant; we set that coefficient to zero. Meanwhile, the across-chip variation can also be fitted by a chip-level second-order polynomial. Statistics show the variation along different columns does not have a significant systematic component, while variation along the different rows displays a significant (half-) parabolic pattern. The simplified approximation is shown in Equation 3.3.

$$D\langle-DWP\rangle = D\langle-DWP\rangle_{AW} + D\langle-DWP\rangle_{AWR}$$

$$D\langle-DWP\rangle_{AW} = a_W X_W^2 + b_W X_W + 0 \times Y_W^2 + d_W Y_W + e_W$$

$$(3.2)$$

$$D\langle T-WP\rangle = D\langle T-WP\rangle_{AC} + D\langle T-WP\rangle_{ACR}$$

$$D\langle T-WP\rangle_{AC} = 0 \times X_C^2 + 0 \times X_C + c_C Y_C^2 + d_C Y_C + e_C$$

$$(3.3)$$

Using RO delay as an example, the fitted coefficients and their 95% confidence intervals are shown in Figure 3.9: and Figure 3.10.

Below, I apply statistical tests *algorithmically* rather than *statistically*: the underlying statistical models do not hold: there is no basis for the assumed probability distribution of the data, and all the null hypotheses are false. The tests do not have their nominal significance levels in this problem; indeed, it is not clear what "significance level" would even mean. Nonetheless, applying statistical tests may provide insight into which components of variation are worth modeling, and may lead to models that make more reliable and useful predictions.

To examine the layout dependence effects on the variation pattern, we used t-statistic to compare the estimates of fitted coefficients from the three layout designs. Use a_W as example, and under the assumption null-hypothesis H_0: $a_{W_{3A}} = a_{W_{4A}}$ is rejected if:

$$t_{a_{W_{3A}}, a_{W_{4A}}} = \frac{\left|\hat{a}_{W_{3A}} - \hat{a}_{W_{4A}}\right|}{\sqrt{SE_{\hat{a}_{W_{3A}}}^2 + SE_{\hat{a}_{W_{4A}}}^2}} > t_{\frac{\alpha}{2}, N-1} \tag{3.4}$$

The estimates of mean and standard error (SE) in the linear regression model (Table 3.1), give

$$t_{a_{W_{3A}}, a_{W_{4A}}} = \frac{|-0.1454 + 0.1423|}{\sqrt{1.903 \times 10^{-6} + 1.867 \times 10^{-6}}} = 1.612$$

$$< t_{\frac{0.05}{2}, 5757} = 1.96 \tag{3.5}$$

Similarly,

$$t_{a_{W_{3A}}, a_{W_{5A}}} = \frac{|-0.1454 + 0.145|}{\sqrt{1.903 \times 10^{-6} + 1.902 \times 10^{-6}}} = 0.224$$

$$< t_{\frac{0.05}{2}, 5757} = 1.96 \tag{3.6}$$

$$t_{a_{W_{4A}}, a_{W_{5A}}} = \frac{|-0.1423 + 0.145|}{\sqrt{1.867 \times 10^{-6} + 1.902 \times 10^{-6}}} = 1.385$$

$$< t_{\frac{0.05}{2}, 5757} = 1.96 \tag{3.7}$$

None of the three pair-wise null hypotheses is rejected. This suggests it may be adequate to take the coefficient a_W to be equal for all three layout designs. Similar

analysis leads us to model the rest of "shape" coefficents b_W and d_W as equal for the three layout designs. On the other hand, the same tests for the intercept coefficient e_W give:

$$t_{e_{W_{3A}}, e_{W_{4A}}} = 19.1 > 1.96 \tag{3.8}$$

$$t_{e_{W_{3A}}, e_{W_{5A}}} = 22.4 > 1.96 \tag{3.9}$$

$$t_{e_{W_{4A}}, e_{W_{5A}}} = 3.47 > 1.96 \tag{3.10}$$

Because these differences are (nominally) statistically significant, we retain differences among the intercept terms e_W for layouts 3A, 4A and 5A. It is also worth noting that the t-statistics between layout 4A and 5A are much smaller than that between either of those layouts and layout 3A.

		3A	4A	5A
a_W	Estimate	-0.145	-0.142	-0.145
	SE^2	1.9E-06	1.87E-06	1.9E-06
b_W	Estimate	0.0174	0.0164	0.0171
	SE^2	5.27E-08	5.17E-08	5.27E-08
d_W	Estimate	0.0111	0.0108	0.0110
	SE^2	1.26E-08	1.23E-08	1.26E-08
e_W	Estimate	1.40	1.31	1.29
	SE^2	1.31E-05	1.29E-05	1.31E-05

Table 3.1: Estimates and standard errors of fitting coefficient a_W, b_W, d_W and e_W for the across-wafer spatial variation patterns of layout 3A, 4A and 5A

The number of pairwise statistical tests of coefficient equality required can grow quickly as more layout designs are in comparison. As an alternative, we simply observe

the trend of confidence intervals (CI) of the fitting coefficients across different layouts. While not a statistically valid test for a difference, we treat these shape coefficients as equal if their CIs have large overlaps.

With this alternative method, we found that most of the layout-dependent effects are accounted for by differences in the intercept terms e_W and e_C. We model the layout-dependent component in this process as an additive component on top of the systematic across-wafer and across-chip component. The large overlap of the confidence intervals of the layout component e_w and e_C between layouts 4A and 5A while layout 3A is far apart is consistent with the fact that layouts 4A and 5A are mirror images while 3A has a different pattern density. Therefore, the layout-dependence differences between them are minimal, while layout 3A with dummy polys on both sides of the gate actually behaves as a slower device in general. This observation is contradictory to the common knowledge that a more regular poly-grating structure will result in a narrower printed poly gate critical dimension (poly CD). Unfortunately, it requires more detailed electrical tests as well as physical examination of the device cross-section to reveal the root cause. Last, a similar conclusion can be drawn if we perform the same experiment on the RO leakage data.
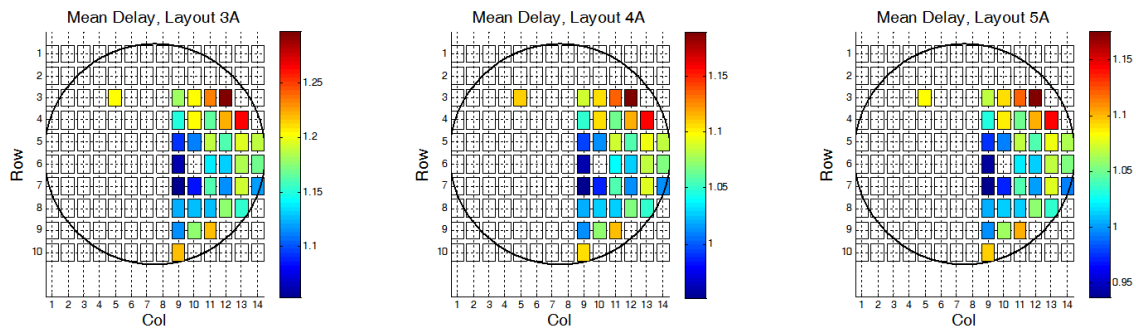

Figure 3.5: Wafer maps of mean RO delay of layouts 3A, 4A, and 5A [54]: $D\langle -DWP\rangle$
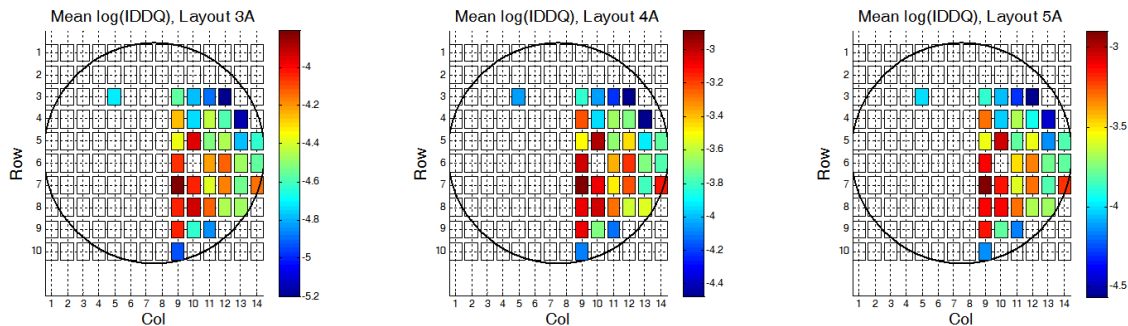
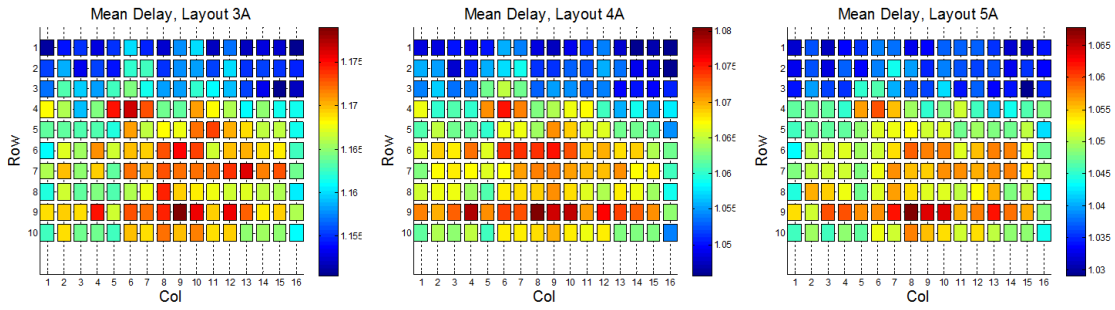Figure 3.6: Wafer maps of the mean RO log(IDDQ) of layouts 3A, 4A, and 5A [54]: $I\langle -DWP \rangle$

Figure 3.7: Chip maps of the mean RO delay of layouts 3A, 4A, and 5A [54]: $D\langle T - WP \rangle$
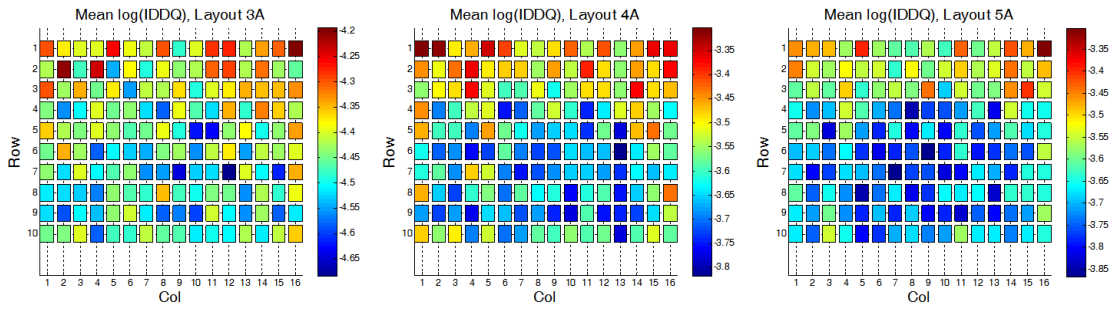

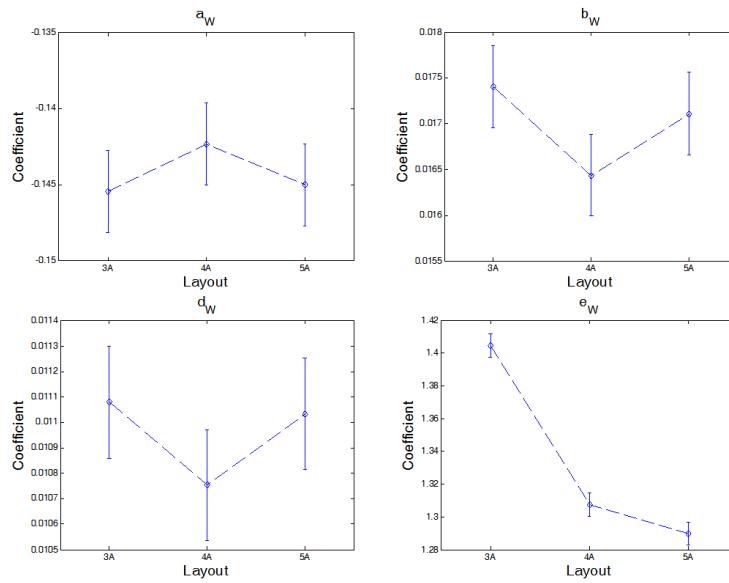Figure 3.8: Chip maps of the mean RO log(IDDQ) of layouts 3A, 4A, and 5A [54]: $I\langle T - WP \rangle$

Figure 3.9: Estimate and confidence interval of across-wafer fitting coefficients: layouts 3A, 4A, and 5A
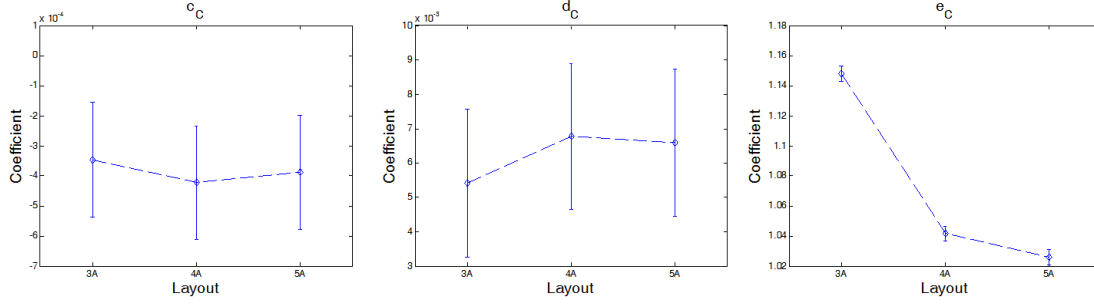
Figure 3.10 Estimate and confidence interval of across-chip fitting coefficients: layouts 3A, 4A, and 5A

Figure 3.11 and Figure 3.12 illustrate the decomposition of across-wafer and across-chip RO delay variations of layout 3A. The fitting residuals after the removal of the systematic across-wafer and across-chip components become much closer to a standard Gaussian distribution, as shown in Figure 3.14.

To test how much the hierarchical variability model improves on the conventional "Global+Local" variability model, Monte Carlo experiments are performed to simulate the distribution of the RO delay of 10,000 chips with 160 test devices per chip. Assume each chip has the exact equal chance to be chosen from the 36 chip locations on the wafer, and each test device has the exact equal chance to be chosen from the 160 tile locations on the chip. Under the simple "Global+Local" model, the delay of each RO device is the sum of two Gaussian random variables. One carries the same variance as the total chip-to-chip variation from the raw measurement, while the other carries the same variance as the total within-chip variance of the raw measurement data. Under the hierarchical variability model, the RO delay is still modeled as the sum of the chip-level component and the within-chip component. However, each component is now composed of a systematic across-wafer/across-chip component in addition to the residual Gaussian random variation. The formula for simulating the distribution is shown in equation 3.11 to 3.13.

RO delay of the kth layout from the jth tile on the ith chip:

$$D_{i,j,k} = Layout_k + Chip_i + Tile_j \qquad (3.11)$$

"Global+Local" variation model:

$$Chip_i \sim N(0, \sigma_{Global})$$
$$Tile_j \sim N(0, \sigma_{Local}) \qquad (3.12)$$

Hierarchical model:

29

$$Chip_i \sim f_{AW}\left(X_{W_i}, Y_{W_i}\right) + N(0, \sigma_{AWR})$$

$$(3.13)$$

$$Tile_j \sim f_{AC}\left(X_{C_j}, Y_{C_j}\right) + N(0, \sigma_{ACR})$$

The normal quantile plots (Figure 3.15) provide direct comparisons of the two models' Monte Carlo experiment results. Both model predictions are fairly close to the raw measurement for the most part within $\pm 2\sigma$. At $\pm 3\sigma$, the hierarchical model starts to show less deviation from the raw measurement than the "Global+Local" model, especially on the fast side. The numerical comparisons of $\pm 3\sigma$ and the median delay of layouts 3A, 4A, and 5A are shown in Table 3.2. The two models are within 0.5% of each other at $+3\sigma$ for all three layouts, while at -3σ, the hierarchical model consistently shows 2% better accuracy than the simple "Global+Local" model.
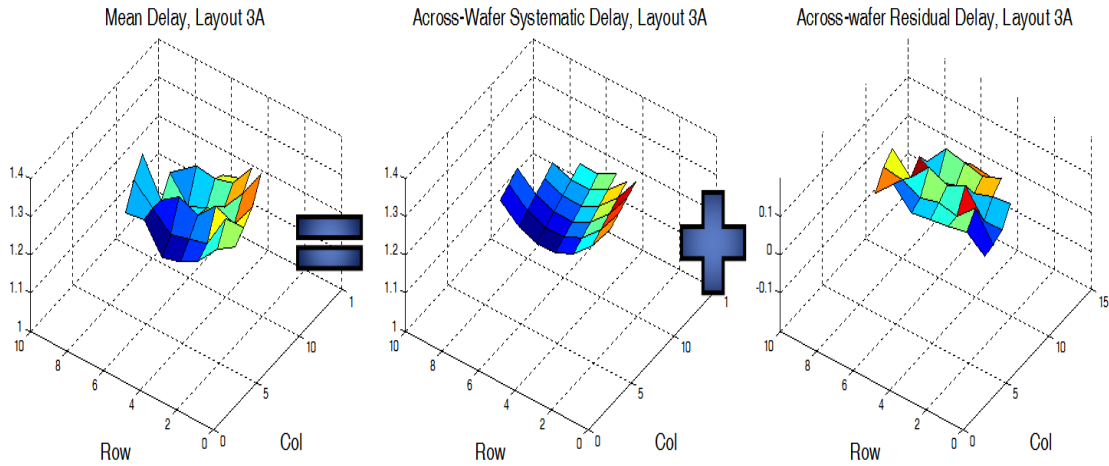

Figure 3.11: Decomposition of wafer-level variation of layout 3A:
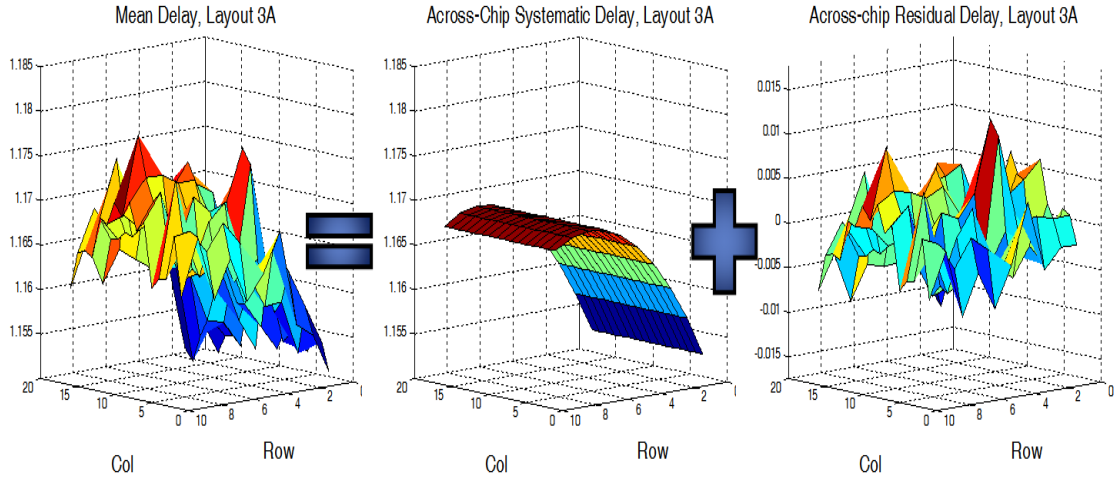$$D\langle-DWP\rangle = D\langle-DWP\rangle_{AW} + D\langle-DWP\rangle_{AWR}$$

Figure 3.12: Decomposition of wafer-level variation of layout 3A:

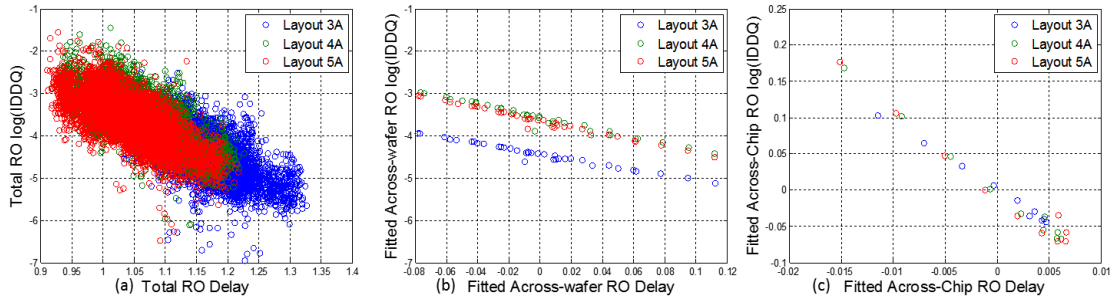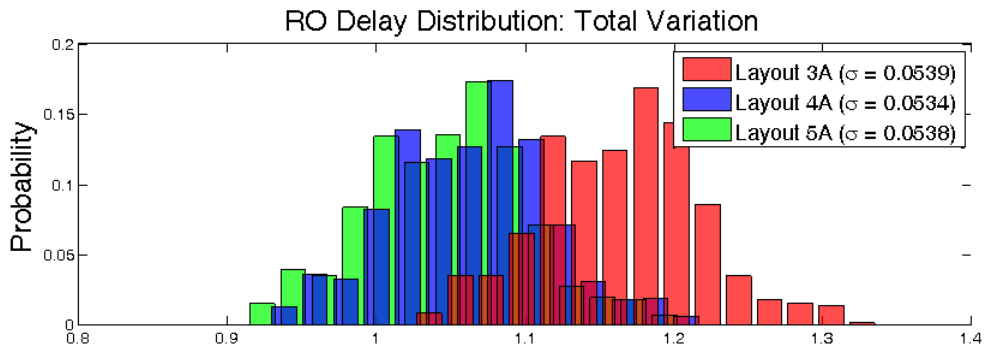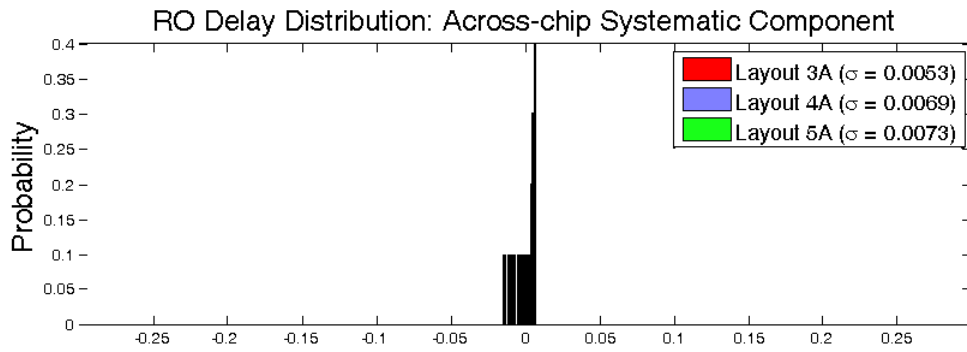$$D\langle T - WP\rangle = D\langle T - WP\rangle_{AC} + D\langle T - WP\rangle_{ACR}$$

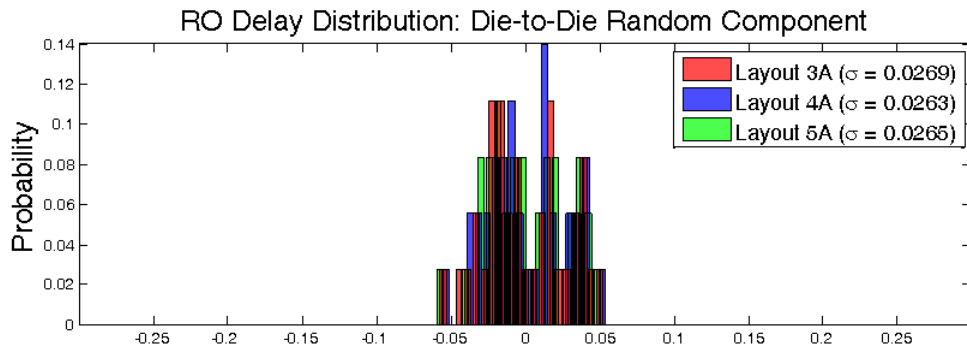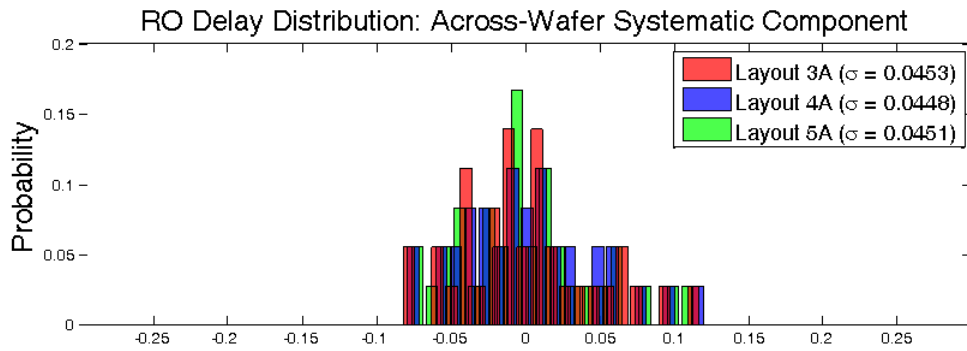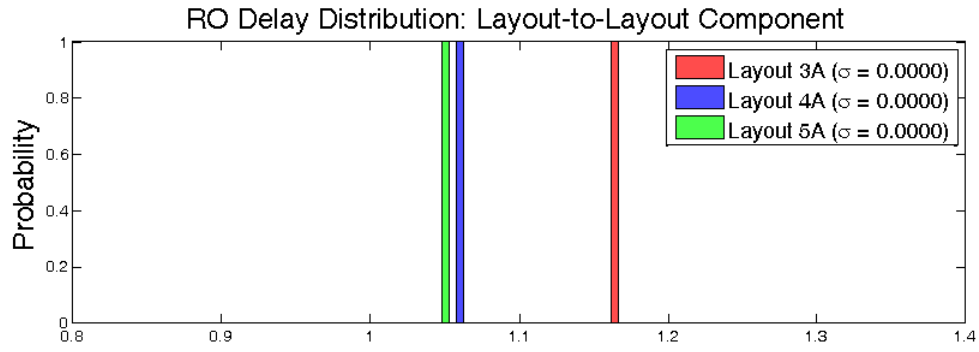
Figure 3.13: Correlation between RO leakage and delay:

(a) $I\langle TDWP\rangle\ vs.D\langle TDWP\rangle$; (b) $I\langle T - WP\rangle_{AW}\ vs.D\langle T - WP\rangle_{AW}$; (c) $I\langle -DWP\rangle_{AC}\ vs.D\langle -DWP\rangle_{AC}$

RO Delay Distribution: Layout-to-Layout Component

Probability

Layout 3A ($\sigma = 0.0000$)
Layout 4A ($\sigma = 0.0000$)
Layout 5A ($\sigma = 0.0000$)

RO Delay Distribution: Across-Wafer Systematic Component

Probability

Layout 3A ($\sigma = 0.0453$)
Layout 4A ($\sigma = 0.0448$)
Layout 5A ($\sigma = 0.0451$)

RO Delay Distribution: Die-to-Die Random Component

Probability

Layout 3A ($\sigma = 0.0269$)
Layout 4A ($\sigma = 0.0263$)
Layout 5A ($\sigma = 0.0265$)

RO Delay Distribution: Across-chip Systematic Component

Probability

Layout 3A ($\sigma = 0.0053$)
Layout 4A ($\sigma = 0.0069$)
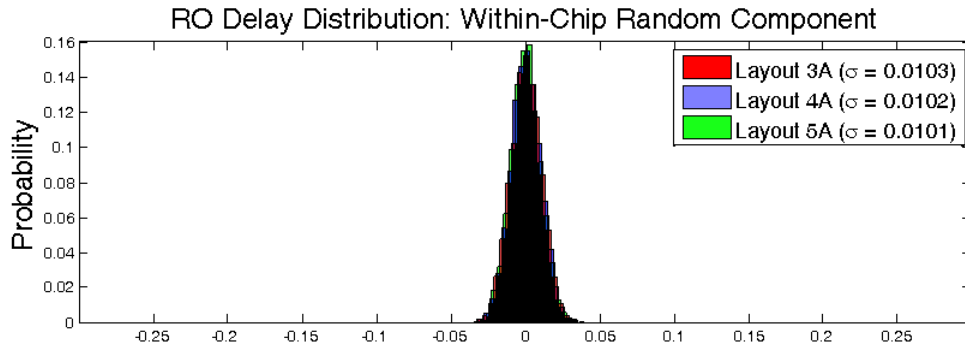Layout 5A ($\sigma = 0.0073$)

Figure 3.14: Histogram of the RO delay distribution as well as the systematic across-wafer, layout-to-layout, random chip-to-chip, systematic across-chip, and random tile-to-tile variability
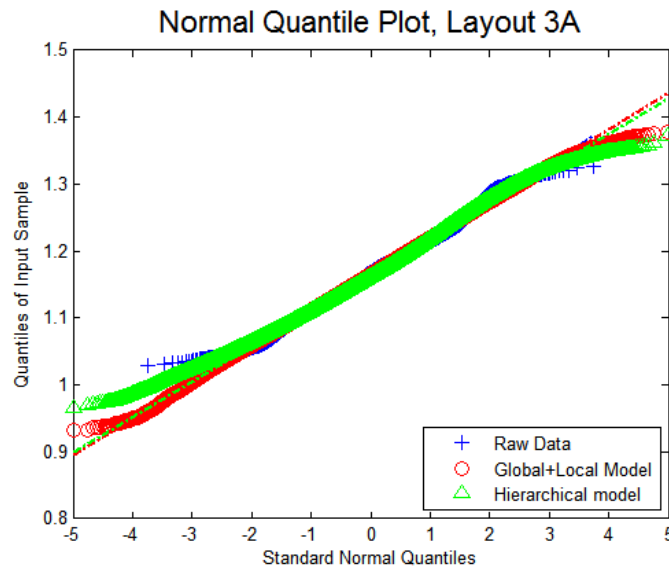

Figure 3.15: Comparing the prediction accuracy of the "Global+Local" model versus the hierarchical model for the RO delay distributions of layout 3A

		Measurement	"Global+Local" Model	Hierarchical Model
Layout 3A	$+3\sigma$	1.317	1.327 (+0.8%)	1.324 (+0.5%)
	Median	1.168	1.165 (-0.2%)	1.162 (-0.5%)
	-3σ	1.036	1.004 (-3.1%)	1.024 (-1.1%)
Layout 4A	$+3\sigma$	1.212	1.225 (+1.1%)	1.226 (+1.2%)

33

	Median	1.066	1.062 (-0.4%)	1.059 (-0.7%)
	-3σ	0.935	0.904 (-3.3%)	0.926 (-1.0%)
Layout 5A	$+3\sigma$	1.198	1.214 (+1.3%)	1.218 (+1.7%)
	Median	1.051	1.048 (-0.4%)	1.045 (-0.6%)
	-3σ	0.919	0.890 (-3.2%)	0.909 (-1.2%)

Table 3.2: Median and +/- 3s of simple "Global+Local" model and the hierarchical variability model in comparison with the measurements (difference to measurement shown in percentages)

3.3 The 45nm Ring Oscillator and SRAM Test Chips

3.3.1 Chip Overview

To further investigate the process dependency of the device and circuit variability, a newer set of 45nm test chip circuitries was designed by Liang-teck Pang et al. [57] and Zheng Guo [59]. The test chips were fabricated using a 45nm low-power strained-Si CMOS process [47], [48], [60], with an array of ROs and corresponding off-state leakage current measurement circuitry, as well as 18 SRAM macros that allow the characterization of SRAM padded-out transistors and the SRAM read/write margins. The die photo is shown in Figure 3.16.

To keep up with the aggressive technology scaling, new fabrication practices and stricter design rules have been introduced to the 45nm technology. Poly spacing can no longer be freely adjusted; instead, only a small continuous range followed by a discrete jump in Poly-Si spacing is allowed. All transistor channels are oriented in the <100> direction, which enhances PMOS mobility and makes it insensitive to stress [61]. Two major sources of stress are introduced both by design and unintentionally in this process: strain caused by the contact-etch stop layer (CESL) and the shallow trench isolation (STI) stress. Subatmospheric chemical vapor deposition oxide (SACVD) largely reduces usually strong compressive STI stress and turns it into a weak tensile one. CESL is formed by intentionally depositing a nitride layer on top of NMOS transistors, which introduces a strong horizontal tensile strain that greatly enhances the electron mobility. Another important feature of the new 45nm test chip fabrication is the different gate-trimming treatment for the two wafers we have, aiming at a nominal 4nm reduction in gate CD from the slower wafer (#1) to the faster wafer (#2). The major features of the 45nm process are summarized in Table 3.3.
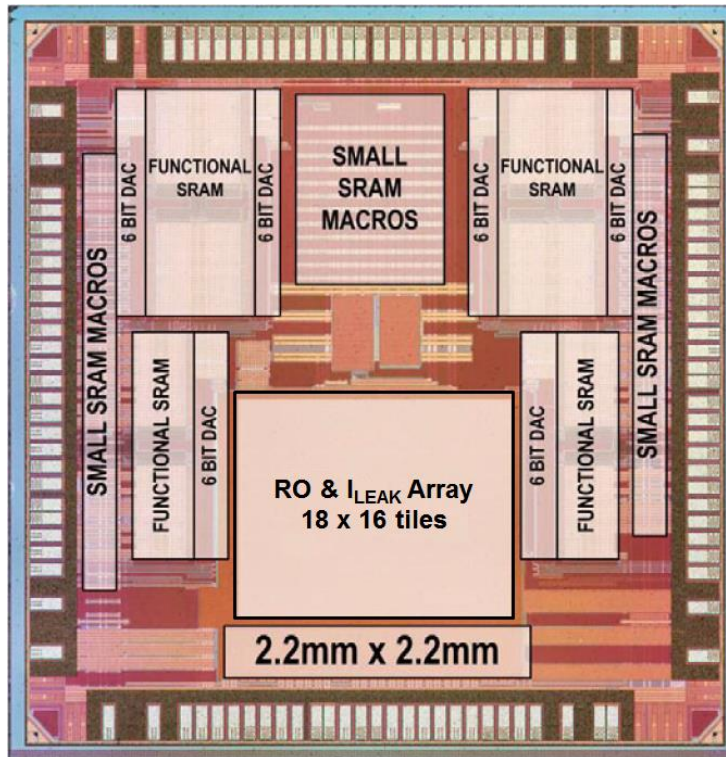
Figure 3.16: Die photo of the 45nm test chip

Process Feature	45nm Process	Effect
Si substrate	[100]-oriented channel	Higher PMOS mobility
Shallow trench isolation (STI)	Sub-atmospheric deposited oxide	Lower STI stress
Contact etch stop layer (CESL)	Nitride layer creating high tensile strain	Higher NMOS mobility
Immersion lithography	NA > 1	Improved resolution
Backend dielectric	Low k ~2.5	Low RC delay

Table 3.3: Summary of the 45nm process

The RO array contains 18×16 identical tiles. Each tile consists of 17 thirteen-stage ROs and 17 pairs of off-state NMOS and PMOS transistors for leakage measurements, each with the same transistor sizing embedded in a different layout pattern. A total of 17 different RO transistor layouts are designed based on the new process and design rules to

capture possible layout-dependent effects, including various Poly-Si gate-dummy pitches, different source/drain areas with and without STI, and orientation of transistor placement. The layouts are presented in Figure 3.17 and Figure 3.18. Note that the pre-OPC patterns depicted in Figure 3.17 are subject to OPC treatment prior to fabrication, the specifics of which remain unknown to us. Measurement circuitry was adopted from the design of the 90nm test chip. The RO delay and corresponding off-state NMOS/PMOS transistor leakage currents were measured in our laboratory after the wafers were diced and the chips were packaged.
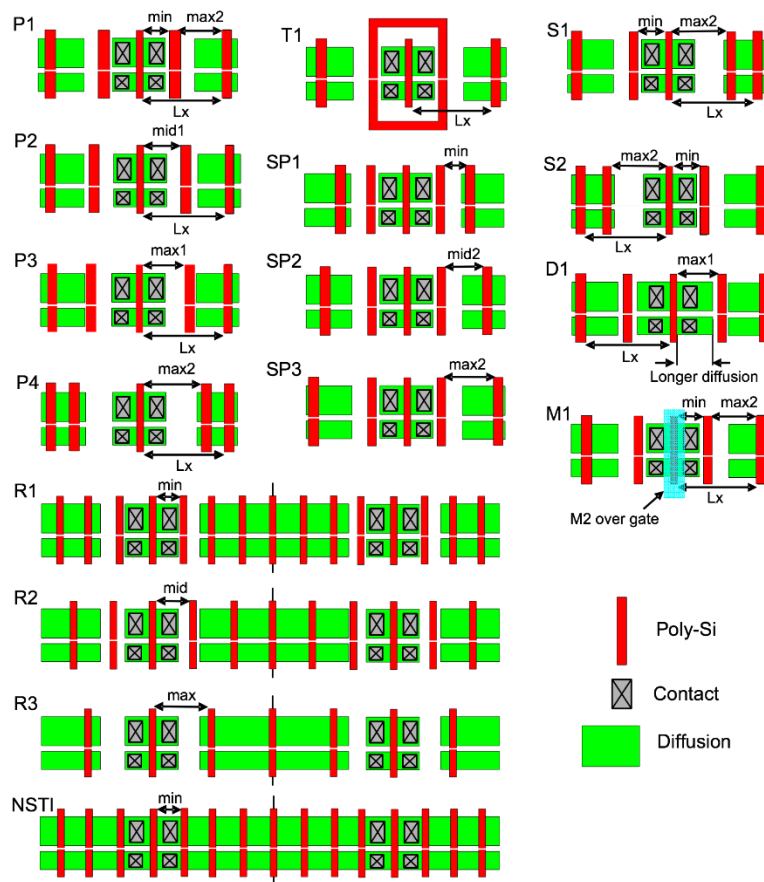

Figure 3.17: Sixteen pre-OPC RO layout configurations in the 45nm test chip, all arranged horizontally (An additional configuration using the same design of layout P1 but arranged vertically is shown separately in Figure 3.18)
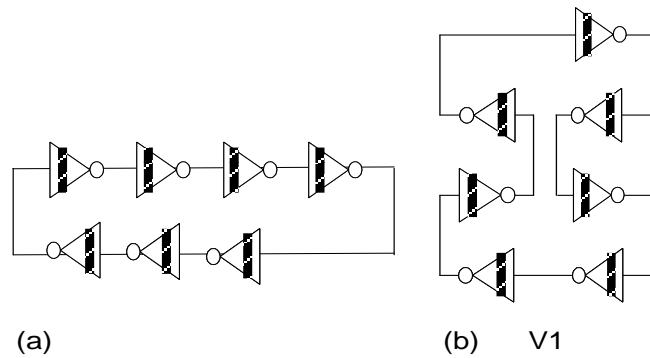
Figure 3.18: Two different RO implementations of the layout pattern P1: (a) horizontal arrangement, (b) vertical arrangement

SRAM is known to be sensitive to process variation, especially threshold voltage variations caused by random dopant fluctuation, line-edge roughness, work function fluctuations and etc. To characterize the variability of SRAM in a modern semiconductor process, SRAM test structures were also incorporated in these 45nm test chips. Each test chip contains 18 SRAM macros, and each macro contains 20 rows $\times$ 40 columns of SRAM cells, as shown in Figure 3.22. Along the diagonal of each macro, 20 bit-cells have all their internal nodes accessible through a switch network (Figure 3.20), thus allowing the automated measurement of SRAM functional metrics as well as the electrical characteristics of each of the 6 individual transistors in a bit cell.

Typical SRAM functional metrics consist of read stability and write stability, which stand for the amount of disturbance bit cells can withstand without accidental change of the data stored during a read cycle or a write cycle, respectively. The read stability is usually characterized by the Read Static Noise Margins (RSNM), which is extracted by measuring a pair of voltage transfer characteristics (VTC), more commonly known as the "butterfly curves" [62]. The RSNM is quantified as the largest square that can fit into the pair of read VTC from the same bit cell. Meanwhile, SRAM write stability can often be represented by the writeability current (I_W), which is extracted from the N-curve for writeability [63]. I_W is defined as the minimum current past the inverter trip point (the sudden drop in current in the N-curve). Figure 3.13 illustrates both setups for characterizing SRAM cell-design margins.
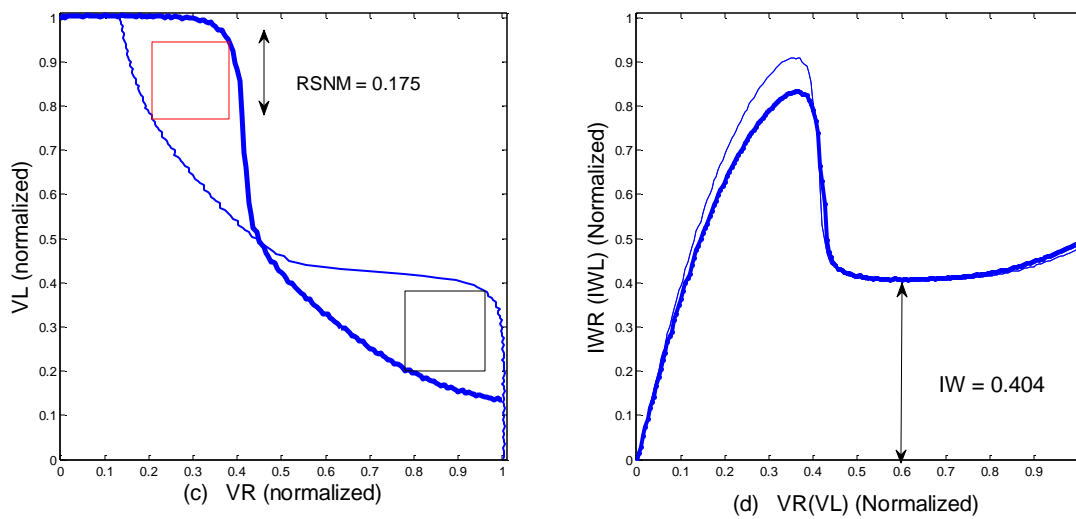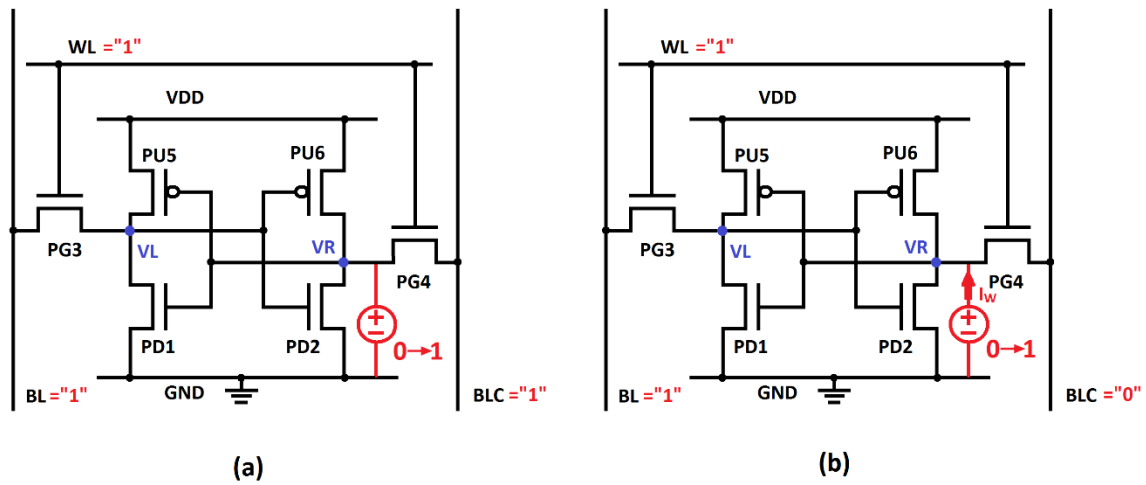
Figure 3.19: (a) Bit cell measurement setup for the "butterfly curve" to extract the Read Static Noise Margin [63] (RSNM), (b) measurement setup for the "N-curve" to extract the writeability current (I_W) [64], (c) butterfly curve with its corresponding measurement highlighted, (d) N-curve with corresponding measurement highlighted
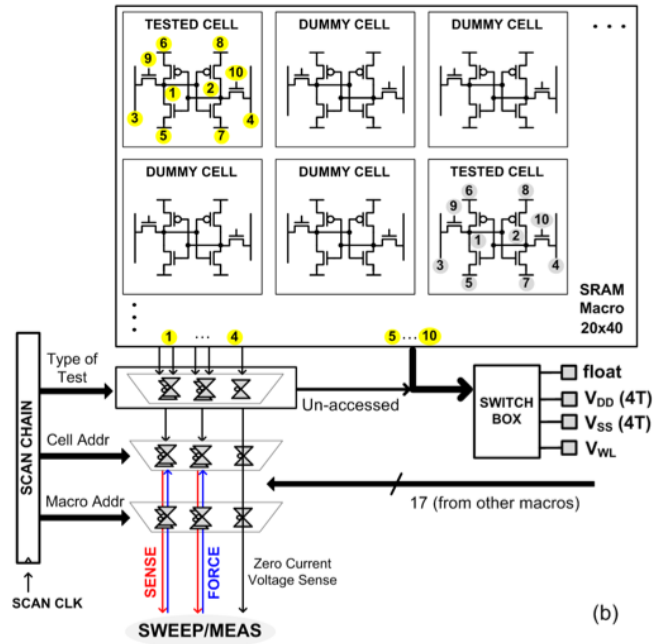
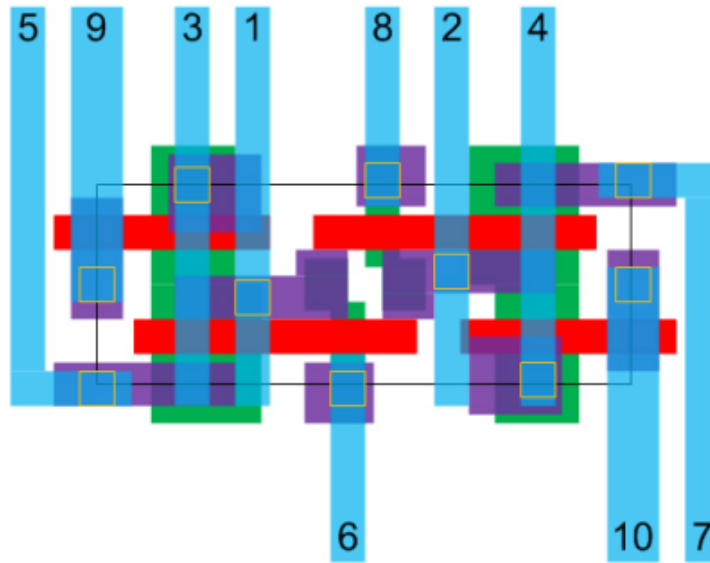Figure 3.20: All-internal-node access scheme in SRAM macros [56]

Figure 3.21: Layout cartoon for a 0.374 μm^2 bit cell with all 10 internal nodes wired out (*Courtesy: Zheng Guo, UC Berkeley*)
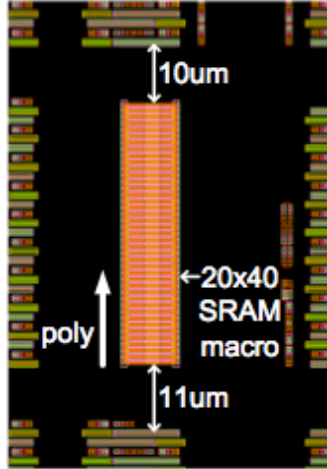
Figure 3.22: Layout view of a 20 ×40 SRAM macro, with poly in the vertical direction, using all internal node access surrounded by a large STI [56]

3.3.2 Ring Oscillator Variability Observation

A total of 76 dies from the 2 wafers are packaged for characterization. Bearing the systematic across-wafer variation profile in mind, more emphasis is put on the dies near the periphery and the center of the wafer for better coverage of the leading and trailing edge of the performance distribution. At least 8 tiles of the 18 ×16 RO array are measured at each die site, while full-array characterization had been done for 15 selected dies, as shown in Figure 3.23.

The within-chip RO variability averaged over 15 fully characterized dies is shown in Figure 3.25. The figure shows that there is no strong systematic across-chip variation. Thus, it is reasonable to estimate full within-chip statistics from a random sample of locations within a die. A simple decomposition of the variability (see Figure 3.31) shows that within-chip variation (~2%) is relatively small compare to wafer-level variations (20~30%). Hence, even a small sample of devices from the die should suffice to estimate the chip median and the across-wafer variability accurately. We chose to measure only 8 sites per chip for the majority of chips, which saves a significant amount of characterization time without noticeably compromising the accuracy of estimates of across-chip variability.

Wafer-level RO delay and leakage (NMOS and PMOS) variation of layout P2 are shown in Figure 3.24. Each data block stands for the mean delay/leakage over the measured tiles, which is noted by symbol $D\langle -DWP \rangle$, $I_{LEAKN}\langle -DWP \rangle$, and $I_{LEAKP}\langle -DWP \rangle$. Similarly, chip-level RO delay and leakage variation are shown in Figure 3.25. Each data

40

block stands for the average value of the 15 fully characterized dies, which is noted by symbol $D\langle T-WP\rangle$, $I_{LEAKN}\langle T-WP\rangle$, and $I_{LEAKP}\langle T-WP\rangle$.

The basic assumptions about the composition of variations are very similar to those described in Section 3.2, used to analyze the 90nm technology. The across-wafer RO delay variation can be approximated adequately by a second-order polynomial, as shown in Equation 3.14. The across-chip variation can be approximated adequately by a linear surface, as shown in Equation 3.15. In modern processes, two major sources contribute to the across-wafer systematic variation. First, during post-exposure-bake (PEB), the wafer temperature is non-uniform during the rapid heating step [65]. Second, during plasma etching, higher temperatures near the center of the wafer typically cause over-etch, leading to faster devices [66]. Both may cause the gate critical dimension (gate CD) to have a bull's-eye pattern across the wafer.

$$D\langle -DWP\rangle = D\langle -DWP\rangle_{AW} + D\langle -DWP\rangle_{AWR}$$

$$D\langle -DWP\rangle_{AW} = a_W(X_W - X_0)^2 + c_W \times (Y_W - Y_0)^2 + e_W$$

$$(3.14)$$

$$D\langle T-WP\rangle = D\langle T-WP\rangle_{AC} + D\langle T-WP\rangle_{ACR}$$

$$D\langle T-WP\rangle_{AC} = 0 \times X_C^2 + 0 \times X_C + 0 \times Y_C^2 + d_C Y_C + e_C$$

$$(3.15)$$

Still using RO delay as an example, the fitted coefficients and their 95% confidence intervals of all 17 layouts are shown in Figure 3.26 and Figure 3.27. As was the case for the 90nm test chip results, the confidence intervals for the "shape parameters" of both wafer-level (a_W, c_W) and chip-level systematic variations (d_C) overlap across all layouts. Again we treat these parameters as equal even without rigorous statistical proof. Most layout-dependent effects are thus captured by the intercept terms e_w and e_C, and we model the layout-dependent component in the 45nm process as an additive term in addition to the systematic across-wafer and across-chip components. For RO delay variability, the devices showing the strongest layout-dependent effects were layout #10 (D1), which features the largest diffusion width, and layout #17, which has the vertical RO placement.

To better understand the underlying mechanisms, we compare the layout effect components from the RO delay analysis as well as from the NMOS and PMOS leakage data. We focus on the intercept term from the within-chip fitting of RO delay and leakages for all 17 layouts, as in Figure 3.28. We can see that the layout dependence of the RO delay and log NMOS leakage are both significant and strongly correlated, while PMOS leakage

shows little layout dependence and does not correlate to the RO delay or NMOS leakage. This suggests that the layout-to-layout gate length variation might not be the actual source of variability since NMOS and PMOS are both subject to gate length related effects. A more plausible explanation is that the threshold voltage depends on the layout pattern. One such mechanism is the STI stress, which causes NMOS Vt to decrease and the mobility to increase with a larger length of diffusion (LOD) and smaller STI width [67], [68]. PMOS, however, is not as sensitive to stress effects due to the <100> channel orientation of this specific 45nm process. This can explain the higher speed and higher NMOS leakage for layout #10 (D1). Further investigation would require access to the internal transistors, which is not possible with this chip.

Overall, the variability of 45nm RO delay (or leakage) can be well summarized as the sum of a strong layout-to-layout-dependent component, a strong across-wafer paraboloid "bowl" (or "dome"), a smaller chip-to-chip Gaussian random noise, and a within-chip site-to-site Gaussian random noise of similar magnitude as the chip-to-chip random noise. The across-chip systematic component is negligible.

The same methodology as described in Section 3.2.3 is applied to compare the simple "Global+Local" model against the hierarchical variability model in this 45nm process. Distributions of delay and leakage from 10,000 chips with 8 tiles each are simulated in accordance with the actual measurement scheme with emphasis on the across-wafer variability and less so on the within-chip variations. Results of the Monte Carlo experiment are shown in Figure 3.32. Due to the strong systematic across-wafer variability, the delay distribution has a long tail on the slower end. The "Global+Local" model does not capture this behavior nearly as well as our hierarchical model: at -3σ, the estimate based on the simple model is as much as 18% lower than the measured delay, while the hierarchical model is consistently within 5% of the measurement at all key quantiles (Table 3.4). This is shows how ignoring systematic variability will bias estimates of the total variation in the process, possibly leading to pessimistic designs.
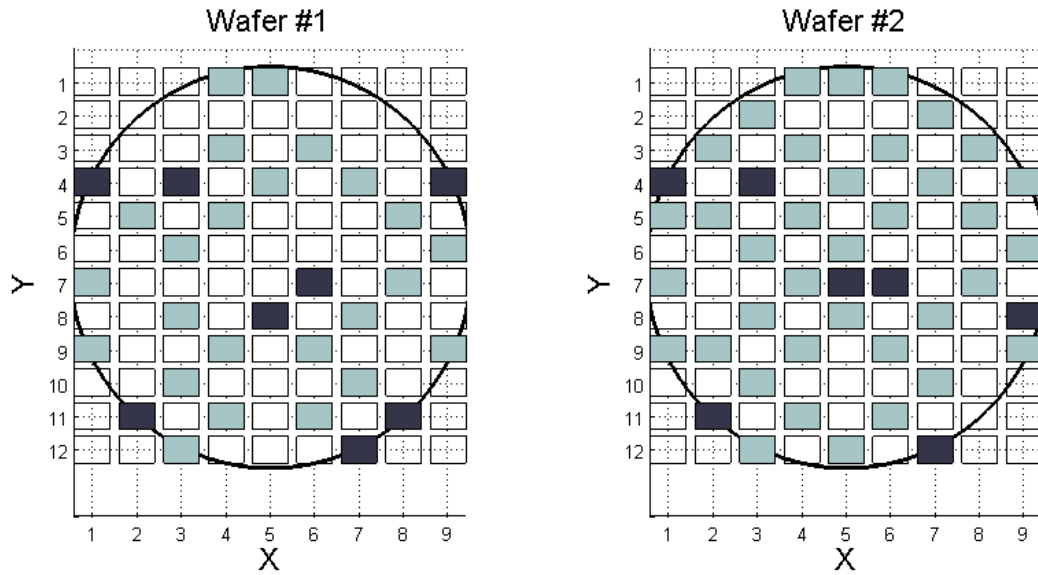
Figure 3.23: Wafer maps of RO delay/leakage measurements: dark tile = full characterization with 288 sites per die per layout; light tile = sparse characterization with 8 sites per die per layout
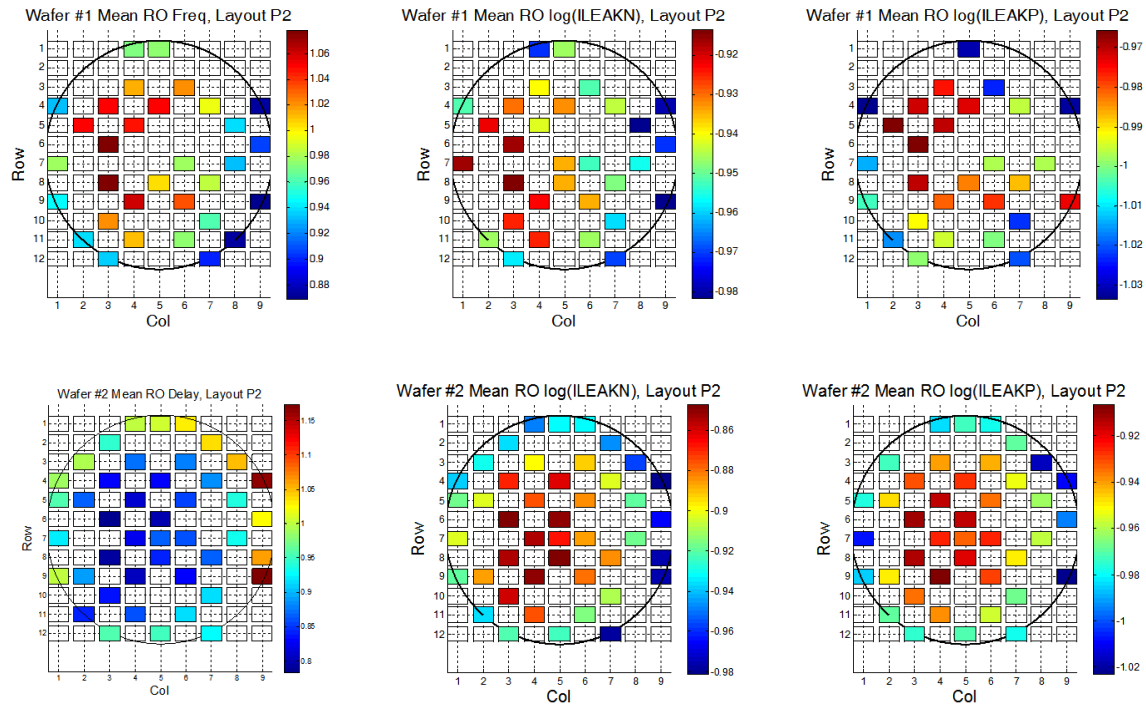
Figure 3.24: Wafer maps of mean RO delay, mean Log($I_{LEAK,N}$), and mean Log($I_{LEAK,P}$) for layout pattern P2: $D\langle -DWP \rangle = D\langle -DWP \rangle_{AW} + D\langle -DWP \rangle_{AWR}$

Figure 3.25: Chip maps of mean RO delay, mean Log($I_{LEAK,N}$), and mean Log($I_{LEAK,P}$) for layout pattern P2: $f\langle -DWP \rangle = f\langle -DWP \rangle_{AW} + f\langle -DWP \rangle_{AWR}$

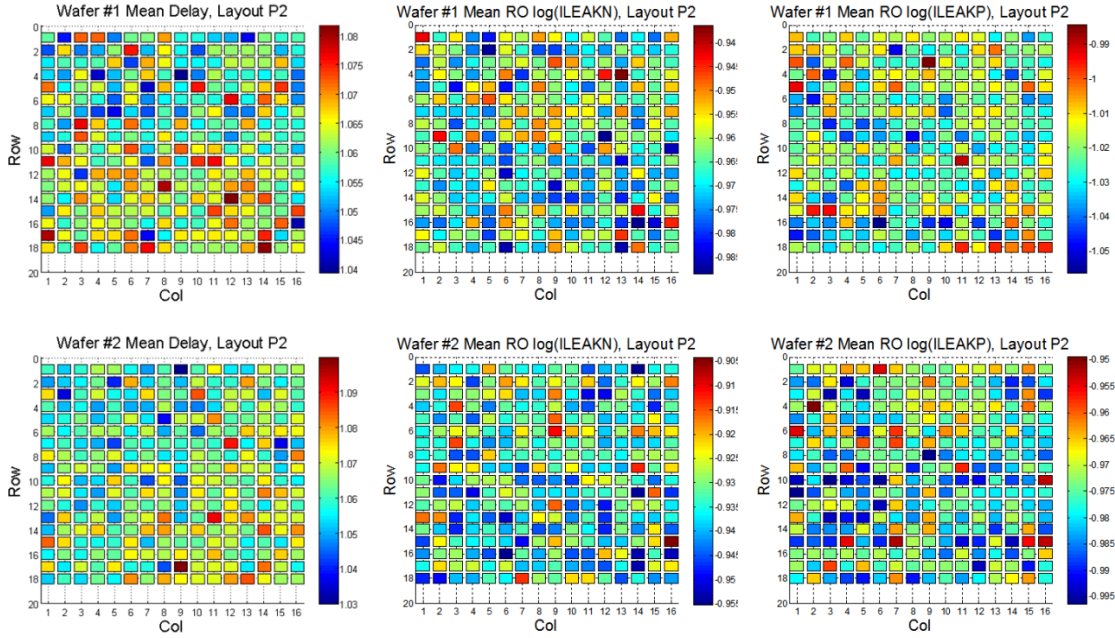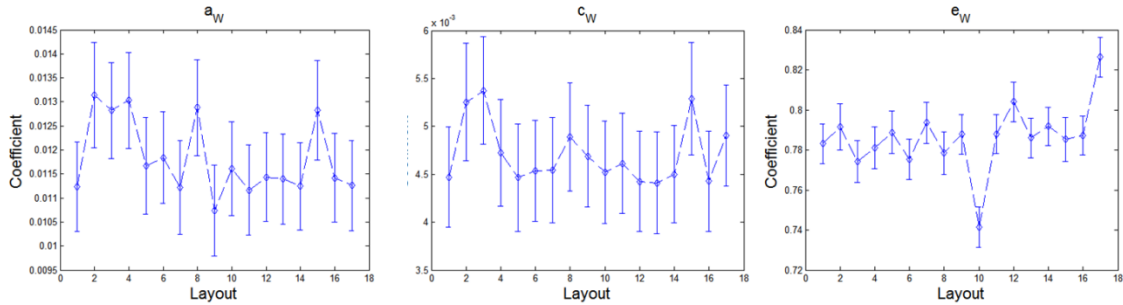Figure 3.26: Estimate and confidence interval of across-wafer fitting coefficients for all 17 layouts on wafer#2
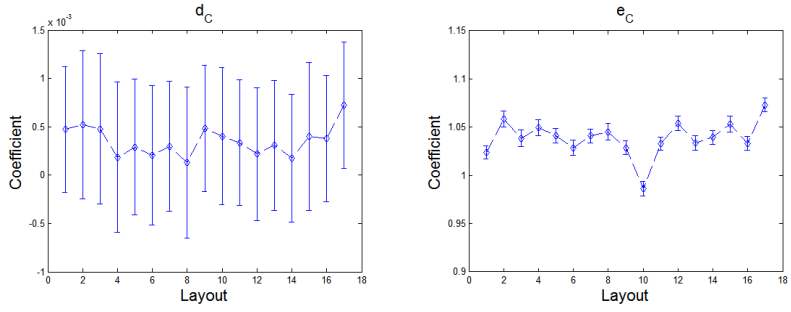
Figure 3.27: Estimate and confidence interval of across-chip fitting coefficients for all 17 layouts on wafer#2
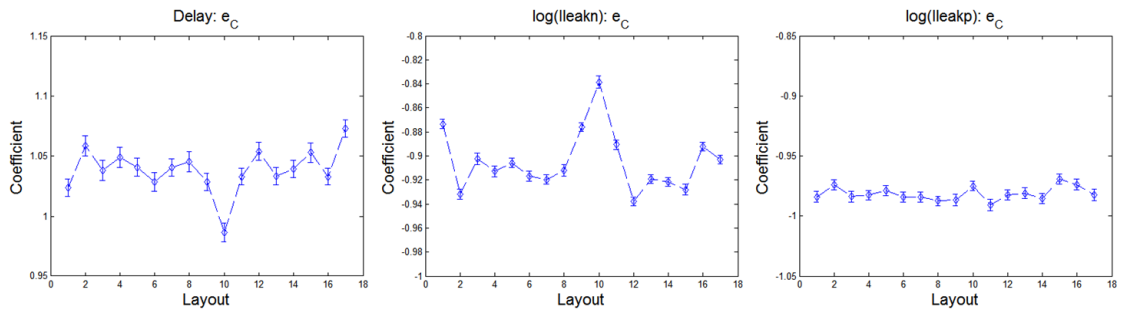

Figure 3.28: Comparison of layout dependence of RO delay, $\log(I_{LEAKN})$, and $\text{Log}(I_{lEAKP})$ on wafer#2
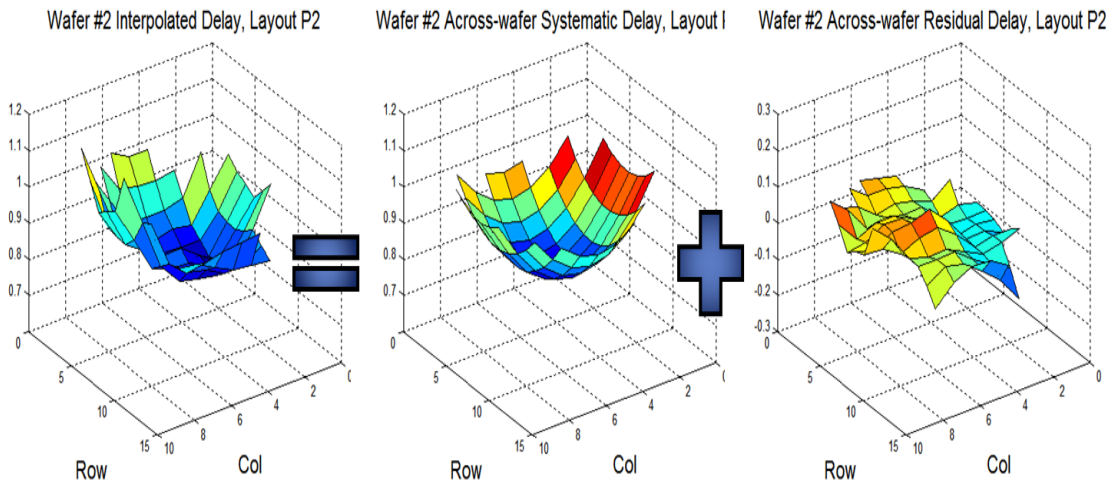

Figure 3.29: Wafer-level RO delay variation decomposition of layout pattern P2 on wafer #2:
$$D\langle -DWP \rangle = D\langle -DWP \rangle_{AW} + D\langle -DWP \rangle_{AWR}$$

Figure 3.30: Chip-level RO delay variation decomposition of layout pattern P2 on wafer #2: $D\langle -DWP \rangle = D\langle -DWP \rangle_{AW} + D\langle -DWP \rangle_{AWR}$
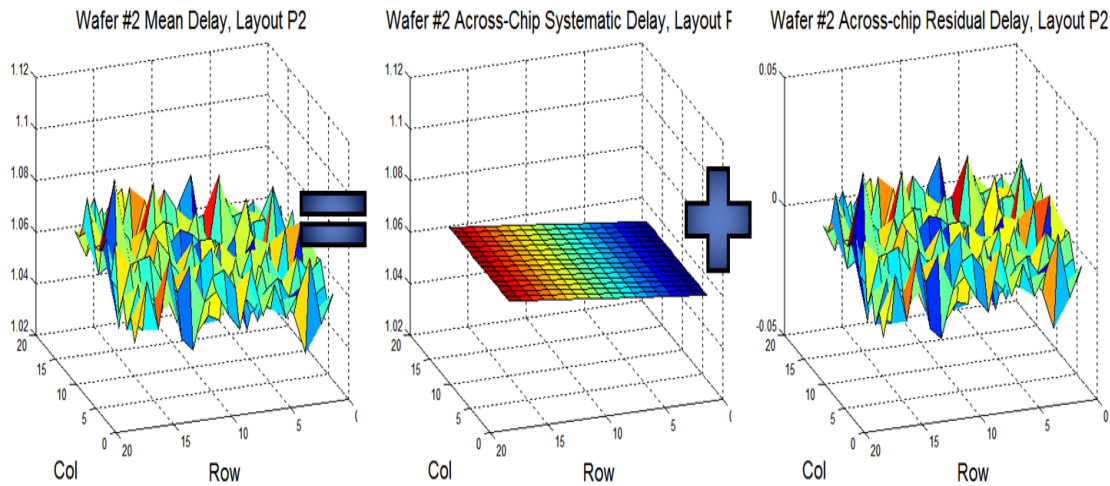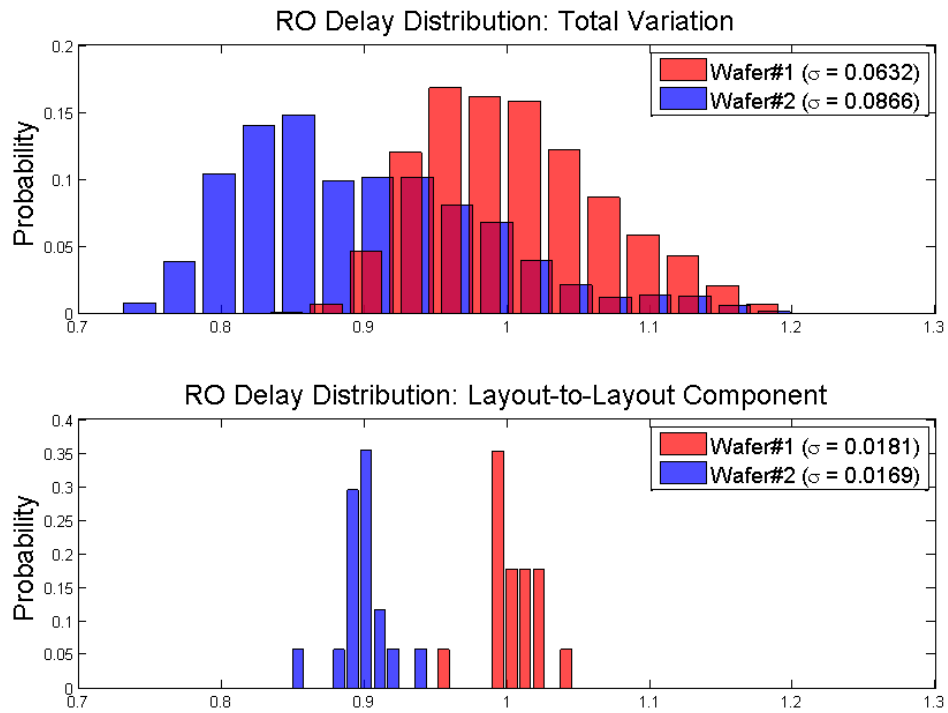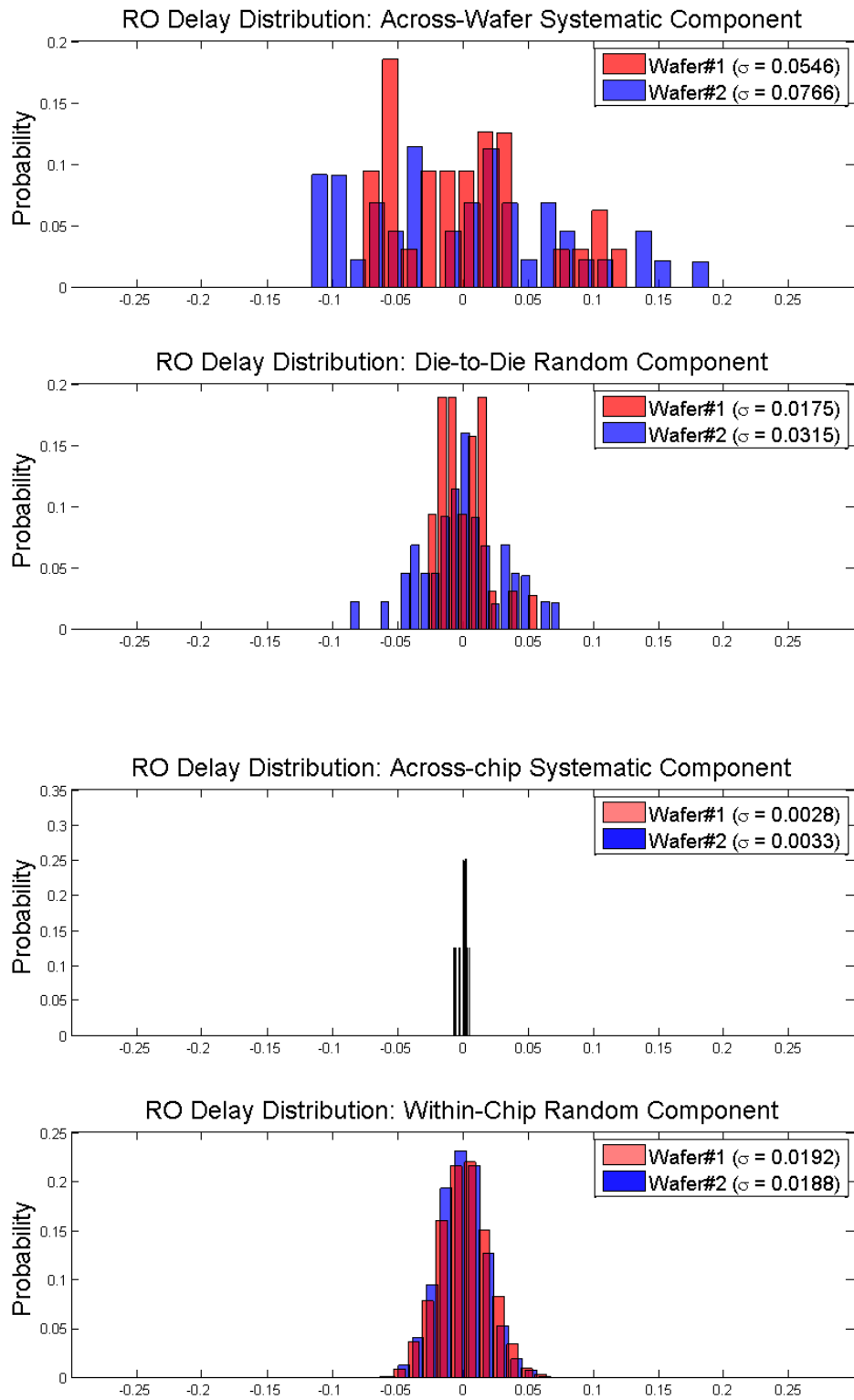
Figure 3.31: Histogram of the RO delay distribution as well as the systematic across-wafer, layout-to-layout, random chip-to-chip, systematic across-chip, and random tile-to-tile variability
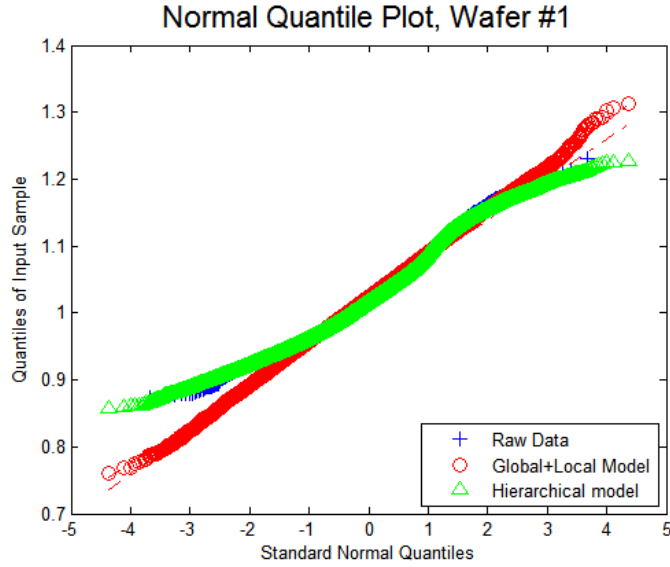
47

Figure 3.32: Comparing the prediction accuracy of the "Global+Local" model versus the hierarchical model for wafer #1 RO delay distribution

		Measurement	"Global+Local" Model	Hierarchical Model
Wafer #1	$+3\sigma$	1.206	1.217 (+0.9%)	1.194 (-1.0%)
	Median	1.017	1.023 (+0.7%)	1.017 (0%)
	-3σ	0.880	0.824 (-6.4%)	0.889 (+1.0%)
Wafer #2	$+3\sigma$	1.210	1.215 (+0.4%)	1.158 (-4.3%)
	Median	0.902	0.917 (+1.7%)	0.915 (+1.4%)
	-3σ	0.753	0.618 (-18.0%)	0.723 (-4.0%)

Table 3.4: Median and +/- 3s of simple "Global+Local" model and the hierarchical variability model in comparison with the raw measurements

3.3.3 SRAM Variability Observation

In addition to the RO variability, this 45nm test chip also provided variability measurements from the SRAM bit cell arrays and the individual padded-out transistors. Full *Id-Vg* and *Id-Vd* curves are collected for each of the 6 transistors as shown earlier in

48

Figure 3.19. The read static noise margin (SNM) and writeability current (I_W) are measured for each bit cell. Quality data were collected from 50 chips from the 2 wafers available. Due to the within-chip stripe pattern (explanations following) and the limited resources, only 20 cells on the top half of the 2 central columns are measured except for 3 chips, where all $18 \times 20 = 360$ SRAM cells are characterized.

Figure 3.33 through Figure 3.38 illustrate the average wafer and chip map of the measured transistor on-current *Idsat*, read static noise margin RSNM, and writeability current I_W. Unlike the RO variability, the SRAM transistors and cells do not show strong across-wafer systematic variations. However, the on-current of the four NMOS transistors consistently show a significant within-chip stripe pattern that is higher in the top and bottom rows but lower in the middle. Similar behavior can be observed in the RSNM and I_W chip map as well (the high/low is flipped for the RSNM). This can be explained by the fact that both the RSNM and I_W are functions of the transistor threshold voltages and the relative strength of the pull-down (PD), pull-up (PU), and pass gates/access transistors (PG). Larger RSNM requires strong pull-down transistors and weak access gates, while to achieve high write stability, one needs strong pull-up PMOS transistors and weak pass gate transistors. Notice the conflicting demand on the driving strength of the access transistors (PG): the within-chip pattern exactly predicts its positive correlation with the writeability current I$_W$ and negative correlation with the RSNM.

In the hierarchical model, the same paraboloid across-wafer systematic variation is included even though it is estimated to be insignificant, while a half-tube shaped variation that changes along the rows is included to model across-chip variation (equations 3.16 and 3.17). The decomposition of the *Idsat* variability shows that the vast majority of the variation comes from the within-chip random component, while the across-chip systematic variability is greater than the across-wafer component (Figure 3.43, Figure 3.44). The fact that the Gaussian random noise dominates the variability of transistor metrics such as on-current as well as bit cell read/write noise margins naturally leads to the result that the overall statistical distribution of these measured characteristics is very close to Gaussian distribution. As shown in Figure 3.45 and Figure 3.46, both the conventional "Global+Local" model and the hierarchical model are equally good in predicting the statistics of the measurement data.

Even though the hierarchical model is no more accurate than a simple model when random variability dominates, decomposing the variability into components still helps reveal some of the underlying mechanisms in the process. The systematic across-chip on-current variation only shows up for NMOS devices (pull-down and pass gates), which is similar to the RO case where the layout-dependence effect is only significant for NMOS leakage. This can be explained by the large STI surrounding the SRAM test block, as

depicted earlier in Figure 3.22. As mentioned in Section 3.3.2, the NMOS transistors are sensitive to stress, in this case exerted by the surrounding STI, while PMOS is insensitive due to the <100> channel direction. This lines up well with the observation that the closer to the edge of the array, the greater the NMOS drive current. The *Vtlin* chip map shown in Figure 3.36 suggests that this drive-current enhancement is not due to the threshold voltage shift, leading to the conclusion that the STI stress effect is playing its role via the mobility enhancement.

$$I\langle -DWP \rangle = I\langle -DWP \rangle_{AW} + I\langle -DWP \rangle_{AWR}$$

$$I\langle -DWP \rangle_{AW} = a_W(X_W - X_0)^2 + c_W \times (Y_W - Y_0)^2 + e_W$$

$$(3.16)$$

$$I\langle T - WP \rangle = D\langle T - WP \rangle_{AC} + D\langle T - WP \rangle_{ACR}$$

$$I\langle T - WP \rangle_{AC} = 0 \times X_C^2 + 0 \times X_C + c_C \times Y_C^2 + d_C Y_C + e_C$$
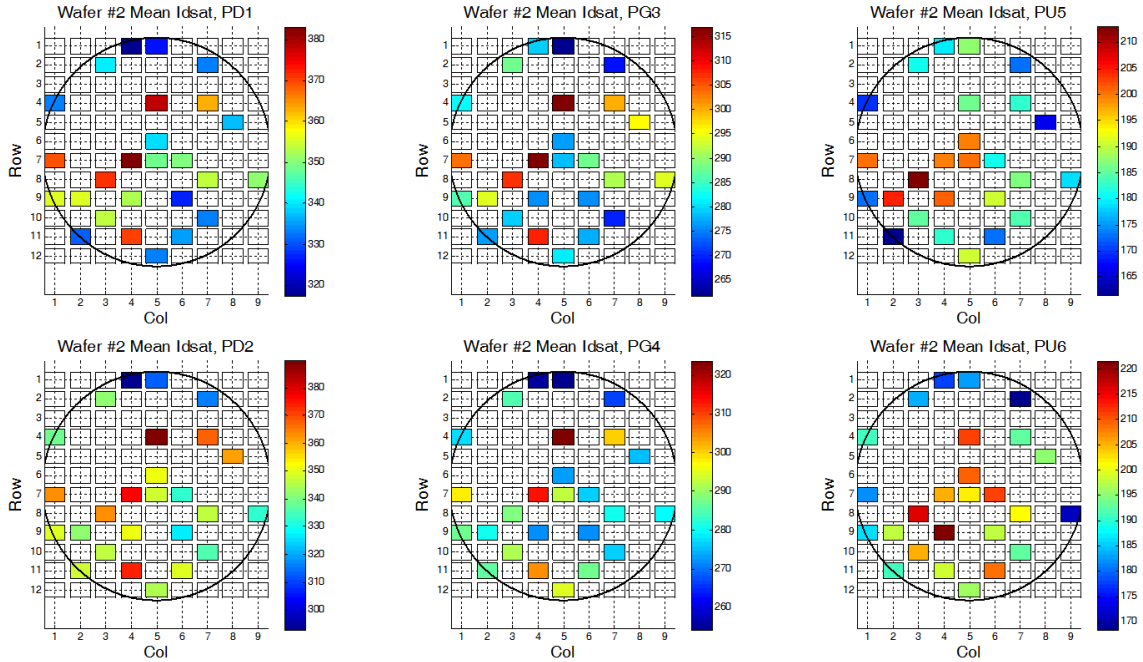
$$(3.17)$$

Figure 3.33: Wafer maps of mean on-current for SRAM padded-out transistors: $Idsat\langle -DWP \rangle$
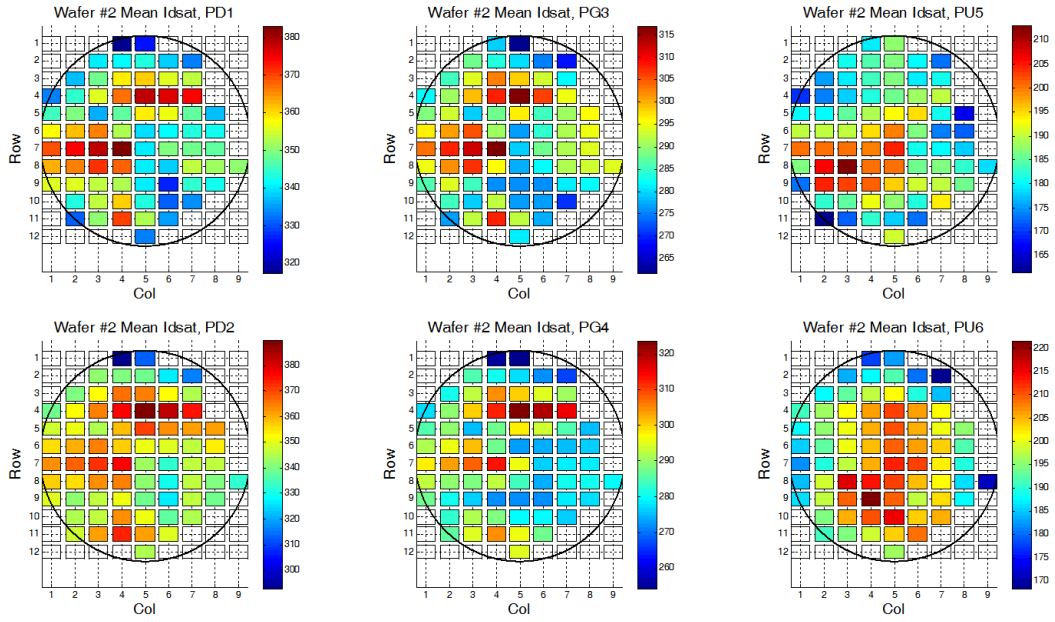
50

Figure 3.34: Interpolated wafer maps of mean on-current Idsat for SRAM padded-out transistors: $Idsat\langle -DWP \rangle$
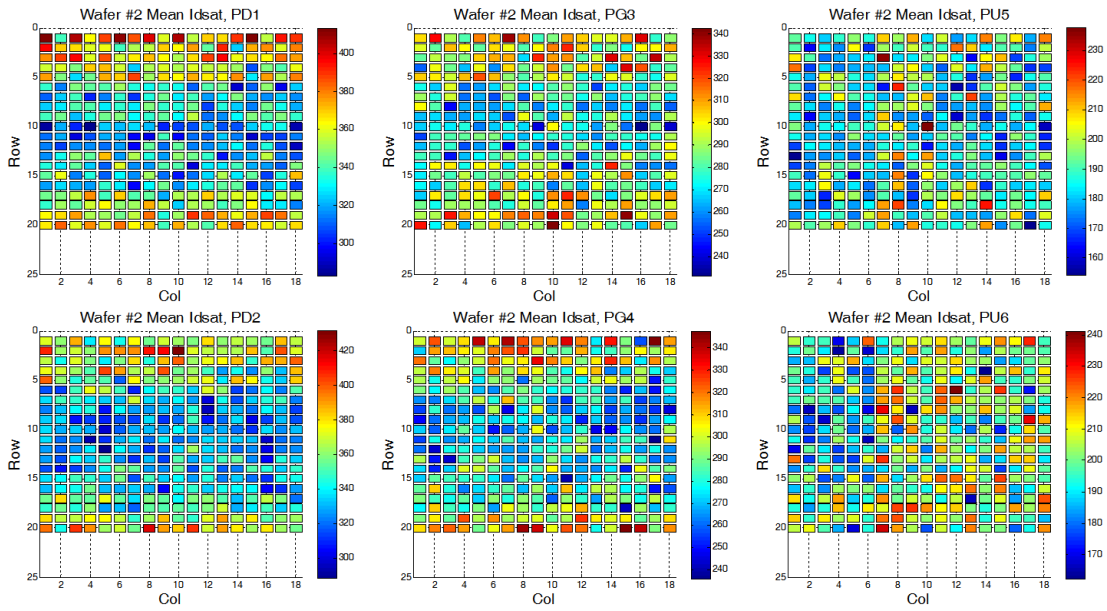
Figure 3.35: Chip maps of mean SRAM padded-out transistor $Idsat\langle T - WP \rangle$
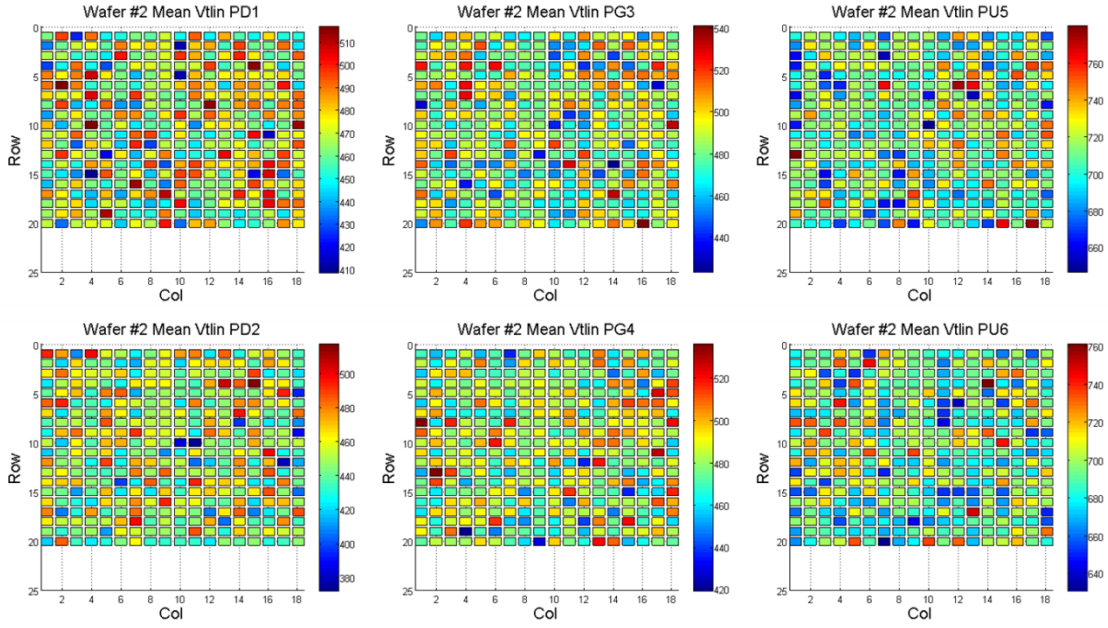
Figure 3.36: Chip maps of mean SRAM padded-out transistor $Vtlin\langle T - WP \rangle$
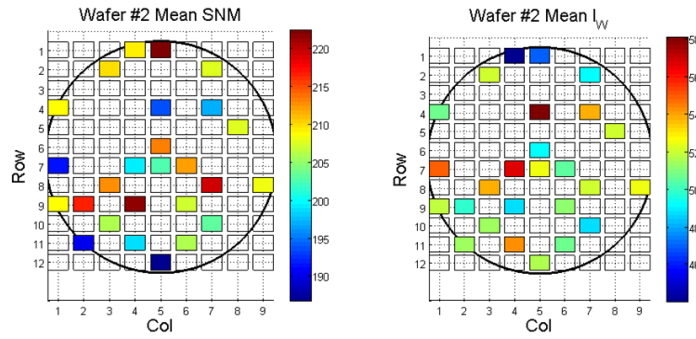
Figure 3.37: Wafer maps of mean SRAM read static noise margin $RSNM\langle -DWP \rangle$ and writeability current $I_W \langle -DWP \rangle$
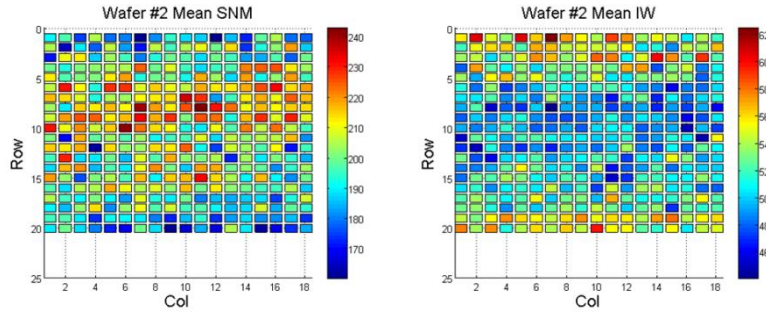
Figure 3.38: Chip maps of mean SRAM read static noise margin $RSNM\langle T - WP \rangle$ and writeability current $I_W \langle T - WP \rangle$

52
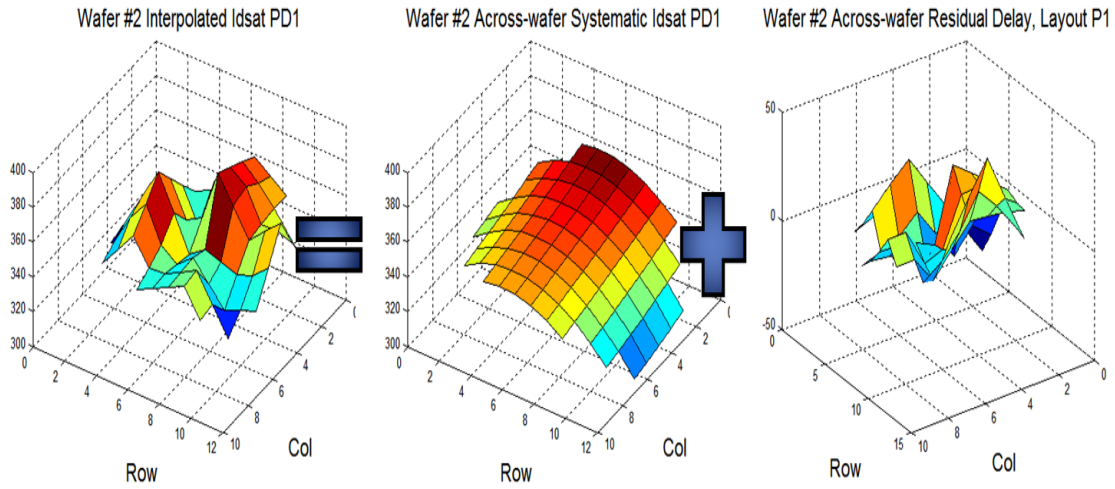
Figure 3.39: Wafer-level *Idsat* variation decomposition for left pull-down transistor on wafer #2:

$$I\langle -DWP \rangle = I\langle -DWP \rangle_{AW} + I\langle -DWP \rangle_{AWR}$$
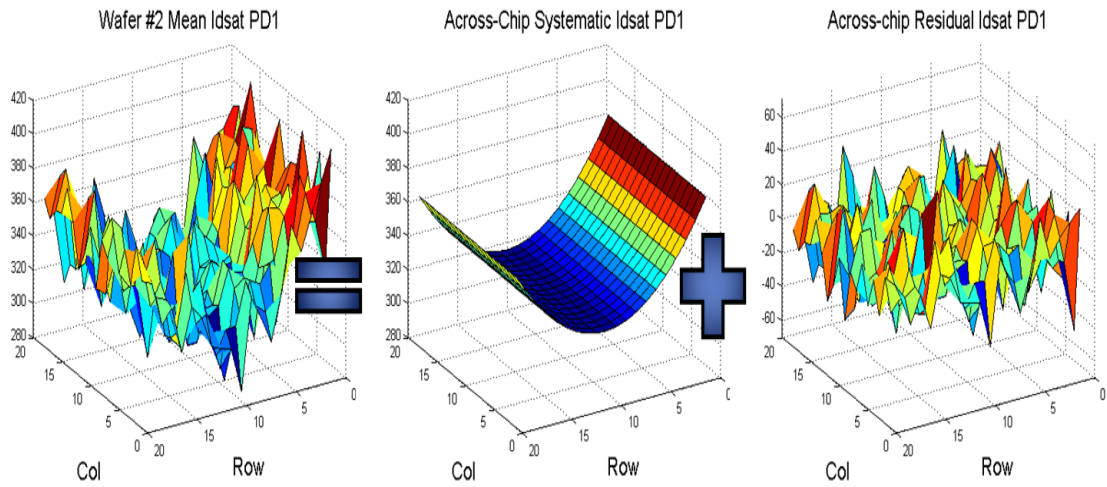
Figure 3.40: Chip-level *Idsat* variation decomposition for left pull-down transistor on wafer #2:

$$I\langle -DWP \rangle = I\langle -DWP \rangle_{AW} + I\langle -DWP \rangle_{AWR}$$

53
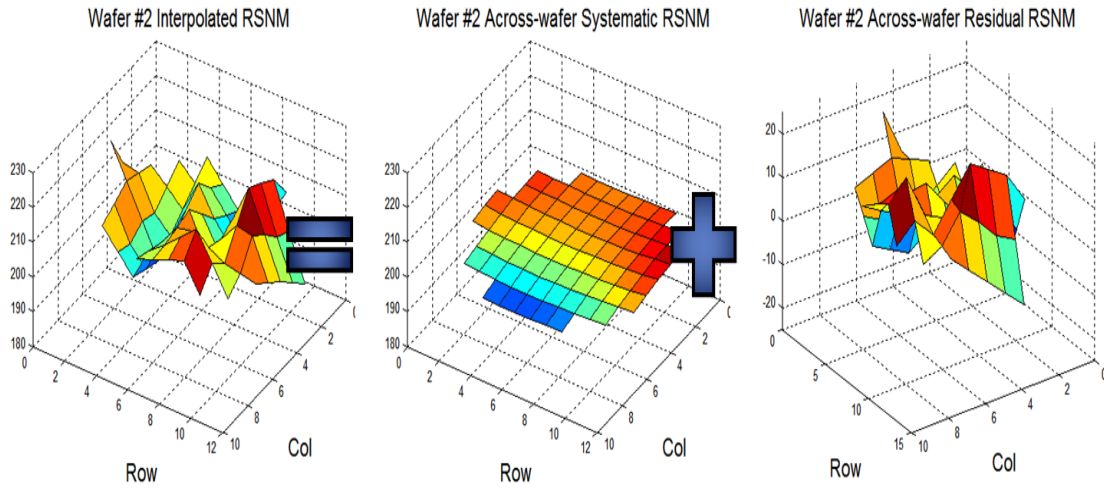
Figure 3.41: Wafer-level *Idsat* variation decomposition for left pull-down transistor on wafer #2:
$$RSNM\langle -DWP \rangle = RSNM\langle -DWP \rangle_{AW} + RSNM\langle -DWP \rangle_{AWR}$$
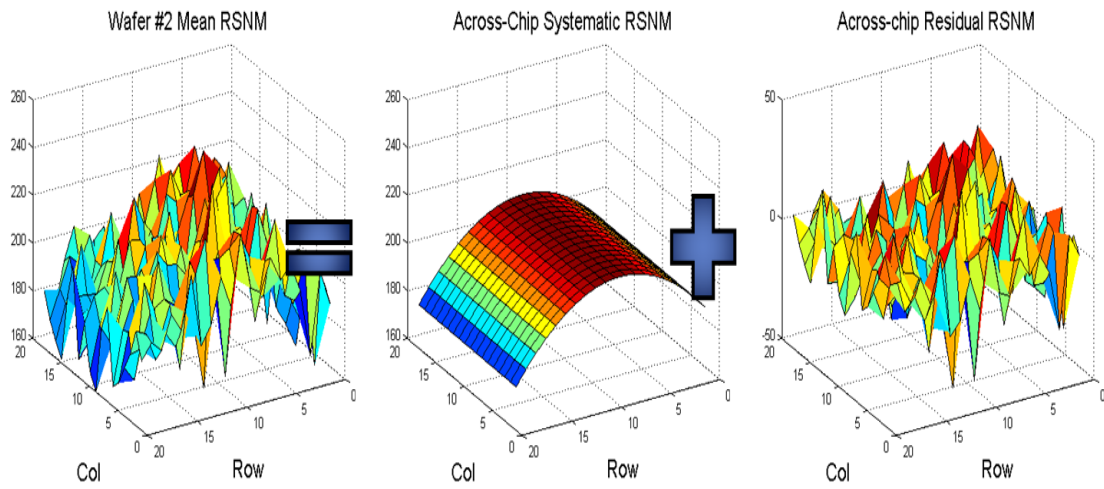
Figure 3.42: Chip-level RSNM variation decomposition for left pull-down transistor on wafer #2:
$$RSNM\langle -DWP \rangle = RSNM\langle -DWP \rangle_{AW} + RSNM\langle -DWP \rangle_{AWR}$$
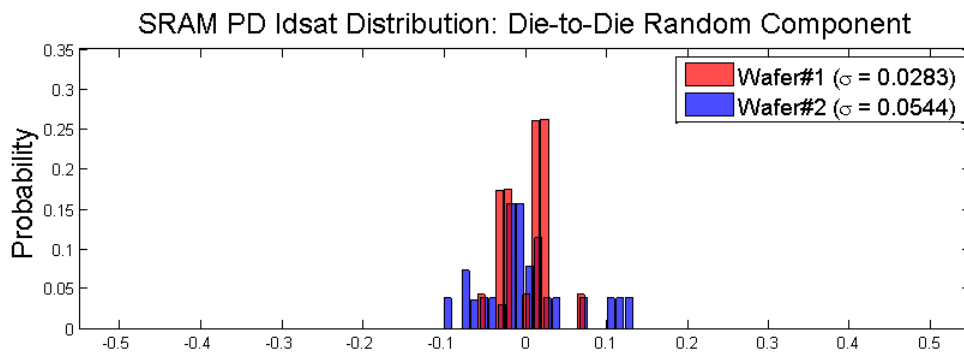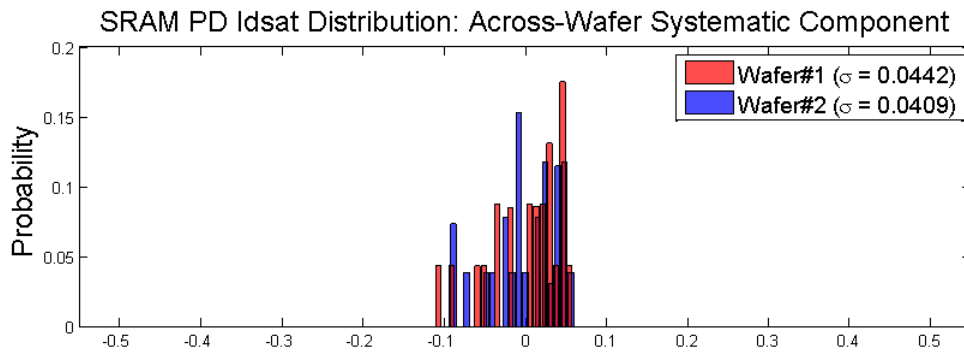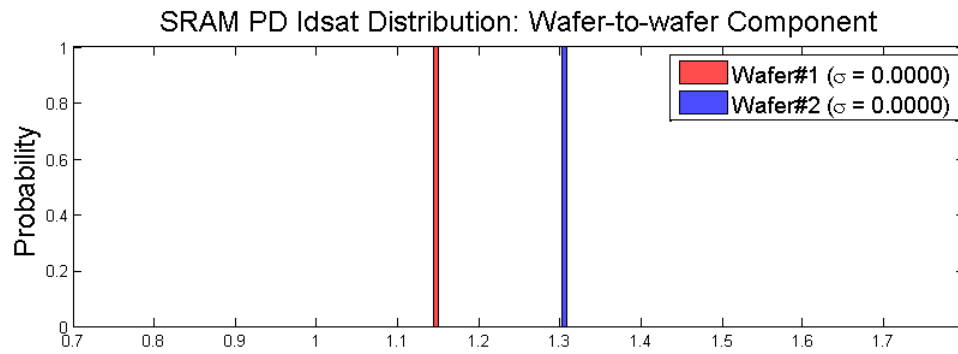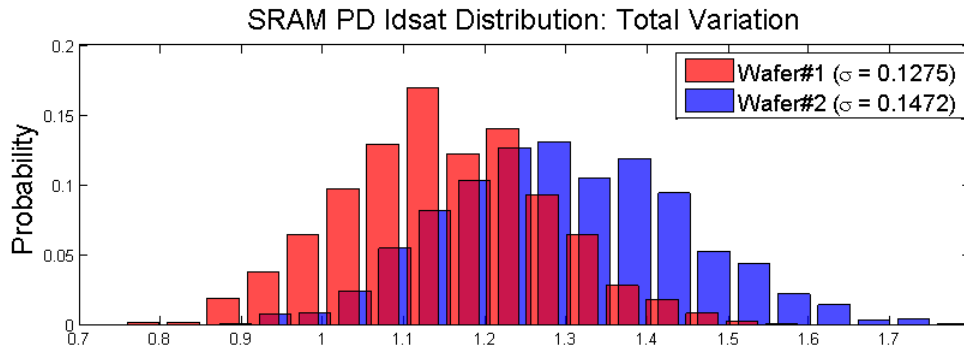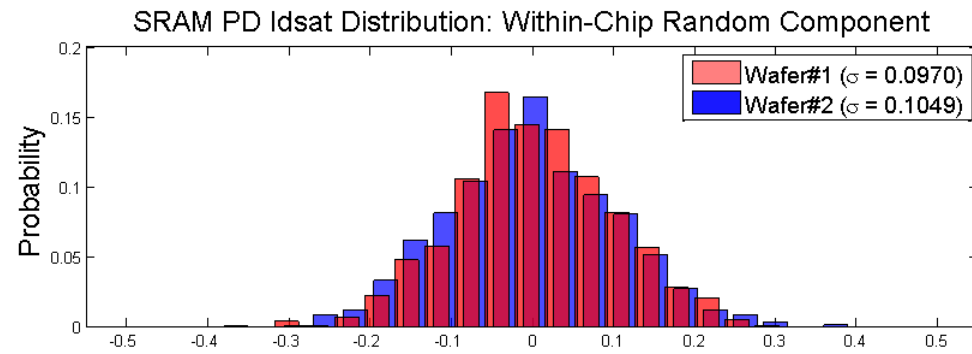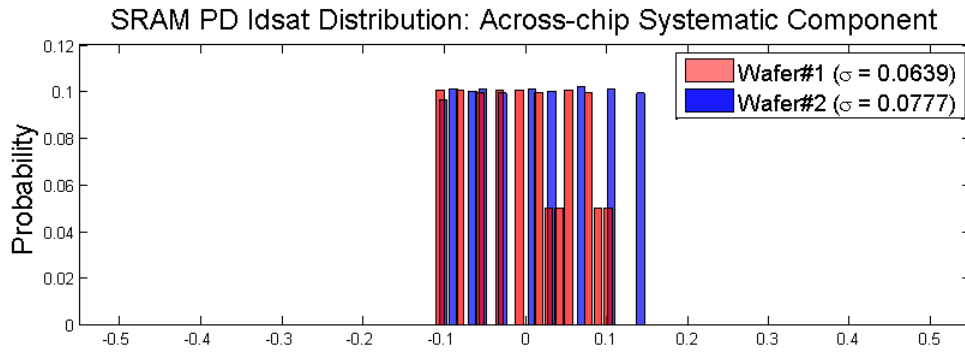
Figure 3.43: Histogram of the SRAM pull-down transistor *Idsat* distribution as well as the systematic across-wafer, layout-to-layout, random chip-to-chip, systematic across-chip, and random tile-to-tile variability

SRAM RSNM Distribution: Wafer-to-wafer Component

SRAM RSNM Distribution: Across-Wafer Systematic Component

SRAM RSNM Distribution: Die-to-Die Random Component

SRAM RSNM Distribution: Across-chip Systematic Component

57
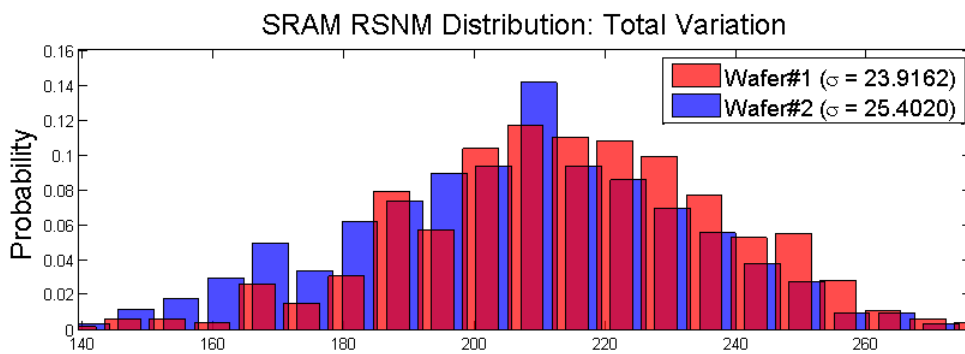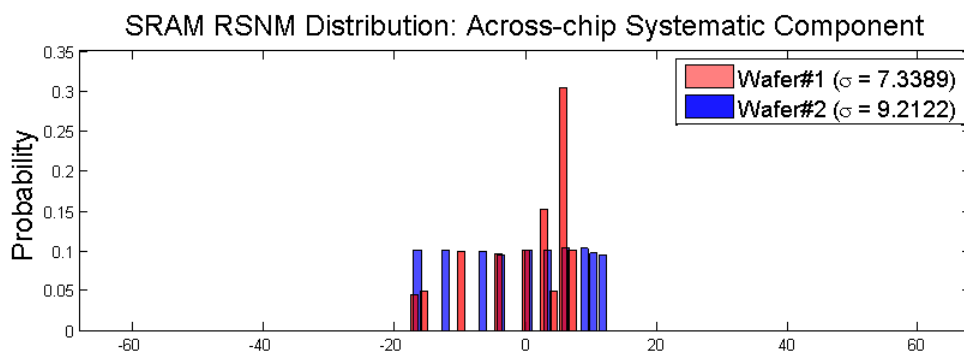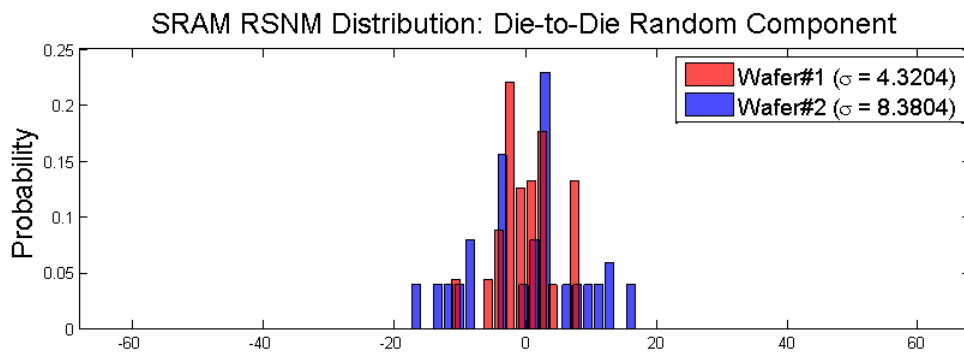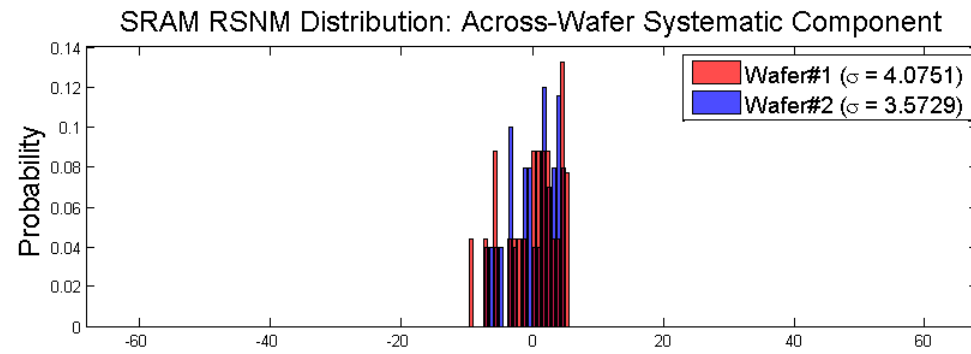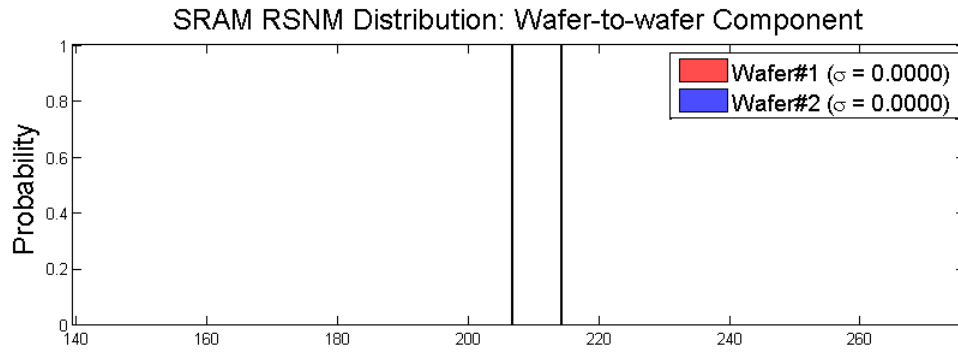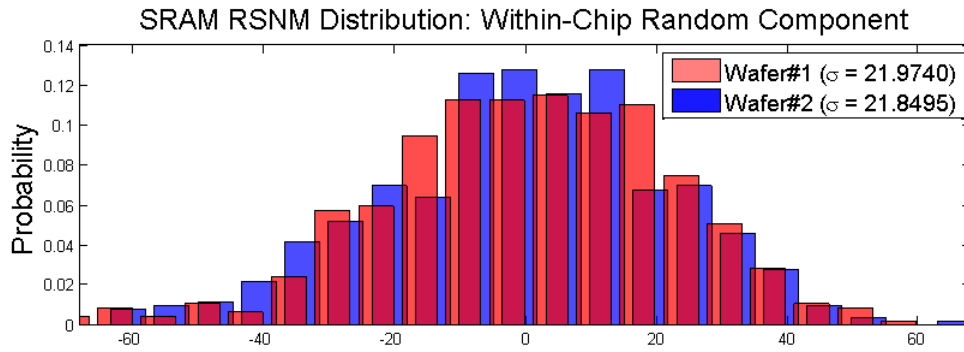
Figure 3.44: Histogram of the SRAM bit cell RSNM distribution as well as the systematic across-wafer, layout-to-layout, random chip-to-chip, systematic across-chip, and random tile-to-tile variability
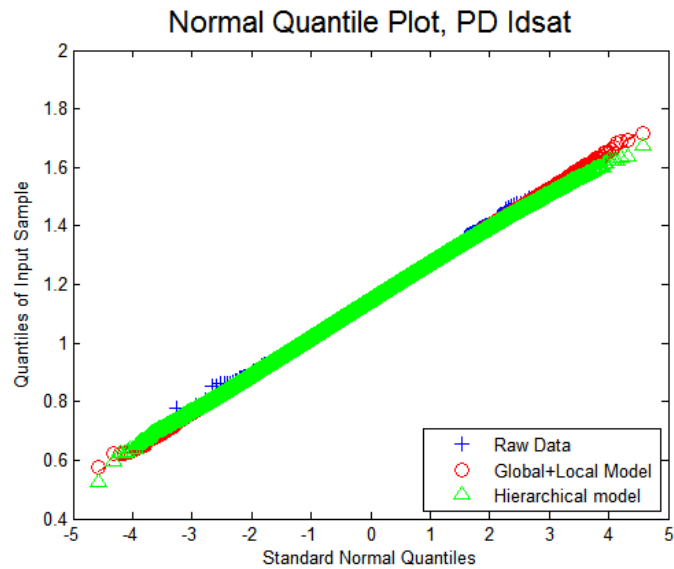
Figure 3.45: Comparing the prediction accuracy of the "Global+Local" model versus the hierarchical model for Wafer #2 pull-down transistor Idsat distribution
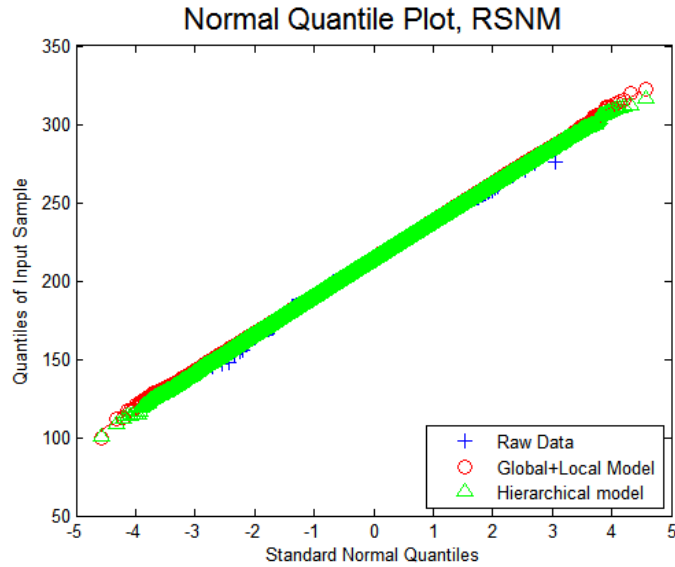
Figure 3.46: Comparing the prediction accuracy of the "Global+Local" model versus the hierarchical model for Wafer #2 pull-down transistor *Idsat* distribution

3.4 Summary

This chapter presents the design and measurement results of 90nm and 45nm technology test chips with variability-sensitized test structures, including ROs, off-state current measurement circuits, SRAM bit cell arrays, and wired-out individual transistors. Sampling and measurement plans are designed based on our hierarchical variability model so that the measurement cost in both time and packaging are minimized, while maintaining the statistical significance of the results.

Ring oscillator delay and leakage analysis on the 90nm and 45nm test chips demonstrate significant systematic across-wafer variations in a dome/bowl-like pattern. Rings with differently designed layout patterns have very similar patterns in their shapes of across-wafer and across-chip variations, with a parallel shift as a result of the layout effects. This allows us to capture the variability of the process accurately using an additive hierarchical model. Compared to the conventional methodology that decomposes total variation into global (chip mean) variation and local (within-chip) variation, the hierarchical model is clearly superior in predicting the extreme quantiles of the distribution of device and circuit performance metrics with the presence of strong systematic variability.

Analysis of the 45nm SRAM bit cell array and its internal transistors shows that the within-chip local variation dominates the variation profile. In this case, a Gaussian

distribution is sufficient to describe the variability, and the conventional "Global+Local" model is just as good as the hierarchical model in predicting the statistical distributions of transistor and SRAM metrics. Nevertheless, the across-chip variation demonstrates a systematic pattern as the distance to the top/bottom rail of the STI changes.

While physical inspection was not possible, a variability analysis of the electrical data still provides some insight into the mechanisms of the randomness. The strong across-wafer systematic variation is most likely related to the gate critical dimension variation across the wafer due to process variation during post-exposure bake or plasma etching. Starting with the 45nm technology, strained silicon plays a significant role in the layout-dependent component, which may be in the form of both threshold shift and mobility enhancement.

Chapter 4

Statistical Compact Model Parameter Extraction

4.1 Introduction

The uncertainty in the manufacturing process introduces statistical variations of MOSFET characteristics, which is a major challenge to process engineers and circuit designers. While many methods are used to reduce process variations, variation will never be completely eliminated. The ability to accurately predict the statistical characteristics of manufactured transistors is the key to optimizing circuit design for performance and parametric yield. Because MOSFET transistor characteristics are always abstracted by compact SPICE models, transistor variability will naturally be translated into compact model parameter variations. A variety of studies have been done to explore the possibility of accurately modeling statistical transistor behavior with compact model parameters [19]. In this chapter, we will leverage the statistical compact model parameter extraction procedure with automated parameter selection and the flexibility of measurement data availability.

The basics of compact model extraction will be introduced first. Then, we will explain the stepwise parameter selection methodology. The effectiveness of this methodology will be examined by applying it to the EKV model as an example of simple one-step extraction and to the industrial standard PSP model as an example of sequential extraction.

4.2 Statistical Compact Model Parameter Extraction

4.2.1 Compact Model Parameter Extraction

Once the required transistor *I-V* data or *C-V* data are acquired, one can perform either analytical regression or numerical optimization to estimate the compact model parameters. This procedure and the full set of compact model parameters are commonly referred to as compact model *parameter extraction* [73] and the "model card" respectively. For each device, one corresponding model card will be extracted from its *I-V* data.

The analytical method uses linear approximations of model equations to represent device characteristics in the limited operation space of devices [73]–[75]. Linear least

squares regression is applied to the linearized equations to estimate the parameters. Parameters estimated via this method usually have a clear physical meaning, as well as strong sensitivity in the specific operation space. As a result, only a few key model parameters can be extracted using the analytical method.

The numerical optimization method, on the other hand, estimates compact model parameters using non-linear least-square optimizations rather than linearization. Given a reasonable set of initial guesses, a set of model parameter values can be estimated by minimizing the error between the model and the measured data. However, the problem is underdetermined: in general, innumerable combinations of parameter values fit the data equally well, and many of those combinations are physically unrealistic. Imposing constraints on the optimization problem can ensure that the results are physically realistic and can reduce, but not eliminate, the indeterminacy.

In practice, a full-fledged compact model is usually generated using a combination of pre-known technological process data, the analytic method, and the numerical optimization method. Model parameters that are directly related to process conditions, such as *Cox* (gate capacitance) and *Xj* (junction depth), will be acquired from the process condition of the technology and remembered for the remainder of model generation. The analytic extraction method, while it is only applied directly to the initial analysis of the most dominant parameters, provides a guideline for a "divide and conquer" approach. Virtually every compact model has a different set of parameters specifically designed to model device behavior in various subsets of device operation space, such as the sub-threshold region, the linear operation region, or the saturated operation region, in the case of MOSFET transistors. Dividing the overall optimization problem into smaller problems and solving each numerically in its smaller parameter space, reduces the computational burden: each subproblem has fewer model parameters and smaller datasets relevant for these parameters. This is especially important for extracting model parameter distributions for a large number of devices, and it tends to produce estimates of the model parameters that are physically more reasonable.

We focus here on improving the automated numerical optimization procedure so that we not only achieve a good *fitting quality*, but also a sound *extraction quality*. Defining the *fitting quality* is simple: the goodness of fit can be quantified by in a standard way, such as sum of the squares of the residuals. The *extraction quality*, on the other hand, is trickier to define: we want the extracted compact model parameters to have statistical distributions centered at physically realistic values; the distributions should not be so dispersed that the extreme quantiles are physically unrealistic; the correlation structure among model parameters shall be as simple as possible; and we want the fewest possible model

parameters to be fitted for each device, as long as both *fitting quality* and *extraction quality* are guaranteed.

4.2.2 Basics of Optimization

Compact model extraction can essentially be established as the following non-linear optimization problem. For the total of n compact model parameters $(p_1, p_2, .., p_n)$ that need to be extracted, we can define a model parameter vector p as:

$$p = [p_1, p_2, \dots, p_n]^T \qquad (4.1)$$

The possible combinations of values for n model parameters is called an n-dimensional parameter space. The compact model equations relate functions defined on the parameter space. Assume $f(p)$ is such a function whose value (either a scalar or a vector) can be physically measured from actual devices. We also have a constant vector, y, which represents such measured characteristics (*I-V* or *C-V* data). Select a nonnegative, real-valued, continuously differentiable *objective function* $F(p)$ to measure the discrepancy between the model function and the data. Then, the optimization problem can be defined as finding an optimal value, p^*, so that $F(p)$ reaches its minimum value, $F(p^*)$.

$$F(p^*) = \min_p F(p) \qquad (4.2)$$

The most popular choice of objective function for model parameter extraction is the sum of squared residuals, which leads to least squares estimation. Suppose the model equation $f(p)$ is an m-dimensional function of the n model parameters. The least-squares objective function, $F(p)$, is then defined as follows:

$$F(p) = \frac{1}{2} \sum_{i=1}^{m} w_i [r_i(p)]^2$$

$$\qquad (4.3)$$

$$= \frac{1}{2} \sum_{i=1}^{m} w_i [f_i(p) - y_i]^2$$

Here, f_i and y_i are the ith component of the model equation and the measurement data, respectively; $r_i = [f_i(p) - y_i]$ is the fitting residual or error function; and w_i is the weighting factor for *the i*th data point. Weighting factors can be increased for specific operation regions in which accurate fitting is especially important.

Assume that first and second partial derivatives exist for all the m components of the n-dimensional objective function $F(p)$. Its first three terms of the Taylor series expansion are:

$$F(p + \Delta p) = F(p) + \sum_{j=1}^{n} \frac{\partial F}{\partial p_j} \Delta p_j + \frac{1}{2} \sum_{j=1}^{n} \sum_{l=1}^{n} \frac{\partial^2 F}{\partial p_j \partial p_l} \Delta p_j \Delta p_l + O(\|\Delta p\|^2)$$

(4.4)

$$= F(p) + \nabla F(p)^{\mathrm{T}} \Delta p + \frac{1}{2} \Delta p^T \nabla^2 F(p) \Delta p + O(\|\Delta p\|^2)$$

Here, $\nabla F(p)$ is the gradient of $F(p)$:

$$\nabla F(p) = \left[\frac{\partial F}{\partial p_1}, \frac{\partial F}{\partial p_2}, \dots, \frac{\partial F}{\partial p_n} \right]^T$$

(4.5)

And $\nabla^2 F(p)$ is the second derivative of $F(p)$, also called the Hessian matrix:

$$H(p) \equiv \nabla^2 F(p) = \begin{bmatrix} \dfrac{\partial^2 F}{\partial p_1 \partial p_2} & \cdots & \dfrac{\partial^2 F}{\partial p_1 \partial p_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 F}{\partial p_n \partial p_1} & \cdots & \dfrac{\partial^2 F}{\partial p_n \partial p_n} \end{bmatrix}$$

(4.6)

For the simple case, in which $w_i = 1$ for all $i = 1, \dots n$:

$$F(p) = \frac{1}{2} \sum_{i=1}^{m} [r_i(p)]^2$$

(4.7)

Thus, by the chain rule,

$$\frac{\partial F}{\partial p_j} = \frac{1}{2} \sum_{i=1}^{m} \frac{2 r_i \partial r_i}{\partial p_j}$$

(4.8)

$$\nabla F(p) = \begin{bmatrix} \dfrac{\partial r_1}{\partial p_1} & \cdots & \dfrac{\partial r_1}{\partial p_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial r_m}{\partial p_1} & \cdots & \dfrac{\partial r_m}{\partial p_n} \end{bmatrix}^T r(p)$$

(4.9)

$$\frac{\partial^2 F}{\partial p_j \partial p_l} = \sum_{i=1}^{m} \frac{\partial r_i}{\partial p_j} \frac{\partial r_i}{\partial p_l} + \sum_{i=1}^{m} \frac{r_i \partial^2 r_i}{\partial p_j \partial p_l} \tag{4.10}$$

The *Jacobian matrix* of $r(p)$ is defined as follows:

$$J(p) = \begin{bmatrix} \dfrac{\partial r_1}{\partial p_1} & \cdots & \dfrac{\partial r_1}{\partial p_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial r_m}{\partial p_1} & \cdots & \dfrac{\partial r_m}{\partial p_n} \end{bmatrix} \tag{4.11}$$

Then, in the vector form, we have the following:

$$\nabla F(p) = J(p)^T r(p) \tag{4.12}$$

$$H(p) = J(p)^T J(p) + \sum_{i=1}^{m} r_i(p) \nabla^2 r_i(p) \tag{4.13}$$

If the residual function $r(p)$ is negligible, we can obtain an approximation of the Hessian matrix virtually for free because its leading term can be calculated simply from the Jacobian matrix:

$$H(p) \approx J(p)^T J(p) \tag{4.14}$$

When the model extraction problem is posed as a non-linear least-squares problem, the most widely used optimization method is the gradient-based method [73], [76]. This method searches for a local minimum along the gradient of $F(p)$ using a finite step size.

We will use Newton's method as an example. Assume the Hessian matrix is positive definite with the second term negligible. The local model around p is then as follows:

$$F(p + \Delta p) = F(p) + \nabla F(p)^T \Delta p + \frac{1}{2} \Delta p^T \nabla^2 F(p) \Delta p \tag{4.15}$$

Taking the derivative over step size Δp for both sides of this equation, we obtain the following:

$$\nabla F(p)^T + \nabla^2 F(p) \Delta p = 0 \tag{4.16}$$

This is the necessary condition for the local minimum with all possible Δp. The local optimal step size Δp can then be calculated as follows:

65

$$\Delta p = -[\nabla^2 F(p)]^{-1} \nabla F(p) \tag{4.17}$$

Or

$$\Delta p = -H(p)^{-1} \nabla F(p) \tag{4.18}$$

With the new parameter vector $p + \Delta p$, we will have a new decreased objective function, $F(p + \Delta p)$. Remember the assumption that $H(p)$ is positive definite:

$$F(p + \Delta p) = F(p) - \frac{1}{2} \Delta p^T \nabla^2 F(p) \Delta p \leq F(p) \tag{4.19}$$

Repeat this procedure until the changes in objective function are smaller than the predetermined tolerance value ϵ. The entire flow can be summarized as follows:

1. Start from an initial parameter, vector p^0.
2. At the kth iteration, calculate the search step, $\Delta p^k = -H(p^k)^{-1} \nabla F(p^k)$.
3. Calculate the next step, p^{k+1} as $p^k + \Delta p^k$.
4. If $|F(p^k) - F(p^{k+1})| > \epsilon$, go to Step 2. Here, ϵ is the predetermined tolerance.
5. Terminate the calculation when $|F(p^k) - F(p^{k+1})| < \epsilon$.

To calculate the step size of Newton's method, $-H(p^k)^{-1} \nabla F(p^k)$, one would need to calculate the inverse of the Hessian matrix, which requires the invertibility of $J(p)^T J(p)$. There are various modified optimization methods, including the Levenberg-Marquardt method [77], [78]. This method regularizes the $J(p)^T J(p)$ matrix by adding a diagonal matrix, to avoid numerical instability. Trust region reflective methods [79], [80], place bounds on step sizes to ensure that the quadratic approximation is accurate at each iteration. We will use non-linear least-squares as described in this section for the rest of the thesis.

4.2.3 Backward Stepwise Parameter Selection

Industrial standard MOSFET compact models such as BSIM [11] and PSP [12] have hundreds of parameters. However, not every parameter is fitted for each device, due to the high computational cost involved in optimization problems with large numbers of variables. Furthermore, the redundancy of model parameters causes numerical instabilities for (unregularized) non-linear least-squares. Therefore, it is helpful to reduce the number of model parameters to be extracted so that only essential parameters are fitted, without sacrificing the goodness of fit.

We adopted a backward stepwise selection procedure. Starting with the n parameters of the compact device model, we fit the measurement data curves by non-linear least-squares. Suppose we have a criterion function that represents the "goodness" of each of the extracted parameters, or the so-called *extraction quality*. As long as the current round of extraction provides decent *fitting quality*, the "worst" parameter will be removed from the extraction and be set to a proper constant value. With the reduced parameter set containing $n-1$ parameters, we repeat the same procedure until the fitting error begins to increase significantly. This procedure is illustrated in Figure 4.1.

Extract all eligible parameters in the current step.

Check *extraction quality.* Find the worst parameter.

Check *fitting quality*. Continue if there is no significant deterioration.

Remove the worst parameter from extraction and set its value to default or the median of its extracted value.
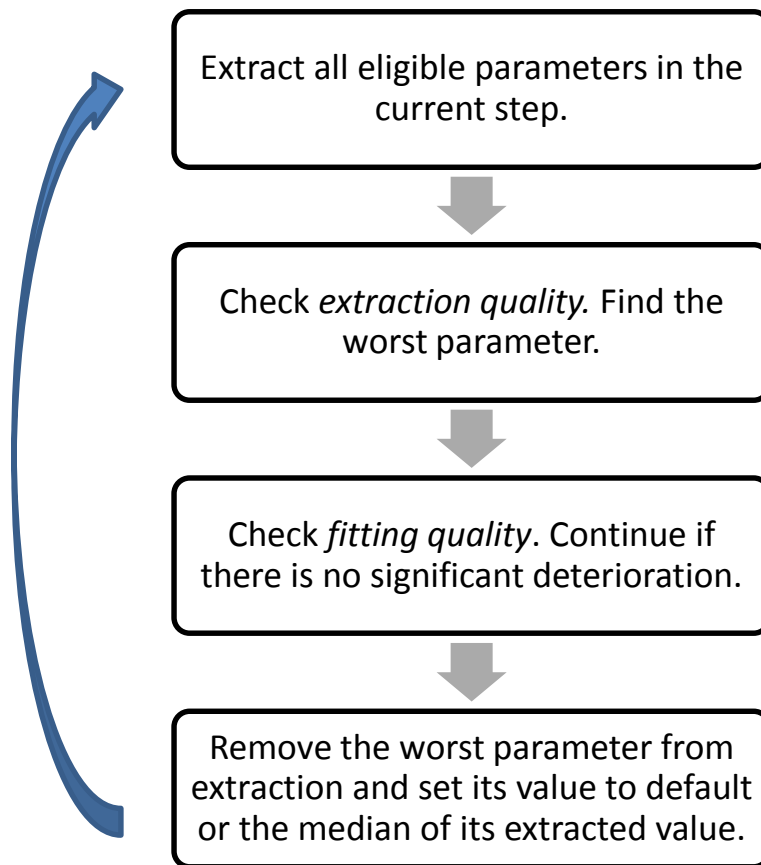
Figure 4.1: Stepwise parameter selection procedure for single-step optimization.

The key to backward parameter selection is the definition of the *extraction quality* criterion. While the *fitting quality* can be defined simply as the sum of squares of the fitting error, it is not as straightforward to define the proper *extraction quality*. Ideally, it should represent how accurate the extracted parameter is. However, in practice, there is no "true value" of model parameters from real devices that can be used for comparison. Instead, we

must define the fitting quality criterion with metrics that can be calculated or observed from the extraction result itself.

From the perspective of statistical simulation, here are a few characteristics of the extraction results to consider: We would like the statistical distribution as well as the correlation structure assigned for each of the model parameters to be as simple as possible. For example, a normally distributed *Vth* with a reasonable median value is preferred over a multimodal distribution. The range of variation should not be too wide either, because it may carry non-realistic values at the extreme quantiles of the distribution. These examples are illustrated in Figure 4.2. A multimodal distribution often indicates that there are two or more distinct device behaviors in the dataset. A wide distribution, on the other hand, indicates that either the sensitivity of the data to this parameter in the operation region is weak; or that the parametrization is deficient. For instance, if key parameters are missing, this particular parameter must carry all the variations that should be accounted for by another parameter. Lastly, we prefer model parameter correlations that can be reasonably captured by a single correlation coefficient rather than strong systematic dependence, as shown in Figure 4.3. The latter indicates strong parameter interactions in the model.
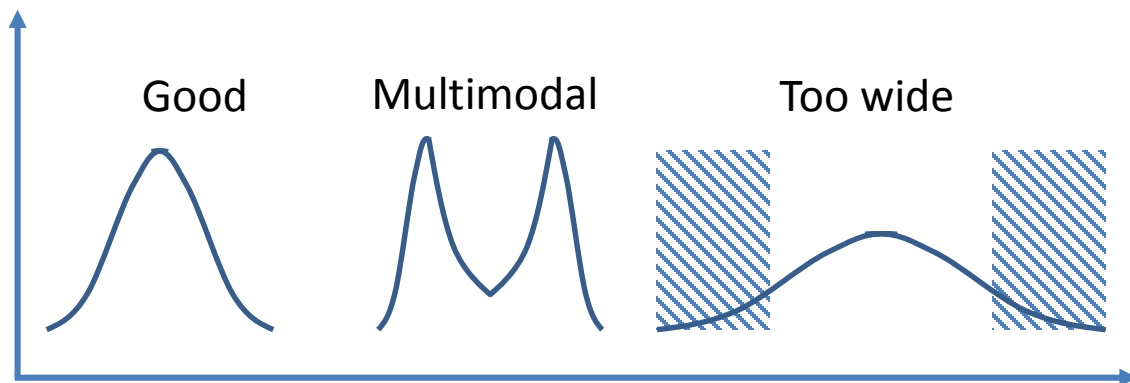
.

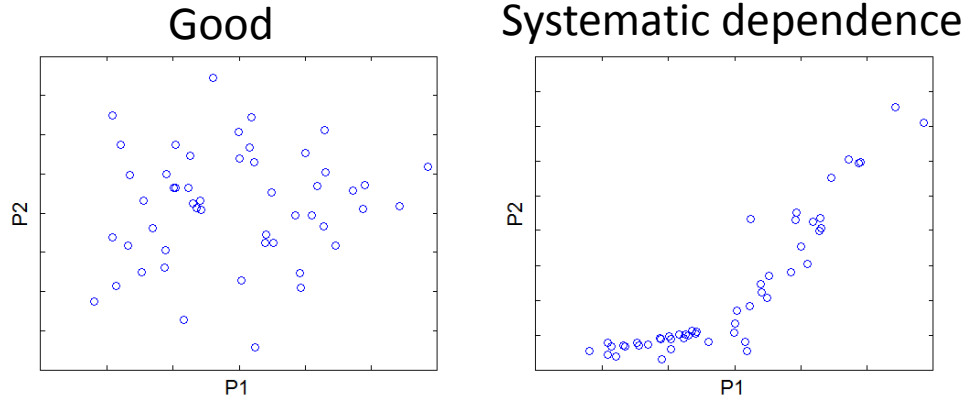Figure 4.2: "Good" vs. "Bad" parameter distribution.

Figure 4.3: "Good" vs. "Bad" parameter correlation.

To automate parameter selection, we need quantitative rather than qualitative measures. By experimenting with parameter extraction using simulated transistor *I-V* data (to be presented in Sections 4.3 and 4.4), we found the normalized confidence interval (CI) of the extracted compact model parameters to be a good proxy for the extraction quality of the parameters. More specifically, for every model parameter, an estimate μ and its half-width CI can be obtained via non-linear optimization for each device under test. We define the normalized confidence interval as CI/μ. If the 95% CI for a parameter does not contain zero, in other words, $|CI/\mu| < 100\%$, then we can reject the hypothesis that the corresponding model parameter is equal to zero at 5% significance (on the assumption that the underlying stochastic model holds). To use normalized CI as the extraction quality criterion for our parameter selection problem, we compare the distribution of the normalized CI of each model parameter at a given quantile, often the median, and remove the parameter with the largest normalized CI from future parameter extraction.

More insight can be obtained by looking at the math. Let us assume that we have the same parameter extraction/optimization setup as described in Section 4.2.2 and that the extracted parameter is $p^* = [p_1^*, p_2^*, \ldots, p_n^*]$. Subject to typical assumptions about the normality and independence of the underlying random variables, the half-width of the (notional, not actual) confidence interval of the ith parameter, p_i, can be estimated [81], [82] by the following:

$$CI_i = t_{1-\frac{\alpha}{2}, N-n} \, \sigma_i \tag{4.20}$$

where $t_{1-\frac{\alpha}{2},N-n}$ is the critical value of Student's t distribution and σ_i is the estimated standard deviation of the extracted parameter. Thus, for a given level of significance (i.e., 0.05) and degrees of freedom, the length of confidence interval is proportional to the parameter standard deviation, which we treat as an alternative extraction quality criterion in the following discussions.

In vector form, we have the following [83]:

$$\sigma^2(p^*) = \sigma^2(r^*) \cdot diag[J(p^*)^T J(p^*)]^{-1} \tag{4.21}$$

Here, r^* is the fitting error and $J(p^*)$ is the Jacobian matrix by the end of the non-linear least-square optimization. If the Jacobian matrix is singular or close to singular, then one or more variances of the extracted parameters will be infinite or unrealistically large, thereby indicating poor extraction quality.

The confidence interval also provides a measure of the residual after the stepwise parameter deletion. To illustrate this, we examine the local Taylor expansion in the final optimization step. Assume that Δp is the optimal step calculated at p, and that J_i is the ith column of the Jacobian matrix, $J(p)$. Then, we have the following linear approximation:

$$y - f(p) = J_1 \Delta p_1 + J_2 \Delta p_2 + \cdots + J_n \Delta p_n + r \tag{4.22}$$

Suppose the nth parameter, p_n, is to be excluded from the extraction. Then, the contribution of the term $J_n \Delta p_n$ must be compensated for by the other $n - 1$ parameters. As long as the Jacobian matrix is non-singular, one can always solve the linear regression problem.

$$[J_1, J_2, \overset{.}{\dots}, J_{n-1}]k = J_n \tag{4.23}$$

Here, $k = [k_1, k_2, \dots, k_{n-1}]^T$ is the fitting coefficient. Assuming that γ_n is the fitting residual, we can re-write Equation 4.22 as follows:

$$y - f(p) = J_1 \Delta p_1 + J_2 \Delta p_2 + \cdots + J_{n-1} \Delta p_{n-1} + \left(\sum_{i=1}^{n-1} k_i J_i + \gamma_n \right) \Delta p_n + r \tag{4.24}$$

$$= J_1 (\Delta p_1 + k_1 \Delta p_n) + J_2 (\Delta p_2 + k_2 \Delta p_n) + \cdots + J_{n-1} (\Delta P_{n-1} + k_{n-1} \Delta P_n)$$
$$+ (r + \gamma_n \Delta P_n)$$

The smaller Δp_n is, the less of an impact from the removal of p_n on the rest of the model parameters; the parameter p_l, whose corresponding J_l is most parallel to J_n (k_l

being the largest of $k_1, k_2, \ldots, k_{n-1}$), will be affected the most. If J_n can be very well approximated by the linear combination of $J_1, J_2, \ldots, J_{n-1}$, then the removal of p_n will only increase the fitting error by a small amount due to the small γ_n. Furthermore, the variance of the extracted parameter, p_i, can be written as follows [14]:

$$\sigma^2(p_i) = \sigma^2(r) \times \frac{1}{SS(\gamma_i)} \tag{4.1}$$

where $SS(\gamma_i)$ is the sum of squares of the Jacobian fitting residual, γ_i. This tells us that the parameter with the largest variance also happens to be the one that can be best replaced by the other parameters. This suggests that the elimination of the parameter with the largest variance is likely to introduce the smallest increase in fitting error. On the other hand, deleting some model parameters may introduce biases to the estimated values of the remaining parameters.

4.2.4 Sequential Extraction

As stated in section 4.2.1, there are numerous benefits to divide the full model parameter extraction into smaller optimization problems. These range from a better representation of physical meanings to less computational cost. Thus, most model extractions are performed sequentially. During sequential extraction, parameters are estimated in a pre-defined series of localized optimization steps, and each step only fits a subset of parameters to a subset of measurement data. The parameter extraction completes as soon as the last optimization step is done.

There are several strategies for combining stepwise parameter selection with sequential extraction. In the conventional setup, the parameters to be extracted at each step are predetermined by a guideline. In the stepwise parameter selection scheme, the parameters to be extracted are to be determined on the fly for each step. Thus, for the model parameters involved in different steps, one has several different options for conducting the extraction. A simple approach, the *greedy algorithm*, only estimates a parameter once, then holds it constant while other parameters are estimated. Greedy algorithms may diverge in some problems unless they are regularized. Instead, we start the parameter selection algorithm by re-fitting all the parameters included in the model at each step. The members of "extractable" parameters and their estimates tend to vary by step. During a given extraction step, whenever a parameter is discarded by the parameter selection algorithm, its estimates will be reset to either previous estimates from the last extraction step where the parameter is selected, or the median estimate from the current parameter selection step. This way, we can keep refreshing the estimates of relevant parameters when new datasets

become available in subsequent extraction steps without destroying the best estimates found previously for the other parameters.

Another inherent issue of sequential extraction is that the fitting quality may deteriorate after a large number of steps. To address this issue, after we go through all the pre-defined extraction steps for the localized optimization problem, we revisit the full, undivided optimization problem. Every parameter that has been selected at least once during the sequential extraction will be re-fitted using the combined datasets from all the smaller extraction steps, using their latest estimates from the sequential extraction as initial guesses. Because each parameter is already optimized for its most relevant operation region, this final optimization tends to be much faster than the same extraction with an un-optimized initial model card. Parameter estimates from this global optimization become the final model cards. This procedure is illustrated in Figure 4.4.
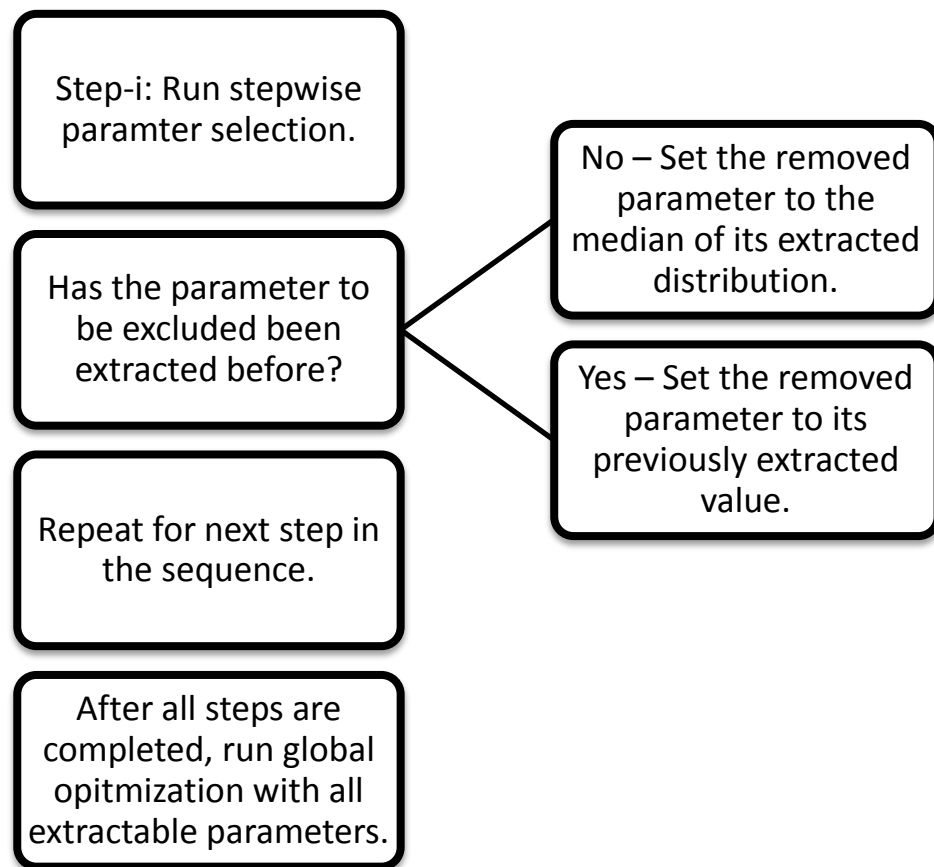

Figure 4.4: Sequential parameter extraction.

4.3 Simulated Experiment with the EKV Model

4.3.1 EKV Model Introduction

The EPFL-EKV MOSFET model is a compact SPICE simulation model built on the fundamental physical properties of MOS transistors [29]. It is relatively lightweight compared to models such as BSIM or PSP, with only a few tens of key parameters to cover the full operation space of MOSFET transistors. We chose to use the EKV V2.6 model as the subject of our single-step parameter selection and extraction study.

The parameters of the EKV model can be divided into several categories: process-related parameters, which are the physical dimensions directly defined by the fabrication process; intrinsic model parameters, which are the electrical properties of the transistors; and parameters that describe specific device physics effects, such as channel length modulation, charge sharing, reverse short-channel effects, impact ionization, temperature dependence, matching, and flicker noise [29]. In our experimental setup, ten parameters were chosen for the purpose of curve fitting. The parameter names, physical meanings, and default values are listed in Table 4.1.

Name	Description	Units	Default
DW	Channel width correction	m	0
DL	Channel length correction	m	0
VTO	Long-channel threshold voltage	V	0.5
GAMMA	Body effect factor	$\sqrt{V}$	1.0
PHI	Bulk Fermi potential ($2\times$)	V	0.7
KP	Transconductance parameter	A/V^2	50.0E-6
E0	Mobility reduction coefficient	V/m	1.0E12
UCRIT	Longitudinal critical field	V/m	2.0E6
LAMBDA	Depletion length coefficient	–	0.5
LETA	Short channel effect coefficient	–	0.1

Table 4.1 Candidate EKV model parameters for extraction [29].

4.3.2 Experiment Setup

We test the effectiveness of our parameter extraction methodology with the following simulation experiment. Assuming we have an ideal EKV model that perfectly captures the behavior of the real world CMOS transistor, the variations of transistor characteristics can then be entirely explained by the variability of the compact model parameters. The goal of capturing the transistor variability with the statistics of the compact model parameter then becomes equivalent to extracting the true values of the compact model parameters that generate the transistor's electrical *I-V* variations.

The first step is to generate random *I-V* data with the EKV model. We selected 10 parameters (as listed in 4.1) to carry all the variation of the transistor. All the other EKV model parameters were set to nominal values appropriate for the technology node this experiment represents. Absent information regarding the correlation among the model parameters, we chose to draw the ten model parameters from a multivariate normal distribution with independent components. The standard deviation of each parameter was taken to be 3% of the parameter's nominal value. A total of 100 model cards (whose parameter distribution/correlation is shown in Figure 4.5) were generated. They will be referred to as the "original model cards." For each EKV model card, a set of *I-V* curves were simulated for NMOS W/L = 0.5/0.15 μm using the HSPICE built-in level-55 EKV model. The transistor terminal bias space was chosen so that the gate and drain bias voltages are equally spaced from 0 to *Vdd* with zero body bias (Figure 4.6). This provides good coverage of the main transistor operating regions, which will suffice for our study of a single-step optimization with the EKV model. The on/off current distributions are shown in Figure 4.8. The 3σ variation of the on current is about 13%, which is a reasonable approximation of the variation profile in a well-controlled modern silicon process.
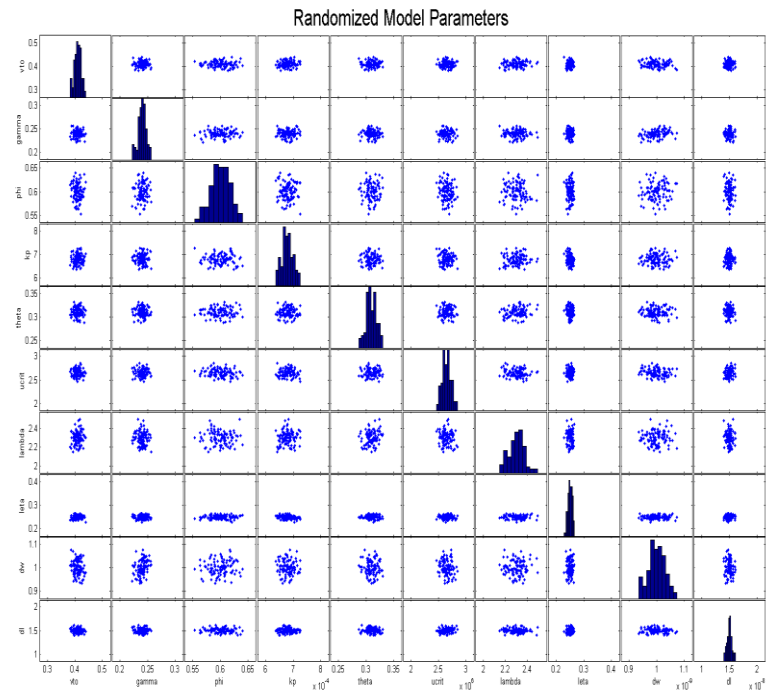
74

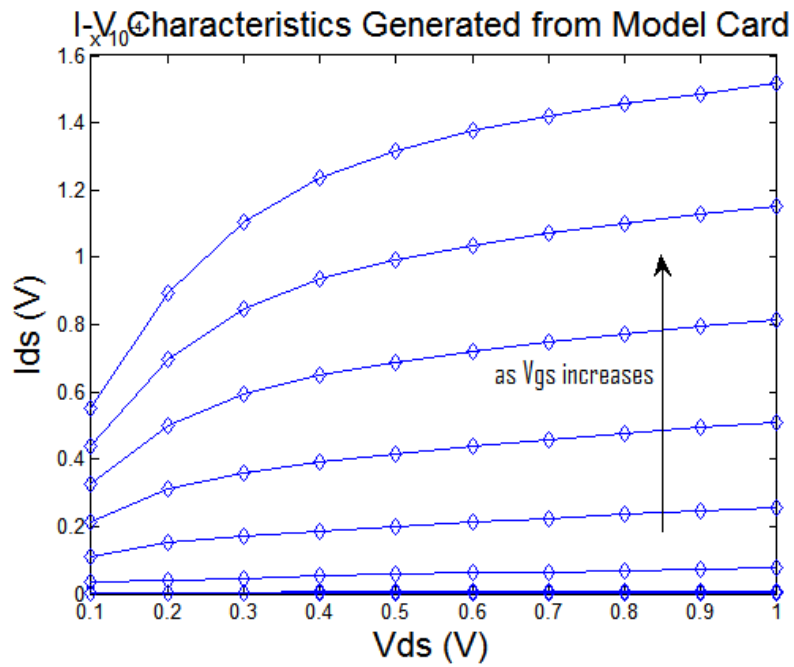Figure 4.5: Randomized EKV model parameters as the base of the transistor *I-V*.

Figure 4.6: Transistor *I-V* characteristics generated from one EKV model card.

75
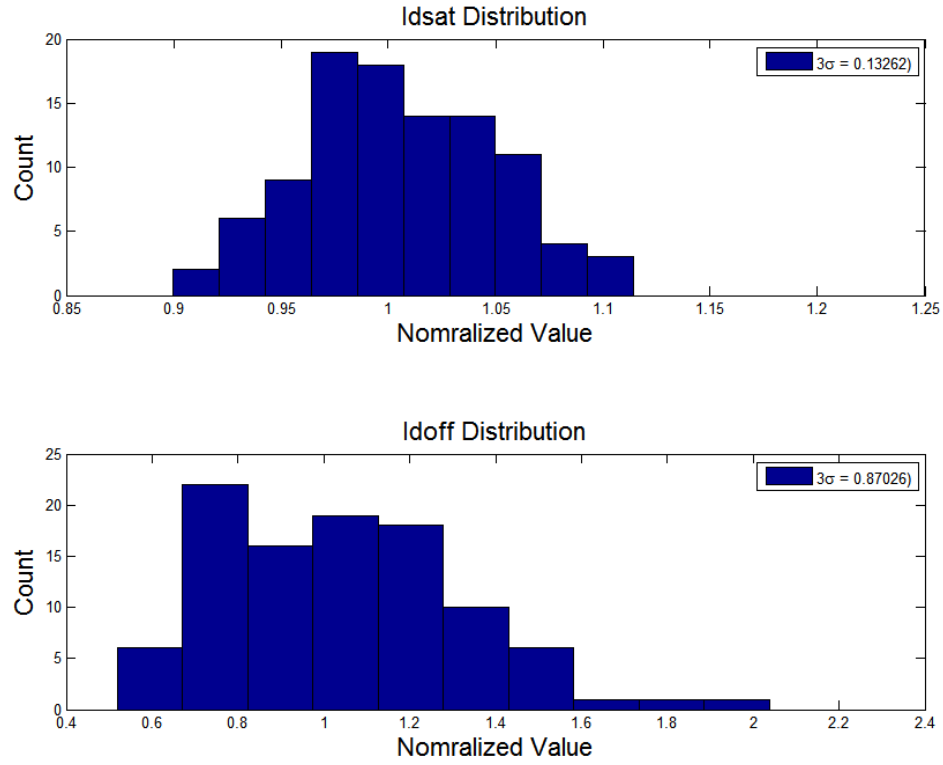
Figure 4.7: Histogram of normalized on/off current of simulated random *I-V* characteristics.
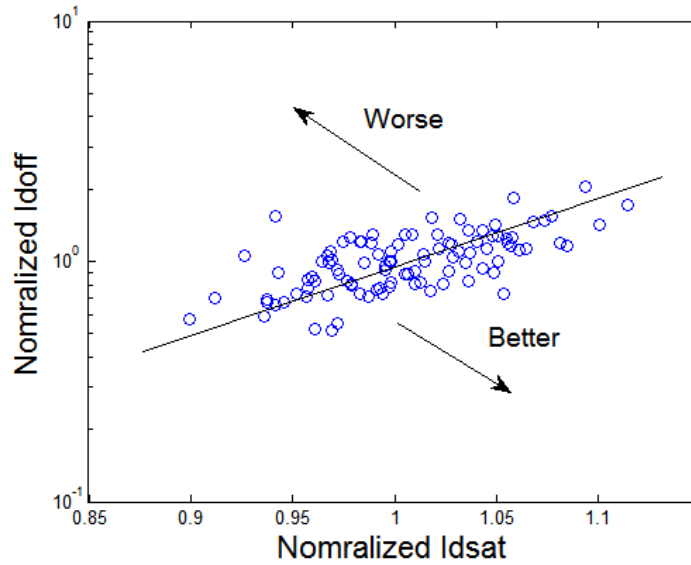

Figure 4.8: Variation in the on/off current space. Transistors in the lower right corner are considered superior to the ones to the top left corner due to their higher on currents and lower leakage.

4.3.3 Stepwise Parameter Selection

The stepwise parameter selection procedure illustrated in Figure 4.1 was applied to the simulated data. Starting with all N = 10 model parameters under consideration, a non-linear least-square optimizer with a trust region reflective algorithm [79], [80], [84] is used to fit the *I-V* curve sets simulated for this experiment. After the first round of extractions, we rank the normalized confidence intervals of the parameter estimates as indicator of the extraction quality of the model parameters. The parameter with largest normalized CI is set to the median estimate from the current fitting step and removed from the next round of curve fitting. By repeating this process, at the *i*th round of extraction, only $N - i + 1$ parameters are fitted to the simulated *I-V* curves. At the end of each round of extraction, the fitting quality is examined, comparing the fitted *I-V* curves to the original simulated data, as shown in Figure 4.12. The stepwise parameter removal process stops when the fitting quality begins to degrade significantly.

Using the simulated *I-V* curves, the described methodology demonstrates very good agreement with visual inspections of the extraction and fitting quality. Table 4.2 lists the 90th percentile of the normalized notional confidence interval after each round of parameter selection and extraction. The removed parameter after each round is labeled with a "-" for all the following steps. As Figure 4.10 shows, with all ten model parameters included, several pairs of model parameter estimates are correlated, notably *gamma* vs. *phi*, *phi* vs. *leta* and *dl*, and *vto* vs. *leta* and *dl*, even though the parameters for the model cards were generated independently. When the stepwise parameter selection procedure continues, these artificial correlations go away once phi, $leta$, and dl are excluded from the extraction. Comparison between extracted compact model parameters and the original data set (Figure 4.10) also confirms the improvement of the accuracy of the extraction. The removal of the first three parameters improves the fit between the remaining model parameters with their values in the original model card, as measured by correlation (without regard for scale). For example, correlation of *gamma* improves from a correlation coefficient of 0.62 to 0.83 after the third round of parameter removal, and *dl* improves from 0.44 to 0.78. By now, the remaining seven parameters all have decent correlations with their true values assigned by the experiment. However, since we also want the number of model parameters used in extraction to be as small as possible, we continue the stepwise parameter removal until only one parameter remains. The removal of *dl* and *gamma* now begins to decrease the accuracy of *theta*, whose correlation coefficient with the original model card drops from 0.995 after round 3 parameter removal to 0.98 after round 4, and 0.915 after round 5. The confidence intervals for all the remaining parameters, however, only begin to increase after the fifth round of parameter removal (*gamma*), as shown in

Figure 4.9. The sum of squares of the fitting error only starts to rise after the elimination of *gamma* (Figure 4.13), which is consistent with visual inspection of the model to *I-V* data comparison, as shown in Figure 4.12. The further removal of model parameters only inflates the confidence intervals of the remaining parameters and reduces their correlation with the original model card, while exponentially increasing the sum of squares of error (SSE).

The simulations suggest that stepwise parameter selection using confidence intervals as measures of extraction quality provides a reasonable approach to determining a small set of parameters that is satisfactory in terms of both extraction quality and fitting quality. In this specific experiment, the optimal parameter group will include *vto, gamma, kp, theta, ucrit,* and *lambda.* Further reduction of model parameters hurts the fitting quality and the extraction quality.

Round	vto	gamma	phi	kp	theta	ucrit	lambda	leta	dw	dl
1	0.49%	7.07%	12.0%	0.43%	0.56%	0.31%	0.33%	8.05%	34.7%	5.14%
2	0.43%	6.31%	10.7%	0.39%	0.52%	0.28%	0.32%	7.29%	-	4.61%
3	0.22%	0.74%	-	0.30%	0.47%	0.18%	0.21%	6.88%	-	2.61%
4	0.02%	0.59%	-	0.24%	0.37%	0.13%	0.14%	-	-	1.96%
5	0.01%	0.22%	-	0.03%	0.18%	0.05%	0.03%	-	-	-
6	0.09%	-	-	0.27%	1.58%	0.42%	0.25%	-	-	-
7	0.11%	-	-	0.22%	-	0.65%	0.42%	-	-	-
8	0.15%	-	-	0.20%	-	-	0.48%	-	-	-
9	0.56%	-	-	0.77%	-	-	-	-	-	-

Table 4.2: 90th percentile of normalized confidence intervals after each round [29].
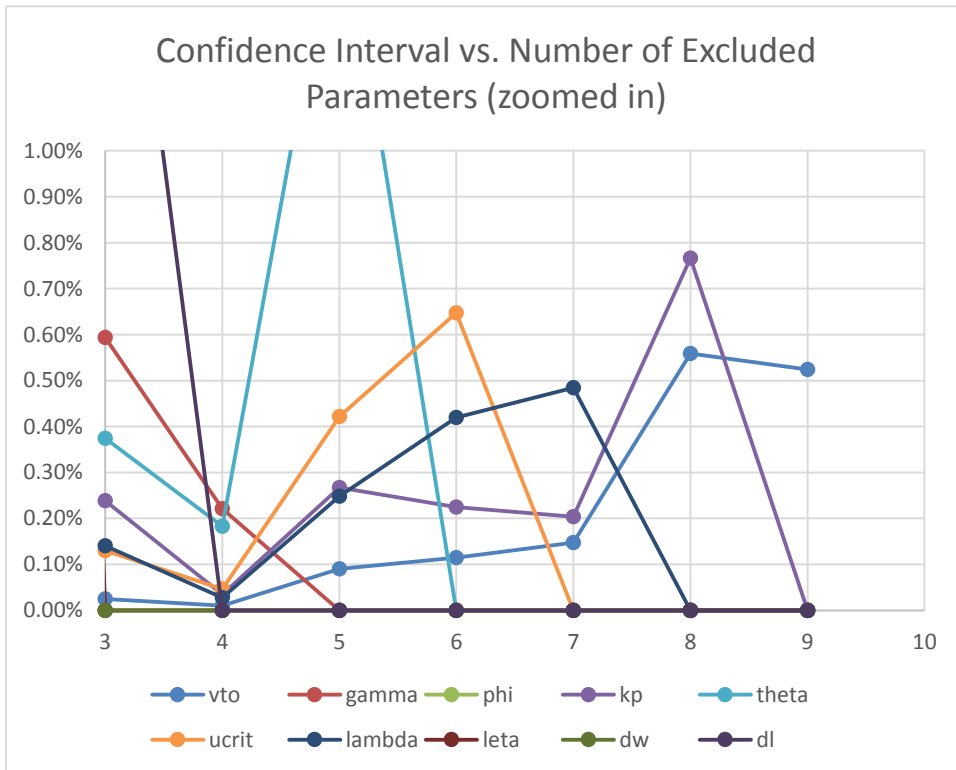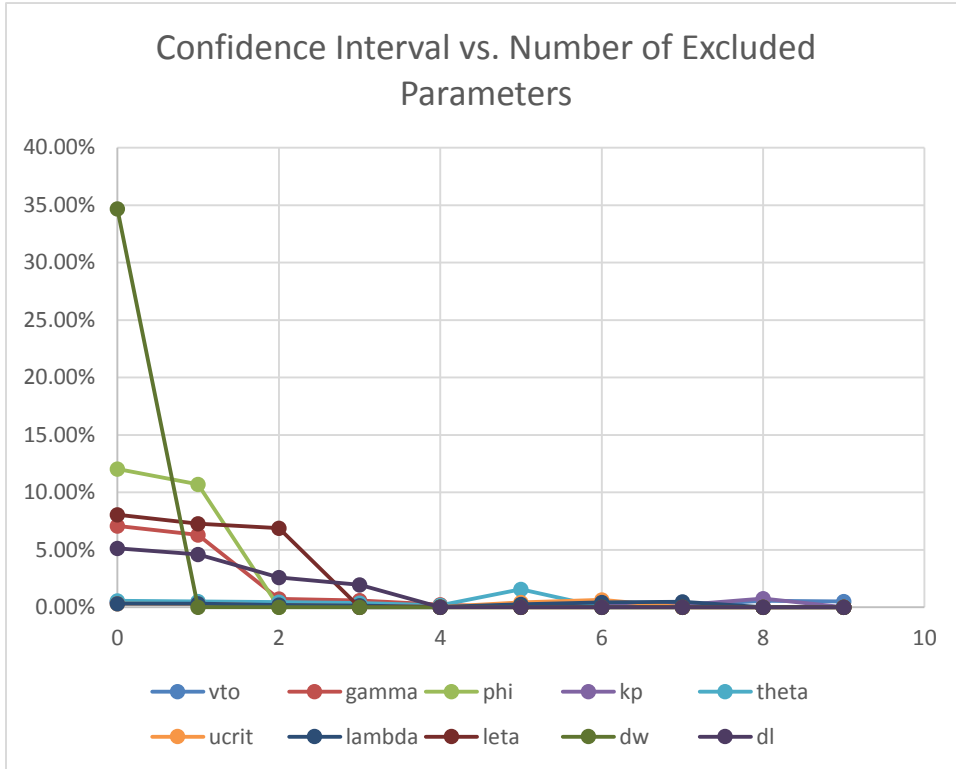
Figure 4.9: Changes of normalized confidence interval for all model parameters after each round of stepwise parameter selection and extraction.
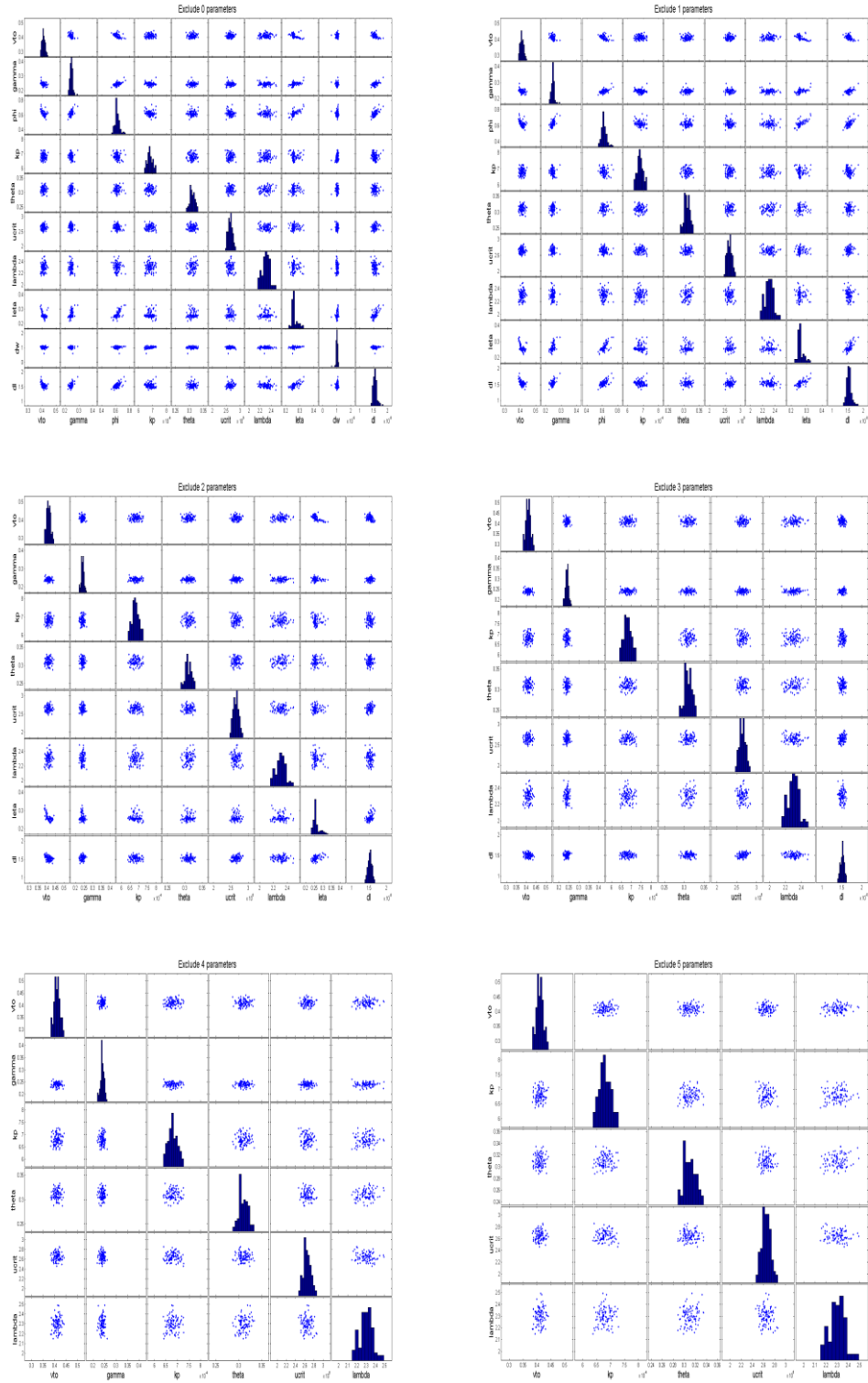
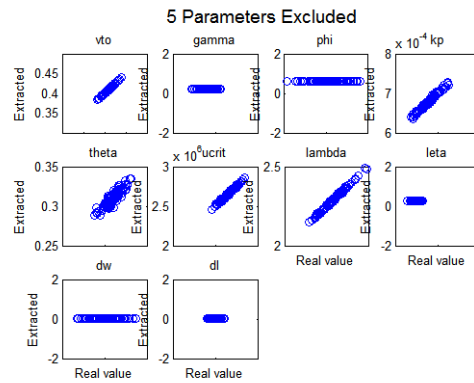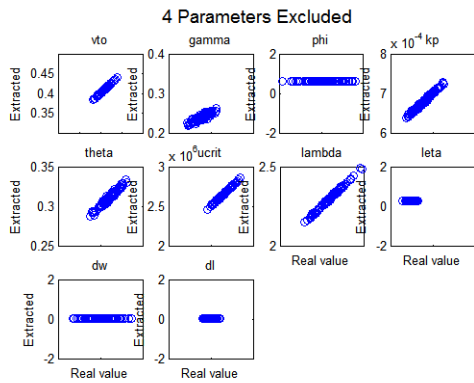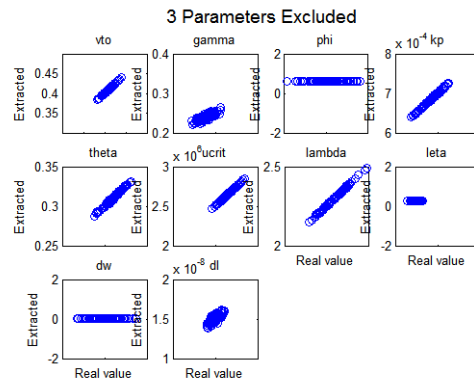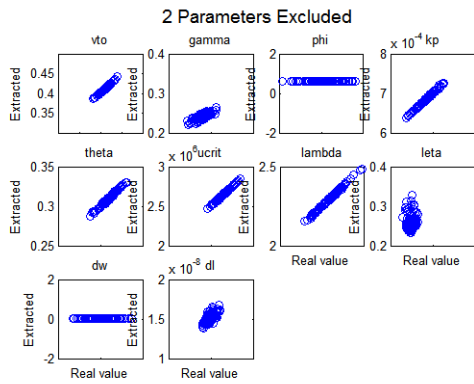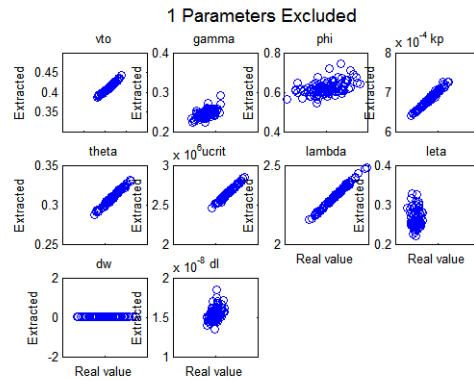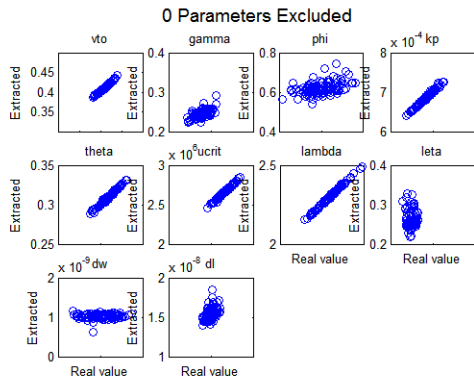Figure 4.10: Statistical distribution of the parameter estimates and their correlation structure after excluding 0 to 5 parameters from the extraction.
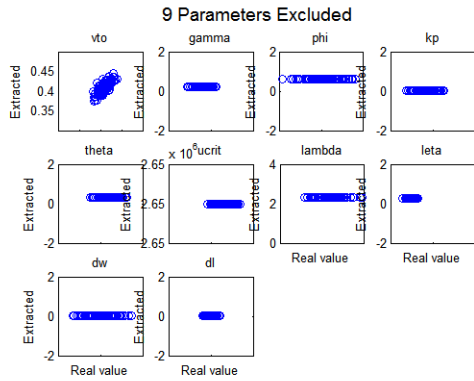
9 Parameters Excluded

Figure 4.11: Parameter estimates vs. original randomized model card after excluding 0 to 5 parameters from the extraction.

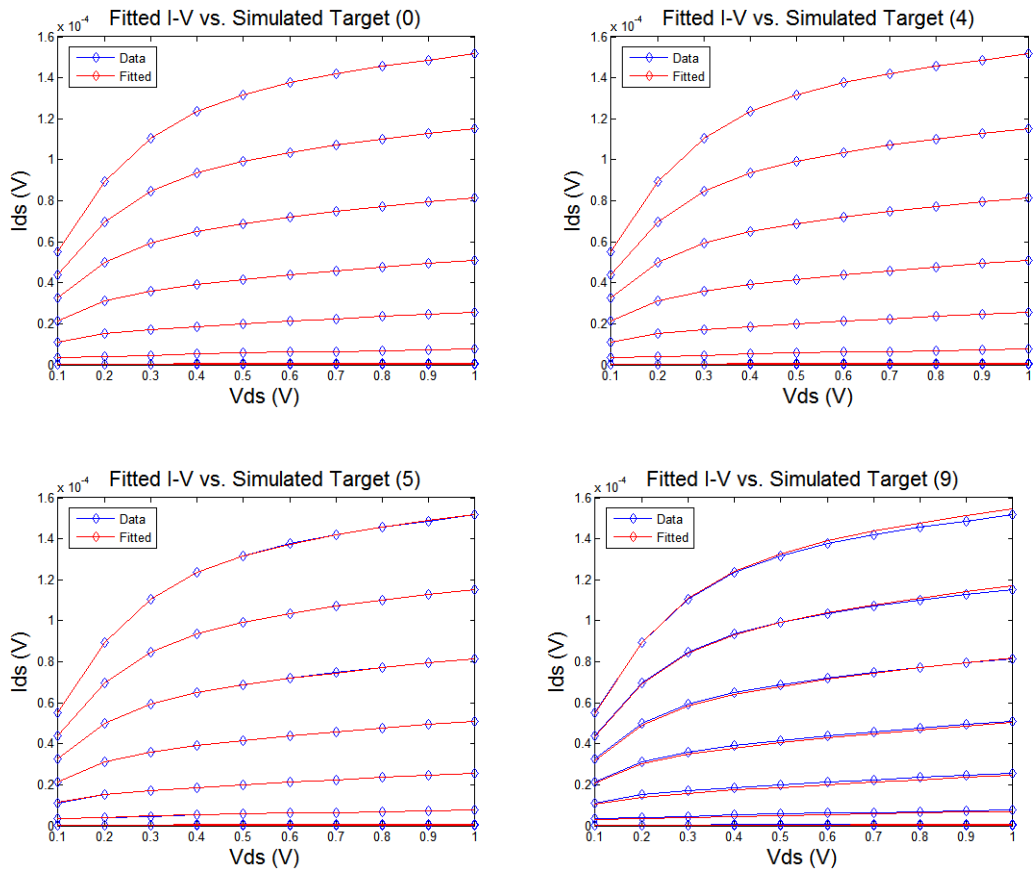
Figure 4.12: Fitted *I-V* vs. simulated targets of a single device with 0, 4, 5 and 9 parameters excluded from the extraction.
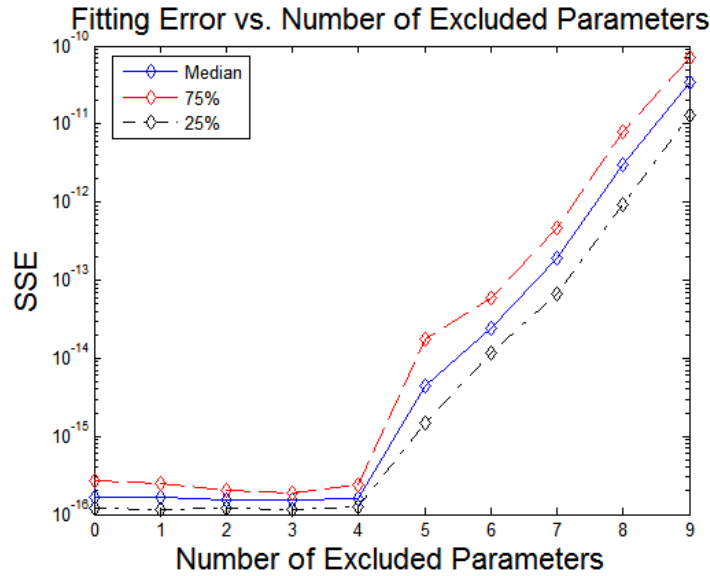
Figure 4.13: Fitting error increases significantly when more than four parameters are excluded from the optimization.

4.4 Simulated Experiment with PSP model

4.4.1 PSP Model Introduction

The PSP model is an advanced surface potential based compact SPICE model [85], [86] and one of the two industrial standard models of today (the other is the long-standing BSIM model [11]). It includes all relevant physical effects, including mobility reduction, velocity saturation, DIBL, gate current, and STI stress to model today's deep sub-micron CMOS technologies [87].

The PSP model has two sets of model parameters: the global-level parameter set, which describes entire space of device geometries, and the local-level parameter set, which models transistors with specific device dimensions. Since we are only extracting parameters of transistors of a single size, we focus on local-level parameters. According to the recommended local parameter extraction procedure in the PSP manual [87] and the *I-V* data available in the experiment, 16 parameters are chosen as candidates for our experiment in parameter extraction. The parameter names and their physical meanings are listed in Table 4.3.

Param.	Description	Param.	Description
vfbo	Geometry-independent flat-band voltage	*cso*	Geometry-independent Coulomb scattering
nsubo	Geometry-independent substrate doping	*xcoro*	Geometry-independent non-universality
dphibo	Geometry-independent offset of ϕ_B	*rsw1*	Source/drain series resistance
cto	Geometry-independent part of interface states factor CT	*thesato*	Geometry-independent velocity saturation
cfl	Length dependence of CT	*alpl*	Length dependence of CLM pre-factor ALP
uo	Zero-field mobility at TR	*alp1l1*	Length dependence of CLM enhancement factor above threshold
xmueo	Geometry-independent mobility reduction coefficient	*alp2l1*	Second-order length dependence of ALP1
themuo	Mobility reduction exponent	*vpo*	CLM logarithmic dependence

Table 4.3: Candidates of EKV model parameters for extraction [87].

4.4.2 Experiment Setup

In addition to re-validating our findings with EKV model extraction, we want to test our sequential parameter extraction procedure. Thus, instead of using all the generated *I-V* data points in one run with all parameter candidates, we instead divide the data into three *I-V* curves: *Id-Vg* linear (*Vds* = 0.1V, *Vgs* = 0,…,1V), *Id-Vd* (*Vgs* = 1V, *Vds* = 0,…,1V), and *Id-Vg* saturation (*Vds* = 1V, *Vgs* = 0,…,1V). As illustrated in Figure 4.4, a full parameter extraction will be performed for each of the three curves in the exact sequence in which they were introduced. The parameter values extracted in the earlier steps are used as the initial values for the next step or are set to a constant if the parameter is excluded later.

As in the EKV simulation experiment, we drew the 16 parameters from a multivariate normal distribution with independent components, taking the standard deviation of each parameter to be 3% of the parameter's nominal value. A total of 50 original model cards were generated, and their parameter distributions/correlations are shown in Figure 4.14.

For each of these randomized model cards, three *I-V* curves, as described above, are simulated (Figure 4.15) for a NMOS transistor with W/L = 0.2/0.055 μm. The electrical simulation is performed using HSPICE with a built-in level-69 PSP model.

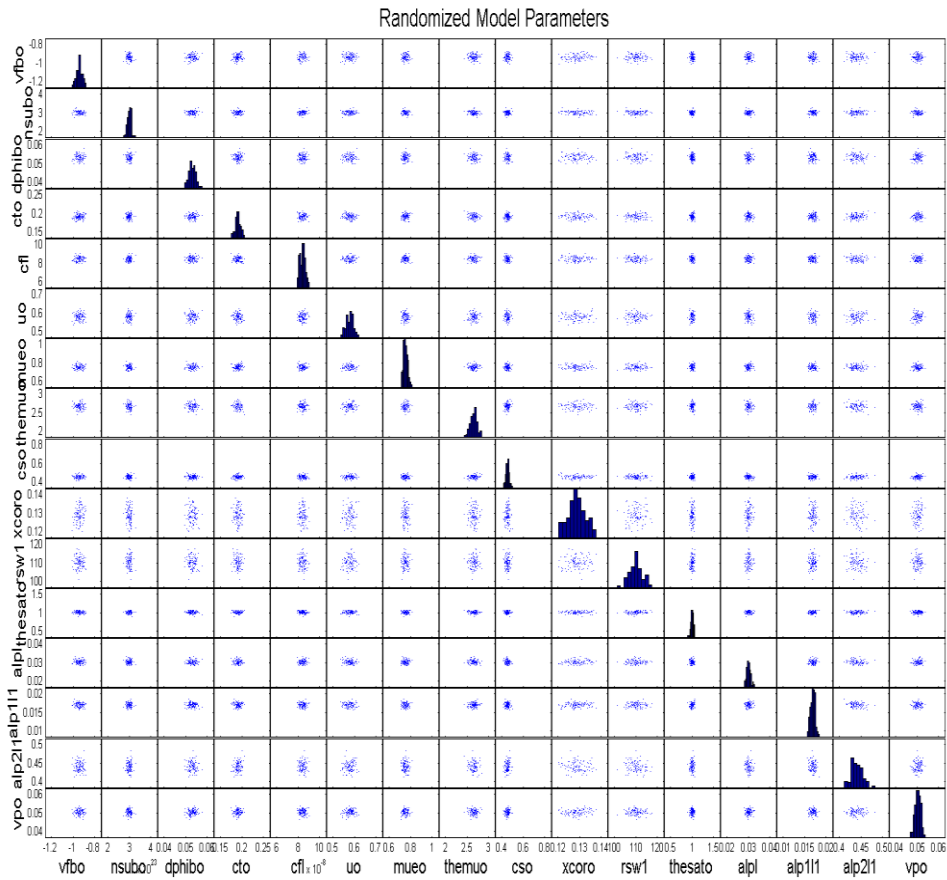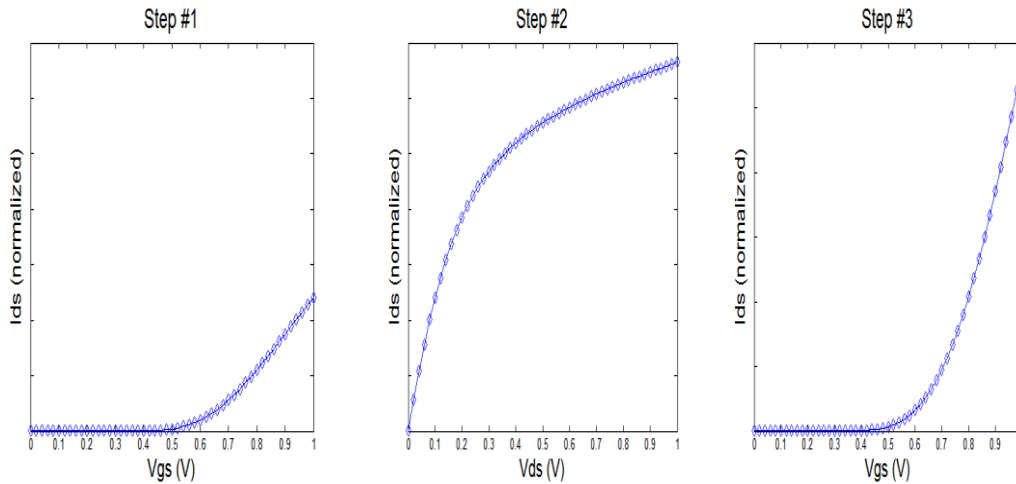Figure 4.14: Randomized PSP model parameters as the base of the transistor *I-V*.

Figure 4.15: Target *I-V* data for the three extraction steps

4.4.3 Stepwise Parameter Selection in Sequential Extraction

For Step #1, the same backwards deletion scheme using the normalized length of confidence intervals as the selection criterion is applied to the simulated Id-Vg (linear) data. Without listing all the details, we show the results from the stepwise parameter selection in Figure 4.16 and Figure 4.21. The stepwise parameter selection procedure located six good parameters for the data fitting in Step #1. Notice that with all 16 parameters in the extraction, one of the final "good" parameters, *cto*, actually correlates poorly with the original model card. However, when the algorithm stops after excluding ten parameters, its correlation is greatly improved. This, again, demonstrates the effectiveness of our parameter selection methodology.

The 50 extracted model cards from Step #1 are then used as the initial values to begin the extraction of Step #2. The same stepwise parameter selection is carried out, except that the removed parameters are now set to their previously extracted values (if they are available). The algorithm stops after removing ten parameters (Figure 4.16). Three of the remaining six parameters have already been extracted from Step #1, namely *vfbo, uo,* and *rsw1*. We then propagate the newly extracted model cards to the Step #3 extraction. Again, the stepwise parameter selection algorithm will provide us with a set of seven parameters to be optimized (Figure 4.17). This time, five out of the seven remaining parameters—including *vfbo, cto, uo, themuo,* and *rsw1*—have already been extracted at least once in the earlier steps. The inclusion of previously optimized parameters is important to obtain an accurate estimation of the newly extracted parameters. Otherwise, the fitting and extraction quality suffer.

86

Finally, combining the simulated *I-V* data from all three steps, we ran a global optimization with all 11 model parameters that were extracted at least once during the three sequential steps. Because each of these parameters was already well-calibrated during the sequential step, the global optimization converged very quickly. Thus, we guaranteed that optimizations occurring later in the sequence do not decrease the fitting quality of earlier steps. The comparison between the final model cards and the cards after the Step 3 extraction is shown in Figure 4.20.
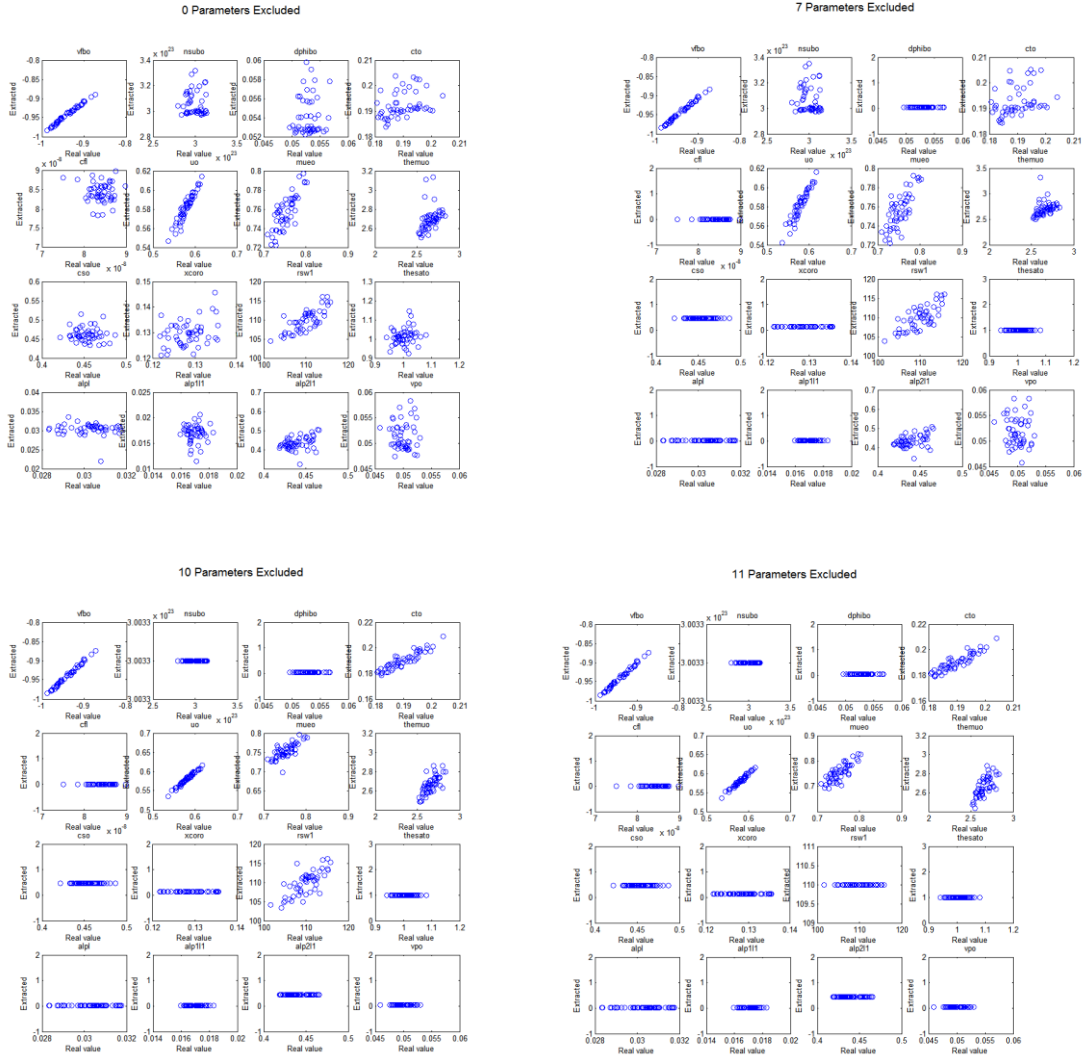

Figure 4.16: Parameter estimates vs. original randomized model card after excluding 0, 7, 10, and 11 parameters from Step #1 extraction.

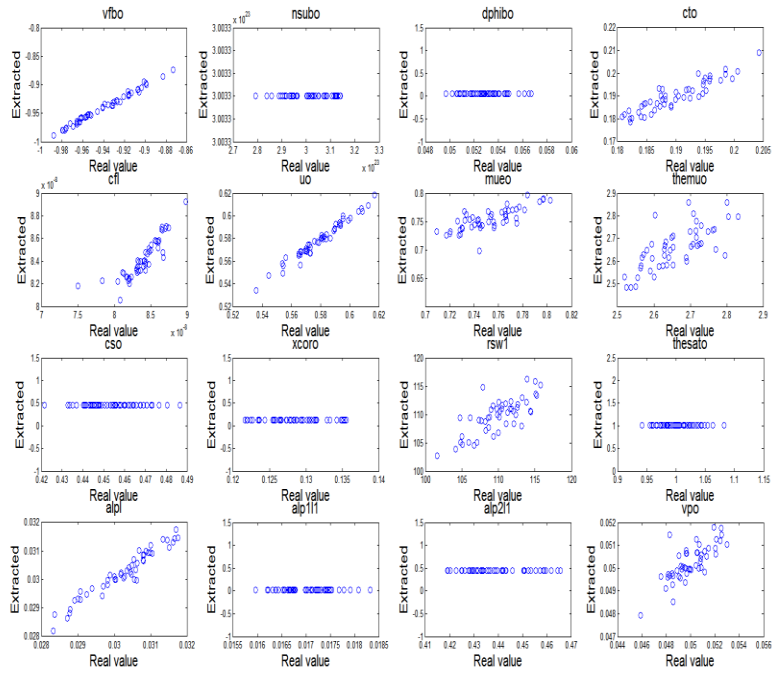Figure 4.17: Parameter estimates vs. original randomized model card after excluding 10 parameters from Step #2 extraction. Three of the six remaining parameters were extracted in Step#1.
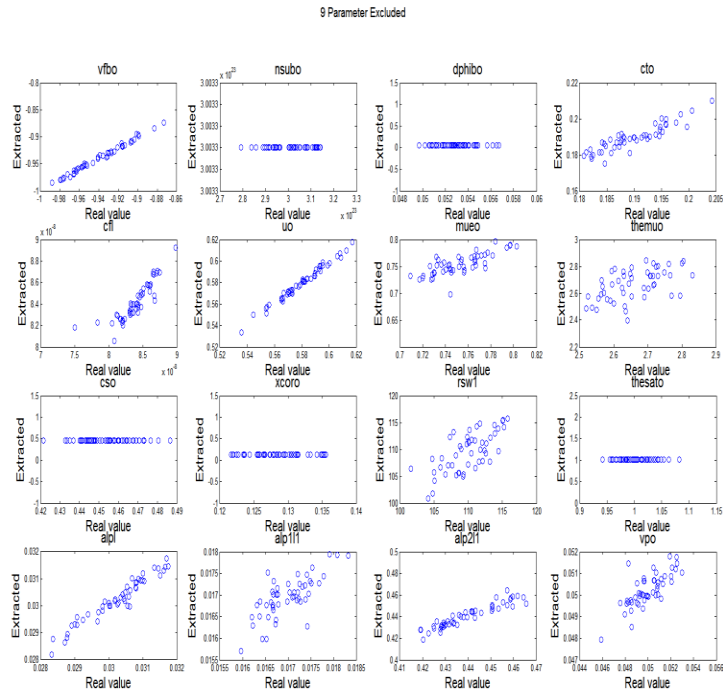
Figure 4.18: Parameter estimates vs. original randomized model card after excluding nine parameters from Step #3 extraction. Five of the remaining seven parameters were extracted (at least once) in Step #1 and Step #2.
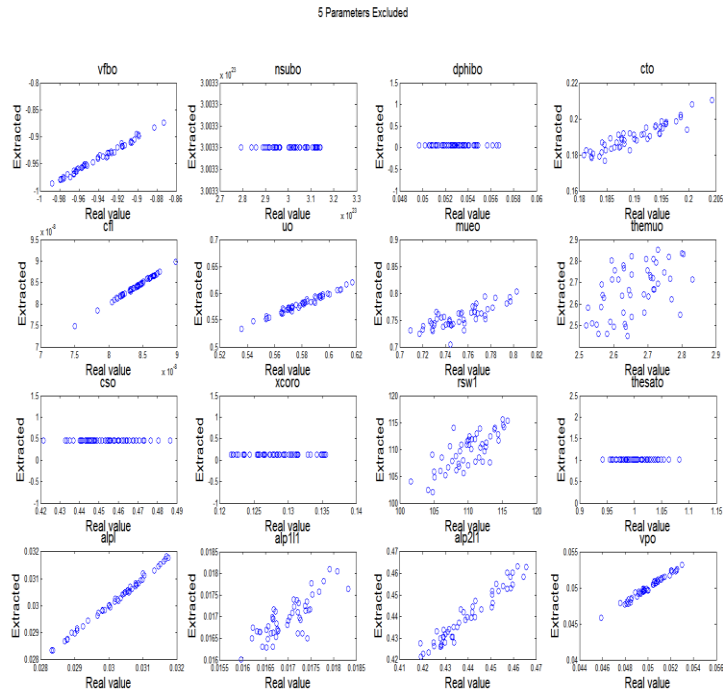
Figure 4.19: Parameter estimates vs. original randomized model card after global optimization with all previously extracted parameters and *I-V* data from all three steps.
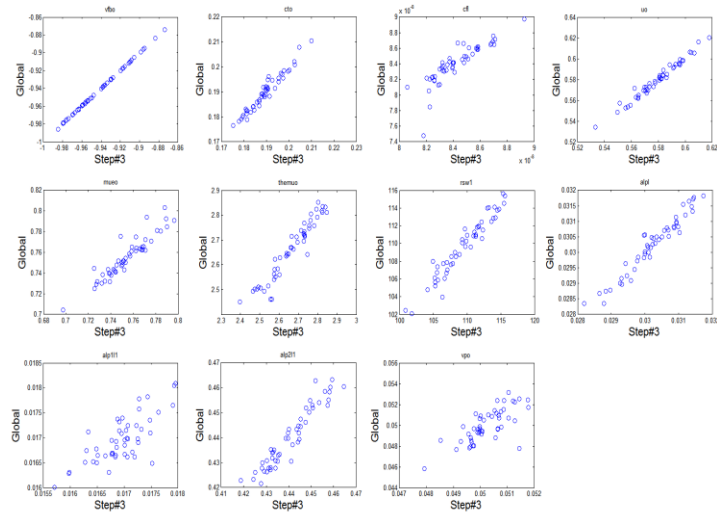
Figure 4.20: Model cards after global optimization vs. model cards after Step #3.
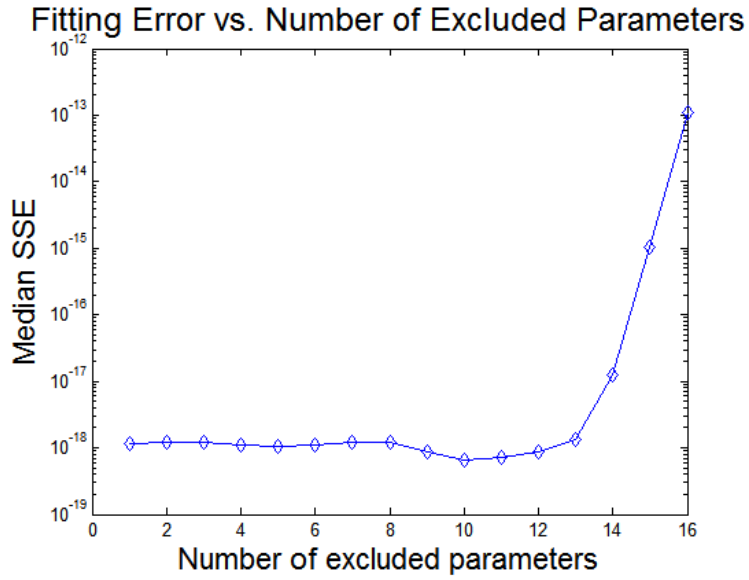
Figure 4.21: Fitting error increases when more than 10 parameters are excluded from Step #1 optimization.
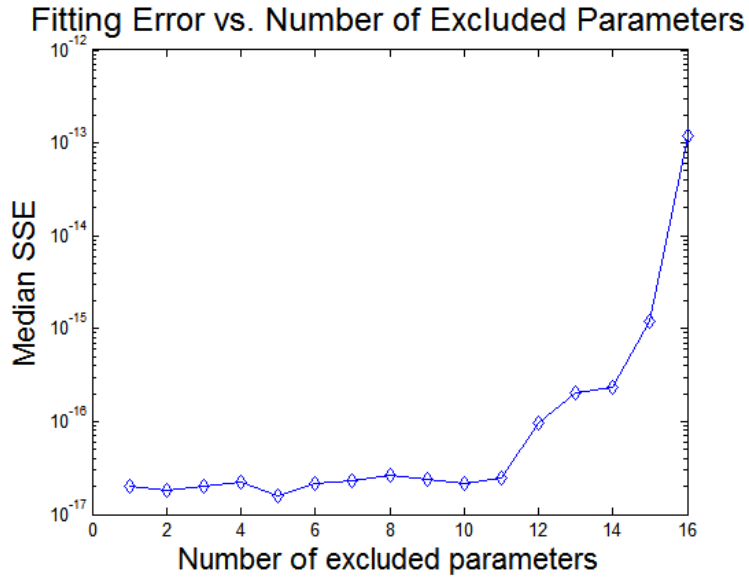
Figure 4.22: Fitting error increases significantly when more than 10 parameters are excluded from Step #2 optimization.
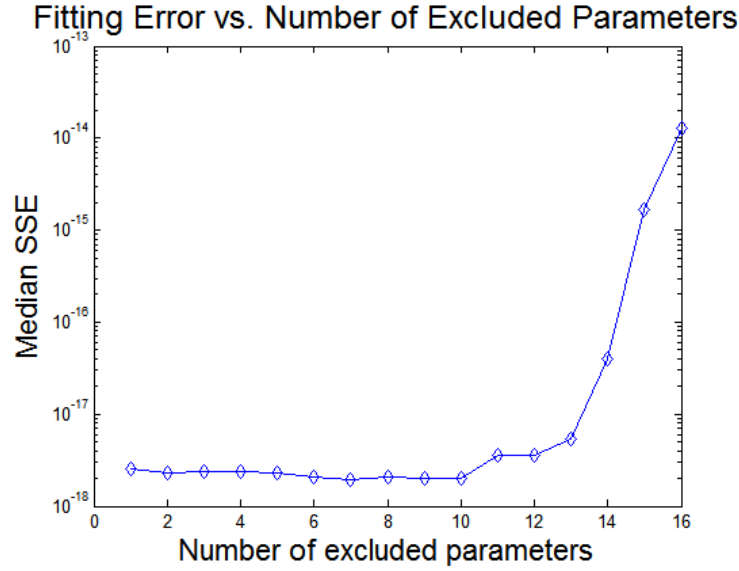
Figure 4.23: Fitting error increases when more than nine parameters are excluded from Step #3 optimization.

4.5 Summary

In this chapter we used non-linear least-squares to estimate compact model parameters. We proposed a backward parameter selection procedure that uses the normalized length of notional confidence intervals as the criterion for parameter removal. Simulated experiments are carried out with an EKV model as an example of single-step extraction, and a PSP model is used as an example of sequential parameter extraction. In simulations, stepwise parameter selection is highly effective when the target *I-V* data can be fully captured by the compact model and it works very well with the existing sequential parameter extraction procedure.

Chapter 5

Statistical Extraction and Modeling with Experimental Silicon Data

5.1 Introduction

The statistical compact model parameter extraction methodology proposed in Chapter 4 is applied to the transistor *I-V* data experimentally collected from one of the 45nm SRAM test chips. The model parameters for extraction will be determined for the EKV model as well as the PSP model by running the stepwise parameter selection procedure over a subset of transistors available for both models. The full model parameter distributions will then be estimated from all the SRAM transistors. The hierarchical variability model is fitted to the parameter estimates to decompose device variability into systematic and random components. The hierarchical variability model is then compared to the conventional "Global+Local" variability model by examining how representative the device electrical metrics generated by each approach compare to the experimental measurement.

5.2 Measurement for Parameter Extraction

As observed in Chapter 3, the SRAM padded-out transistors do not have much across-wafer systematic variation; in that situation, the hierarchical model essentially reduces to the conventional "Global+Local" model, in which chip-to-chip variation is represented by a normal distribution (or log normal, in the case of leakage). On the other hand, there is clear evidence of systematic across-chip variation in both the transistor electrical metrics (such as *Idsat*) and the SRAM bit-cell read/write noise margins (such as *RSNM* and I_W). As a result, we decided to use the measured data from a single chip in our statistical compact model parameter extraction methodology and to demonstrate the hierarchical model.

The bit-cell transistors chosen for parameter extraction come from the SRAM test array on a chip near the central position of wafer #2, as shown in Figure 3.23. The SRAM test array contains 360 bit cells arranged into 18 columns by 20 rows. For each of the six transistors of a bit cell, a set of *I-V* curves are experimentally measured with *Vgs* ranging from 0 to 1V at a step size of 0.02V and *Vds* ranging from 0 to 1V at a step size of 0.1V (a finer *Vds* step size of 0.02V is enforced for *Vgs* = 1V). Figure 5.2 shows an example of the *I-V* characteristics of one of the pull-down transistors (PD) on the test chip. Depending on

the extraction flow for each compact model, either the complete set of measured *I-V* or a subset of the data will be used for the model parameter extraction. The measured current in the subthreshold region [88] is much higher than expected because the off-state leakage current from the many switching network transistors also contribute to the measurement. This is a known weakness of our experiment, and it will significantly limit the accuracy of the model extraction, especially in the subthreshold operation regime.
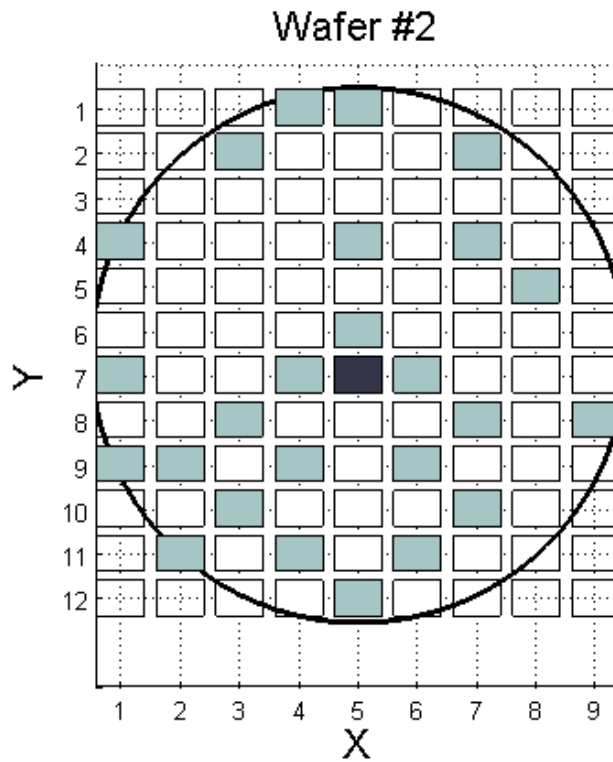

Figure 5.1: Wafer maps of SRAM *I-V* and read/write margin measurements: light tile – chips measured; dark tile – chips used for statistical parameter extraction.
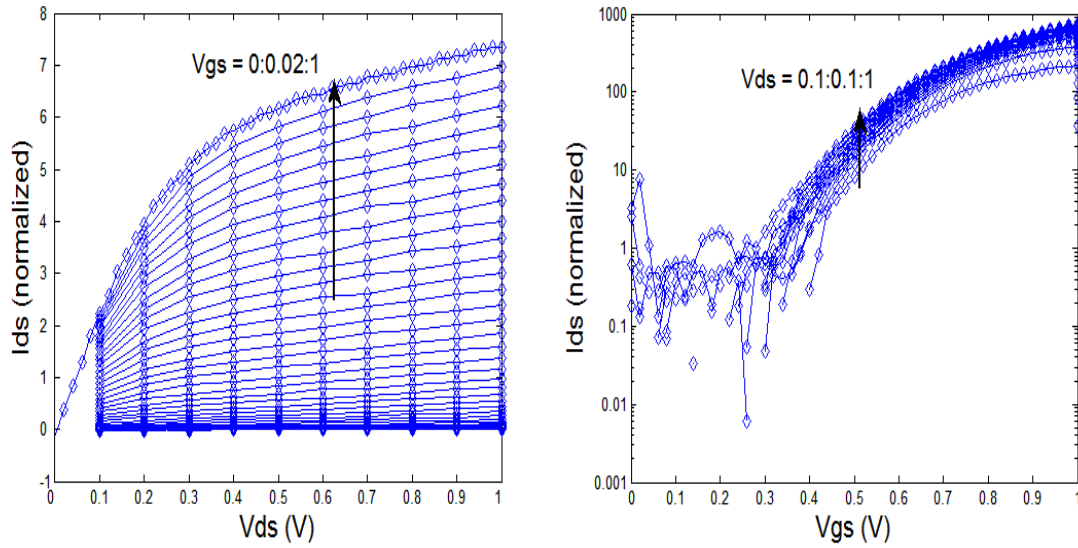
Figure 5.2: *I-V* measurement data from a sample pull-down transistor on the selected die. The left plot shows the *Id-Vd* with stepped *Vgs*, and the right plot shows the *Id-Vg* with stepped *Vds*.

5.3 Parameter Extraction with EKV Model

5.3.1 Parameter Extraction

As described in Chapter 4, we chose ten major parameters from the EKV V2.6 model as candidates for parameter extraction, as listed in Table 5.1. From the prior simulations, we learned that DL and DW are among the first to be excluded. Thus, to save computing time, we chose to exclude these two parameters from the extraction first. The remaining eight parameters will go through the stepwise parameter selection procedure as previously demonstrated using the relative confidence interval (confidence interval normalized by the corresponding extracted parameter value) as the criterion for extraction quality and parameter selection. The target function is a subset of the complete *I-V* data measured, with *Vds* ranging from 0.1V to 1V and *Vgs* from 0V to 1V with a step size of 0.1V (Figure 5.3).

The initial extraction results with eight parameters (or two excluded parameters) are shown in the top half of Figure 5.4, Figure 5.5 and Figure 5.6. Each figure is combining the left and right copies of the same type of transistor; that is, pull-down, pass-gate, or pull-up transistors together, because they are physically very similar devices and highly correlated to each other due to the close physical placement. As we can see, not only do most parameters have a wide distribution and hitting the optimization boundaries

frequently, many of them are also have large correlations with each other, resulting in estimated values that lie in two or more clusters, in the case of *UCRIT* and *LAMBDA*, and/or complex correlation structure such as between *GAMMA* and *PHI*. As the stepwise parameter selection procedure goes on, however, the parameter estimates with strong correlations with others tend to be removed from the optimization, and the *extraction quality* improves without much sacrifice in *fitting quality*. While the fitting error does start to increase when more than six parameters are excluded for all six SRAM bit cell transistors, the extracted compact model parameters have a lot less interactions or dependencies when seven parameters are excluded from extraction. The normalized notional confidence interval lengths show that by excluding six parameters (or including four parameters), the parameter with the worst extraction quality will have the 90th percentile of its confidence interval the in the neighborhood of 100%, which suggests poor reliability. This improves to roughly 10% when only three parameters are included. Therefore, the optimal set of parameters for parameter extraction includes three parameters for all types of transistors, namely *VTO, KP, and LAMBDA*.

Name	Description	Units	Default
DW	Channel width correction	m	0
DL	Channel length correction	m	0
VTO	Long-channel threshold voltage	V	0.5
GAMMA	Body effect factor	$\sqrt{V}$	1.0
PHI	Bulk Fermi potential (2x)	V	0.7
KP	Transconductance parameter	A/V^2	50.0E-6
E0	Mobility reduction coefficient	V/m	1.0E12
UCRIT	Longitudinal critical field	V/m	2.0E6
LAMBDA	Depletion length coefficient	-	0.5
LETA	Short channel effect coefficient	-	0.1

Table 5.1: Candidate of EKV model parameters for extraction [29]
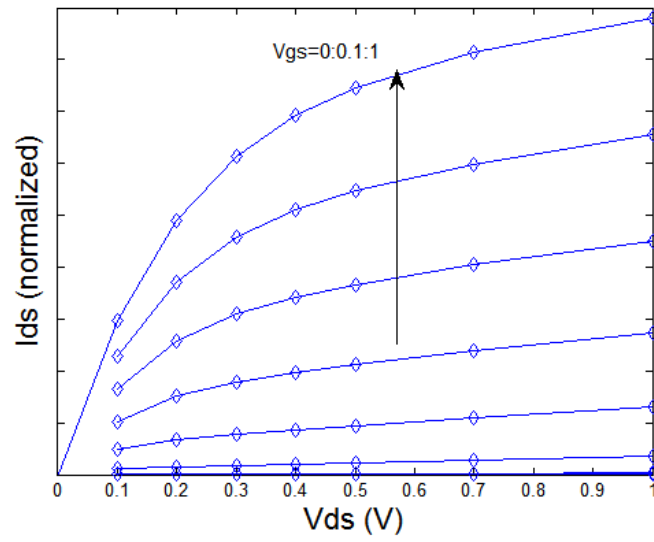
Figure 5.3: Target *I-V* data for EKV model parameter extraction.

Exclude 2 parameters
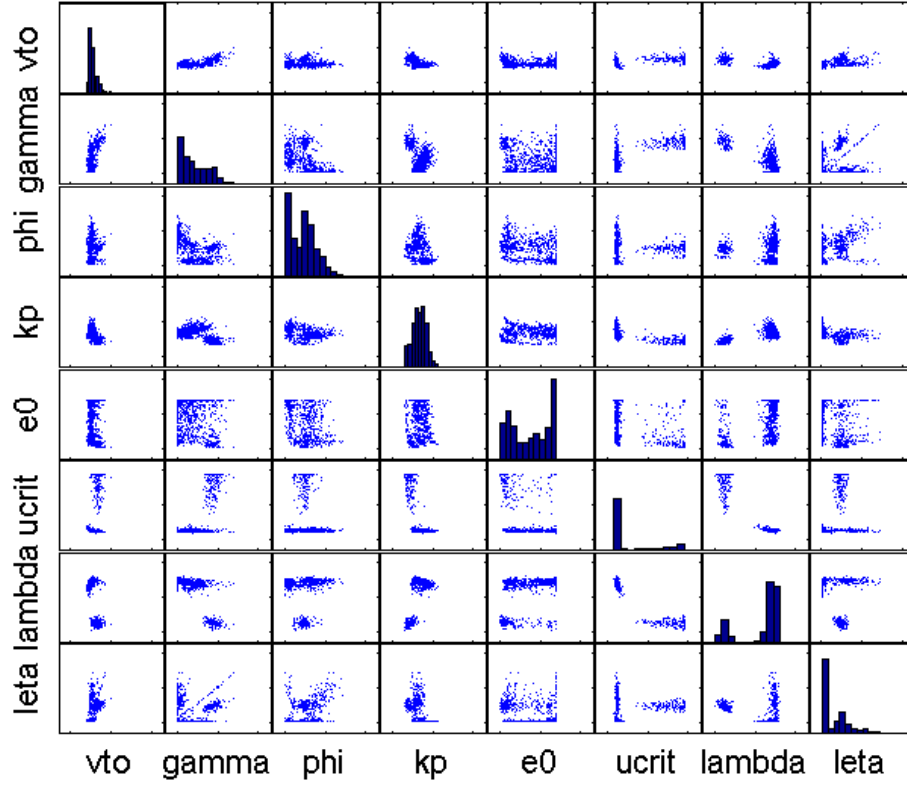
Exclude 6 parameters
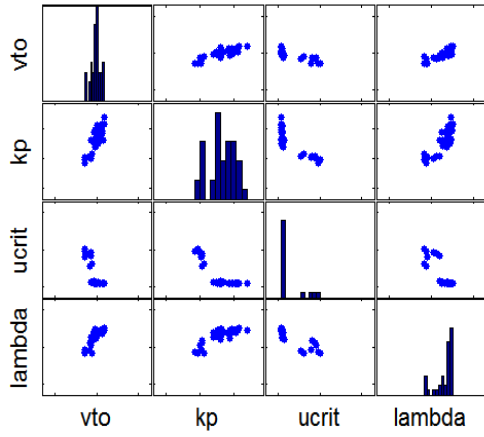
Exclude 7 parameters

Figure 5.4: Stepwise parameter selection results for the pull-down transistors (PD1/PD2). Subplots showing the initial extracted parameters without any exclusion, the final optimal parameter set, and the change in normalized confidence interval length and sum of squared fitting errors (SSE) after each round of selection.

Figure 5.5: Stepwise parameter selection results for the pass-gate transistors (PG3/PG4). Subplots showing the initial extracted parameters without any exclusion, the final optimal parameter set, and the change in normalized confidence interval length and sum of squared fitting errors (SSE) after each round of selection.

Exclude 2 parameters

Exclude 6 parameters

Exclude 7 parameters
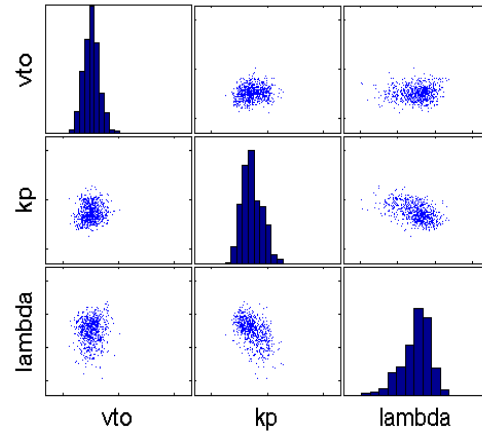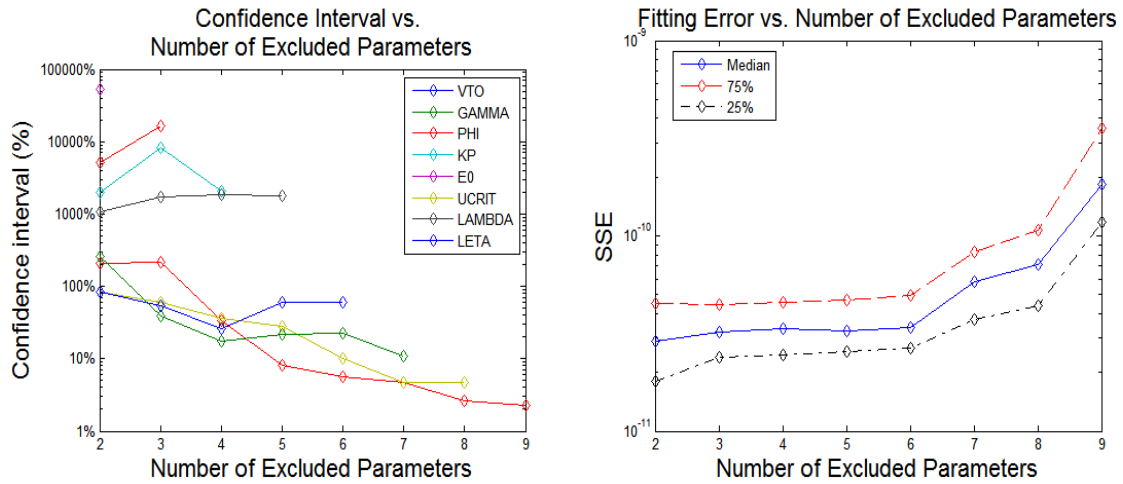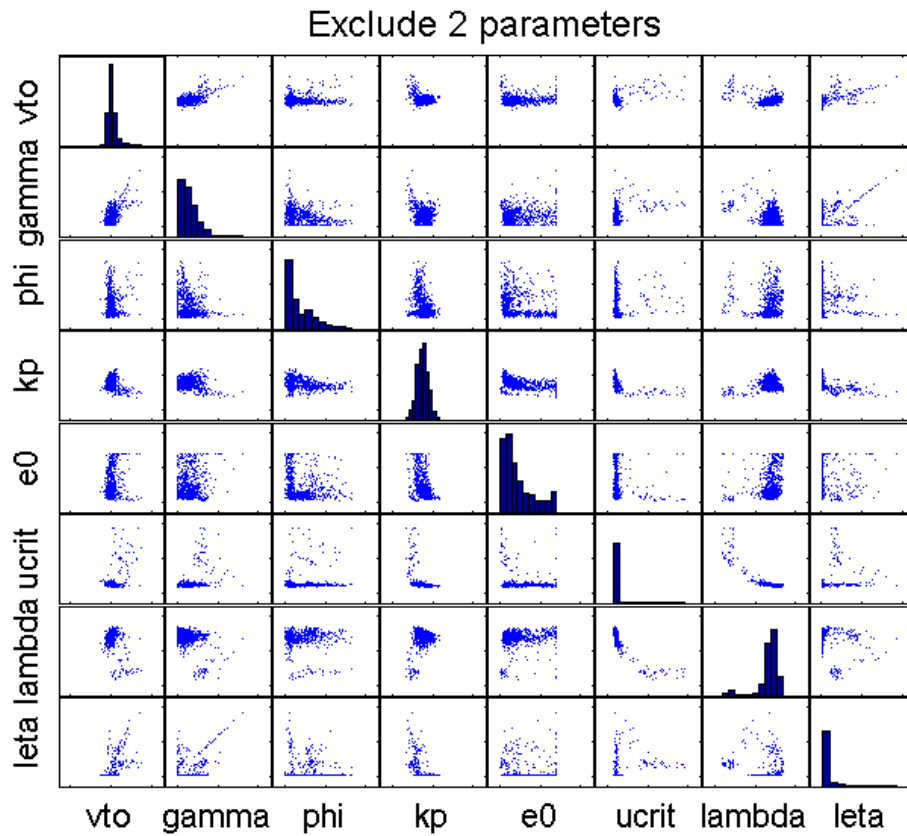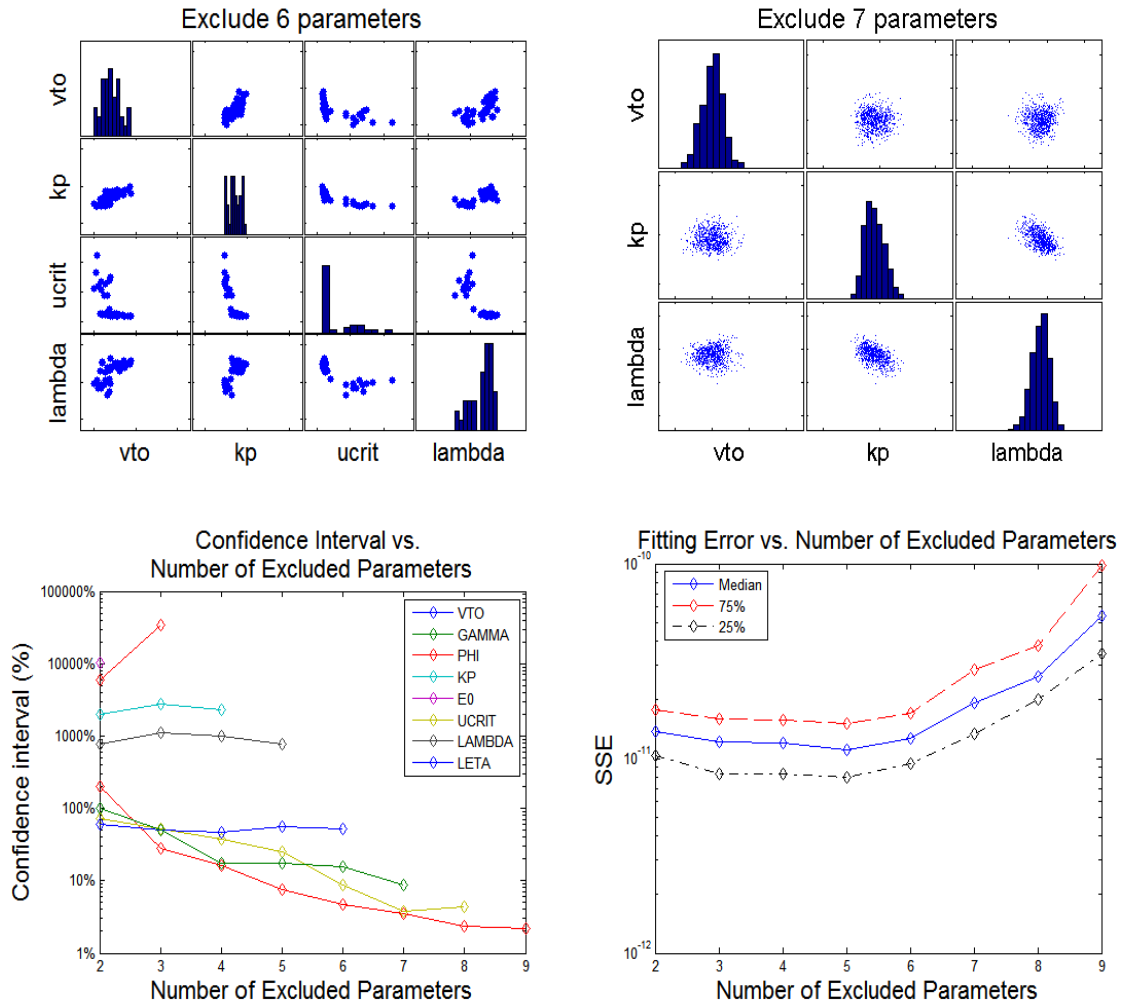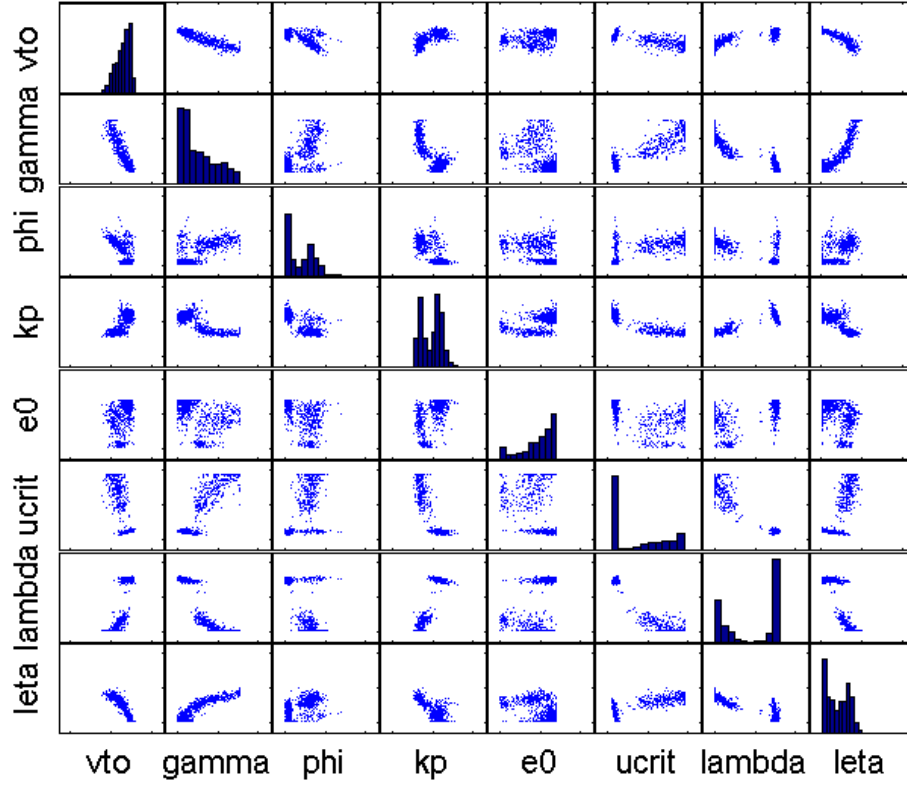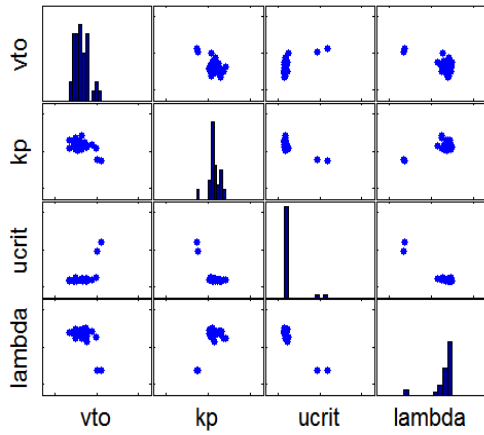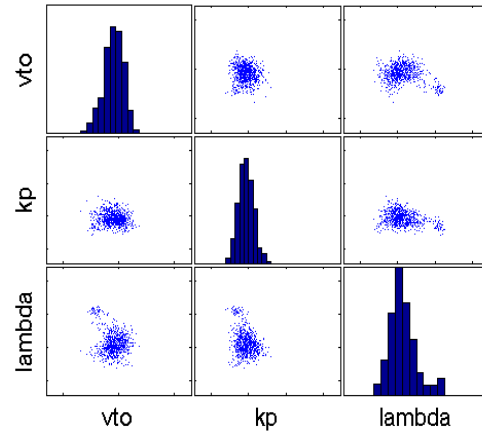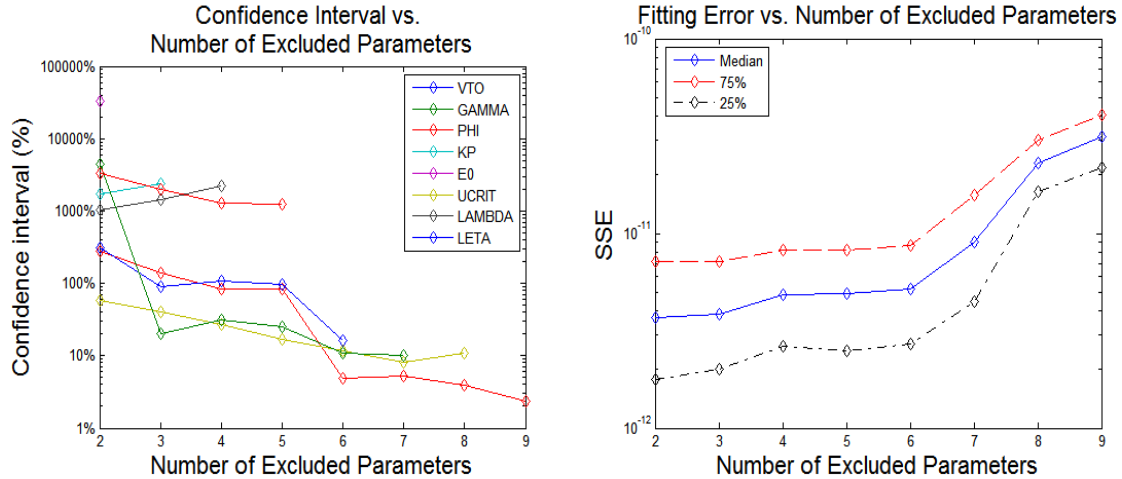
Figure 5.6: Stepwise parameter selection results for the pull-up transistors (PU5/PU6). Subplots showing the initial extracted parameters without any exclusion, the final optimal parameter set, and the change in normalized confidence interval length and sum of squared fitting errors (SSE) after each round of selection.

5.3.2 Parameter Variability Modeling

The within-chip spatial pattern of the extracted parameters (*VTO, KP,* and *LAMBDA*) of the three types of SRAM transistors are shown in Figure 5.7, Figure 5.8 and Figure 5.9, respectively. The threshold parameter *VTO* does not show any significant across-chip pattern, which is in line with the fact that threshold voltage variation is mainly the result of random dopant fluctuation and is largely dominated by the random components. On the other hand, the parameters *KP* and *LAMBDA* both show a clear across-chip pattern that varies along the rows of the SRAM array for all the NMOS transistors (pull-down transistors and pass-gates), while PMOS does not show such a systematic pattern. These across-chip patterns relate closely to the spatial pattern we see in the measured SRAM bit-cell and transistor electrical metrics in Chapter 3.

In the same way we decomposed the variability in the measured electrical device metrics, we apply our hierarchical variability model to the extracted parameters *KP* and *LAMDA* for the two PD transistors and the two PG transistors (Equation 5.1). Parabolic surfaces along chip rows are fitted to the extracted compact model parameters, as shown in Figure 5.10 and Figure 5.11. Figure 5.12 and Figure 5.13 show normal quantile plots of the original extracted parameters, the fitted across-chip systematic component, and the residuals of the fitted model parameters. The original extracted values of both *KP* and *LAMBDA* clearly deviate from Normal distributions for both types of NMOS transistors. After the removal of the fitted across-chip systematic component, the distribution of

residuals of the parameter *KP* is approximately Gaussian. However, the same cannot be said for parameter *LAMBDA*, whose distribution has such long tails on the lower end that even after removing the systematic component, the residuals still do not appear normal. This is illustrated in Figure 5.15: the standard deviation of the residual of *LAMBDA* after fitting the across-chip systematic pattern also has a systematic across-chip pattern. The residual variance is larger at the top and bottom rows and smaller in the center. For that reason, we fit a systematic across-chip function to the standard deviation of the across-chip residual of *LAMBDA* ($LAMBDA_{ACR}$). The variance of $LAMBDA_{ACR}$ within each row is also approximately quadratic in row position Y_C, as stated in Equation 5.2, which is incorporated into the hierarchical variability model.

$$LAMBDA\langle T - WP \rangle = LAMBDA\langle T - WP \rangle_{AC} + LAMBDA\langle T - WP \rangle_{ACR}$$

(5.1)

$$LAMBDA\langle T - WP \rangle_{AC} = 0 \times X_C^2 + 0 \times X_C + c_C \times Y_C^2 + d_C Y_C + e_C$$

$$LAMBDA\langle T - WP \rangle_{ACR} \sim N\left(0, \sigma^2(Y_C)\right)$$

(5.2)

$$\sigma(Y_C) = 0 \times X_C^2 + 0 \times X_C + s_C \times Y_C^2 + t_C Y_C + r_C$$
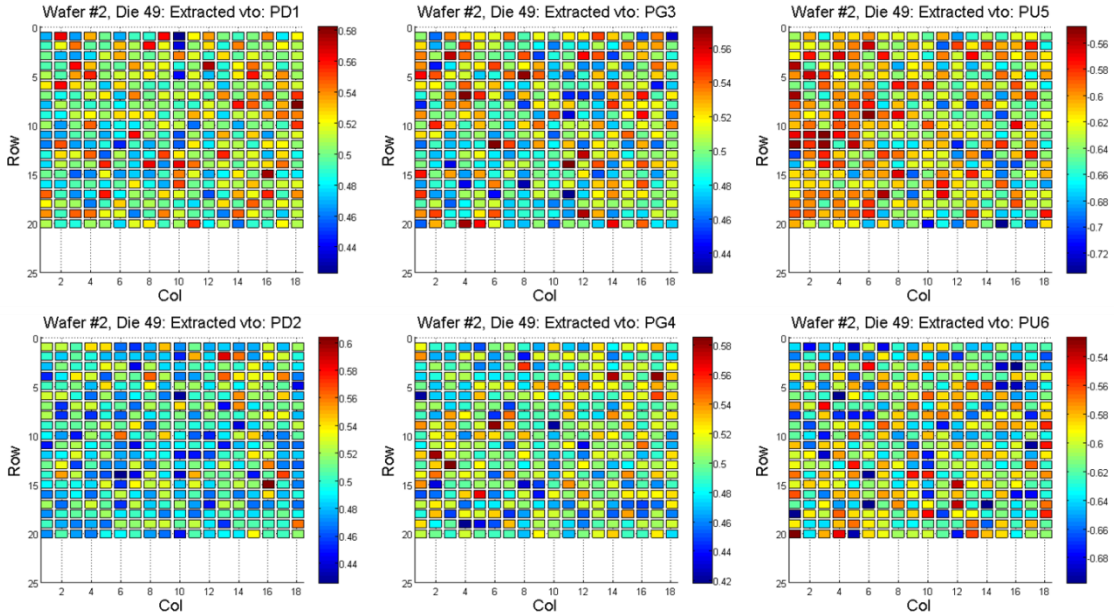
Figure 5.7: Chip maps of extracted compact model parameters *VTO*.
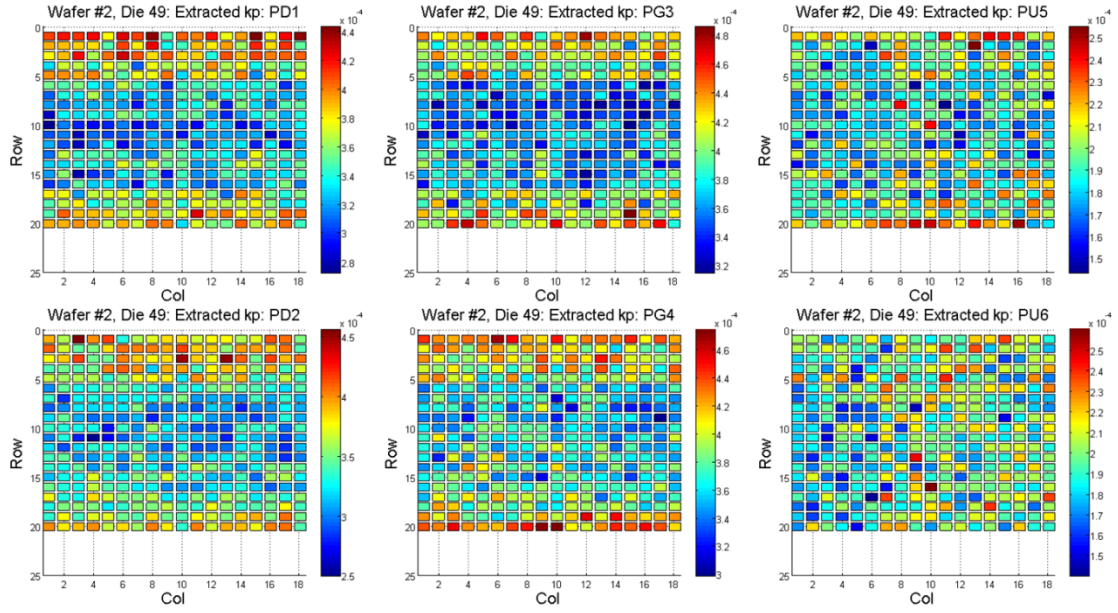
Figure 5.8: Chip maps of extracted compact model parameters *KP*.
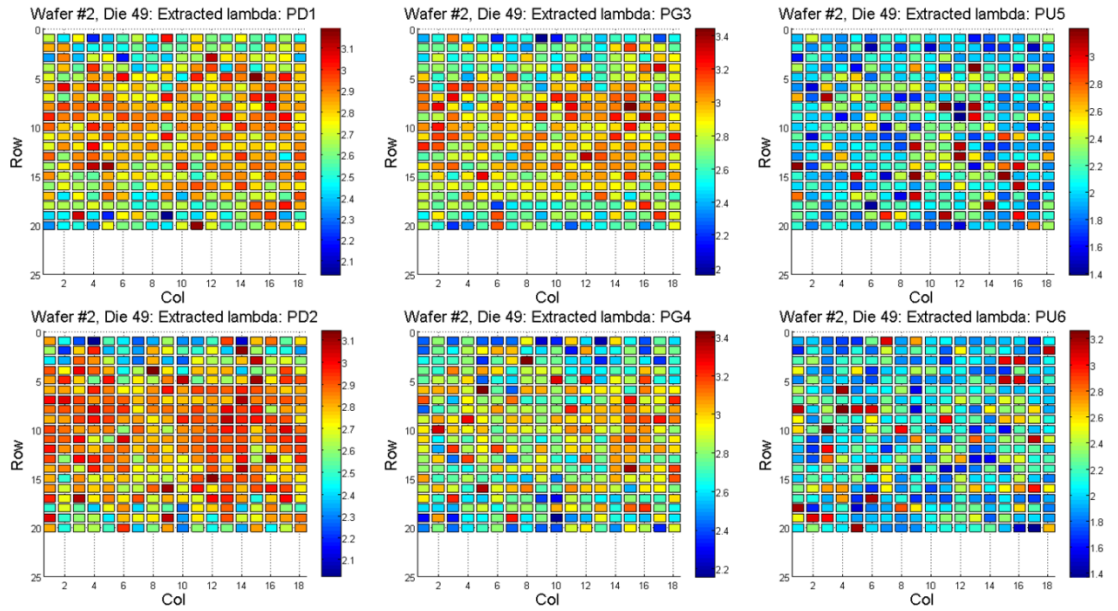
Figure 5.9: Chip maps of extracted compact model parameters *LAMBDA*.
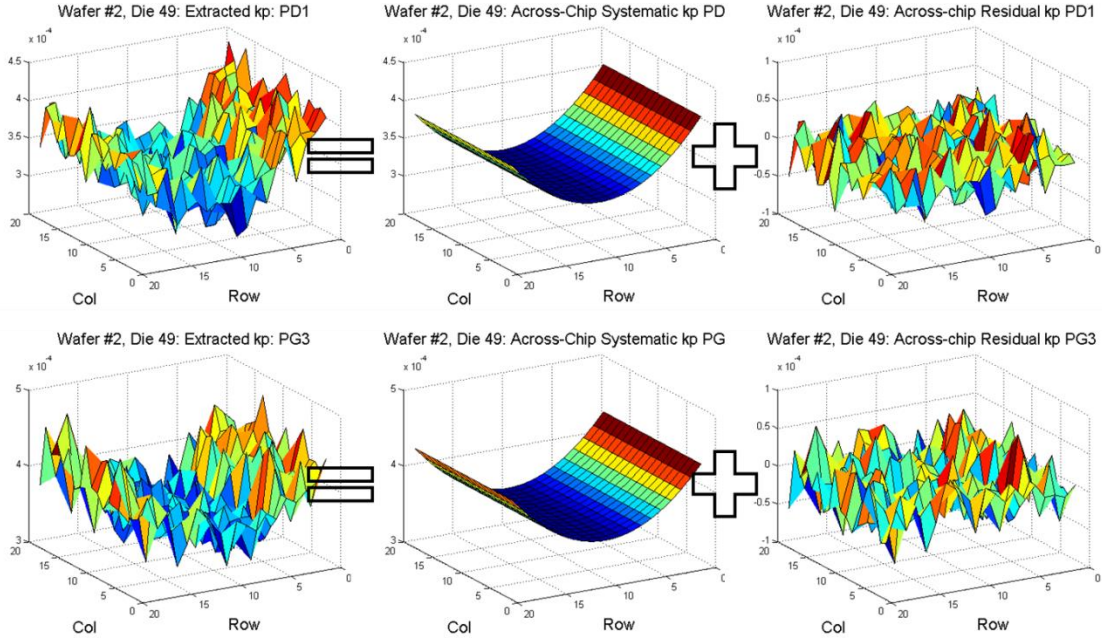
Figure 5.10: Chip level variation decomposition for *KP* extracted from the left pull-down transistor and pass-gate: $KP\langle -DWP\rangle = KP\langle T - WP\rangle_{AC} + KP\langle T - WP\rangle_{ACR}$.
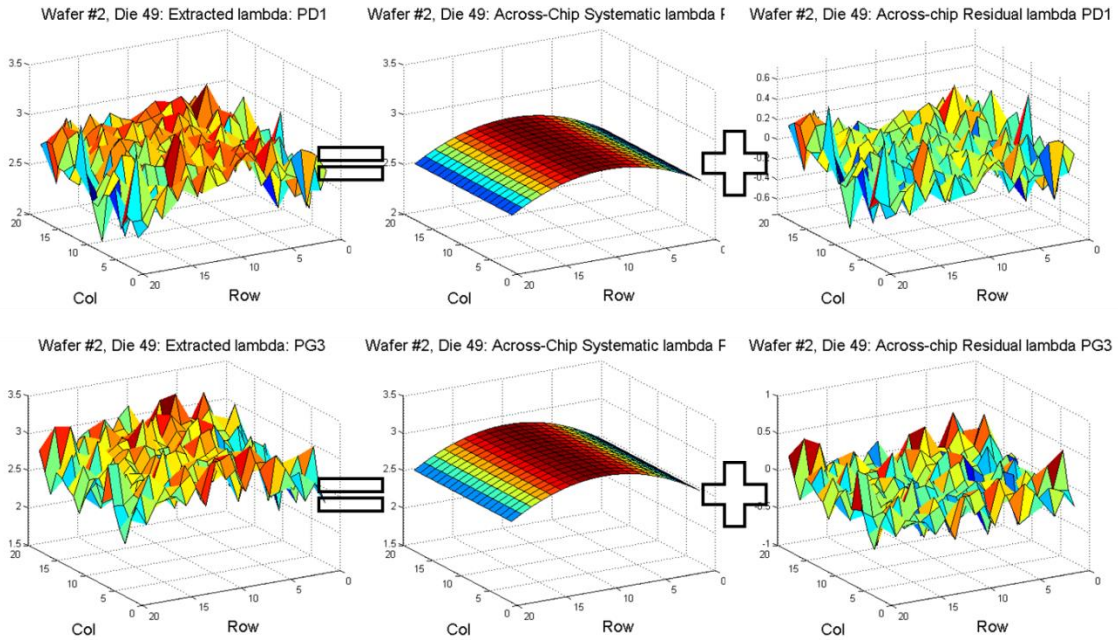
Figure 5.11: Chip level variation decomposition for *LAMBDA* extracted from the left pull-down transistor and pass-gate: $LAMBDA\langle T - WP\rangle = LAMBDA\langle T - WP\rangle_{AC} + LAMBDA\langle T - WP\rangle_{ACR}$.
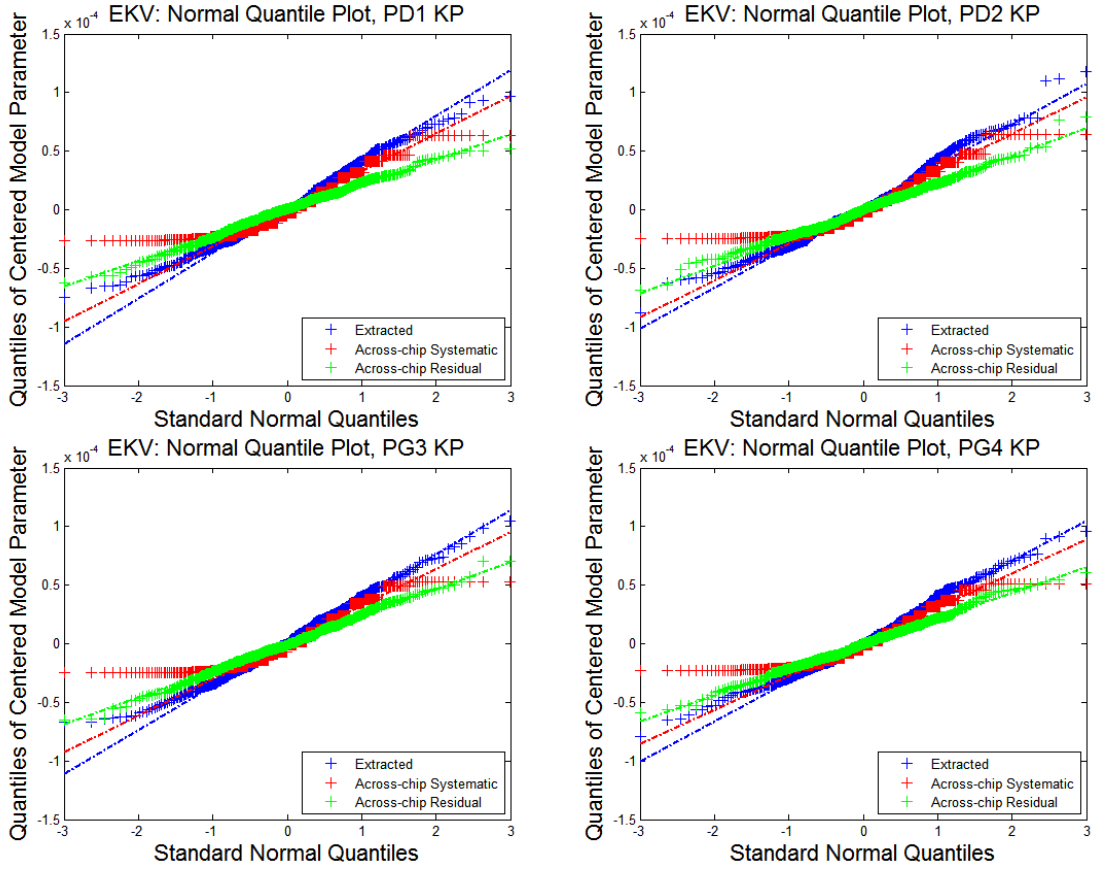
Figure 5.12: Comparison of the distributions of the extracted parameter *KP* and its corresponding systematic and random components for pull-down and pass-gate transistors.
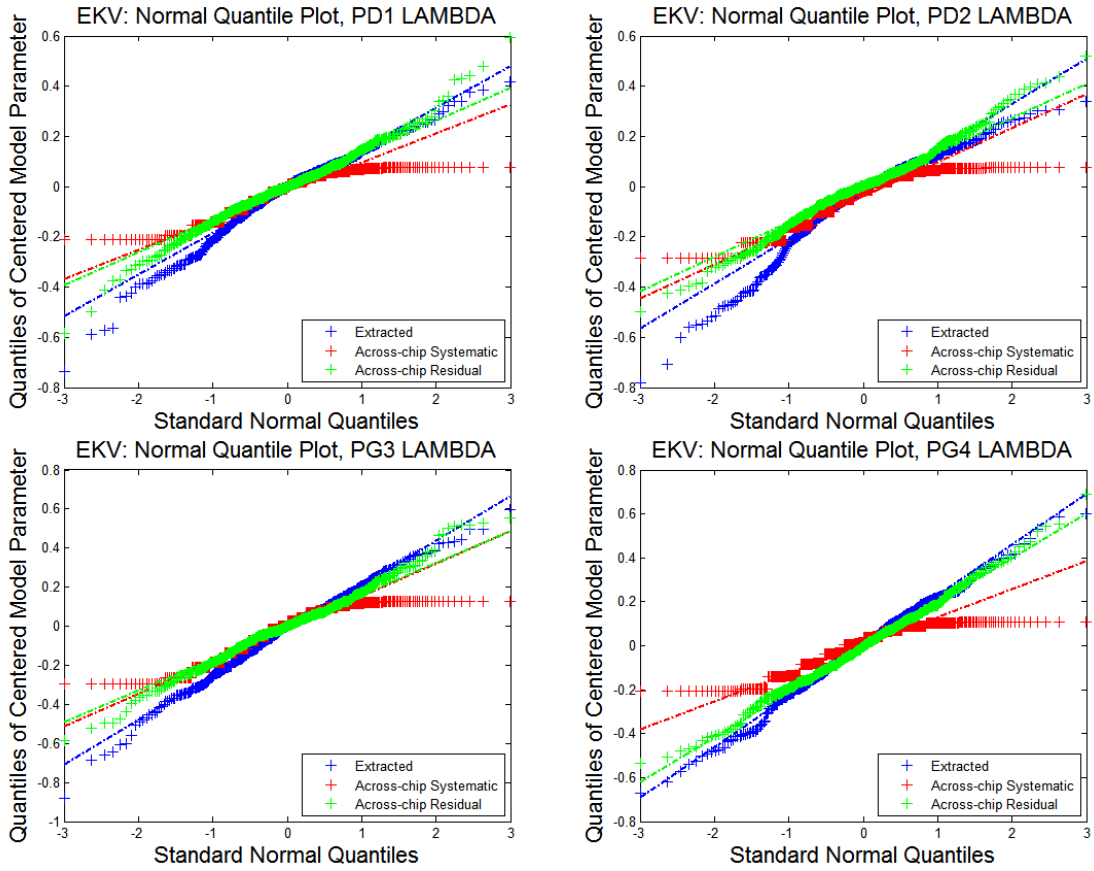
Figure 5.13: Comparison of the distribution of the extracted parameter *LAMBDA* and its corresponding systematic and random components for pull-down and pass-gate transistors.
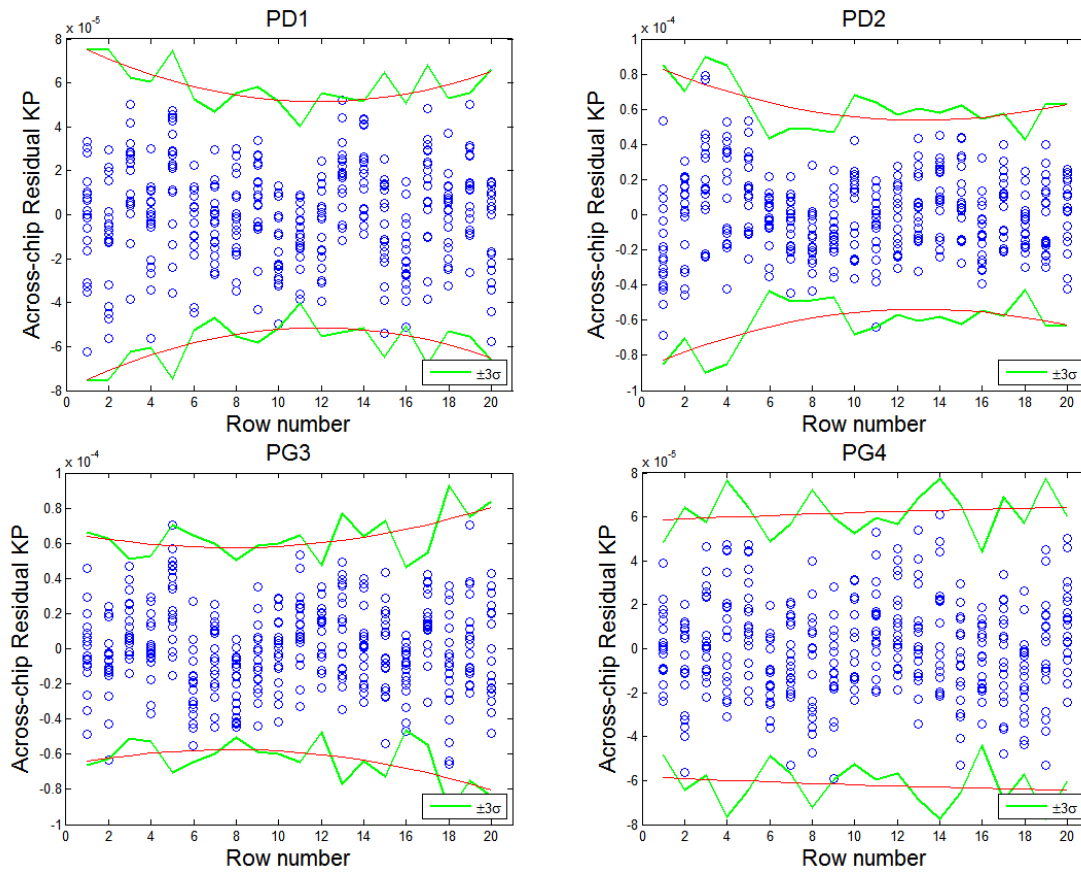
Figure 5.14: Across-chip systematic pattern of the within-row variance of the across-chip residual of parameter *KP*.
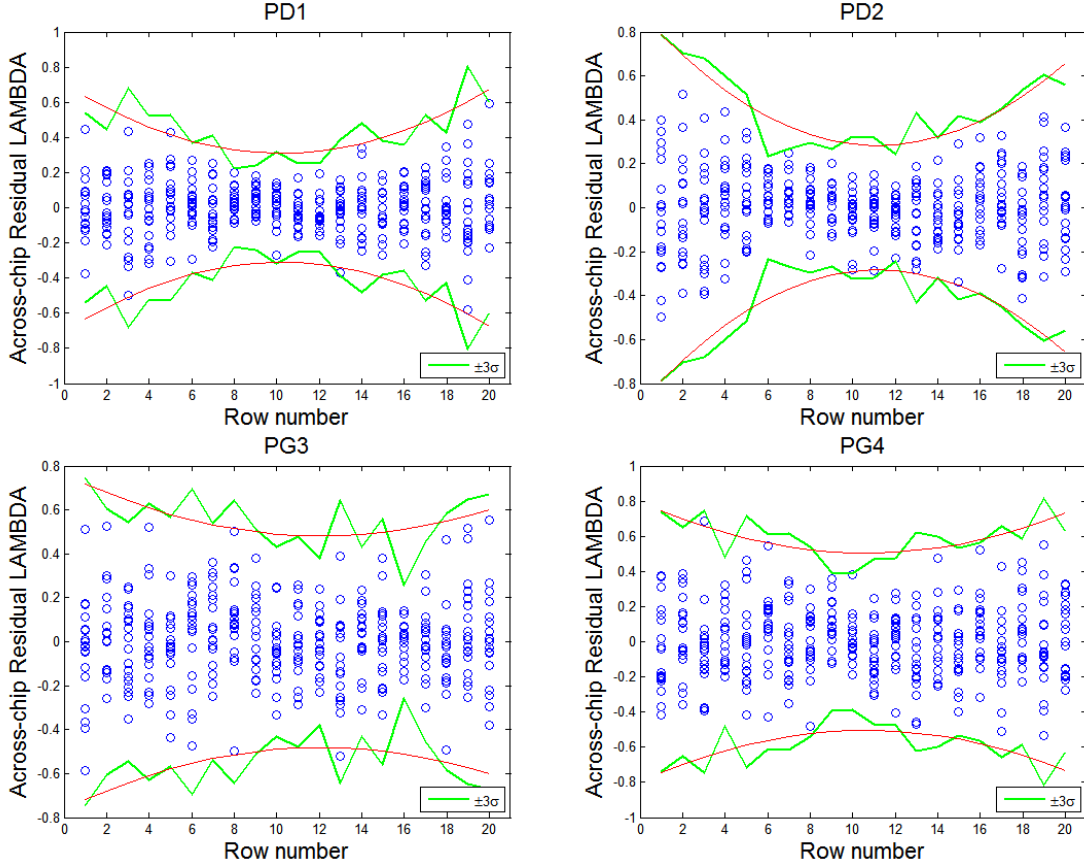
Figure 5.15: Across-chip systematic variation of the within-row variance of the across-chip residual of parameter *LAMBDA*.

5.3.3 Parameter Variability Reconstruction

With the above decomposition of parameter variability, it becomes possible to simulate the distributions of the extracted parameters. As in Chapter 3, we evaluate two variability models: the conventional "Global+Local" model and our hierarchical variability model. Now instead of assuming a constant variance for the local variation (as in the "Global+Local" model) or the local residual variation (as in the hierarchical model), we consider the correlations among parameters, and the correlation among parameters of different devices in general. A two parameter case will be used as an example to illustrate the reconstruction process. Assume we have the i_1th parameter of the j_1th device p_{i_1,j_1} and the i_2th parameter of the j_2th device p_{i_2,j_2}. The indices i_1 and i_2 may be equal and the indices j_1 and j_2 may be equal, but not both at the same time. Under the conventional model, p_{i_1,j_1} and p_{i_2,j_2} are assumed to be correlated Normal variables with a correlation matrix estimated from the extracted distributions of the two compact model parameters, as described by Equation 5.3. Under the hierarchical model, each parameter is the sum of its

corresponding systematic component function, f_{AC}, and a random component, r. The random component is generated in the same way as the "Global+Local" variation model, but replacing the original extracted parameter values with the residuals after removal of the systematic component as described by Equation 5.4.

"Global+Local" variation model:

$$(p_{i_1,j_1}, p_{i_2,j_2}) \sim N(\mu_1, \mu_2, \Sigma_{Local})$$

$$\Sigma_{local} = cov(p_{i_1,j_1}, p_{i_2,j_2})$$

$$(5.3)$$

Hierarchical model:

$$p_{i_1,j_1} = f_{i_1 j_1, AC}(X_C, Y_C) + r_{i_1,j_1}$$

$$p_{i_2,j_2} = f_{i_2 j_2, AC}(X_C, Y_C) + r_{i_2,j_2}$$

$$(r_{i_1,j_1}, r_{i_2,j_2}) \sim N(0, 0, \Sigma_{Local}^2)$$

$$\Sigma_{local} = cov(r_{i_1,j_1}, r_{i_2,j_2})$$

$$(5.4)$$

A total of ten chips of model cards (360x10=3600) are generated with this methodology for both the "Global+Local" model and the hierarchical model. A comparison among the original estimates of extracted parameters, parameters simulated from the "Global+Local" model, and those simulated from the hierarchical model, is shown in Figure 5.16 and Figure 5.17. The hierarchical model is capturing the non-Gaussian behavior of the original extracted parameter distribution fairly well, especially at the lower end of the tails, while the conventional "Global+Local" model strictly follows the Normal distributions thus deviating from the original extraction.

To quantify the difference of the distributions of the original estimates of extracted compact model parameters and those simulated with the "Global+Local" model as well as the hierarchical model, we compare quantiles across models. For example, the 1% quantile of *KP* predicted by the "Global+Local" model can be as much as 6% lower than the original estimates of extracted parameters, and 5% lower in the case of parameter *LAMBDA*. As comparison, the 1% quantile of parameters predicted by the hierarchical model is generally within 3% of that from the original distribution of parameter estimates. The accuracy of the hierarchical model tends to increase when examining even more extreme quantiles. A detailed list of the extreme quantiles of the original and reconstructed parameters can be found in Table 5.2.
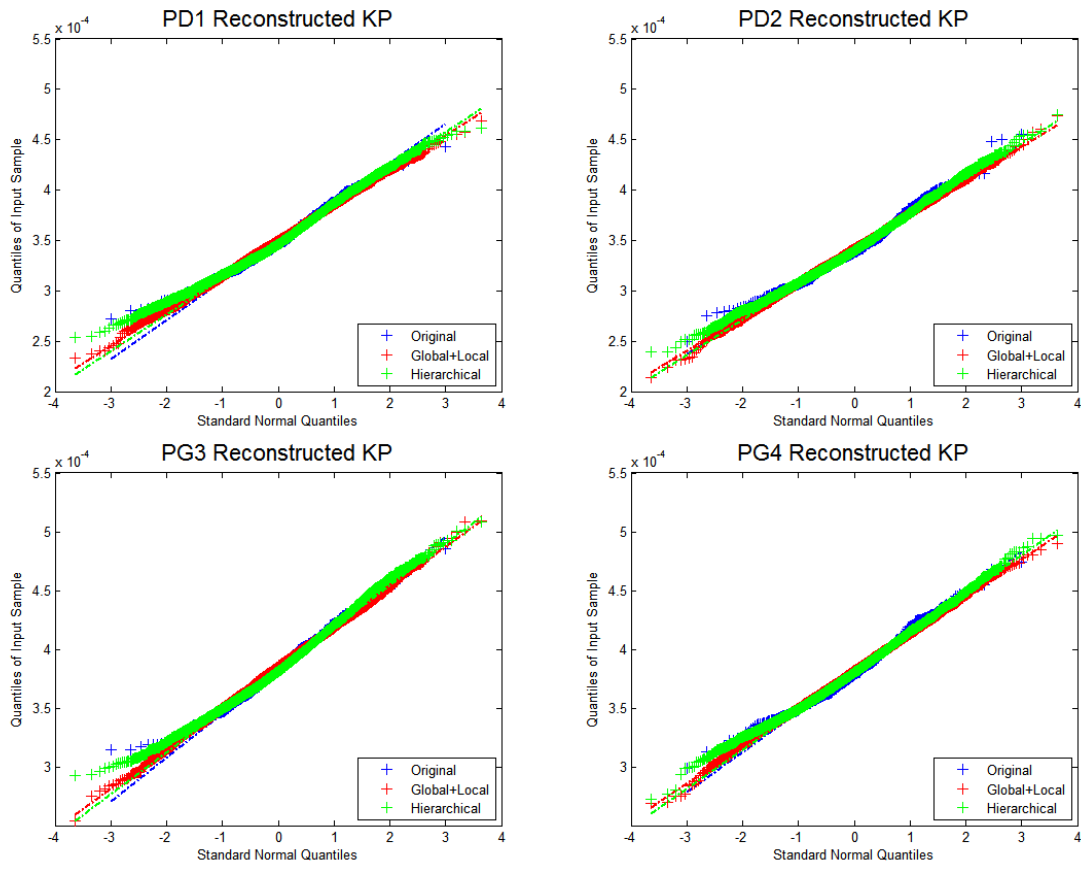
Figure 5.16: Comparison of the distribution of the extracted parameter *KP* and the reconstructed distributions using the "Global+Local" model and the hierarchical model.
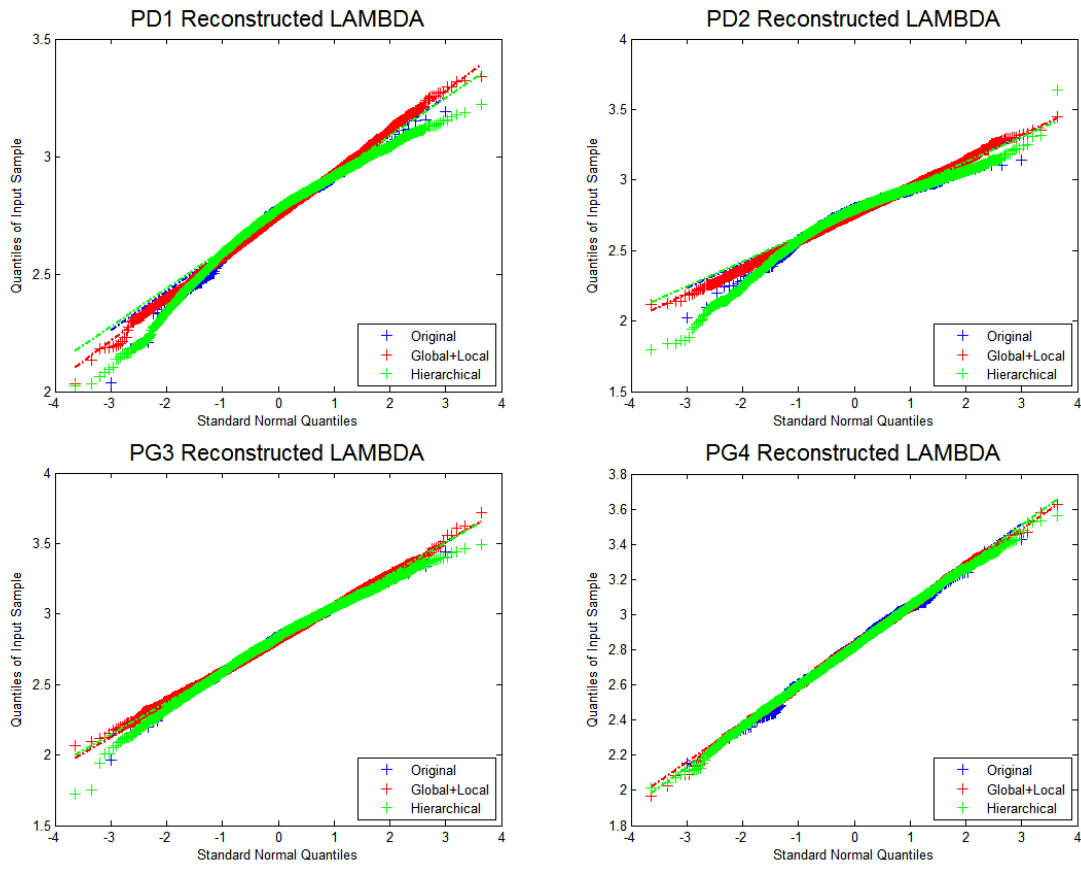
Figure 5.17: Comparison of the distribution of the extracted parameter *LAMBDA* and the reconstructed distributions using the "Global+Local" model and the hierarchical model.

Parameter	Device	Percentile	original	"Global+Local" Model		Hierarchical Model	
			Value	Value	Error	Value	Error
KP	PD1	1%	2.82E-04	2.70E-04	-4%	2.82E-04	0%
		99%	4.28E-04	4.29E-04	0%	4.35E-04	2%
	PD2	1%	2.79E-04	2.64E-04	-5%	2.71E-04	-3%
		99%	4.17E-04	4.27E-04	2%	4.28E-04	3%
	PG3	1%	3.18E-04	3.00E-04	-6%	3.15E-04	-1%
		99%	4.66E-04	4.65E-04	0%	4.72E-04	1%

	PG4	1%	3.17E-04	3.06E-04	-3%	3.13E-04	-1%
		99%	4.55E-04	4.54E-04	0%	4.58E-04	1%
LAMBDA	PD1	1%	2.221	2.328	5%	2.276	2%
		99%	3.114	3.185	2%	3.093	-1%
	PD2	1%	2.243	2.309	3%	2.170	-3%
		99%	3.101	3.206	3%	3.111	0%
	PG3	1%	2.202	2.294	4%	2.200	0%
		99%	3.287	3.375	3%	3.291	0%
	PG4	1%	2.285	2.305	1%	2.285	0%
		99%	3.350	3.353	0%	3.319	-1%

Table 5.2: 99% and 1% quantiles of original and reconstructed parameter distributions.

5.4 Parameter Extraction with PSP Model

5.4.1 Parameter Extraction

The PSP model extraction is carried out with the same set of SRAM bit cell transistor *I-V* data used for the EKV model. The extraction setup is inherited from the simulation in Chapter 4, with the same group of model parameters and the same stepwise parameter selection method in a three-step sequential parameter extraction flow. The difference is that we include the printed gate length deviation *DL* as an additional parameter, even though it is not one of the local PSP model parameters, and that one type of imagined NMOS transistor is replaced by six real transistors from the SRAM bit cells. The list of PSP model parameters as candidates for extraction is provided in Table 5.3.

During every round of parameter selection, each pair of the mirror-imaged transistors of the same type, pull-down, pass-gate, and pull-up transistors, is grouped together due to that pair's highly similar physical nature. The three-step sequential extraction is carried out in their respective operation region, as shown in Figure 5.18. The stepwise parameter selection results are illustrated in Figure 5.19 through Figure 5.23 using Step#1 (linear region *Id-Vg*) as an example. As shown in these plots, the fitting error increases when fewer

than four parameters are included in the model extraction, while the length of the normalized confidence interval of the extracted parameters drops below 100% when three parameters or fewer are included. Combined with the observations regarding the distribution and correlation structure of the extracted parameters, we decided that three parameters, *VFBO*, *UO,* and *RSW1,* are an adequate set of model parameters for Step#1 curve fitting. This combination of parameters offers well-bounded, reasonable distributions with minimal loss of fit. The number of "extractable" parameters is significantly less than the simulations suggest, largely because real silicon devices do not act exactly like ideal model devices, and measurement data noise will make it hard to extract parameters that have little influence on performance. The second and third steps of the optimization go through the same stepwise parameter selection procedure, with the initial model card of each step inherited from the previous step and the excluded parameter set to its extracted value attained in the previous step (if it is previously extracted). In this way, we add more extractable parameters as we go through more optimization steps, while keeping the good parameter values extracted in earlier optimization steps but not in the later steps. For PD and PG transistors, Step#2 will extract *VFBO*, *UO*, *MUEO* and Step#3 will extract *VFBO* and *UO*; for PU transistors, Step#2 will extract *VFBO, CFL,* and *UO,* and Step#3 will extract *VFBO* and *UO*.

After all three optimization steps are complete, we perform one final global optimization step, which uses the complete measurement data (including all the data points from all three steps) to fit all the optimized parameters as the initial candidates for parameter extraction. Again we apply the parameter selection scheme to this global optimization step, and we are able to reduce the extractable parameters down to three for each type of transistors: *VFBO, UO,* and *RSW1* for the PD and PG transistors and *VFBO, CFL,* and *UO* for the PU transistors. The initial extraction results with all four extractable parameters from the sequential extraction and the final extraction results with three extractable parameters are shown in Figure 5.20, Figure 5.22 and Figure 5.24.

Param.	Description	Param.	Description
vfbo	Geometry-independent flat-band voltage	*cso*	Geometry-independent Coulomb scattering
nsubo	Geometry-independent substrate doping	*xcoro*	Geometry-independent non-universality
dphibo	Geometry-independent offset of ϕ_B	*rsw1*	Source/drain series resistance

cto	Geometry-independent part of interface states factor CT	thesato	Geometry-independent velocity saturation
cfl	Length-dependence of CT	alpl	Length-dependence of CLM pre-factor ALP
uo	Zero-field mobility at TR	alp1l1	Length-dependence of CLM enhancement factor above threshold
xmueo	Geometry-independent mobility reduction coefficient	alp2l1	Second order length-dependence of ALP1
themuo	Mobility reduction exponent	vpo	CLM logarithmic dependence

Table 5.3: Candidate of PSP model parameters for extraction [87].
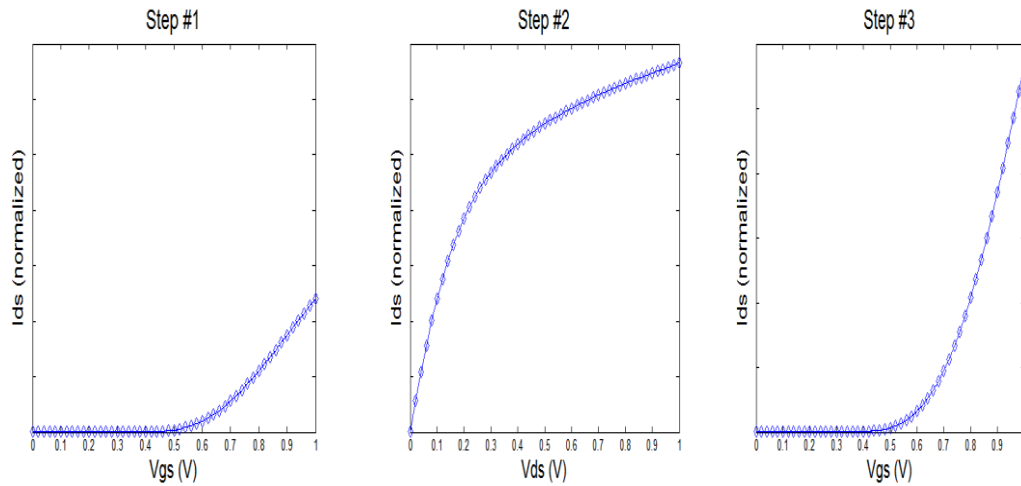
Figure 5.18: Target *I-V* data for the three extraction steps. From left to right: *Id-Vg* with Vds=0.1V, *Id-Vd* with *Vgs*=1V, and *Id-Vg* with Vds=1V.

Exclude 0 parameters

Exclude 13 parameters

Exclude 14 parameters
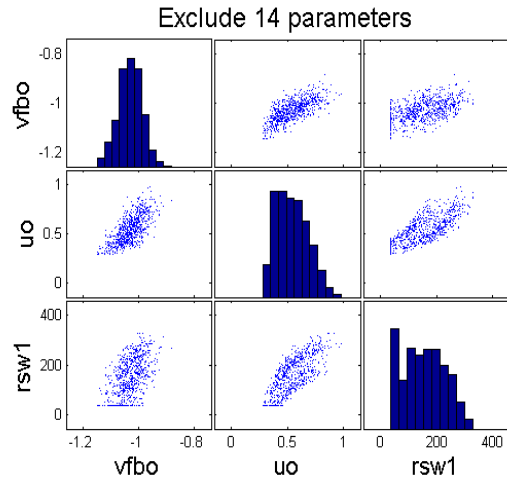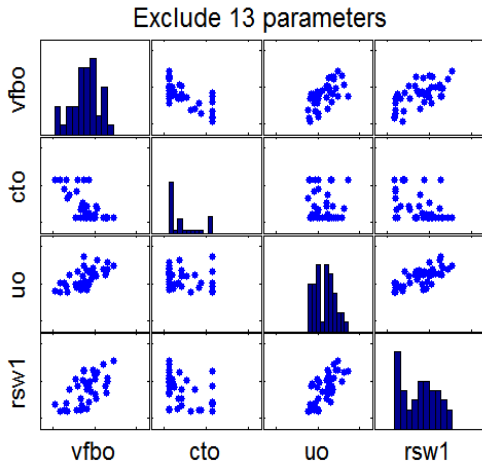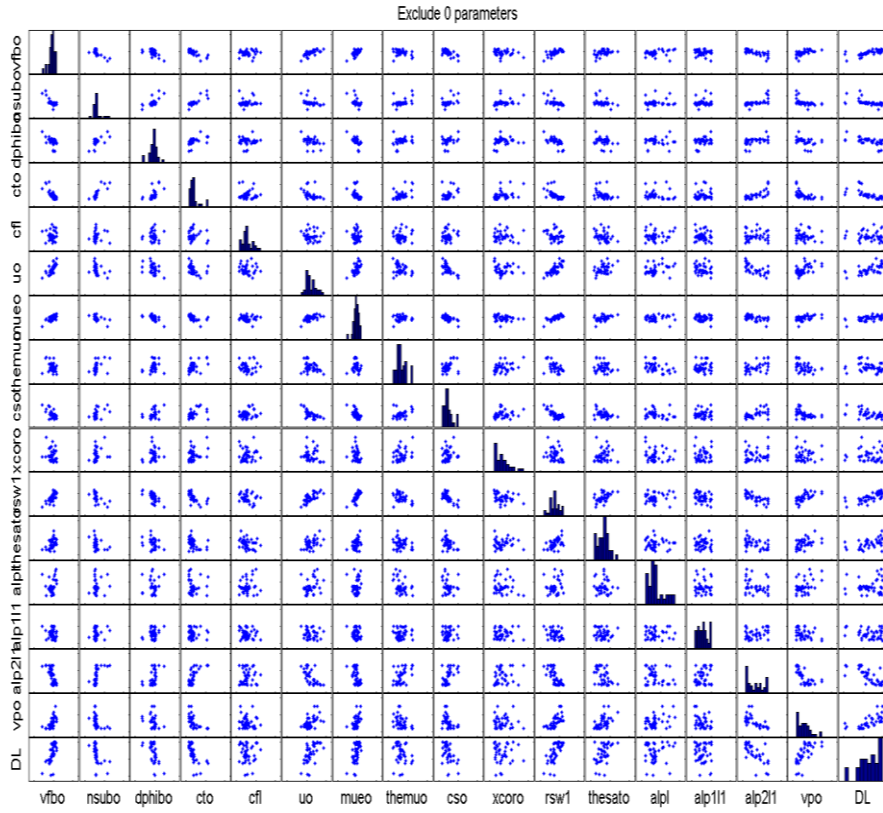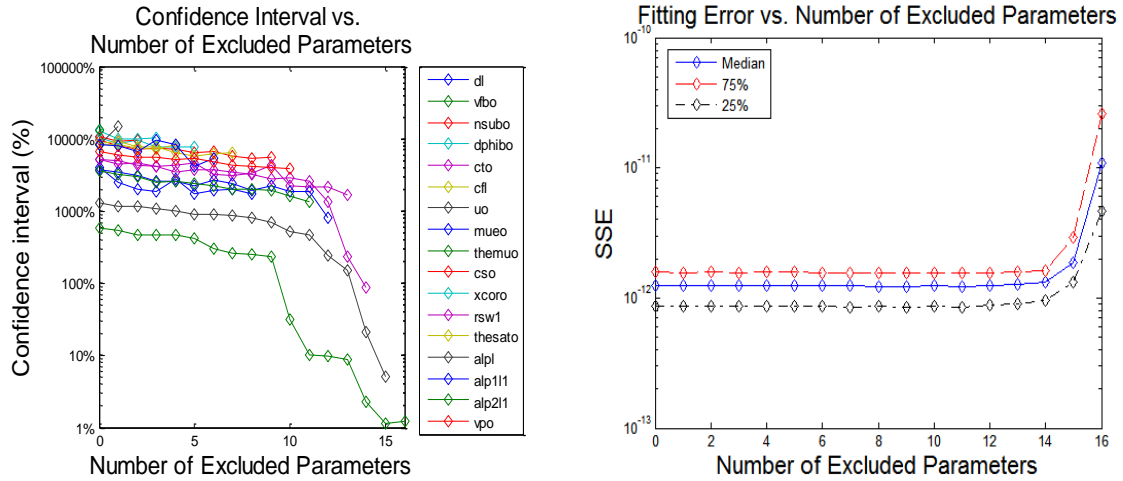
Figure 5.19: Stepwise parameter selection results from Step#1 for the pull-down transistors (PD1/PD2). Subplots showing the initial extracted parameters without any exclusion, the final optimal parameter set, and the change in normalized confidence interval and sum of squares of fitting error (SSE) after each round of selection.
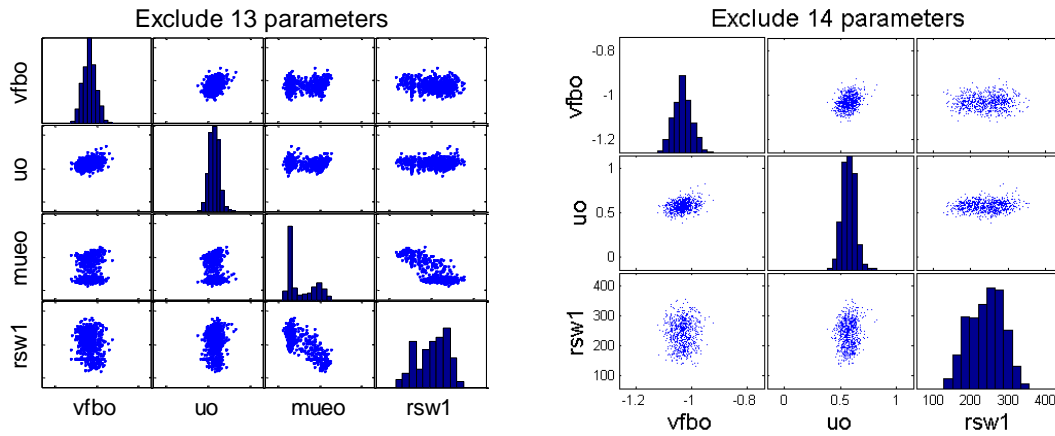
Figure 5.20: Initial and final extracted values after global optimization for the pull-down transistors (PD1/PD2).

Exclude 0 parameters

Exclude 13 parameters

Exclude 14 parameters

118
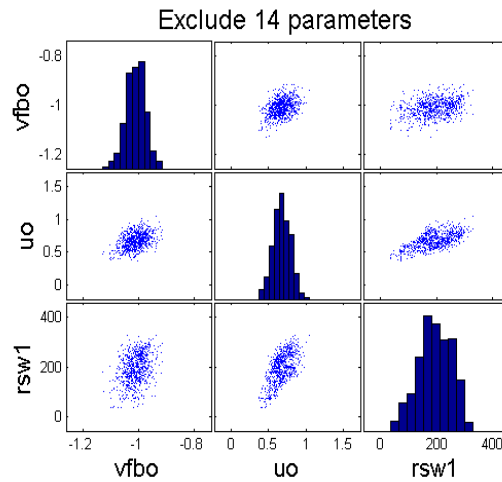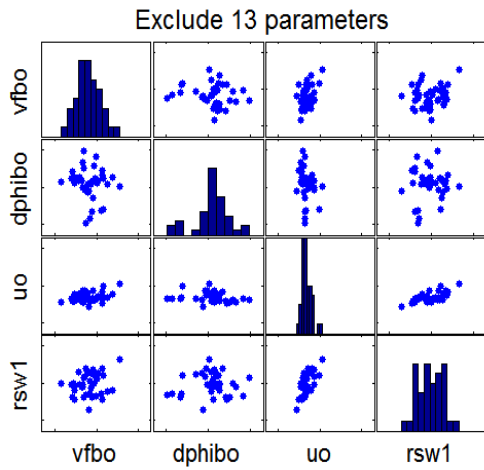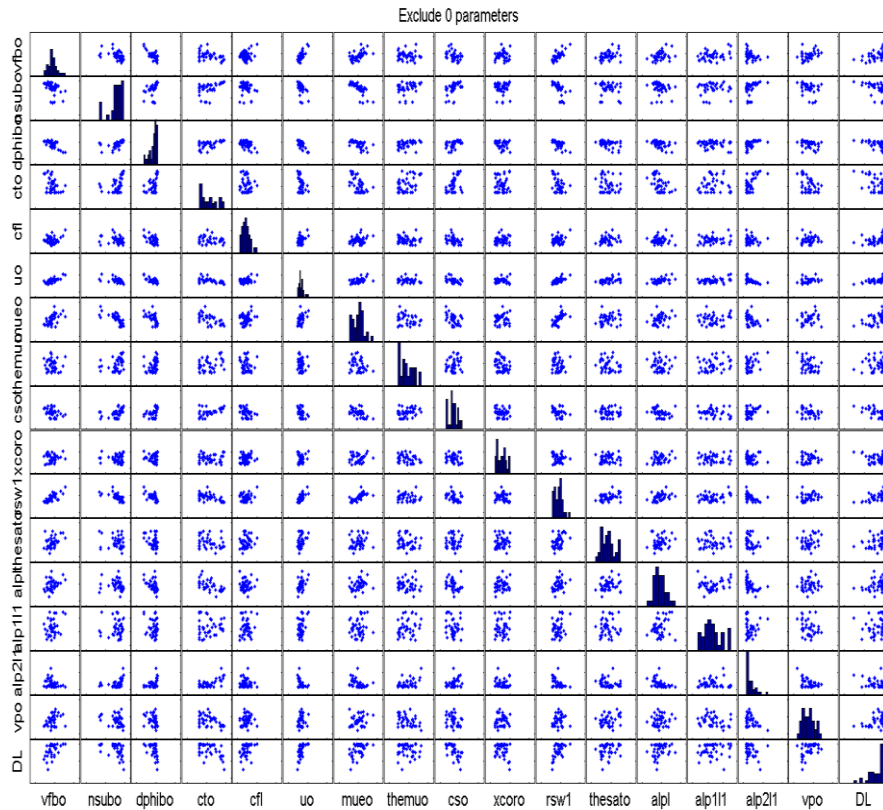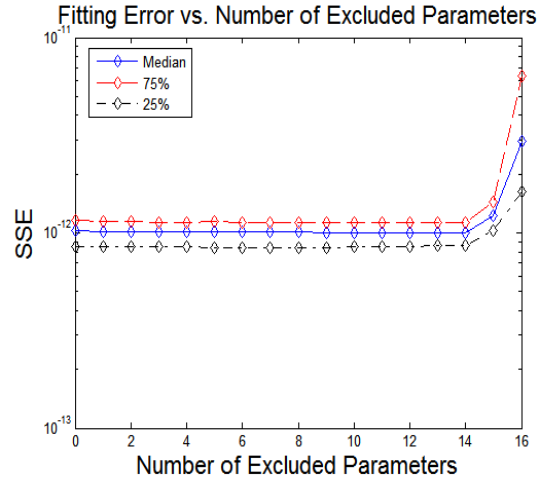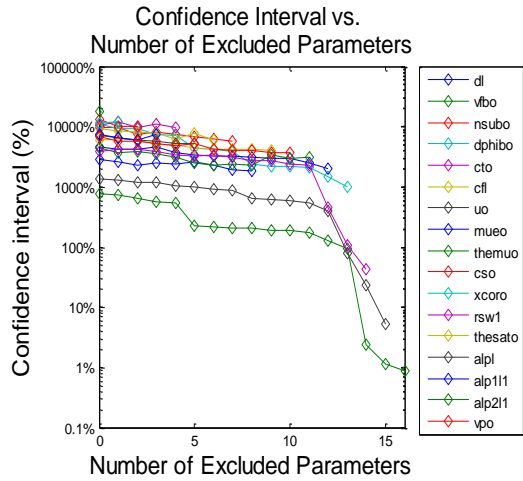
Figure 5.21: Stepwise parameter selection results from Step#1 for the pass-gate transistors (PG3/PG4). Subplots showing the initial extracted parameters without any exclusion, the final optimal parameter set, and the change in normalized confidence interval and sum of squares of fitting error (SSE) after each round of selection.
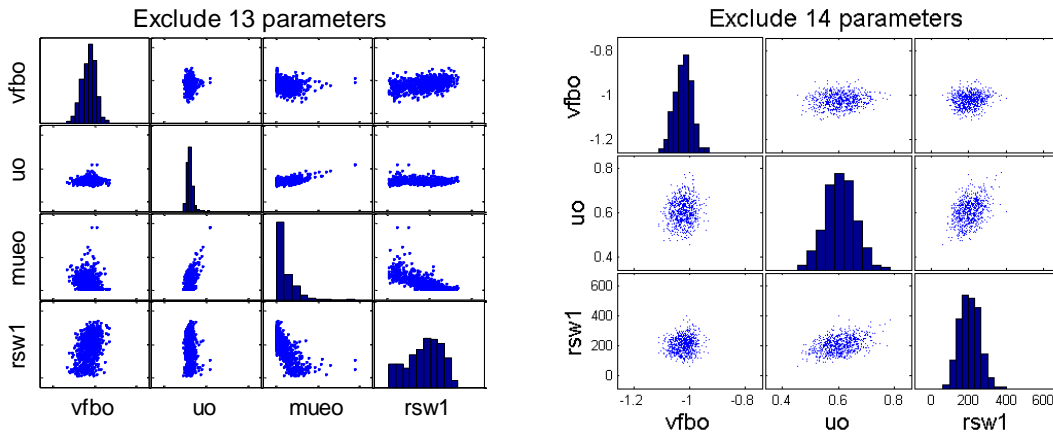
Figure 5.22: Initial and final extracted value after global optimization for the pass-gate transistors (PG3/PG4).

Exclude 0 parameters
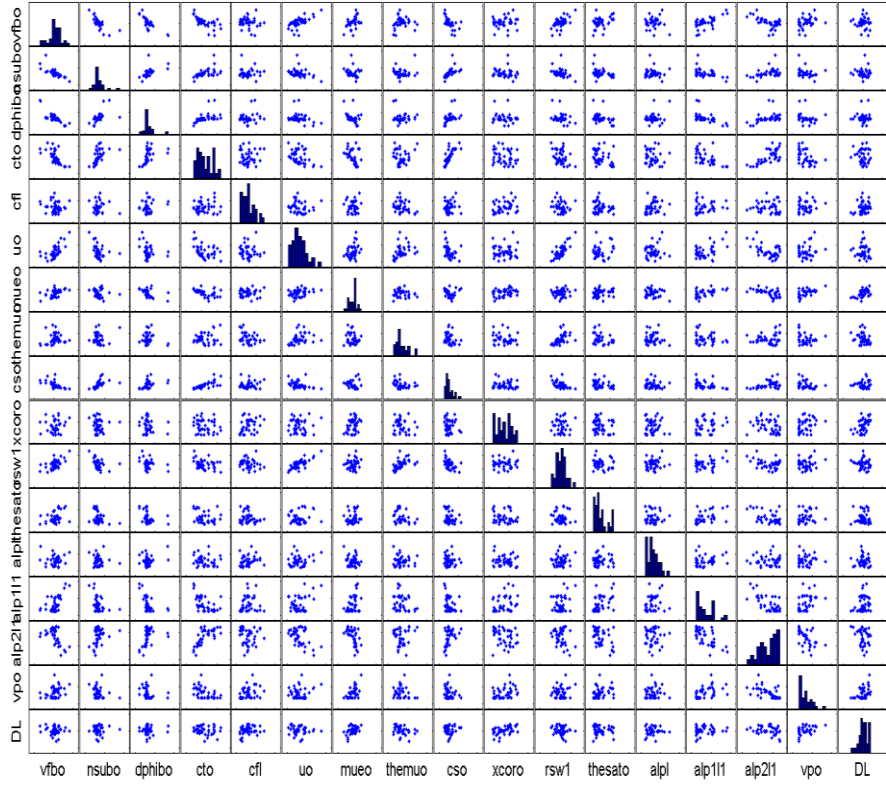
Exclude 13 parameters
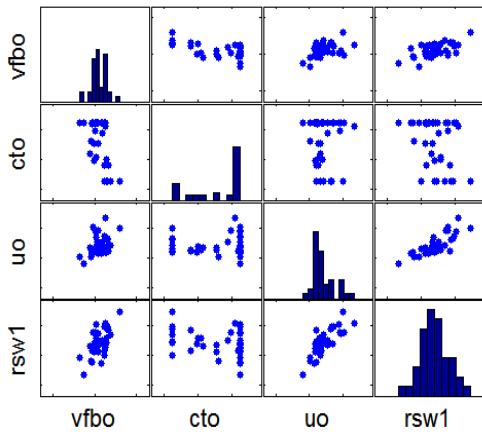
Exclude 14 parameters
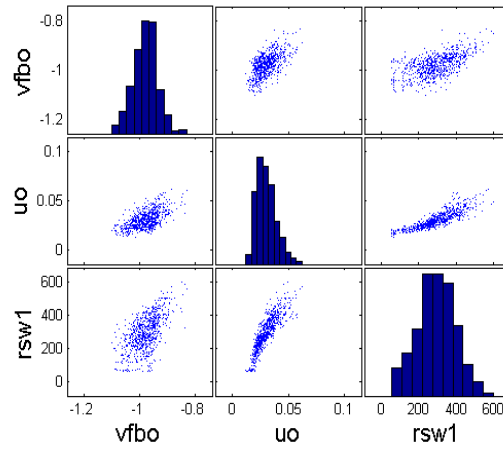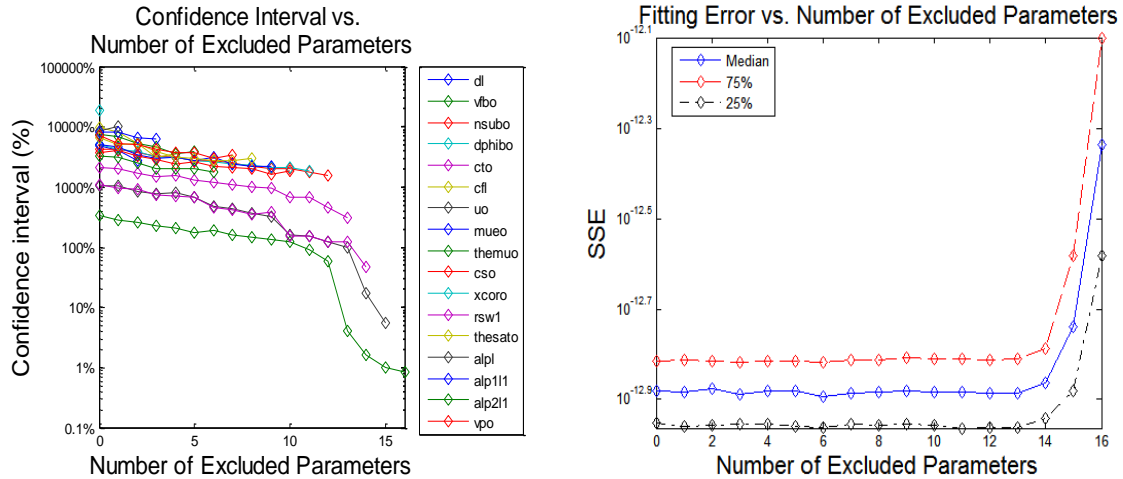
Figure 5.23: Stepwise parameter selection results from Step#1 for the pull-up transistors (PU5/PU6). Subplots showing the initial extracted parameters without any exclusion, the final optimal parameter set, and the change in normalized confidence interval and sum of squares of fitting error (SSE) after each round of selection.
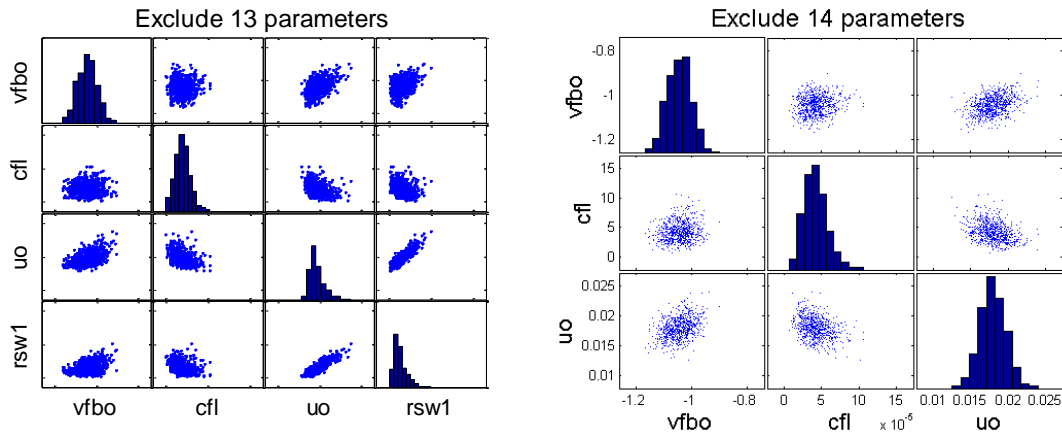
Figure 5.24: Initial and final extracted values after global optimization for the pull-up transistors (PU5/PU6).

5.4.2 Parameter Variability Modeling

The within-chip spatial pattern of all the extracted parameters for the three types of SRAM transistors are shown in Figure 5.25, Figure 5.26 and Figure 5.27. Again, the flat-band voltage parameter, *VFBO*, whose value shift is equivalent to that of the threshold voltage, does not show any systematic variation, nor does the zero-field mobility parameter *UO*. The source/drain resistance parameter has a strong across-chip pattern along the rows for the PD and PG transistors (it is not extractable for PU transistors). Lastly, parameter

CFL, which is only extractable for PU transistors, does not illustrate any systematic patterns.

Using the same practice as the EKV model, we apply the hierarchical variability model to the extracted parameter, *RSW1*, of the two PD transistors and the two PG transistors (Equation 5.5). A parabolic surface along the rows of the test chip is fitted to the extracted parameters, as shown in Figure 5.28. Figure 5.29 shows the normal quantile plots of the original extracted parameters, the fitted across-chip systematic component, and the fitting residuals of parameter *RSW1*. The original extracted values of *RSW1* clearly deviate from a normal distribution at extreme quantiles for PD transistors, and at the lower tails for PG transistors. After the removal of the fitted across-chip systematic component, the distribution of residuals of the parameter *RSW1* is closer to a normal distribution for the PD transistors, but deviates from a normal distribution in the upper tail for the PG transistors. This could also be explained by the systematic across-chip pattern of the residual variance, as illustrated in Figure 5.30. A systematic across-chip function is fitted to the standard deviation of the across-chip residual of *RSW1* (*RSW1$_{ACR}$*). The standard deviation of *LAMBDA$_{ACR}$* within each row varies quadratically with its row position Y_C, as stated in Equation 5.2.

$$RSW11\langle T - WP \rangle = RSW11\langle T - WP \rangle_{AC} + RSW11\langle T - WP \rangle_{ACR}$$

$$RSW11\langle T - WP \rangle_{AC} = 0 \times X_C^2 + 0 \times X_C + c_C \times Y_C^2 + d_C Y_C + e_C$$

(5.5)

$$RSW11\langle T - WP \rangle_{ACR} \sim N\left(0, \sigma^2(Y_C)\right)$$

$$\sigma(Y_C) = 0 \times X_C^2 + 0 \times X_C + s_C \times Y_C^2 + t_C Y_C + r_C$$

(5.6)
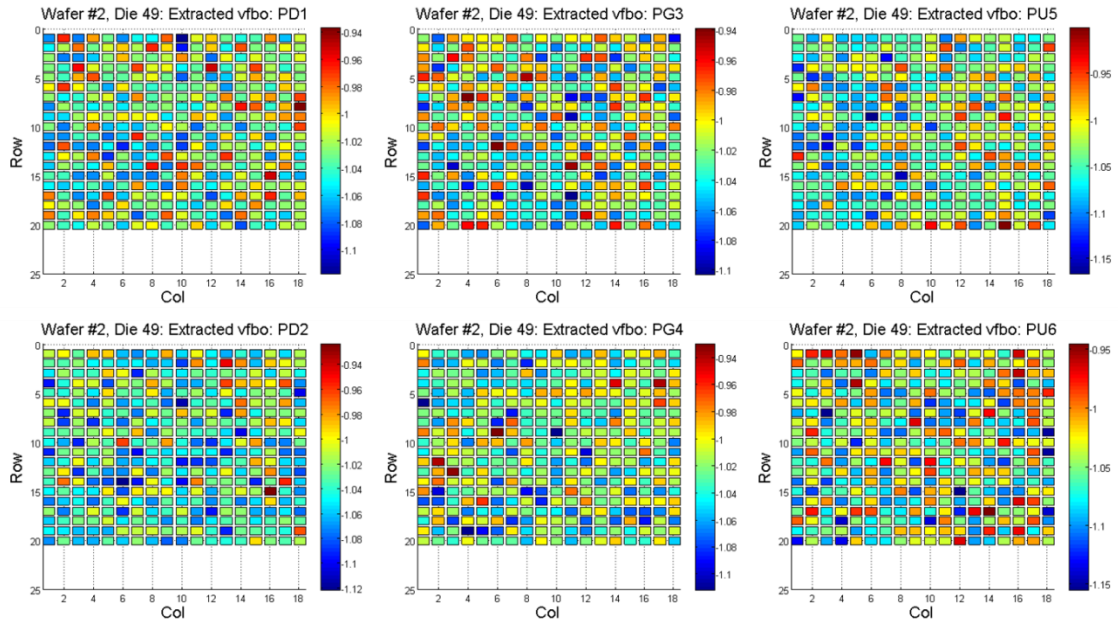
Figure 5.25: Chip maps of extracted compact model parameters *VFBO*.
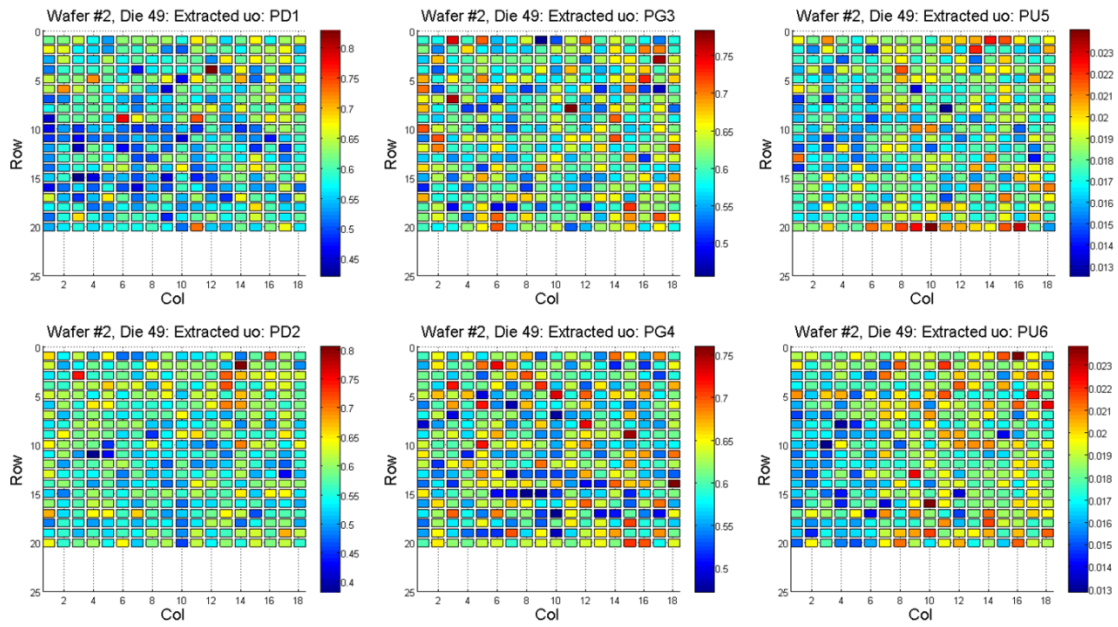
Figure 5.26: Chip maps of extracted compact model parameters *UO*.
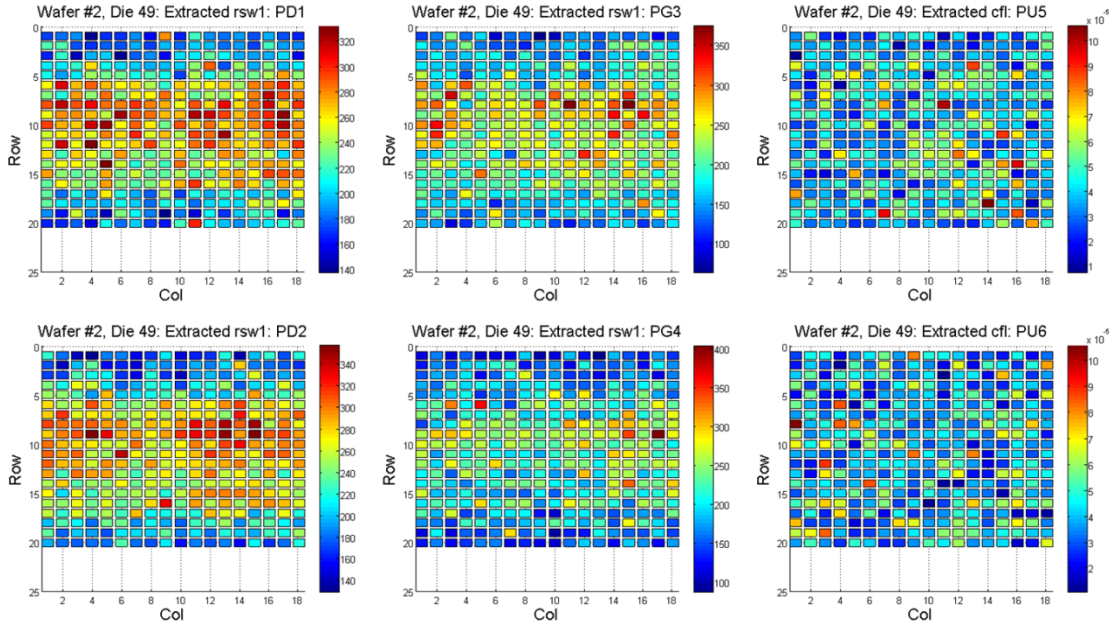
Figure 5.27: Chip maps of extracted compact model parameters *RSW1* (PD/PG) and *CFL* (PU).
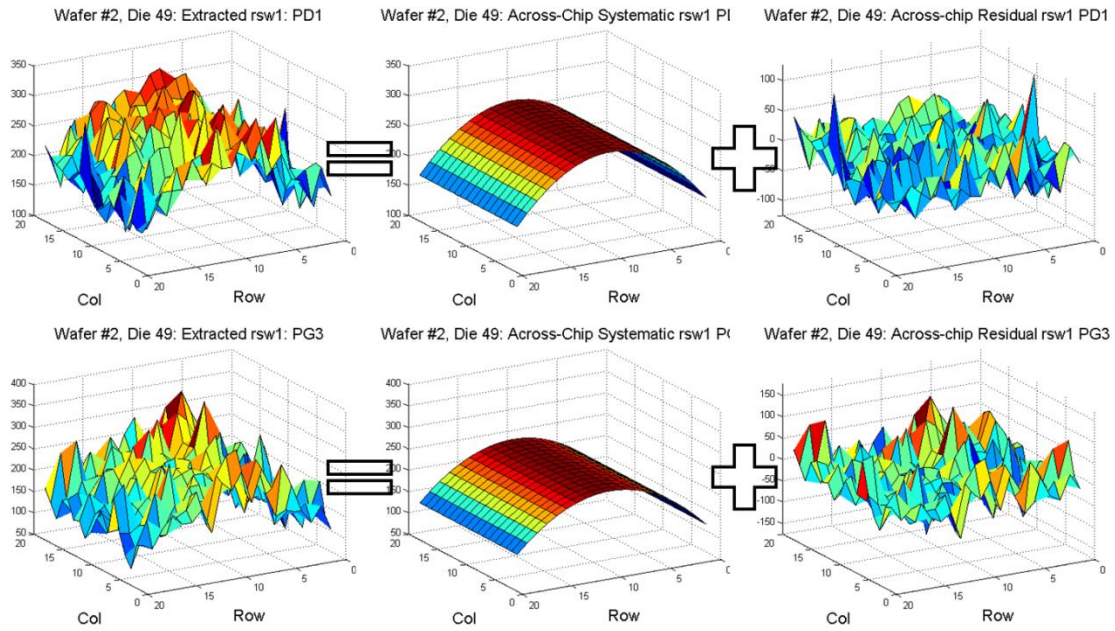
Figure 5.28: Chip-level variation decomposition for *LAMBDA* extracted from the left pull-down transistor and pass-gate: $RSW11\langle T - WP \rangle = RSW11\langle T - WP \rangle_{AC} + RSW11\langle T - WP \rangle_{ACR}$.
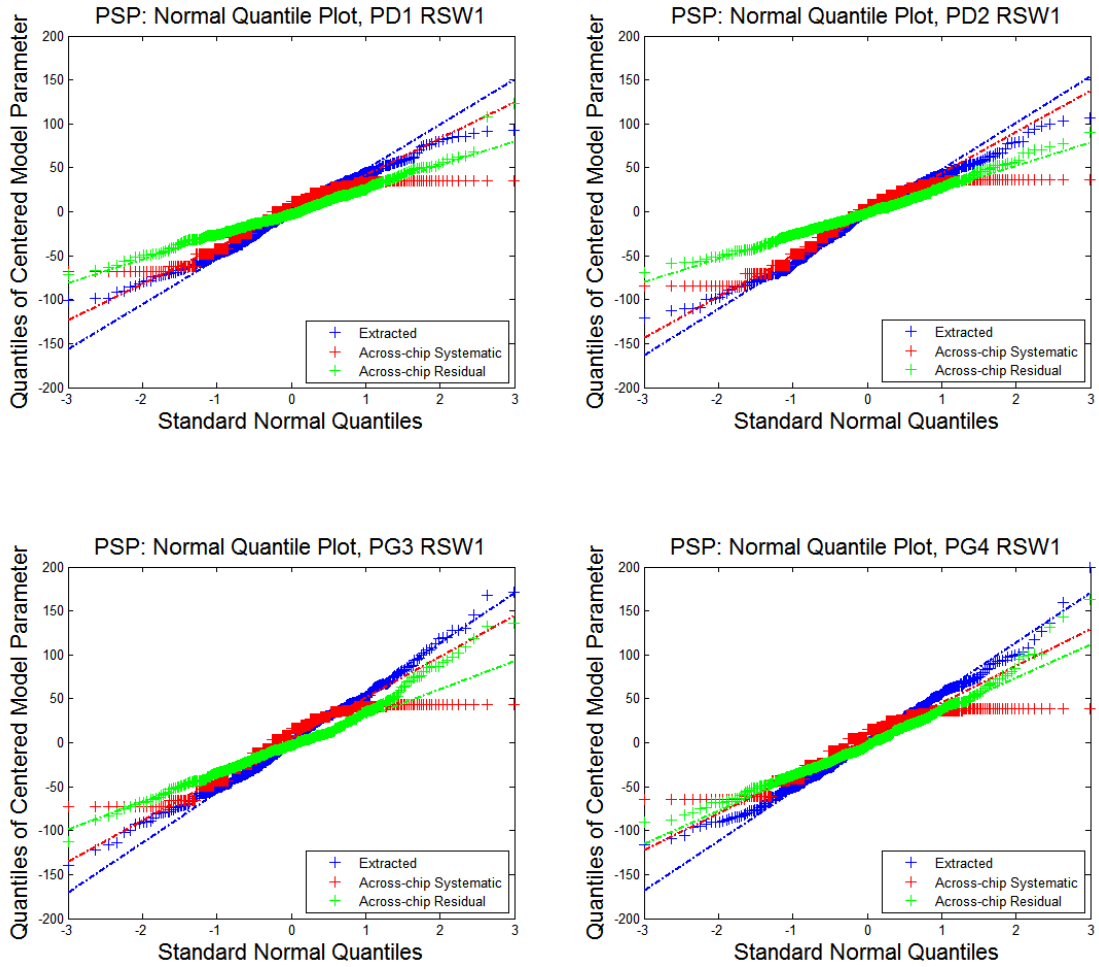
Figure 5.29: Normal quantile plots of the extracted values *RSW1*, the across-chip systematic component *RSW1_{AC}*, and the across-chip residual *RSW1_{ACR}*. All components are centered at zero.
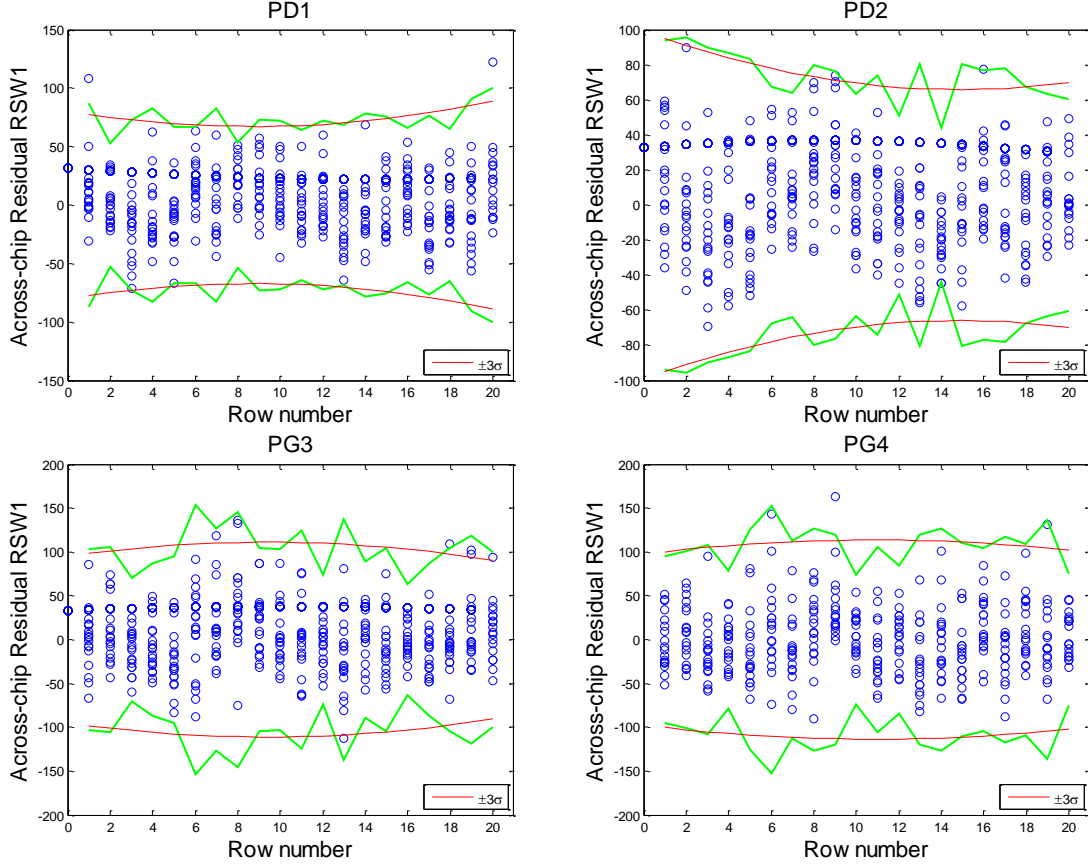
Figure 5.30: Across-chip systematic pattern of parameter *RSW1* variance.

5.4.3 Parameter Variability Reconstruction

We now simulate the distributions of the extracted *RSW1* with both the conventional "Global+Local" model and our hierarchical variability model. The method of modeling the parameter variability is exactly the same as with the EKV model. The underlying assumptions are captured by Equations 5.7 and 5.8.

"Global+Local" variation model:

$$(p_{i_1,j_1}, p_{i_2,j_2}) \sim N(\mu_1, \mu_2, \Sigma_{Local}^2)$$
$$\Sigma_{local} = cov(p_{i_1,j_1}, p_{i_2,j_2}) \tag{5.7}$$

Hierarchical model:

$$p_{i_1,j_1} = f_{i_1 j_1, AC}(X_C, Y_C) + r_{i_1,j_1} \tag{5.8}$$

$$p_{i_2,j_2} = f_{i_2 j_2,AC}(X_C, Y_C) + r_{i_2,j_2}$$

$$(r_{i_1,j_1}, r_{i_2,j_2}) \sim N(0,0,\Sigma^2_{Local})$$

$$\Sigma_{local} = cov(r_{i_1,j_1}, r_{i_2,j_2})$$

Again, ten chips of PSP model cards (360x10=3600) are generated with the "Global+Local" model and the hierarchical model. A comparison between the original and the reconstructed parameter distributions is shown in Figure 5.31. The hierarchical model generally captures the non-Gaussian distribution behavior of the original extracted parameter distributions better than the conventional "Global+Local" model does, especially in the lower tail. The gap between reconstructed parameter distributions and the original extracted parameter distributions at the 1% and 99% quantiles can be found in Table 5.4. The hierarchical model can be up to 10% better than the conventional model at these quantiles.
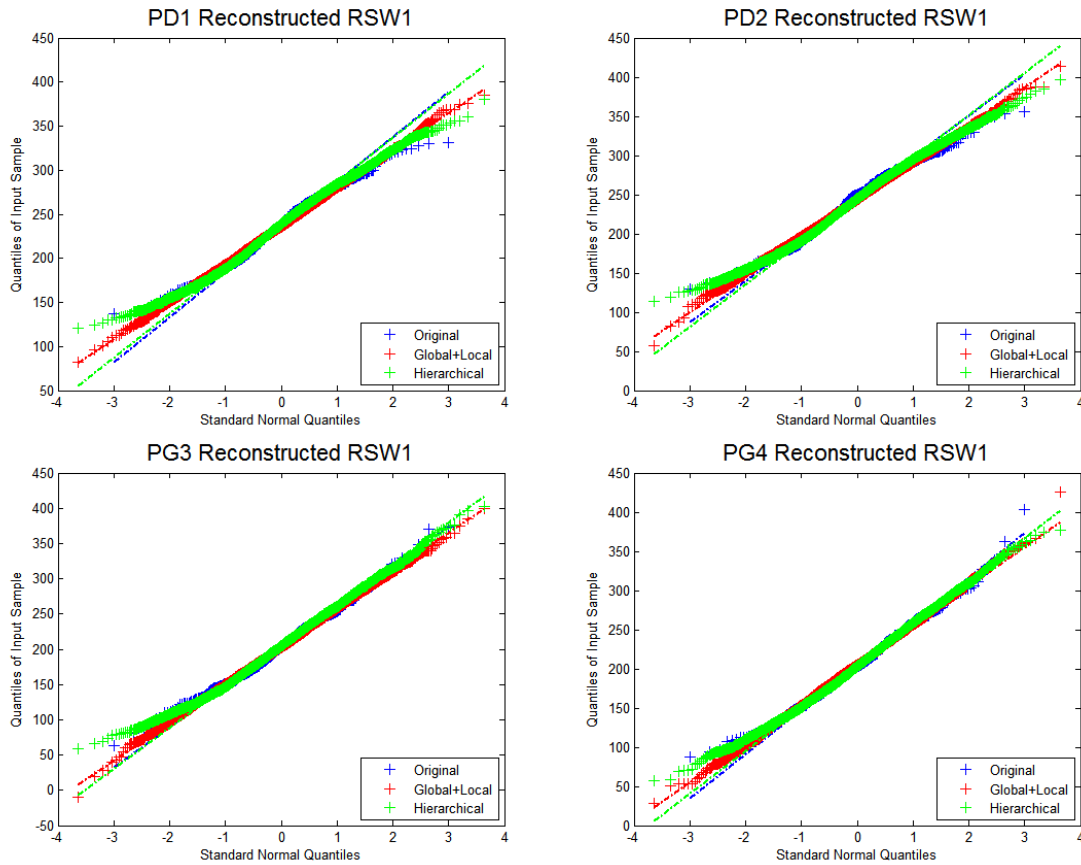
Figure 5.31: Comparison of the distribution of the extracted parameter *RSW1* and the reconstructed distributions using "Global+Local" model and the hierarchical model.

Parameter	Device	Percentile	original	"Global+Local" Model		Hierarchical Model	
			Value	Value	Error	Value	Error
RSW1	PD1	99%	324.0	336.2	4%	335.7	4%
		1%	146.6	136.7	-7%	145.6	-1%
	PD2	99%	346.5	349.2	1%	347.6	0%
		1%	140.2	133.3	-5%	143.9	3%
	PG3	99%	332.4	324.4	-2%	331.5	0%
		1%	89.9	79.8	-11%	96.3	7%
	PG4	99%	329.0	328.0	0%	328.5	0%
		1%	107.1	86.3	-19%	98.1	-8%

Table 5.4: 99% and 1% quantiles of the original and the reconstructed parameter distributions.

5.5 Hierarchical Model Application for Extracted Parameters

As shown in Sections 5.3 and 5.4, across-chip hierarchical models are fitted to the NMOS model parameters *KP* and *LAMBDA* of the EKV model and *RSW1* of the PSP model, respectively. After the removal of the systematic component, the residual of the compact model parameters of the six bit cell transistors are more nearly Gaussian. Their residual variance can be further normalized by fitting a systematic function of the standard deviation across the chip. Consequently, we are able to estimate the distributions of the random component of the parameters by simulating model parameters as correlated Gaussian variables, using the mean and covariance matrix estimated from the normalized across-chip residual of the hierarchical model. The extracted systematic across-chip component is then added onto the generated random component by uniform sampling over all possible locations on the chip, so that we can recreate the full picture of the variability of model parameters. For comparison, we also generated parameter distributions with the

conventional "Global+Local" model. The details of the re-construction process are described in Sections 5.3.3 and 5.4.3 for the EKV model and the PSP model, respectively.

Ten chips of model cards (360x10=3,600) are generated from both the hierarchical variability model and the conventional "Global+Local" model. The accuracy of the two variability models can then be evaluated by simply simulating the electrical performance of the device and circuits with the 3,600 model cards. Due to the fact that the variability model is based on the extracted-parameter distributions, the best a variability model can do is to faithfully regenerate the electrical performance distributions predicted by the originally extracted parameters. Nevertheless, we put the experimentally measured device and circuit metrics against those simulated with the extracted parameters and the reconstructed model cards using the EKV model and PSP models and the hierarchical and conventional "Global+Local" variability model. A selected comparison of such electrical metrics is shown in Figure 5.32 through Figure 5.36, including the SRAM Read Static Noise Margin (*RSNM*), the writeability current (I_W), and the on-current (*Idsat*) of the PD1, PG3, and PU5 transistors.

For the SRAM bit-cell read/write margins *RSNM* and I_W, we found that even the original extracted parameters of either compact model (EKV or PSP) cannot accurately predict the distribution from the experimental measurement. This may be explained by the fact that SRAM operations are highly sensitive to transistor behavior around the threshold voltage, exactly where our electrical measurements were hampered by large parasitic leakage currents. Thus in Figure 5.32 and Figure 5.33, the simulated distributions are normalized so that their median matches that of the measured statistics. The rest of electrical metrics are compared as-is. For applications that require a high yield, we should look at the extreme quantiles of the statistical distributions. As a simple example, we evaluated the 1% and 99% quantiles of the measured, extracted, and re-constructed electric metrics of transistor and SRAM bit cells, as listed. The full extraction results from both the EKV model and PSP model at the tails are generally within 1~2% of the measurement data, except for the *RSNM* distributions where the EKV model can have up to a 4% error margin and the PSP model can have up to a 6% error margin on the higher end. This indicates that our extraction methodology is reasonably accurate in terms of capturing the silicon device behavior.

On the other hand, there are gaps between the raw extraction and the conventional and hierarchical variability models as well. For the EKV model, the hierarchical model always fit the full extraction results 1 to 2% better than the "Global+Local" model, and up to 4% better when predicting the bottom 1% of the writeability current I_W. The case of the PSP model is very similar. In most cases, the differences between the extreme quantiles of conventional and hierarchical model is within 1%, while at 99% of PD1 *Idsat* the

129

hierarchical model is up to 2% better than the "Global+Local" model. The differences may look small, but they could have a large impact on yield estimates if the parametric-yield threshold is high. Figure 5.37 shows an example with PD1 *Idsat* distributions. Assume we use 99% of the real silicon measurements as the upper limit of *Idsat* (for illustration purposes, not a real life criterion), and extrapolate that into the predicted normal quantile plot of the "Global+Local" model and hierarchical model prediction; it would correspond to the 99.1% of the hierarchical model and the 99.4% of the "Global+Local" model. In this case, the "Global+Local" model will predict a 0.6% fail rate while the hierarchical model will predict a 0.9% fail rate, compared to the true fail rate of 1%. In this sense, the conventional model is underestimating the failure rate by 40% while the hierarchical model is only underestimating it by 10%. At more extreme distribution quantiles, this gap can be even more significant and make the hierarchical model far superior for yield estimation.
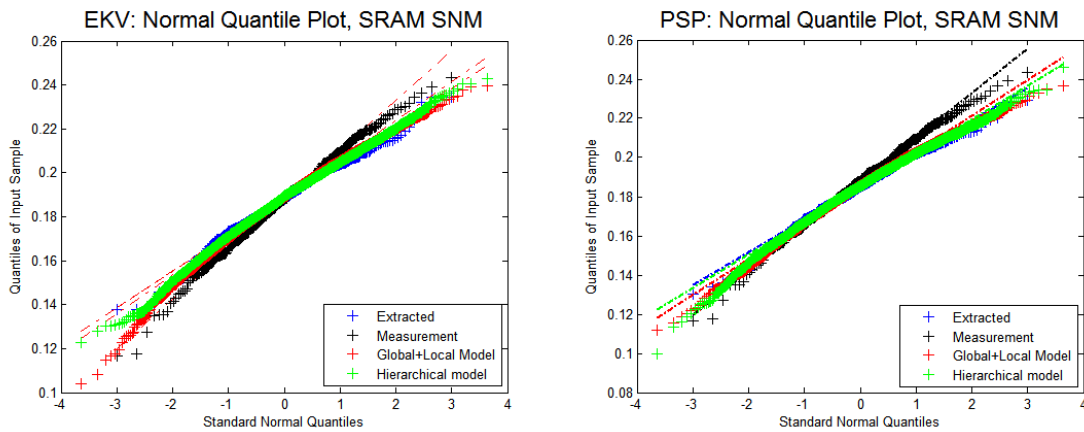
Figure 5.32: Comparing prediction accuracy of the "Global+Local" model vs. the hierarchical model for the distributions of SRAM read static noise margin *SNM*. The model parameters were extracted using the EKV and PSP Models.
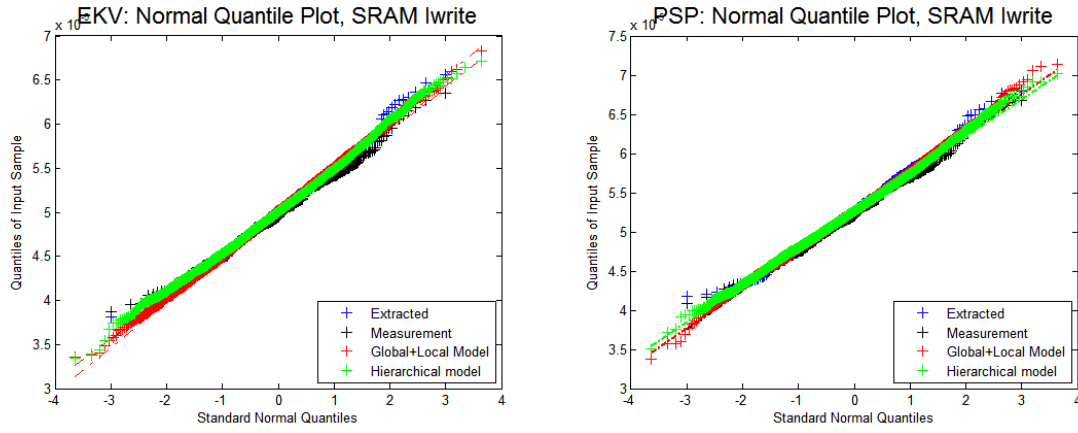
Figure 5.33: Comparing prediction accuracy of the "Global+Local" model and the hierarchical model for the distributions of SRAM writeability current I_W. The model parameters were extracted using the EKV and PSP Models.
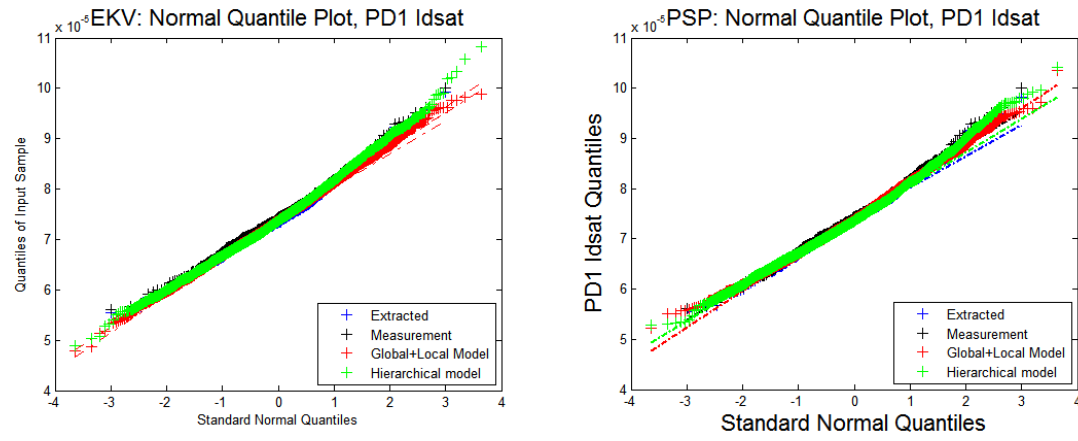
Figure 5.34: Comparing prediction accuracy of the "Global+Local" model and the hierarchical model for the distributions of PD1 *Idsat*. The model parameters were extracted using the EKV and PSP Models.

131
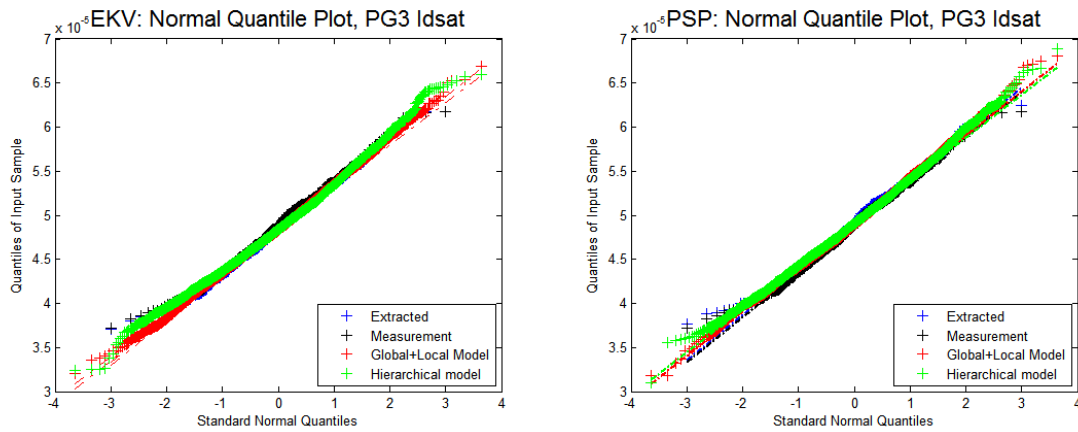
Figure 5.35: Comparing prediction accuracy of the "Global+Local" model vs. the hierarchical model for the distributions of PG3 *Idsat*. The model parameters were extracted using the EKV and PSP Models.
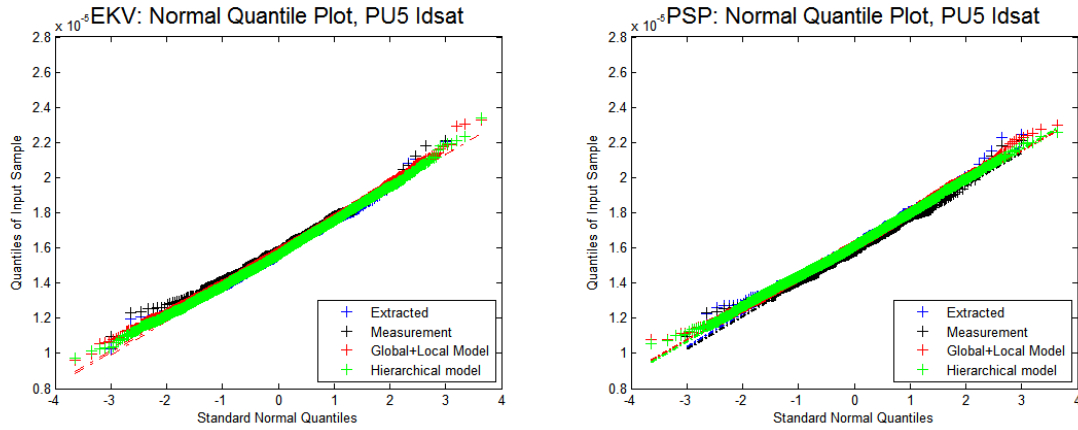
Figure 5.36: Comparing prediction accuracy of the "Global+Local" model and the hierarchical model for the distributions of PU5 *Idsat*. The model parameters were extracted using the EKV and PSP Models.

	Model	PD1 Idsat (µA)		PG3 Idsat (µA)		RSNM (V)		I_W (µA)	
		1%	99%	1%	99%	1%	99%	1%	99%
Measurement	–	59.1	93.6	38.9	61.2	0.135	0.234	42.8	64.7
Full Extraction	EKV	58.6	93.1	38.6	60.9	0.141	0.224	42.6	66.1
	PSP	58.9	93.5	39.2	61.3	0.138	0.219	42.8	65.7
"Global+Local" Model	EKV	57.7	91.6	37.1	60.4	0.141	0.225	40.6	65.4
	PSP	59.1	91.7	37.9	61.5	0.138	0.221	42.6	64.9
Hierarchical Model	EKV	58.1	92.9	37.8	61.6	0.142	0.225	42.0	65.3
	PSP	59	93.3	38.2	61.5	0.137	0.221	42.6	64.8

Table 5.5: 99% and 1% quantiles of transistor and SRAM electrical metrics from: experimental measurement, simulation using extracted parameters, and simulation using reconstructed compact model cards.

	Model	PD1 Idsat (μA)		PG3 Idsat (μA)		RSNM (V)		I_W (μA)	
		1%	99%	1%	99%	1%	99%	1%	99%
"Global+Local	EKV	-2%	**-2%**	**-4%**	-1%	0%	1%	**-5%**	-1%
" Model	PSP	0%	**-2%**	-3%	0%	0%	1%	0%	-1%
Hierarchical	EKV	-1%	**0%**	**-2%**	1%	1%	1%	**-1%**	-1%
Model	PSP	0%	**0%**	-3%	0%	-1%	1%	0%	-1%

Table 5.6: 99% and 1% quantiles of transistor and SRAM electrical metrics: error between simulated distributions from full extraction and re-constructed model cards with the "Global+Local" model and hierarchical model.
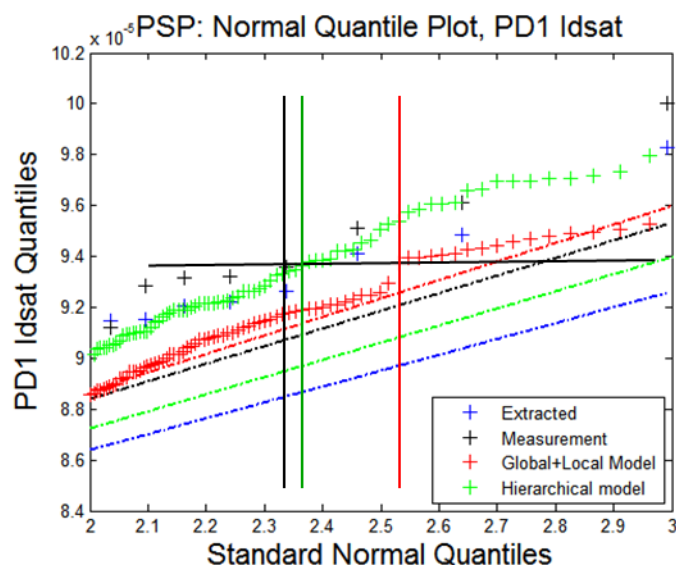
Figure 5.37: Error in yield estimation with the "Global+Local" model and the hierarchical variability model.

5.6 Summary

The compact model parameter extraction methodology is applied to experimentally collected *I-V* data from the SRAM bit cell transistors on the 45nm test chips. The stepwise parameter selection procedure is carried out first to determine a good set of model parameters for variability extraction for both the EKV model in a one-step parameter optimization and the PSP model in a sequential style optimization. With the full parameter

distributions extracted for one full chip of data, we applied the hierarchical variability model to the resulting compact model parameters, and compared its accuracy with that of the conventional "Global+Local" variability model by re-constructing the model parameter distributions under their respective assumptions. In most cases the hierarchical model performs slightly better than the conventional method in predicting extreme quantiles, and up to 4% better in which the distribution strongly deviates from normal. The prediction errors that are small in absolute magnitude can produce large errors in yield estimation when the yield threshold is set high. The accuracy of the directly-extracted model is also limited due to the lack of accuracy in the subthreshold regime in the electrical test. Thus, the predicted SRAM read/write margin specs are even further off from raw measurements due to their high sensitivity to the threshold voltage changes. Nonetheless, our study shows a scalable parameter extraction framework that is capable of dealing with complex problems and may prove especially valuable when there are strong systematic components of variability.

Chapter 6

Conclusion

Variability characterization and analysis have been performed for two sets of customized test chips manufactured using an advanced CMOS process. The devices under test (DUT) included ring oscillators from one wafer on a 90nm process as well as ring oscillators, SRAM bit cells, and the individual transistors within the SRAM cells from two wafers on a 45nm process. Each DUT was repeated tens to hundreds of times in an on-chip array, while tens of chips with good spatial coverage were measured over each wafer. This allowed a hierarchical breakdown of the device variations into wafer-level and die-level systematic and random components as well as the identification of layout-dependent effects. With the newly proposed parameter-extraction methodology, two sets of compact model parameters were extracted for the padded-out transistors in the SRAM bit cells using the EKV and PSP models. These extracted parameters were subsequently fed into a hierarchical variability model, which successfully reproduced the variability profile of the SRAM cells and its internal transistors.

6.1 Key Contributions

This work provides two key contributions. First, the comprehensive methodology can capture the systematic and random variation components in the early stages of the advanced 90nm and 45nm CMOS processes. The characterization and analysis were carried out with two sets of customized test chips with arrays of small test circuits, such as ring oscillators and SRAM bit cells. Following the careful breakdown of the wafer-level and die-level variability using a hierarchical variability model, we successfully identified several significant systematic variations, including across-wafer ring oscillator delay variability, across-chip SRAM read/write margin and transistor drive current variations, and layout-dependent effects, among ring oscillators with different layout pattern densities. We illustrated how the systematic variation components are crucial in achieving high confidence for predicting the extreme quantiles of device performance distributions.

Second, we designed a compact model parameter-extraction framework that intelligently selects model parameter candidates for numerical data extraction so that the extracted compact model parameters are physically reasonable, with minimal artificial correlation between the parameters, and fit the data adequately. This methodology was first

validated through simulations using the EKV and PSP models and then applied to real silicon transistor *I-V* data that we had previously characterized from the 45nm SRAM test arrays. When there are large variations, only a handful of core model parameters can be extracted with high credibility without hurting the fitting and extraction quality. Nevertheless, by applying our hierarchical variability model to the extracted compact model parameters, this selected set of model parameters still effectively captures the systematic across-chip variations that we observed in the SRAM read/write margins and transistor drive currents. The parametric yield estimation in the top and bottom 1% quantiles showed a clear advantage in accuracy compare to conventional methods.

6.2 Future Work

In the ring oscillator variation analysis, we identified significant across-wafer systematic variation as well as layout-dependent effects. However, the lack of companion transistors (individually measurable transistors with the same layout design as those in the ring oscillator) makes it difficult to pin down the exact reason behind such variations. A new test-chip design could incorporate such companion transistors, which might reveal the underlying physical mechanics of the systematic variations with the help of the statistical parameter extraction methodology we developed.

The SRAM test circuitry can be redesigned so that the individual padded-out transistors are less vulnerable to the off-state leakage from the switching networks. This will help improve the *I-V* measurement accuracy, particularly in the subthreshold region. It is critical for the accurate prediction of the SRAM bit cell read/write noise margins, as they are extremely sensitive to the threshold voltages of the transistors.

The model parameter-extraction methodology can be further expanded by experimenting with the ordering of the different steps when performing sequential parameter extraction. This may help with finding the optimal sequence of extraction steps, avoiding iterations, and making extraction more robust. There are also a variety of standard forward and backward selection algorithms from the statistical literature that could be explored.

Lastly, the compact model parameter-extraction framework shall also be applied to the BSIM model, which is the most widely used industrial standard compact model. Collaboration with model developers, utilizing the newly obtained knowledge about the robustness of various model parameters under statistical extraction, could result in improved reference extraction flow and even improvement in the model equations themselves.

Bibliography

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Proc. IEEE*, vol. 86, pp. 82–85, 1998.

[2] A. B. Kahng and Y. C. Pati, "Subwavelength lithography and its potential impact on design and EDA," in *Proceedings of the 36th ACM/IEEE conference on Design automation conference - DAC '99*, 1999, pp. 799–804.

[3] B. Wu and A. Kumar, "Line width roughness and its control on photomask," in *SPIE Photomask Technology*, 2013, p. 888004.

[4] J. Croon, G. Storms, S. Winkelmeier, I. Pollentier, M. Ercken, S. Decoutere, W. Sansen, and H. Maes, "Line-edge roughness: characterization, modeling and impact on device behavior," in *Electron Devices Meeting, 2002. IEDM'02. Digest. International*, 2003, pp. 307–310.

[5] R. W. Keyes, "Effect of randomness in the distribution of impurity ions on FET thresholds in integrated electronics," *IEEE J. Solid-State Circuits*, vol. 10, no. 4, pp. 245–247, Aug. 1975.

[6] P. J. Timans, W. Lerch, J. Niess, S. Paul, N. Acharya, and Z. Nenyei, "Challenges for ultra-shallow junction formation technologies beyond the 90 nm node," in *11th IEEE International Conference on Advanced Thermal Processing of Semiconductors. RTP 2003*, 2003, pp. 17–33.

[7] I. Ahsan, N. Zamdmer, O. Glushchenkov, R. Logan, E. Nowak, H. Kimura, J. Zimmerman, G. Berg, J. Herman, E. Maciejewski, A. Chan, A. Azuma, S. Deshpande, B. Dirahoui, G. Freeman, A. Gabor, M. Gribelyuk, S. Huang, M. Kumar, K. Miyamoto, and D. Mocuta, "RTA-Driven Intra-Die Variations in Stage Delay, and Parametric Sensitivities for 65nm Technology," in *2006 Symposium on VLSI Technology, 2006. Digest of Technical Papers.*, 2006, pp. 170–171.

[8] T. T. B. Hook, J. Brown, P. Cottrell, E. Adler, D. Hoyniak, J. Johnson, and R. Mann, "Lateral ion implant straggle and mask proximity effect," *IEEE Trans. ...*, vol. 50, no. 9, pp. 1946–1951, Sep. 2003.

[9]   S. Nassif, "Delay variability: sources, impacts and trends," in *2000 IEEE International Solid-State Circuits Conference. Digest of Technical Papers (Cat. No.00CH37056)*, 2000, pp. 368–369.

[10]   a. J. Bhavnagarwala and J. D. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE J. Solid-State Circuits*, vol. 36, no. 4, pp. 658–665, Apr. 2001.

[11]   B. J. Sheu, D. L. Scharfetter, P.-K. Ko, and M.-C. Jeng, "BSIM: Berkeley short-channel IGFET model for MOS transistors," *IEEE J. Solid-State Circuits*, vol. 22, no. 4, pp. 558–566, Aug. 1987.

[12]   G. Gildenblat, X. Li, W. Wu, H. Wang, A. Jha, R. Van Langevelde, G. D. J. G. D. J. Smit, A. J. A. J. a. J. Scholten, D. B. M. D. B. M. Klaassen, S. Member, and R. Van Langevelde, "PSP: An Advanced Surface-Potential-Based MOSFET Model for Circuit Simulation," *IEEE Trans. Electron Devices*, vol. 53, no. 9, pp. 1979–1993, Sep. 2006.

[13]   G. S. May, C. J. Spanos, and J. Wiley, *Fundamentals of semiconductor manufacturing and process control*. Wiley Online Library, 2006.

[14]   a. Asenov, a. R. Brown, J. H. Davies, S. Kaya, and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs," *IEEE Trans. Electron Devices*, vol. 50, no. 9, pp. 1837–1852, Sep. 2003.

[15]   B. Cheng, S. Roy, a. R. Brown, C. Millar, and a. Asenov, "Evaluation of intrinsic parameter fluctuations on 45, 32 and 22nm technology node LP N-MOSFETs," *ESSDERC 2008 - 38th Eur. Solid-State Device Res. Conf.*, no. c, pp. 47–50, 2008.

[16]   B. Cheng, S. Roy, G. Roy, F. Adamulema, and a Asenov, "Impact of intrinsic parameter fluctuations in decanano MOSFETs on yield and functionality of SRAM cells," *Solid. State. Electron.*, vol. 49, no. 5, pp. 740–746, May 2005.

[17]   A. Asenov, S. Kaya, and J. H. Davies, "Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations," *IEEE Trans. Electron Devices*, vol. 49, no. 1, pp. 112–119, 2002.

[18]   A. Asenov, B. Cheng, D. Dideban, U. Kovac, N. Moezi, C. Millar, G. Roy, A. Brown, and S. Roy, "Modeling and simulation of transistor and circuit variability

and reliability," in *Custom Integrated Circuits Conference (CICC), 2010 IEEE*, 2010, pp. 1–8.

[19]   B. Bindu, B. Cheng, G. Roy, X. Wang, S. Roy, and a. Asenov, "Parameter set and data sampling strategy for accurate yet efficient statistical MOSFET compact model extraction," *Solid. State. Electron.*, vol. 54, no. 3, pp. 307–315, Mar. 2010.

[20]   D. Boning, J. Panganiban, K. Gonzalez-Valentin, S. Nassif, C. McDowell, A. Gattiker, and F. Liu, "Test structures for delay variability," *Proc. 8th ACM/IEEE Int. Work. Timing issues Specif. Synth. Digit. Syst. - TAU '02*, p. 109, 2002.

[21]   J. Panganiban, "A ring oscillator based variation test chip," *Master's thesis, Massachusetts Inst. Technol. Cambridge, Mass.*, 2002.

[22]   E. Chang, B. Stine, T. Maung, R. Divecha, D. Boning, J. Chung, K. Chang, G. Ray, D. Bradbury, O. S. Nakagawa, S. Oh, D. Bartelink, S. Nakagawa, M. I. T. Eecs, C. Ma, H. P. Co, and P. Alto, "Using a statistical metrology framework to identify systematic and random sources of die- and wafer-level ILD thickness variation in CMP processes," *Proc. Int. Electron Devices Meet.*, pp. 499–502, 1995.

[23]   D. S. D. Boning and J. J. E. Chung, "Statistical metrology: Understanding spatial variation in semiconductor manufacturing," in *Microelectronic Manufacturing 1996*, 1996, vol. 2874, no. 1, pp. 16–26.

[24]   Sang-Hoon Lee, Dong-Yun Lee, Tae-Jin Kwon, Joo-Hee Lee, Young-Kwan Park, Bum-Sik Kim, Jeong-Taek Kong, S. Lee, D. Lee, T. Kwon, J. Lee, Y. Park, B. Kim, and J. Kong, "An efficient statistical model using electrical tests for GHz CMOS devices," *2000 5th Int. Work. Stat. Metrol. (Cat.No.00TH8489*, pp. 72–75, 2000.

[25]   J. C. Chen, C. Hu, C. P. Wan, P. Bendix, and A. Kapoor, "ET based statistical modeling and compact statistical circuit simulation methodologies," in *Electron Devices Meeting, 1996., International*, 2002, vol. 00, no. 510, pp. 635–638.

[26]   M. Miyama, S. Kamohara, K. Okuyama, and Y. Oji, "Parametric yield enhancement system via circuit level device optimization using statistical circuit simulation," *2001 Symp. VLSI Circuits. Dig. Tech. Pap. (IEEE Cat. No.01CH37185)*, pp. 163–166, 2001.

[27] K. Takeuchi and M. Hane, "Statistical Compact Model Parameter Extraction by Direct Fitting to Variations," *IEEE Trans. Electron Devices*, vol. 55, no. 6, pp. 1487–1493, Jun. 2008.

[28] B. Cheng, N. Moezi, D. Dideban, G. Roy, S. Roy, and a. Asenov, "Benchmarking the Accuracy of PCA Generated Statistical Compact Model Parameters Against Physical Device Simulation and Directly Extracted Statistical Parameters," *2009 Int. Conf. Simul. Semicond. Process. Devices*, pp. 1–4, Sep. 2009.

[29] M. Bucher, C. Lallement, C. Enz, F. Théodoloz, F. Krummenacher, I. Revision, and R. Ii, "The EPFL-EKV MOSFET model equations for simulation," *EPFL Lausanne Switzerland, Tech. Rep.*, 1998.

[30] K. Bernstein, D. Frank, A. Gattiker, W. Haensch, B. Ji, S. Nassif, E. Nowak, D. Pearson, and N. Rohrer, "High-performance CMOS variability in the 65-nm regime and beyond," *IBM J. Res. Dev.*, vol. 50, no. 4.5, pp. 433–449, 2010.

[31] D. Varghese, D. Saha, S. Mahapatra, K. Ahmed, F. Nouri, and M. Alam, "On the dispersive versus arrhenius temperature activation of nbti time evolution in plasma nitrided gate oxides: measurements, theory, and implications," *IEEE Int. Devices Meet. 2005. IEDM Tech. Dig.*, vol. 00, no. D, pp. 684–687, 2005.

[32] W. J. Poppe, J. Holwill, L.-T. Pang, P. Friedberg, Q. Liu, L. Alarcon, and A. Neureuther, "Transistor-based electrical test structures for lithography and process characterization," *Proc. SPIE*, vol. 6520, p. 65203N–65203N–11, 2007.

[33] T. W. Kim and E. S. Aydil, "Effects of Chamber Wall Conditions on Cl Concentration and Si Etch Rate Uniformity in Plasma Etching Reactors," *J. Electrochem. Soc.*, vol. 150, no. 7, p. G418, Jul. 2003.

[34] Y.-M. Sheu, K.-W. Su, S. Tian, S.-J. Yang, C.-C. Wang, M.-J. Chen, and S. Liu, "Modeling the Well-Edge Proximity Effect in Highly Scaled MOSFETs," *IEEE Trans. Electron Devices*, vol. 53, no. 11, pp. 2792–2798, Nov. 2006.

[35] R. Bianchi and G. Bouche, "Accurate modeling of trench isolation induced mechanical stress effects on MOSFET electrical performance," in *Electron Devices Meeting, 2002. IEDM'02. Digest. International*, 2003, pp. 117–120.

[36] R. W. Keyes, "The effect of randomness in the distribution of impurity atoms on FET thresholds," *Appl. Phys.*, vol. 8, no. 3, pp. 251–259, Nov. 1975.

[37]   P. Friedberg, J. Cain, and C. Spanos, "Modeling Within-Die Spatial Correlation Effects for Process-Design Co-Optimization," *Sixth Int. Symp. Qual. Electron. Des.*, no. 510, pp. 516–521, 2005.

[38]   C.-H. Lin, M. V. Dunga, D. D. Lu, A. M. Niknejad, and C. Hu, "Performance-Aware Corner Model for Design for Manufacturing," *IEEE Trans. Electron Devices*, vol. 56, no. 4, pp. 595–600, Apr. 2009.

[39]   M. Kanno, A. Shibuya, M. Matsumura, K. Tamura, H. Tsuno, S. Mori, Y. Fukuzaki, T. Gocho, H. Ansai, and N. Nagashima, "Empirical Characteristics and Extraction of Overall Variations for 65-nm MOSFETs and Beyond," vol. 2, pp. 88–89, 2007.

[40]   H. J. Levinson, *Principles of Lithography*. SPIE Press, 2005, p. 423.

[41]   L. T.-N. Wang, "Design and Measurement of Parameter-Specific Ring Oscillators," EECS Department, University of California, Berkeley, 2010.

[42]   D. A. Steele, A. Coniglio, C. Tang, B. Singh, S. Nip, and C. J. Spanos, "Characterizing post-exposure bake processing for transient- and steady-state conditions in the context of critical dimension control," *Proc. SPIE*, vol. 4689. pp. 517–530, 2002.

[43]   S.-D. Kim, H. Wada, and J. C. S. Woo, "TCAD-Based Statistical Analysis and Modeling of Gate Line-Edge Roughness Effect on Nanoscale MOS Transistor Performance and Scaling," *IEEE Trans. Semicond. Manuf.*, vol. 17, no. 2, pp. 192–200, May 2004.

[44]   A. Asenov, "Statistical Nano CMOS Variability and Its Impact on SRAM," in *Extreme Statistics in Nanoscale Memory Design SE - 3*, A. Singhee and R. A. Rutenbar, Eds. Springer US, 2010, pp. 17–49.

[45]   a. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm MOSFET's: A 3-D 'atomistic' simulation study," *IEEE Trans. Electron Devices*, vol. 45, no. 12, pp. 2505–2513, 1998.

[46]   S. E. Thompson, M. Armstrong, C. Auth, M. Alavi, M. Buehler, R. Chau, S. Cea, T. Ghani, G. Glass, T. Hoffman, C.-H. Jan, C. Kenyon, J. Klaus, K. Kuhn, Z. Ma, B. Mcintyre, K. Mistry, a. Murthy, B. Obradovic, R. Nagisetty, P. Nguyen, S. Sivakumar, R. Shaheed, L. Shifren, B. Tufts, S. Tyagi, M. Bohr, and Y. El-Mansy,

"A 90-nm Logic Technology Featuring Strained-Silicon," *IEEE Trans. Electron Devices*, vol. 51, no. 11, pp. 1790–1797, Nov. 2004.

[47]   E. Josse, S. Parihar, O. Callen, P. Ferreira, C. Monget, A. Farcy, M. Zaleski, D. Villanueva, R. Ranica, M. Bidaud, D. Barge, C. Laviron, N. Auriac, C. Le Cam, S. Harrison, S. Warrick, F. Leverd, P. Gouraud, S. Zoll, F. Guyader, E. Perrin, E. Baylac, J. Belledent, B. Icard, B. Minghetti, S. Manakli, L. Pain, V. Huard, G. Ribes, K. Rochereau, S. Bordez, C. Blanc, A. Margain, D. Delille, R. Pantel, K. Barla, N. Cave, and M. Haond, "A Cost-Effective Low Power Platform for the 45-nm Technology Node," *2006 Int. Electron Devices Meet.*, pp. 1–4, Dec. 2006.

[48]   C. Le Cam, F. Guyader, C. De Buttet, P. Guyader, G. Ribes, M. Sardo, S. Vanbergue, F. Bœuf, F. Arnaud, E. Josse, M. Haond, C. Le Cam, C. de Buttet, and F. Buf, "A Low Cost Drive Current Enhancement Technique Using Shallow Trench Isolation Induced Stress for 45-nm Node," *2006 Symp. VLSI Technol. 2006. Dig. Tech. Pap.*, no. August 2004, pp. 82–83, 2006.

[49]   K. Su, Y. Sheu, C. Lin, and C. H. Diaz, "A scaleable model for STI mechanical stress effect on layout dependence of MOS electrical characteristics," in *Proceedings of the IEEE 2003 Custom Integrated Circuits Conference, 2003.*, 2003, pp. 245–248.

[50]   B. E. Stine, D. S. Boning, and J. E. Chung, "Analysis and decomposition of spatial variation in integrated circuit processes and devices," *IEEE Trans. Semicond. Manuf.*, vol. 10, no. 1, pp. 24–41, 1997.

[51]   J. P. Cain and C. J. Spanos, "Electrical linewidth metrology for systematic CD variation characterization and causal analysis," *Proc. SPIE*, vol. 5038, pp. 350–361, May 2003.

[52]   M. Orshansky, L. Milor, and C. Hu, "Characterization of Spatial Intrafield Gate CD Variability, Its Impact on Circuit Performance, and Spatial Mask-Level Correction," *IEEE Trans. Semicond. Manuf.*, vol. 17, no. 1, pp. 2–11, Feb. 2004.

[53]   B. E. Stine, D. S. Boning, J. E. Chung, D. Ciplickas, and J. K. Kibarian, "Simulating the impact of poly-CD wafer-level and die-level variation on circuit performance," *1997 2nd Int. Work. Stat. Metrol.*, no. June, pp. 24–27, 1997.

[54] B. Nikolic and L. Pang, "Measurements and analysis of process variability in 90nm CMOS," *2006 8th Int. Conf. Solid-State Integr. Circuit Technol. Proc.*, pp. 505–508, 2006.

[55] L.-T. Pang and B. Nikolic, "Impact of Layout on 90nm CMOS Process Parameter Fluctuations," *2006 Symp. VLSI Circuits, 2006. Dig. Tech. Pap.*, vol. 00, no. c, pp. 69–70, 2006.

[56] Z. Guo, A. Carlson, L.-T. Pang, K. T. Duong, T.-J. K. Liu, and B. Nikolic, "Large-Scale SRAM Variability Characterization in 45 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 11, pp. 3174–3192, Nov. 2009.

[57] L.-T. Pang, K. Qian, C. J. Spanos, and B. Nikolic, "Measurement and Analysis of Variability in 45 nm Strained-Si CMOS Technology," *IEEE J. Solid-State Circuits*, vol. 44, no. 8, pp. 2233–2243, Aug. 2009.

[58] L. T. L. T. Pang, *Measurement and Analysis of Variability in CMOS circuits*. University of California, Berkeley, 2008.

[59] Z. Guo, "Large-Scale Variability Characterization and Robust Design Techniques for Nanoscale SRAM," *techreports.lib.berkeley.edu*, 2009.

[60] B. Le Gratiet, P. Gouraud, E. Aparicio, L. Babaud, K. Dabertrand, M. Touchet, S. Kremer, C. Chaton, F. Foussadier, F. Sundermann, J. Massin, J.-D. Chapon, M. Gatefait, B. Minghetti, J. de-Caunes, and D. Boutin, "Process control for 45 nm CMOS logic gate patterning," *Proc. SPIE*, vol. 6922, p. 69220Z–69220Z–11, 2008.

[61] a. V.-Y. Thean, L. Prabhu, V. Vartanian, M. Ramon, B.-Y. Nguyen, T. White, H. Collard, Q.-H. Xie, S. Murphy, J. Cheek, S. Venkatesan, J. Mogab, C. H. Chang, Y. H. Chiu, H. C. Tuan, Y. C. See, M. S. Liang, and Y. C. Sun, "Uniaxial-biaxial stress hybridization for super-critical strained-si directly on insulator (SC-SSOI) PMOS with different channel orientations.," *IEEE Int. Devices Meet. 2005. IEDM Tech. Dig.*, vol. 00, no. c, pp. 509–512, 2005.

[62] A. Kranti and G. A. Armstrong, "6-T SRAM cell design with nanoscale double-gate SOI MOSFETs: impact of source/drain engineering and circuit topology," *Semicond. Sci. Technol.*, vol. 23, no. 7, p. 075049, Jul. 2008.

[63] E. Seevinck, F. J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE Journal of Solid-State Circuits*, vol. 22. pp. 748–754, 1987.

[64] A. Carlson, Z. Guo, S. Balasubramanian, L. T. L. -t. Pang, T. J. K. T. -j. T. J. K. Liu, and B. Nikolic, "FinFET SRAM with Enhanced Read / Write Margins," in *2006 IEEE international SOI Conferencee Proceedings*, 2006, vol. 10, no. 510, pp. 105–106.

[65] Q. Zhang, K. Poolla, and C. J. Spanos, "Across Wafer Critical Dimension Uniformity Enhancement Through Lithography and Etch Process Sequence: Concept, Approach, Modeling, and Experiment," *IEEE Trans. Semicond. Manuf.*, vol. 20, no. 4, p. 488, 2007.

[66] S. Kanno, G. Miya, J. Tanaka, T. Masuda, K. Kuwahara, M. Sakaguchi, a Makino, T. Tsubone, and T. Fujii, "Controlling gate-CD uniformity by means of a CD prediction model and wafer-temperature distribution control," *Thin Solid Films*, vol. 515, no. 12, pp. 4941–4944, Apr. 2007.

[67] G. Scott, J. Lutze, M. Rubin, F. Nouri, and M. Manley, "NMOS drive current reduction caused by transistor layout and trench isolation induced stress," *Int. Electron Devices Meet. 1999. Tech. Dig. (Cat. No.99CH36318)*, 1999.

[68] A. B. Kahng, P. Sharma, and R. O. Topaloglu, "Exploiting STI stress for performance," *2007 IEEE/ACM Int. Conf. Comput. Des.*, pp. 83–90, Nov. 2007.

[69] J. C. Chen, B. W. McGaughy, and D. Sylvester, "An on-chip, attofarad interconnect charge-based capacitance measurement (CBCM) technique," *Int. Electron Devices Meet. Tech. Dig.*, vol. 00, no. 510, pp. 69–72, 1996.

[70] A. Walton, *Microelectronic test structures*. Citeseer, 1998.

[71] W. Versnel, "Analysis of the Greek cross, a Van der Pauw structure with finite contacts," *Solid. State. Electron.*, vol. 22, no. 11, pp. 911–914, Nov. 1979.

[72] S. Enderling, C. L. S. Smith, M. H. Dicks, J. T. M. Stevenson, M. Mitkova, M. N. Kozicki, and a. J. Walton, "Sheet Resistance Measurement of Non-Standard Cleanroom Materials Using Suspended Greek Cross Test Structures," *IEEE Trans. Semicond. Manuf.*, vol. 19, no. 1, pp. 2–9, Feb. 2006.

[73]   N. Arora, "Mosfet modeling for VLSI simulation: theory and practice," *Chem. &*, p. 605, 2007.

[74]   G. T. W. A. N. I. H. M. A. Gaffur, "Dependences of Short and Narrow MOSFET ' s," no. 7, pp. 1240–1245, 1985.

[75]   M. F. F. Hamer and B. Sc, "First-order parameter extraction on enhancement silicon MOS transistors," *IEE Proc. I Solid State Electron Devices*, vol. 133, no. 2, p. 49, 1986.

[76]   K. Doganis, D. L. Scharfetter, and I. Introduction, "General optimization and extraction of IC device model parameters," *IEEE Trans. Electron Devices*, vol. 30, no. 9, pp. 1219–1228, Sep. 1983.

[77]   J. More, "The Levenberg-Marquardt algorithm: implementation and theory," *Numer. Anal.*, no. x, pp. 105–116, 1978.

[78]   J. Zhang and L. Chen, "Nonmonotone Levenberg–Marquardt algorithms and their convergence analysis," *J. Optim. Theory Appl.*, vol. 92, no. 2, pp. 393–418, 1997.

[79]   Y. Li, "Centering, Trust Region, Reflective Techniques for Nonlinear Minimization Subject to Bounds," Sep. 1993.

[80]   T. F. Coleman and Y. Li, "An Interior Trust Region Approach for Nonlinear Minimization Subject to Bounds," *SIAM J. Optim.*, vol. 6, no. 2, pp. 418–445, May 1996.

[81]   J. S. Alper and R. I. Gelb, "Standard errors and confidence intervals in nonlinear regression: comparison of Monte Carlo and parametric statistics," *J. Phys. Chem.*, vol. 94, no. 11, pp. 4747–4751, May 1990.

[82]   M. Communications and M. Ben, "Confidence regions and intervals in nonlinear regression * ´ †," vol. 2, pp. 71–76, 1997.

[83]   R. M. R. M. O'brien, "A caution regarding rules of thumb for variance inflation factors," *Qual. Quant.*, vol. 41, no. 5, pp. 673–690, Mar. 2007.

[84]   "http://www.mathworks.com/help/optim/ug/fmincon.html#f605285." [Online]. Available: http://www.mathworks.com/help/optim/ug/fmincon.html#f605285. [Accessed: 03-Apr-2014].

[85] W. Wu, X. Li, G. Gildenblat, G. Workman, S. Veeraraghavan, C. McAndrew, R. van Langevelde, G. D. J. Smit, a. J. Scholten, D. B. M. Klaassen, and J. Watts, "PSP-SOI: A Surface Potential Based Compact Model of Partially Depleted SOI MOSFETs," *2007 IEEE Cust. Integr. Circuits Conf.*, no. Cicc, pp. 41–48, 2007.

[86] W. Wu, X. Li, G. Gildenblat, G. Workman, S. Veeraraghavan, C. Mcandrew, R. Vanlangevelde, G. Smit, a Scholten, and D. Klaassen, "PSP-SOI: An advanced surface potential based compact model of partially depleted SOI MOSFETs for circuit simulations," *Solid. State. Electron.*, vol. 53, no. 1, pp. 18–29, Jan. 2009.

[87] PSP Group, "PSP Manual V102p3," *http://pspmodel.asu.edu/downloads/psp102p3_summary.pdf*.

[88] S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, vol. 2006. John Wiley & Sons, 2006, p. 832.