Title

Semiconductor Yield Assessment: An Integrated Modelling of Defect Limited and Parametric-Shift Limited Yield Mechanisms

ISBN

9798190924664

Author

Karthik Sankaran, FNU

Publication Date

2026-09-08

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Semiconductor Yield Assessment: An Integrated Modelling of Defect Limited and Parametric-Shift Limited Yield Mechanisms

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy in

Materials Science and Engineering

by

Karthik Sankaran

2026

ABSTRACT OF THE DISSERTATION

Semiconductor Defect Limited and Parametric-Shift Limited Yield Modeling, a data-based approach

by

Karthik Sankaran

Doctor of Philosophy in Materials Science and Engineering

University of California, Los Angeles, 2026

Professor Ali Mosleh, Chair

The Semiconductor industry is a multi-billion dollar industry and Semiconductor Yield is a fundamental factor driving device cost. Development cost, time-to-market & yield are the most important factors of interest to the industry. Two major components are associated with Semiconductor Yield: 1. Defect Limited Yield 2. Parametric-Shift limited Yield. Parametric Yield is influenced by process variations for a given design and process node.

Several models from literature which predict Defect-Limited yield based on primarily two parameters: Defect Density and Area of the Die. A “yield calculator” is developed that incorporates all Defect-limited models and allows the user to compare yield estimates among multiple models, or inversely estimate Defect Density, and calculate model parameters based on Foundry-provided Defect Density values and Wafer Sort (WS) test yield data. The methodology also includes a Bayesian formulation for aggregating yield estimates from various models in the form of an uncertainty distribution. This study also proposes a Process Adjustment Factor (PAF) to the

models to better estimate the variation of yield across various process nodes. PAF is estimated using a geometric probability approach using process-based factors such as Metal Pitch, Metal Width etc. PAF is used to adjust yield, and the results are compared with 4 different Application Specific Integrated Circuits (ASICs) at the 40nm and 65nm process nodes.

Many factors influence Parametric-Shift, with the most influential being process variations for a given process technology. Parametric shift is often estimated by circuit simulations which is both time consuming and costly. A simpler method is proposed which uses a machine learning + copula approach to model the complex relations contributing to parametric shift using Wafer Acceptance Test (WAT) data and WS Data (screened for yield loss due to parametric shift failures only). While Foundries do not share process-related data, they provide WAT data which are obtained from test coupons placed at specific locations on the wafer. WAT data records fundamental electrical parameters such as Oxide Thickness, Threshold Voltage, Saturation Current etc. and these recordings provide a window for observing process fluctuations. Trained Gaussian Copulas are used to generate Synthetic data for WAT parameters of a device and use trained machine learning Models (LGBM and XGBoost) to estimate yield with training accuracies of up to 94%.

Process node data is used in conjunction with Physics-informed models from SPICE (Simulated Program with Integrated Circuit Emphasis) for accomplishing this. Model generalization is attempted by including a Design Factor (Transistor Density) to account for design complexity and Process Factor (using Fab Cp/CpK which is associated with WAT parameter variation) to account for process node.

A combined yield metric is proposed which estimates the Total Yield as a function of defect yield and parametric-shift yield by treating them as independent failure events. The yield calculator is further enhanced with a feature for estimating yield loss due to parametric shift. Along with pre-

existing Defect-Limited Yield Models the calculator will output the Total Yield based on the combined yield metric.

The dissertation of Karthik Sankaran is approved.

Enrique A. Lopez Droguett

Amartya Banerjee

Alexander Balandin

Ali Mosleh, Committee Chair

University of California, Los Angeles

2026

Dedication

To my parents, my friends, and colleagues, who made this possible.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

List of Abbreviations

ASIC	Application Specific Integrated Circuit
BSIM	Berkeley Short-channel IGFET Model
CDF	Cumulative Distribution Function
CMOS	Complementary Metal-Oxide Semiconductor
CMP	Chemical Mechanical Polishing
FinFET	Fin Field-Effect Transistor
GBC	Gradient Boosting Classifier
LGBM	Light Gradient Boosting Machine
PCM	Process Control Monitoring
PDF	Probability Distribution Function
PTM	Predictive Technology Model
SHAP	Shapely Additive Explanations
SPICE	Simulated Program with Integrated Circuit Emphasis
TSMC	Taiwan Semiconductor Manufacturing Company
WAT	Wafer Acceptance Test
WS	Wafer Sort
XGBoost	Extreme Gradient Boosting

Acknowledgements

I would like to express my gratitude to everyone who supported me during my time as a graduate student at UCLA. First, I would like to thank my graduate advisor, Prof. Ali Mosleh, for his support and guidance, and for providing me the opportunity to work as a researcher at the B. John Garrick Institute for the Risk Sciences (GIRS). I would also like to extend my gratitude to Prof. Mosleh for his financial support during my time at UCLA. I would like to thank Kamran Hakim from Teradyne for acting as a second guide and for his mentorship throughout my PhD adventure. I would like to thank Prof. Enrique Lopez Droguett, Prof. Amartya Banerjee, and Prof. Alexander Balandin for their service on the doctoral committee as well as Prof. Qibing Pei and Prof. Ya Hong Xie for their service on Qualifying Examination committee.

Additionally, I would like to thank all my colleagues at the institute for their help and support throughout the years. I would like to thank Dr. Mihai Diaconeasa for introducing me to GIRS and Dr. Marilia Ramos for their support and guidance while I was still determining the direction of my studies as a PhD student. I would like to thank my colleagues and friends Lixian Huang, Bineh Ndefru, Theresa Stewart, Tingting Cheng, Camila Correa, and Reza Iranpour for their help on various modeling and technical problems, as well as providing a supportive group environment. I would like to extend my thanks to Joselyne Saldana for her amazing help and support during my time at GIRS.

I would also like to thank several peers at Teradyne including Kumar Kesavan for his important coding help, Kinh Han and his team for helping us welcome into the Teradyne facility and providing important data and Jin Yu for his guidance on machine learning approaches. I would also like to thank the following engineers at Teradyne: Peter Kim, Shih-Tun Chen, Paul

Vandervagt, Earl Shaw and Mike Fung for providing valuable suggestions and feedback on the project.

Finally, I would like to thank my mom and dad who have always believed in me and all my friends who supported and encouraged me during my journey as a Ph.D. student.

Vita

Education

2017 B.Tech. Metallurgical Engineering and Materials Engineering GPA: 3.3/4.0

Indian Institute of Technology, Bombay

Mumbai, Maharashtra, India

Research Experience

- Internship at Bhabha Atomic Research Center, Mumbai, India. Simulated the melting behaviors of metals and metal oxides using FORTRAN language based DFTB+ software, by performing Molecular Dynamic Simulations on the optimized structures of the material. Carried out geometry optimizations using the Conjugate Gradient Algorithm to determine structures of various metals and optimized for required melting points. Analysis provides basis for research into improving cement used in nuclear reactors
- Internship at Monash University, Melbourne, Australia. Performed a 3 step Monte Carlo simulation on large-scale randomly generated lattice mesh of hydrogel network. Outputs of the model provided framework to be used in mechanical experimentation for hydrogel properties. Simulation was optimized for total energy and distance distributions and to contribute for experimental validation
- Research in UCLA, Los Angeles, USA. Performed comprehensive reliability analysis on industrial systems, applying Weibull estimation based on Maximum Likelihood Estimates and Bayesian assessment methods. Analyzed 14 distinct failure modes to conduct cost-benefit evaluations, identifying root causes for operational viability. Analyzed plant generation failure risk using Fault Tree Analysis on 150+ nodes

- Research in UCLA, Los Angeles, USA. Developed a COVID-19 community response and spread model along with two other graduate students. This model tracked a population of individuals as they made decisions to participate in several categories of public and private events which could lead to exposure to the COVID-19 virus. The potential effects of different vaccine and quarantine policies on community spread were modeled for various hypothetical scenarios.

Publications

- W. Chalgham, T. Stewart, K. Sankaran, A. Mosleh, "Risk Assessment of the Feedwater Heaters at Diablo Canyon Power Plant", A Collaborative Research and Development Agreement between The B. John Garrick Institute for the Risk Sciences at UCLA and Pacific Gas and Electric Company, 2026
- K. Sankaran, A. Mosleh, K. Hakim, "Semiconductor Parametric Yield Prediction: A Combination of Machine Learning and Physics-Informed Approach," 2025 Annual Reliability and Maintainability Symposium (RAMS), Destin, FL, USA, 2025
- B. Ndefru, K. Sankaran, T.M. Stewart, A. Mosleh, A. Earthperson, N. Zawalick, 2022, "Risk-Informed Decision-Making Tool for Covid-19 Community Behavior and Intervention Scenario Assessment," *Proceedings of the International Conference on Probabilistic Safety Assessment and Management*, Honolulu HI
- B. Ndefru, K. Sankaran, T.M. Stewart, A. Mosleh, A. Earthperson, N. Zawalick, 2022, "Risk-Informed Decision-Making Tool for Covid-19 Community Behavior and Intervention Scenario Assessment," *Proceedings of the Annual Meeting of the Society for Risk Analysis*, Tampa FL

- Ramos MA, Sankaran K et. al. The need for and conceptual design of an AI model-based Integrated Flight Advisory System. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, 2022
- K. Sankaran, A. Mosleh and K. Hakim, "A Physics-Informed Data-Driven Framework for Semiconductor Parametric-Shift-Limited Yield Modeling Using Copulas and Machine Learning", *IEEE Transactions on Semiconductor Manufacturing,* 2026 (under review)
- K. Sankaran, A. Mosleh, K. Hakim, "A Comprehensive Review and Proposed Methodology for Improving Semiconductor Defect-Limited Yield Estimation", *IEEE Transactions on Semiconductor Manufacturing,* 2026 (under review)

1 Introduction

1.1 Motivation

The semiconductor industry involves the design and manufacturing of integrated electronics/devices and is growing by leaps and strides every year. The industry had an estimated revenue of $681.05 billion [1] in the year 2022 and it is estimated to grow almost 3 times the amount by the year 2032 with a projected Compound Annual Growth Rate of almost 10.06% [2].

The large-scale manufacturing of these semiconductor chips requires precise control of the yield during manufacturing, which with even a fraction of a percentage difference in prediction of 100s of millions of units can lead to a difference in a projection of several millions of dollars in cost. The ability to estimate yield is the key driving factors for companies in their decision-making processes during product development. Yield modeling in the semiconductor industry is primarily defined as Eq. 1.1.

$$Yield = \frac{total\ \#\ working\ devices}{total\ \#\ devices\ manufactured} \tag{1.1}$$

Fab yield is estimated as the product of the Line Yield and Electrical-Test (E-Test) yield [3] also known as WAT (Wafer Acceptance Test) and PCM (Process Control Monitor). Line yield represents the number of wafers which survive the manufacturing line prior to E-test, whereas the E-Test yield represents the number of surviving wafers which are good enough to continue. E-Test yield can be further divided into two categories: Systematic and Random. The systematic yield represents nonrandom defects affecting every die in some region of the wafer. Systematic defects are caused by systematic variations during manufacturing such as chemical mechanical polishing (CMP), photoresist patterning, and other processes. This is not the focus of this research work. The focus of this research lies in developing a plausible model for random yield caused due to

randomly occurring defects in the process or random shifts in electrical parameters. These are not systematic and are always present in semiconductor processes.

1.2 Background

Random yield can be divided into two types: Defect-limited Yield and Parametric-shift-limited Yield.

1.2.1 Defect Limited Yield

Defect-limited yield models are used to represent the classification of failures caused by random defects in the Fab or on the assembly line such as those caused by external intrusive particles on the wafer or extra/missing material-based defects. Random defects are unpredictable in nature such as failures caused by embedded particles during photolithography, scratches in the metal layer, and pinholes in oxide layers, etc. Several of these yield models are discussed in section 3 and an adjustment factor based on process node is further discussed in section 5.

1.2.2 Parametric-Shift Limited Yield

Parametric-shift limited yield (or simply parametric yield) refers to the shifts in device parameters. This can be caused due to multiple factors such as design margins, voltage fluctuations, temperature variation and defects. However, processing variation is often considered as the most influential cause that can cause the device to function outside the operational limits. Parametric yield is considered as the more difficult of the two sources of yield to quantify. This is due to the large number of process variables and due to the increasingly complex nature of semiconductors as more transistors are being squeezed into an already small space at lower nodes.

Parametric yield models are far fewer in comparison to defect-based yield models. Unlike defect-based yield models, most of the research is focused on improving parametric yield detection. Their

goal is to prevent parametric yield fluctuations with the help of image recognition, parametric estimation, or utilization of digital twins [4] [5] [6]. To estimate the parametric yield due to random defects, there has been research that estimates yield using circuit-based simulations. David Gibson et. al. [7]. have explored the simulation of a multinormal distribution with oxide thickness and voltage to estimate parametric deviation. Bartlomiej et. al. [8] explores simulation of AlGaN High Electron Mobility Transistors (HEMTs) to estimate yield. However, these approaches can be very time-consuming due to the inherent nature of circuit simulations which can take days or weeks to finish. Hence, this research focuses on attempting a multi-step data-based approach to parametric yield which can perform estimations within a few minutes. The modeling methodology is discussed in section 7 and the results are discussed in sections 9, 10, 11 and 12.

2 Research Objectives and Contributions

2.1 Problem Statement

Yield is conventionally classified into two components — Defect-Limited Yield, caused by random particles and material defects in the fab or assembly line, and Parametric-shift Limited Yield\, caused by process variations that push device parameters outside operational limits — and each suffers from a distinct modeling gap

On the defect-limited side, the literature offers roughly fifteen statistical yield models (Poisson, Murphy, Bose-Einstein, Negative Binomial, and others), but each was developed under specific assumptions about defect distribution, and no single model performs uniformly well across process nodes, die sizes, or circuit types. Critical model parameters such as the Bose-Einstein factor n and the clustering factor α are not directly extractable from foundry data and must be back-calculated each time the process or design changes. Furthermore, as devices scale below 65 nm, classical

models become increasingly insensitive to process node — a 130 nm device and a 16 nm device with identical defect density would receive the same yield estimate from a Poisson model, despite the latter being demonstrably more vulnerable to the same particle population.

On the parametric-shift side, the modeling gap is more severe. Parametric yield is dominated by complex, correlated process variations whose interactions are typically captured only through full circuit simulation (e.g., SPICE-based Monte Carlo studies). These simulations can take days or weeks per device and require extensive access to proprietary process and design information, making them impractical for rapid iterations required in real-world product development, especially without access to proprietary process related data. Existing data-driven parametric yield work has focused primarily on detection (image recognition, digital twins) rather than on producing rapid, simulation-free yield estimates from the routinely collected WAT and WS datasets that every foundry already provides. Hence, it is imperative to develop models which can predict the yield with sufficient accuracy without consuming much time.

2.2 Research Objectives

To address the gaps identified in Section 2.1, this research pursues the following objectives:

2.2.1 Defect-Limited Yield

1. **Consolidate and evaluate existing defect-limited yield models** to enable direct model comparisons for any given combination of die area and defect density such as in an Excel-based software or similar to provide easier access to yield estimation models.
2. **Quantify uncertainty in defect-limited yield estimates** to determine the uncertainty limits that can be expected due to estimating defect yield with different models.

3. **Address Process Node dependency of defect yield** to improve scaling of existing yield models.

2.2.2 Parametric-Shift Limited Yield

1. **Develop a non-simulation based parametric-shift model** that can predict parametric-shift yield without the requirement for costly SPICE simulations. The proposed methodology will use a data-based approach using WAT data and WS data. They will be analyzed using a Gaussian copula to model dependencies between WAT parameters and Machine Learning (ML) to map WAT vectors to pass/fail outcomes.

2. **Address generalization of parametric-shift limited yield** by incorporating Design and Process node based factors similar to defect limited yield.

2.2.3 Yield Calculator and Combined Yield Metric

1. **Introduce a combined Total Yield metric** which involves both defect limited yield and parametric-shift limited yield as a factor to estimate the total yield.

2. **Consolidate total yield estimation in a single yield calculator.** This will be done using an Excel-based tool with Python-backed ML services, supporting both forward yield prediction and reverse parameter estimation.

2.3 Research Contributions

The principal contributions of this dissertation are as follows:

2.3.1 Defect-Limited Yield

- **The Process Adjustment Factor (PAF)** — a novel correction factor derived from a geometric probability model using Design Rule parameters for Process nodes and ISO fab-class particle distributions. PAF demonstrably improved Bose-Einstein yield estimates for three out of four evaluated 40 nm ASICs and provides the first principled way to introduce process-node sensitivity into classical defect-limited yield models without requiring access to proprietary foundry data.

2.3.2 Parametric-Shift Limited Yield

- **A parametric shift yield model** using a 3 step process that combines (i) Berkeley short-channel IGFET model (BSIM) Level 2 physics-informed WAT estimation, (ii) Gaussian copula modeling of WAT joint structure, and (iii) tree-based ML classifiers achieving up to 94% test accuracy on combined 40 nm and 16 nm datasets — which can produce yield outputs in seconds.

- **Empirical evidence for cross-node generalizability** of the parametric pipeline: correlation tables from the Gaussian copula were shown to be largely consistent in 40 nm and 16 nm devices, and Shapely Additive Explanations (SHAP)-based feature importance analysis identified the shared dominant WAT parameters, justifying mixed-device training and supporting downstream root-cause analysis of parametric failures.

- **Generalization factors** — Pelgrom's Law and Cp/CpK — integrated into the model as lightweight modifiers to the copula's synthetic dataset, enabling the trained ML models to respond to changes in transistor geometry and process maturity without retraining, and

providing the first explicit yield-impact lever for design and fab capability decisions within the framework.

2.3.3 Yield Calculator and Combined Yield Metric

- **A unified defect-limited yield calculator** implementing all fifteen literature models in a single environment, supporting both forward yield estimation and reverse extraction of defect density and model parameters from measured production data. It also contains a Bayesian non-homogeneous uncertainty framework for aggregating yield estimates across models, using a truncated lognormal posterior and a multiplicative-error likelihood implemented with the help of R-DAT software.
- **A unified Total Yield metric** combining defect-limited and parametric-shift yield under a probabilistic OR-gate independence assumption.
- **A Yield Calculator** integrating all of the above into an Excel front-end with Python-backed ML services, designed for use by engineers in product development and capable of eventual deployment as a Software Application within an ATE toolbox.

DEFECT LIMITED YIELD MODELING

3 Defect Yield: Existing Yield Models

Existing defect-limited yield models mainly rely on two key parameters;

- Average defect density D_0
- Area of the die A

This leads to the calculation of the total number of defects in a wafer as $D_0 * A$

Assuming the occurrence of defects as a random Poisson process, and X as the number of defects

$$P(X = x) = \frac{e^{-D_0 A} * (D_0 A)^x}{x!},\ x = 0,\ 1,\ 2, \ldots \tag{3.1}$$

Where P(X=x) is the probability of number of defects equaling to a certain amount x. This leads to the creation of the Poisson model which can simply be calculated by substituting x as 0 as the definition of a working die is the probability that the die has no defects.

$$Y = P(X = 0) = e^{-AD_0} \tag{3.2}$$

where Y is the yield. However, this assumes a fully random distribution of defects and does not capture any "spatial" distribution of these defects. Sometimes the defect density may not be uniformly distributed through the wafer, and we need to consider the possibility of the yield being calculated with a defect density distribution (f(D)). This leads to:

$$Y = \int_0^\infty P(X = 0) f(D) dD = \int_0^\infty e^{-DA} f(D) dD \tag{3.3}$$

Eq. 1.4 leads to the variety of yield models in the literature based on varying f(D), as discussed below.

3.1 Poisson Model

The Poisson Yield model was proposed by Moore et al. [9], based on the assumption of a constant value of defect density across the wafer, which leads to the following form:

$$Y = e^{-AD_0} \tag{3.4}$$

The Poisson model has found popular usage in the industry due to its simplicity since the distribution of defect density is generally not known. The Poisson Yield model works best for smaller die areas (A<=0.25cm^2**)**.

3.2 Murphy Model

Murphy Model was proposed by Murphy et al. [10] In their experiments, they found the distribution of defect density to follow more of a Triangular distribution:

$$Y = \left(\frac{1-e^{-AD_0}}{AD_0}\right)^2 \tag{3.5}$$

However, it has been suggested that the reason for using a triangular distribution is due to the difficulty integrating [11] a Gaussian distribution which their observations resembled more.

Seeds Model: Seeds et.al. [12] found that most defects were caused by large populations of low defect densities and proceeded to use an exponential defect distribution.

$$Y = \left(\frac{1}{1+AD_0}\right) \tag{3.6}$$

This formula unfortunately overestimates the yield even according to Seeds's [13] own findings.

3.3 Bose-Einstein Model

The Bose-Einstein Yield Model is one of the popular models in the industry which was also proposed by Seeds [14]. It also refers to an extra factor in the model called the number of critical processes (n) that are susceptible to defects: the total number of independent critical mask layers, or the total number of critical processing steps used. This allows for further tuning of the yield models with this extra factor.

$$Y = \left(\frac{1}{1+AD_0}\right)^n \tag{3.7}$$

One of the issues of this model is the unavailability of process data to the user for determining the true value of n which even by its own definition can be ambiguous. Typically, nominal values of n are suggested by the fabrication plant to help estimate yield.

3.4 Dingwall Model

Dingwall et al. [15] proposed a variation of the Murphy model based on statistical reasoning. He mentioned the lower bound of yield as the Poisson equation and the upper bound as a mixture of Poisson distributions and Price and Stapper models.

$$Y = \left(1 + \frac{AD_0}{3}\right)^{-3} \tag{3.8}$$

Dingwall's yield model has seen very little use, and its accuracy is debated [16].

3.5 Moore Model

After proposing the Poisson model, Moore also proposed the Moore model based on statistical inference. However, similar to the Dingwall Model, it has also not seen much usage outside of theoretical postulation.

$$Y = e^{-\sqrt{AD_0}} \tag{3.9}$$

3.6 Price Model

Price [17] proposed a modification to the Bose-Einstein model by having individual defect densities (D_1-D_n vs. D_0) for each critical process. This model can also be used with individual defect density values for each process resulting in Eq. 3.10:

$$Y = \left(\frac{1}{1+AD_1}\right)\left(\frac{1}{1+AD_2}\right)\ldots\left(\frac{1}{1+AD_n}\right) \tag{3.10}$$

The Bose-Einstein model continues to be more popular due to lack of process related data to determine values of individual defect densities. The challenge for the user, however, is to obtain multiple D values for this model from the foundry.

3.7 Okabe Model

Okabe [18] propsed the usage of a gamma distribution for the defect distribution as opposed to the simplification done by Murphy. This results in a factorfrom the gamma distribution which can further tune the yield model:

$$Y = \left(1 + \frac{AD_0}{n}\right)^n \tag{3.11}$$

However, Okabe's usage of the gamma factor was considered to be flawed since he treated the gamma factor to be the same as the negative binomial factor which accounted for a number of critical processes [16].

3.8 *Negative Binomial Model*

The negative binomial model is another popularly used model proposed by Stapper et al. [19]This is similar to the Okabe model however, it approximates the factor α instead, which is defined as the clustering factor. A higher value of α correlates to a lower probability of clustering.

$$Y = \left(1 + \frac{AD_0}{\alpha}\right)^{-\alpha} \tag{3.12}$$

3.9 Other Models

The following models are very niche in use or are purely theoretical, hence not having any notable applications.

3.9.1 Half-Gaussian Model:

This is a purely theoretical model proposed by Raghavachari et al [20]. It uses a Half-Gaussian distribution to estimate yield.

$$Y = \exp\left[\left(\frac{\pi}{4}\right) AD_0^2\right] erfc\left(\frac{\pi^{\frac{1}{2}} D_0 A}{2}\right) \tag{3.13}$$

3.9.2 Inverse-Gaussian Model:

This model is also purely theoretical in nature and was proposed by Raghavachari et al [20]. who uses an Inverse-Gaussian distribution instead.

$$Y = \exp[\phi(1 - \left(1 + \left(\frac{2AD_0}{\phi}\right)\right)^{\frac{1}{2}}] \tag{3.14}$$

Both Half-Gaussian and Inverse-Gaussian interpret spatial dependancies of defects but are considered to not be sufficient since they are unable to describe clustering on a wafer as both models simulate individual dies. This requires the denoting of areas on the wafer where there are zones with uniform defect density and other zones with mixed distributions which makes it hard to use for general use cases.

3.9.3 Rayleigh Model:

The Rayleigh Model tries to estimate the yield using a Rayleigh distribution [21]. This is also theoretical in nature.

$$Y = 1 - AD_0 e^{\left(\frac{(AD_0)^2}{\pi}\right)\left(erfc\left(\frac{AD_0}{\pi^{\frac{1}{2}}}\right)\right)} \tag{3.15}$$

3.9.4 Hofstein and Heimann Model:

Hofstein and Heimann [22] proposed a model where the defects are only affecting transistors. This results in a model similar to the Poisson model but with transistor parameters.

$$Y = e^{-NA_gD} \tag{3.16}$$

Where N is the number of transistors and Ag is the area of the transistor. A major issue with this model is the fact that defects can be caused in various ways and often cause catastrophic issues. The model is hence not considered universally valid but can be used for simple approximations (areas A<=0.25cm^2).

3.9.5 Berglund 2 Factor Model:

The Berglund 2 factor model [23] assumes large defect sizes and estimates the yield caused by a difference between the area of the die and the defect. The simplicity of the defect size-based multiplier results in higher error rates when evaluating yield, hence leading to its relatively low usage.

$$Y = e^{-LWD_0} . e^{-D_0\left((L+W)s_0+\left(\frac{\pi}{2}\right)s_0^2\right)} \tag{3.17}$$

Where L and W are the length and breadth of the die respectively and So is the average defect diameter.

Table 1: A Summary of defect-based yield models

Model name	Equation	Extra Parameters
Poisson	$Y = e^{-AD_0}$	N/A
Murphy	$Y = \left(\frac{1 - e^{-AD_0}}{AD_0}\right)^2$	N/A
Seeds	$Y = \left(\frac{1}{1 + AD_0}\right)$	N/A
Bose-Einstein	$Y = \left(\frac{1}{1 + AD_0}\right)^n$	Bose Einstein factor = n
Dingwall	$Y = \left(1 + \frac{AD_0}{3}\right)^{-3}$	N/A
Moore	$Y = e^{-\sqrt{AD_0}}$	N/A
Price	$Y = \prod_{i=1}^{n}\left(\frac{1}{1 + AD_i}\right)$	Number of processes = n
Negative Binomial	$Y = \left(1 + \frac{AD_0}{\alpha}\right)^{-\alpha}$	Clustering factor = alpha
Okabe	$Y = \left(1 + \frac{AD_0}{n}\right)^n$	Bose Einstein factor = n
Uniform	$Y = \frac{1 - e^{-2AD_0}}{2AD_0}$	N/A
Half-Gaussian	$Y = exp\left(\phi\left(1 - \left(1 + \left(\frac{2AD_0}{\phi}\right)\right)^{\frac{1}{2}}\right)\right)$	Shape parameter = ϕ
Inverse-Gaussian	$Y = \left(\frac{1}{1 + AD_0}\right)^n$	Bose Einstein factor = n
Rayleigh	$Y = 1 - AD_0 e^{\left(\frac{(AD_0)^2}{\pi}\right)\left(erfc\left(\frac{AD_0}{\frac{1}{\pi^2}}\right)\right)}$	N/A
Hofstein and Heiman	$Y = e^{-NA_g D}$	N/A
Berglund 2 factor	$Y = e^{-LWD_0} . e^{-D_0\left((L+W)s_0 + \left(\frac{\pi}{2}\right)s_0^2\right)}$	L, W = die dimensions, s_0 = average defect diameter

Some of these 15 models listed in Table 1 have been widely used for decades, such as Bose-Einstein, Negative Binomial, Murphy, and Poisson models. With only a few inputs (D, A, n, α, etc.), these statistical models aim to provide fast and reasonable yield estimation from empirical data. However, drawbacks exist when applying these models in the ASIC yield estimation:

Some parameters are not straightforward (cannot be directly extracted from tests during fabrication processes), and are therefore hard to acquire from the foundry. For example, n and α will need back calculation every time the process and design change; the foundry provides process control

monitoring (PCM) yield, but this yield is related to the parametric measurements done on the dies and not on the devices used for yield control.

As the device scale continuously shrinks, the complexity of design increases and the specifications are stringent these statistical models no longer provide realistic yield estimates. Additional parameters should be required to cover these aspects. For example, when everything is the same, the device in the 130 nm technology node should be less sensitive to the same level of defects if compared with that in the 32 nm node; with everything else the same, devices with stricter failure criteria should have lower defect-limited yield. However, neither of the above cases can be represented using the above models.

Circuits and ASICs with different functionalities can have different defect-limited yield estimations. Regarding the above issues, a new tool for defect-limited yield estimation is needed which is the one of the main objectives of this research work. Section 13.2 will present a calculator which can be used to evaluate different models.

4 Defect Yield: Scope of the model

This section defines the scope, inputs, outputs, and architectural boundaries of the yield estimation framework developed in this dissertation, establishing what the framework is designed to do, what data it requires, and where the limits of its applicability lie before the methodology chapters introduce each component in detail. The framework is organized around two parallel modeling methodologies that converge into a single yield output: A) A defect-limited model that estimates yield loss from random particles and material defects using fifteen statistical models from the literature, augmented by B) A Process Adjustment Factor that introduces process-node sensitivity (Section 5).

4.1 Model Inputs

4.1.1 Inputs related to Defect-Limited Yield

The defect-limited stream operates on a small set of geometric and statistical inputs that are either directly available from foundry documentation or routinely measured during production:

- **Die area (A)** — derived from die height and width
- **Wafer geometry** — wafer diameter, edge exclusion width, notch dimensions, and scribe line width, used by the De-Vries equations to estimate gross die count per wafer
- **Average defect density (D_0)** — provided by the foundry or back-calculated from measured WS yield
- **Model-specific parameters** — Bose-Einstein factor n, clustering factor α, average defect diameter s_0, and number of critical processes, depending on which of the fifteen models is selected

- **Circuit block area (optional)** — area fractions for sub-regions in a die (eg: IO, Logic, Memory, Pure Analog, and High-speed Digital) to measure yield corresponding to the sub-regions

4.1.2 Inputs related to Process Adjustment Factor

PAF requires publicly available process node parameters drawn from Design Rule documentation, deliberately avoiding any dependence on proprietary foundry data:

- **Process node** (e.g., 250 nm, 180 nm, 130 nm, 65 nm, 40 nm, 16 nm) : uses preset values
- **Process node parameters (optional):** The following parameters may be provided if the process is different to preset values: Metal pitch and metal size for each metal layer, Via width and via height for each via level, Number of metal layers and metal layer heights
- **ISO fab cleanliness class** (ISO 1 through ISO 6), which determines the particle size distribution used in the probability calculation for the particle causing a failure (used to estimate PAF as detailed in section 5).

4.2 Model Outputs

4.2.1 Defect-Limited Yield Estimate

For any combination of die area, defect density, and selected model, the calculator produces a point estimate of defect-limited yield (Y_D). When multiple models are selected, the calculator returns a comparative table and graph showing how predictions diverge across models for the same inputs, which is useful for engineers selecting the most appropriate model for a given device or process.

4.2.2 Reverse-Calculated Parameters

The calculator supports inverse operation: given the defect-limited yield obtained from WS data, it back-calculates either the effective defect density D_0 or a chosen set of model parameters (Bose-Einstein n, clustering α, average defect diameter s_0, etc.). Closed-form solutions are used where available (Poisson); numerical solvers (Excel's solver tool) are used for models with more complex parameter relationships.

4.2.3 Bayesian Yield Uncertainty Distribution

Beyond point estimates, the framework produces a posterior uncertainty distribution of yield by aggregating estimates across all fifteen defect-limited models using a non-homogeneous Bayesian framework with a truncated log-normal posterior and multiplicative-error likelihood. The output is a posterior Probability Distribution Function (PDF) and CDF, together with 5th and 95th percentile bounds, providing a defensible uncertainty band rather than a single value.

4.3 Scope Boundaries and Assumptions

- **In scope:** Random defect-limited yield and process-node sensitivity (discussed more in Section 5)
- **Out of scope:** Consideration of Systematic yield (repeatable, non-random defects affecting every die in a wafer region due to CMP, photoresist patterning, etc.), Design Complexity effects beyond transistor density (which would require full circuit simulation), and post-packaging Final Test failures

5 Defect Yield: Process Adjustment Factor (PAF) for Defect Limited Yield

5.1 PAF Definition

The general understanding is that yield is influenced by two major factors: one related to Process Node Complexity and the other related to Design Complexity. In this research we introduce an approach to tackle "Process Node Complexity" in the form of PAF. Accounting for design complexity requires consideration of high variations in electrical circuitry between different devices, posing a formidable challenge requiring further effort and investigation and is considered to be beyond the scope of this research.

The version of PAF that will be discussed first is one that is deemed to be used as a multiplier to yield numbers from any of the yield models.

$$Yield_{defect} = Model\ Yield_{estimate} \times PAF \tag{5.1}$$

Where Model Yield is the output from yield models listed in section 3. This equation is valid for every model in the calculator with the exception of the Hofstein and Heiman model which already accounts for the particle size hence rendering some of the considerations in PAF redundant. In this way, the statistical variations captured by the yield models discussed in section 3 will be maintained and PAF can be modeled by focusing solely on process node data. Due to the proprietary nature of the process and design, the approach for evaluating this factor is based on publicly available data. Therefore, the main variations from one process node to another are based on the "Design Rules" (not to be confused with "Design Complexity") of the processes and with Figure 1 highlighting their influence on PAF.

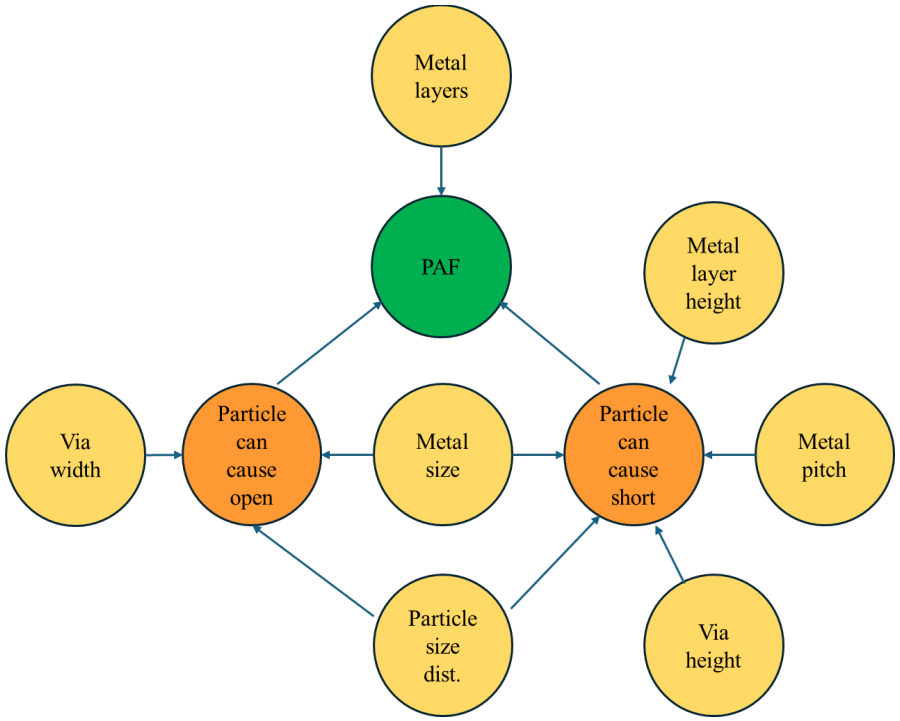


Figure 1: Diagram showing the relation of various process-based factors to the proposed process adjustment factor (PAF). The middle layers: "Particle can cause open", and "Particle can cause short" are for the types of faults that can cause a failure.

Design Rules refer to the restrictions based on a series of parameters provided by manufacturers that are set to account for variability in a manufacturing process. These parameters include various characteristics associated with metal layers:

- Pitch: Center to center distance between the same layers of metal
- Size: Cross-sectional dimension of metal line (breadth and height) on corresponding metal layer
- Via width: Center to center distance between the same layers of via
- Via height: Center to center vertical distance between consecutive layers of vias
- Number of layers: Number of metal layers starting from lowest (transistors) to outermost metal layer
- Layer Height: Center to center vertical distance between consecutive metal layers

A short circuit will occur if the particle is large enough to bridge two metal lines that are either adjacent to one another or reside on the top of the other. An open circuit will occur if the particle

is large enough to fully block a metal line. The final piece of data represents the distribution of particles in a fab. For estimating the distribution of particles, the worst-case scenario is used, and the particle sizes are distributed by the ISO standards of a respective fab.

Each of these factors determines the probability of a particle landing in a way that can cause either a short circuit or an open circuit. This is estimated with the help of using the device geometry associated on a given metal layer and the corresponding via that connects to the next metal layer above. Figure 2 shows the representative diagram used to demonstrate the various interactive elements. Using this, a PAF is estimated based on the geometrical parameters that are mentioned in the Bayesian network.

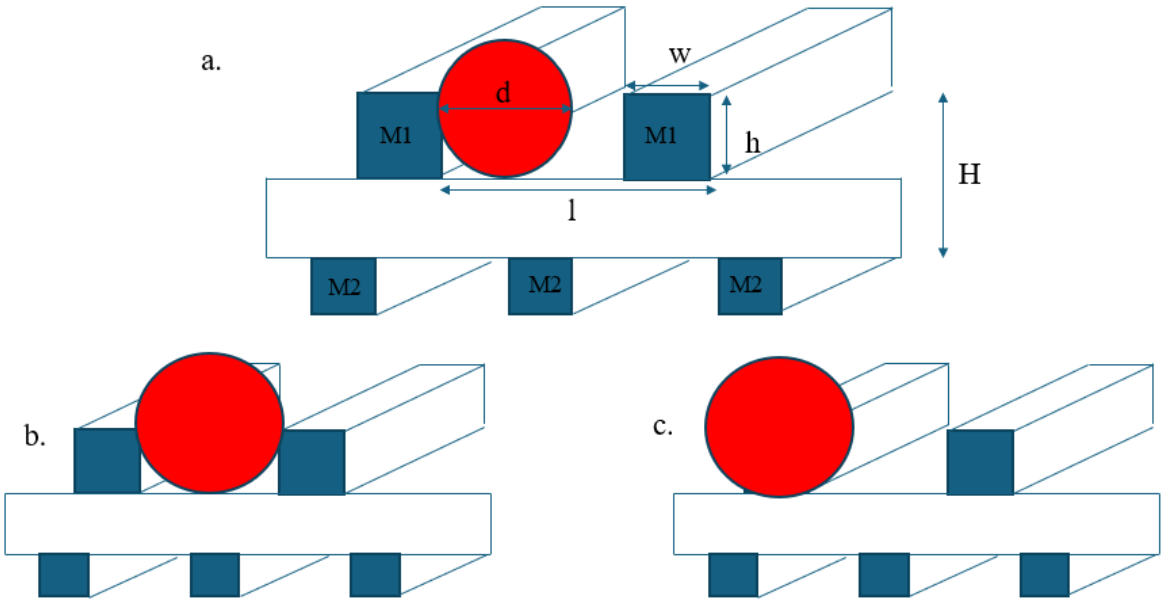


Figure 2: Representative diagram of a particle landed on a metal layer with different scenarios a. particle causes no issues, b. particle causes a short between two lines, and c. particle causes an open by blocking a metal line

5.2 PAF Estimation

The probability of causing harmful interaction is estimated for each layer-via combination defined in a process. For example, a 130nm process node which has 6 metal layers and 5 sets of vias will require 24 sets of calculations to estimate all the possible ways a particle can cause issues.

Assuming a single particle landing in any one of the layers within a device, causing an issue, the overall probability of failure for a particle size can be represented as an OR gate with each metal layer's probability of failure as a part of the OR gate. Moreover, since the failure, size and landing probabilities are independent events from each other, the corresponding probabilities can be calculated with Baye's theorem. With the help of these rules, a model of PAF, as given by Eq. 5.2, is constructed that assumed a particle of a certain size landing on the die which can cause defects by either a short circuit or an open circuit.

$$PAF = 1 - \sum(P_{fail} * P_{size} * P_{landing}) \tag{5.2}$$

Where P_{fail} is the probability of layer or via causing a short or open at a given ISO class, P_{size} is the probability of certain size particles existing at a given ISO class, and $P_{landing}$ is the probability of particles landing on the ASICs. P_{fail} is calculated for a given particle size and process node. The probability of the size of the particles are randomly sampled based on the worst-case measurements provided by ISO standards. The probability of landing is estimated with the help of real data from 40nm process node. An example calculation for the factor for the 40nm node is shown in Table 2 where the probability of a bad die is estimated.

Table 2: Evaluating probabilities of a short or open occurring for a particle of size 100nm. If the probability is 0, in that case the event does not occur for this particle. The final probability of failure is estimated based on the likelihood that the particle can land on any one of these layers

Process Node	40 nm									Particle size	100 nm			Two events happening together		
										Shorts		Opens				
Details	Pitch (nm)	Thickness	Width (nm)		Via name	Via height	Via width	H	Metal layer	Layer short probability (vertical) P1	Layer short prob. (horizontal) P2	Layer open prob. P3	Via open prob. P4	P1 ∩ P2	P3 ∩ P4	Probability of no fault in this layer
Metal 1 (M1)	140	125	70		Via1	54	72	179	Metal 1	0	0.21428571	0.2142857	0.1056604	0	0	0.0267116
Metal 2 (Mx1)	140	145	70	V	Via2	63	63	208	Metal 2	0	0.21428571	0.2142857	0.1298246	0	0	0.0279198
Metal 3 (Mx2)	140	145	70	i	Via3	63	63	208	Metal 3	0	0.21428571	0.2142857	0.1298246	0	0	0.0279198
Metal 4 (Mx3)	140	145	70	a	Via4	63	63	208	Metal 4	0	0.21428571	0.2142857	0.1298246	0	0	0.0279198
Metal 5 (Mx4)	140	145	70	s	Via5	63	63	208	Metal 5	0	0.21428571	0.2142857	0.1298246	0	0	0.0279198
Metal 6 (Mx5)	140	145	70		Via6	63	63	208	Metal 6	0	0.21428571	0.2142857	0.1298246	0	0	0.0279198
Metal 7 (Mx6)	140	145	70		Via7	63	63	208	Metal 7	0	0.21428571	0.2142857	0.1298246	0	0	0.0279198
Metal 8 (My1)	280	145	140		Via8	324	306	469	Metal 8	0	0	0	0	0	0	0
Metal 9 (Mz1)	800	850	400		Via9	324	306	1174	Metal 9	0	0	0	0	0	0	0
Metal 10 /(Mu)	3000	3500	1000		Via11	1800	1800	5300	Metal 10	0	0	0	0	0	0	0
Metal 11	3000	3500	1800						Metal 11	0	0	0	0.0153846	0	0	0.0007692
														Prob. Of bad die		0.1942304

Table 3: ISO Fab classes and particulate concentrations. These numbers are used to estimate the partial fraction of each particle size and hence estimate the probability of a particular size landing. ISO 7,8 and 9 are not used in PAF calculations due to unknown concentrations

ISO Class	FED Class	> 0.1 micron (no./m^3)	> 0.2 micron (no./m^3)	> 0.3 micron (no./m^3)	> 0.5 micron (no./m^3)	> 1 micron (no./m^3)	> 5 micron (no./m^3)
ISO 1		10	0	0	0	0	0
ISO 2		100	24	10	0	0	0
ISO 3	Class 1	1000	237	102	35	0	0
ISO 4	Class 10	10000	2370	1020	352	83	0
ISO 5	Class 100	100000	23700	10200	3520	832	0
ISO 6	Class 1000	1000000	237000	102000	35200	8320	293
ISO 7	Class 10000	very high	very high	very high	352000	83200	2930
ISO 8	Class 100000	very high	very high	very high	3520000	832000	29300
ISO 9	Room Air	very high	very high	very high	35200000	8320000	293000

ISO Plant particulate matter limits (as shown in Table 3) are used to estimate the probability of a particular particle size. Using the estimates from Table 2, the PAF factor is evaluated for various process node and ISO plant combinations as shown in Table 4. The probability of a defect causing particle being present goes up for smaller process nodes while also going up for less strict ISO

standards. Additional factors must be considered in this model in order to moderate the fast yield drop at lower nodes (40nm and 16nm). PAF modifies the yield as a multiplier to existing models.

Table 4: Final probabilities of success for different process nodes and their respective fab classes, given a particle lands.

Process nodes	250 nm	180 nm	130 nm	65 nm	40 nm	16 nm
Fab class						
ISO 1	0.999	0.997	0.996	0.998	0.806	0.745
ISO 2	0.998	0.996	0.984	0.923	0.767	0.604
ISO 3	0.996	0.992	0.980	0.909	0.752	0.594
ISO 4	0.995	0.989	0.975	0.904	0.749	0.591
ISO 5	0.995	0.989	0.975	0.904	0.749	0.590
ISO 6	0.995	0.989	0.975	0.904	0.749	0.590

Since the PAF is designed to be a multiplier to the yield models, this results in an inherent bottleneck of the maximum possible value of yield being restricted to the model's maximum mathematical output. This would result in the calculation of the yield of larger process nodes to have a value of the PAF greater than 1 which should not be possible because PAF is being calculated as a probability-based factor. PAF can also be reformulated by making it affect the Area of the die into a Critical Area (A_c) which is an estimate of the available die area for a particle to cause an issue. This can be done since PAF is already calculated as a geometrical probability of failure.

$$\mathrm{Y_{defect}} = \int e^{-A_c D}(f(D)dD) \tag{5.3}$$

$$where\ A_c = PAF \times A \tag{5.4}$$

Table 5: Sort Yield value comparison for 4 ASICs in 40nm technology node and percentage difference in accuracy between the models, using PAF as a multiplier and PAF as a modifier to critical area

ASIC	Actual Yield	BE Model yield	% difference	Model yield with PAF as Ac	% difference	Model yield with PAF mult	% difference
ASIC A (40nm)	84.50%	90.64%	7.27%	88.52%	4.76%	82.00%	2.96%
ASIC B (40nm)	65.72%	73.34%	11.59%	68.11%	3.64%	66.43%	1.08%
ASIC C (40nm)	85.00%	76.21%	10.34%	71.42%	15.98%	69.03%	18.79%
ASIC D (65 nm)	60.40%	63.87%	5.75%	60.94%	0.89%	57.85%	4.22%

To validate PAF, an evaluation is performed on the Bose-Einstein model using the Foundry provided Bose-Einstein parameter (n) and Defect Density D_0. Table 5 shows the comparison of accuracy data of yield estimation with and without using PAF (as both a multiplier and with critical area). For ASIC A, B, and D, Bose-Einstein model overestimates the sort yield, whilst this PAF correction improves the estimation results. However, for ASIC C, where BE model underestimates the yield value, the multiplication of PAF can increase discrepancy between the estimation value and test data.

5.3 Summary

This section introduced the Process Adjustment Factor (PAF) as a novel multiplicative correction factor, designed to augment the existing defect-limited yield models with explicit process node dependence. By decomposing yield estimation into a statistical component (captured by the underlying model — Poisson, Bose-Einstein, Negative Binomial, etc.) and a geometry-driven process component (captured by PAF), the approach preserves the well-established mathematical foundations of classical yield models while addressing one of their primary shortcomings: insensitivity to process node scaling.

The PAF was constructed from publicly available Design Rule parameters — metal pitch, metal size, via dimensions, metal layer count, and metal layer height — combined with ISO fab-class particle size distributions and empirically estimated particle landing probabilities. A geometric

probability framework was used to compute the probability of a particle of a given size causing a fault on any layer–via combination within a device stack.

Evaluation across six process nodes (250 nm to 16 nm) and six ISO fab classes confirmed the two expected trends: yield-success probability decreased monotonically as feature size shrinks, and decreased as fab cleanliness standards relax. Validation against four 40 nm ASICs demonstrated that PAF improved Bose-Einstein yield estimates in three out of four cases (ASICs A, B, and D), narrowing the gap between predicted and measured Sort Yield. For ASIC C — where the baseline Bose-Einstein model already underestimated the yield — applying PAF widened the discrepancy, highlighting that PAF is most effective when paired with models that tend to overestimate yield, and that its corrective direction is not universally appropriate.

The PAF module has been integrated into the defect-limited yield calculator, where users select a process node and ISO fab class to retrieve the corresponding multiplier (Table 4) and apply it to any compatible yield model. The current formulation, however, is inherently geometry- and particle-centric and does not capture Design Complexity, circuit-type sensitivity, or critical-area effects. A natural extension is to reformulate PAF as a modifier to the Critical Area of the die rather than as an end-stage multiplier and the results are captured in Table 5 showing lower sensitivity to yield changes.

PARAMETRIC-SHIFT LIMITED YIELD MODELING

6 Parametric-Shift Yield: Datasets

6.1 Wafer Acceptance Test (WAT)

Wafer Acceptance Test (also known as E-test or Process Control Monitoring (PCM)) is a wafer-level test performed by the Foundry to check the health of the fabrication process by measuring inline and post passivation electrical parameters. This is done to ensure that the specifications of the devices on a microscopic level (such as transistors, vias etc.) on the wafer lie within acceptable limits of the process technology. WAT testing enables timely detection of process deviations for the fabrication plant based on by monitoring whether electrical parameters fall outside their expected output range. To conserve valuable wafer area, test structures are restricted to strategic locations – typically up to 9 sites – within the scribe lines. This allows for testing for certain parameters which wafer sort testing cannot provide. Test structures can measure various electrical parameters based on functionality available on the structure. These include measuring sheet and contact resistances, measuring capacitances and dielectric integrity, estimating diode junction properties and leakage currents, and transistors properties like drain-source currents, threshold voltages and breakdown voltages. Furthermore, test structures can also be used to measure defectivity. This is accomplished by testing for shorts and opens using various geometrical designs such as those shown in Figure 3 [24]. A snake structure consists of a single metal line which winds like a snake and can detect opens if the line detects a point of very high resistance. A comb structure on the other hand consists of two interwoven metal lines placed in close proximity without touching.. A short can be detected when there is a leakage current between the combs. Note that due to the closely guarded IP nature of semiconductor processing, figures shown here about the testing process are generic in nature.

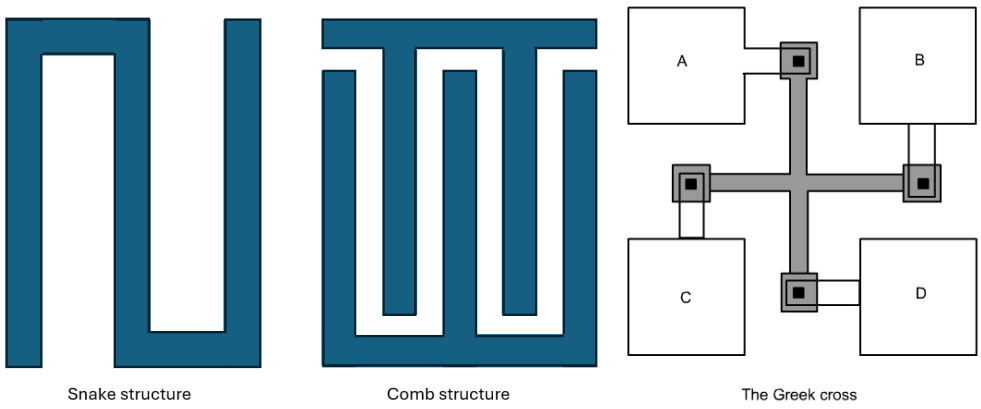


Figure 3: WAT test structures: Snake and Comb (Left) structures are used to detect opens and shorts. Greek cross structure on the right is used to measure sheet resistances of conductive layers.

Since WAT data is measured on scribe lines, data from each WAT coupon has been correlated with its nearest neighbors, to make sure a tighter relationship between process variation and die performance. This data for the case of this study includes up to 4 nearest neighbors. Since this is step is performed at the End of Line in a semiconductor manufacturing process where electrical parameters are measured, WAT parameters will be used to gain more insight into parametric shift since they capture the state of the device that has been manufactured.

6.2 Wafer Sort (WS) Test

Wafer Sort (WS) testing (interchangeably termed as Wafer Functional Test, WFT) is done after receiving the wafers from the fabrication plant after back end of line and before the chip packaging. It is recognized as the key outcome for yield from the fabrication process. The tests are performed with the help of a wafer prober (Teradyne model J750Ex-HD was used in this study) [25]. The wafer prober consists of a probe card which contains a large array of microscopic needles/probers which can measure several dies on a wafer at the same time and runs through a variety of tests defined as a test vector. Multiple wafers can be loaded on to the tester which contains a vacuum

sealed wafer-chunk to keep it in place and can be unloaded for wafer swapping. The tests conducted are considered to be “stop at first fail” which implies tests are stopped the moment a failure is detected and the rest of the tests will not be performed for the die.

The tests range from simple parametric tests like continuity checks to more complex (WPTs like die functionality tests. These tests are grouped into bins based on tested functionality (e.g., memory scan, clock). Test results are stored in STDF files and wafer maps, which use these bin categories to display each die’s pass/fail status. These test results are later used during packaging phase where only dies which have passed all tests will undergo packaging. While WS is used to screen out defective dies on the wafer, ensuring good dies are assembled, Final Test (FT) verifies package integrity, high-frequency performance, and full system functionality. It applies at-speed functional vectors and parametric testing under varied thermal conditions. For this reason, the dataset used in this exercise is solely from WS testing since it is the closest step to WAT tests and also does not include any possible issues resulting from Wafer packaging.

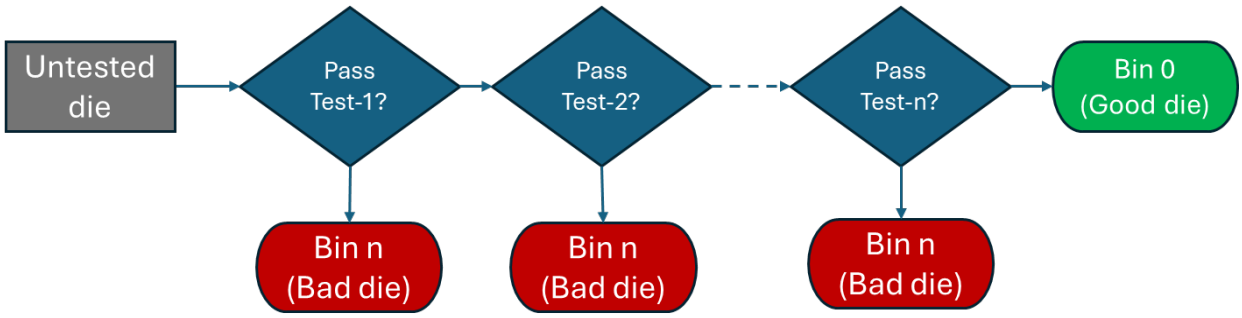


Figure 4: Wafer and die production flowchart [44]

Both WS and WAT tests contain information on spatial relationships which will lend itself well to this analysis. Since WAT measures fundamental physical parameters, it can provide insights into the outputs of WS tests. In order to accomplish this, the two tests’ results are combined based on

the corresponding physical address of each test data. These include the Wafer lot ID, Wafer number, and Die Site IDs. Since WAT is only performed on up to 9 sites and 4 nearest neighbors for each, up to 36 data rows per wafer will be used for the analysis.

7 Parametric-Shift Yield: Data Analysis methodologies

Due to the complex nature of the problem as discussed, two data correlation techniques have been used in the present work: Copulas and Machine Learning (ML) models, to estimate the parametric shifts. WS data contains failure signatures to the corresponding WAT data. Using trained statistical models (such as the ones used for defect-limited yield) can provide a way to generate more instantaneous results which can help decision making process more fluid. This section discusses the two data correlation techniques used to estimate parametric yield. While simulation software like LT SPICE can be used to estimate the performance of devices, these simulations take long times which can be hard to use for decision making.

The proposed model for parametric shift can be estimated with a three-step methodology as shown in Figure 5. This section will discuss the datasets involved (WAT data and WS data) and the methodologies involved in modeling (copulas and machine learning)

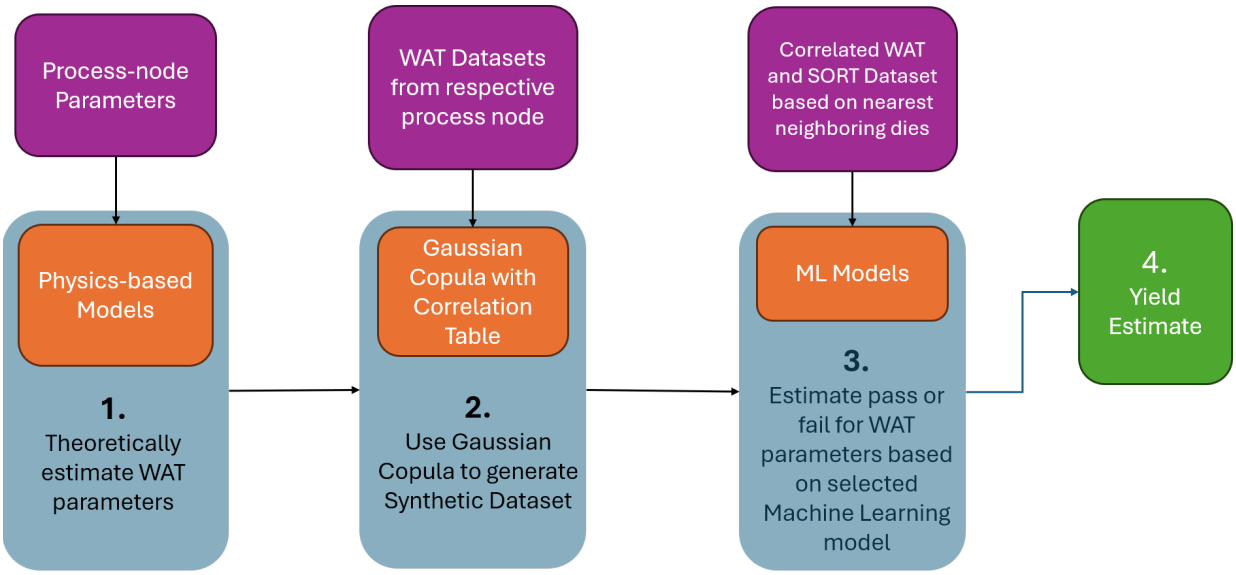


Figure 5: Flowchart for modeling of parametric shift. The methodology in each step is highlighted in blue, the datasets and inputs used to train in the model are highlighted in purple, and orange describes the underlying 'tool' used to process a particular step.

7.1 Estimation of Theoretical WAT Parameters

WAT parameters are evaluated to provide a baseline estimate for feeding into the machine learning model so as to provide the user a way to not rely on foundry data. WAT parameters are estimated using various levels of model complexities with the process node parameters provided by the foundry. Some WAT parameters such as Sheet resistances and Via resistances are estimated with the help of relatively simple equations as they are sufficiently close enough to the real average values of the WAT parameters. The percentage errors of the mean value are compared to the theoretical calculations and are discussed in Section 9.

However, for saturation current and threshold voltage, the estimation required several levels of model complexity. As a baseline the equation 7.1 was used for threshold voltage and 7.2 for saturation current:

$$V_t = V_{FB} + 2\phi_f + \left(\frac{\sqrt{2\epsilon_{Si} q N_A (2\phi_f)}}{C_{ox}} \right) \tag{7.1}$$

$$where\ \phi_f = \frac{kT}{q} \ln\left(\frac{N_A}{n_i}\right), C_{ox} = \frac{\epsilon_{SiO2}}{t_{ox}}, Q_0 = C_{ox}.A_g$$

$$I_{D_{sat}} = \frac{\mu_n C_{ox}}{2} . \frac{W}{L} (V_{GS} - V_t)^2 \tag{7.2}$$

Where k is the Boltzmann constant, T is the temperature (taken as room temperature), q is the charge of an electron, ϕ_f is the fermi potential, V_{FB} is the flat-band voltage, N_A substrate doping, n_i intrinsic carrier concentration, C_{ox} the oxide capacitance per unit area, W is the channel width, L is the channel length, V_{GS} is the gate-source voltage and V_t is the threshold voltage.

While the first step uses the basic model for both threshold voltage and saturation current, several additional phenomenological factors are added as factors in order to estimate all of the WAT parameters. For threshold voltage, the factors considered are metal-semiconductor work function difference, correction factor for parasitic capacitance and adjustment factors for channel length and channel width. For saturation current, updated mobility models [26] [27], maximum drift velocity factor and Vgs factors are considered.

While these factors reduced overall percentage error, the total difference was still too high to be acceptable. Further models were sourced from BSIM. This is a collection of physics-based compact models developed by UC Berkeley for simulating MOSFET in the software SPICE. BSIM models have several levels of complexity starting at level 1 and increasing with higher levels. While higher levels of complexity may provide more accurate results, it often becomes difficult to use the models since it requires several niche factors which can be hard for a user to understand or have the knowledge to input. Hence, the first 4 levels of complexity were considered

and the best results were obtained at level 2. The phenomena accounted for in Level 2 BSIM model are the following [28]:

- Non-uniform charge distribution in depletion layer
- Sub-threshold conduction for gate-source voltages less than threshold voltage
- Carrier mobility variation due to electric fields
- Narrow channel effects
- Non-linear capacitance between gate and source

This form of the threshold voltage comes out as:

$$V_{th} = V_{TH0} + K_1 \left(\sqrt{\phi_s - V_{bseff}} - \sqrt{\phi_s}\right) - K_2\, V_{bseff} - \Delta V_{th,SCE} - \Delta V_{th,DIBL} \tag{7.3}$$

where $V_{TH0}$is the long-channel zero-bias threshold voltage, $\phi_s$the surface potential, and $V_{bseff}$the effective body bias, Vertical non-uniform doping modeled via body-effect coefficients K_1(first-order) and K_2(second-order), or equivalently Short-channel effects & DIBL threshold roll-off through a characteristic length $l_t$and the built-in source/drain voltage V_{bi}.

The level 2 modification for saturation current using A_{bulk} factor comes out as:

$$I_{D_{sat}} = \frac{\mu_{eff} C_{ox}}{2A_{bulk}} . \frac{W}{L} (V_{GS} - V_t)^2 \tag{7.4}$$

With velocity-saturation limit (short channels), defined through $E_{sat} = 2v_{sat}/\mu_{eff}$:

$$I_{Dsat} = W\, v_{sat}\, C_{ox}\, (V_{GS} - V_{th} - A_{bulk} V_{Dsat}) \tag{7.5}$$

$$where\ V_{Dsat} = \frac{E_{sat} L\, (V_{GS} - V_{th})}{E_{sat} L + (V_{GS} - V_{th})},$$

and the Field-dependent mobility being defined as:

$$\mu_{eff} = \frac{\mu_0}{1 + \theta\, (V_{GS} - V_{th})} \tag{7.6}$$

where $\theta$is the mobility-degradation coefficient and $v_{sat} \approx 10^7$cm/s is the saturation velocity.

7.2 Copulas

Copulas are a robust mathematical framework in probability theory and statistics used to evaluate the joint multivariate distribution of correlated parameters. They can be useful for modeling parameters such as threshold voltage and saturation current which cannot be treated as independent variables.

Copulas function by converting individual failure probabilities from a finite probability space to an infinite space which allows for the application of basic mathematical operations such as addition, multiplication, etc. The basis of copulas comes from Sklar's theorem [29] which states that every multivariate cumulative distribution function can be estimated in terms of its marginals and a copula C

$$H(x_1, \dots, x_d) = \Pr[X_1 \leq x_1, \dots, X_d \leq x_d] \tag{7.7}$$

Where H is the conversion function, Pr is the probability of the events occuring and $X_1, X_2, \dots X_d$are marginal continuous distribution functions

It also states that for a given H, copula C is unique on the cartesian product of the ranges of the marginal's Cumulative Distribution Function (CDF). Several copulas can be defined thus, so based on different CDFs such as those in Table 6 [30] [31] [32] [33] [34] [35] [36]. It lists the copula functions C, as well as the range of the correlation variable α. Generally, if α has a value of 0, it refers to a condition of full independence between two variables. This is also listed in Table 6 as the independence condition for copulas and the copula will simply be the product of two variables which is the same as the representation of independent variables through Baye's theorem.

Table 6: A Summary of popularly used copulas showcased for two random variables U and V (s and w in the respective integral expression). α is the correlation parameter of each copula, with a value of 0 indicating independence between the variables.

Copulas	$\boldsymbol{C(u,v,\alpha)}$	$\boldsymbol{\alpha \in \Omega}$
Gauss [30]	$\iint_{\infty}^{u,v} \frac{1}{2\pi\sqrt{1-\alpha^2}} \exp\left(2\alpha sw - s^2 - \frac{w^2}{2(1-\alpha^2)}\right) dsdw$	$\alpha \in [-1,1]$
T [31]	$\iint_{\infty}^{u,v} \frac{1}{2\pi\sqrt{1-\alpha^2}} (1 + \frac{(2\alpha sw - s^2 - w^2)}{2(1-\alpha^2)})\text{^}(-\frac{v+2}{2})\, dsdw$	$\alpha \in (-1,1)$
Clayton [32]	$\max\left([U^{-\alpha} + V^{-\alpha} - 1]^{\frac{1}{\alpha}}, 0\right)$	$\alpha \in (-1,\infty)\backslash\{0\}$
Frank [33]	$-\frac{1}{\alpha}\ln\left(1 + \frac{(e^{-\alpha U}-1)(e^{-\alpha V}-1)}{e^{-\alpha}-1}\right)$	$\alpha \in (-\infty,\infty)\backslash\{0\}$
Gumbel [34]	$\exp\left\{-[(-lnU)^{\alpha} + (-lnV)^{\alpha}]^{\frac{1}{\alpha}}\right\}$	$\alpha \in [1,\infty)$
Ali-Mikhail-Haq [36]	$\frac{uv}{1-\alpha(1-u)(1-v)}$	$\alpha \in [-1,1]$
Independence	uv	$\alpha = 0$

Using copulas can allow us to generate a "Synthetic Dataset" which only needs the mean, standard deviation and correlation parameters to generate the data. The usage of copulas in this manner allows for generation of several WAT parameter "vectors" which can closely mimic real WAT data without the help of a simulation.

7.3 Machine Learning methodologies

Output from WS data provides failure signatures of various failure types. The data contains two types of information: Failure due to defects and failure due to parametric shift. To conduct this evaluation, Sort data is screened to separate tests which are related to parametric shift. This allows us to analyze specifically for parametric-shift failure data by combining with the WAT data. WAT data is combined with the screened sort data by corresponding die wafer addresses (with fields such as lot number, wafer ID, die coordinates etc.). This tasked was done with the help of a script which also converted WS data from a tester output in ".txt" format to an excel table format. Once combined, WAT data is used to predict failure signatures. For a more generalized model, the pass-fail signature is taken from WS data as '1's for pass and '0' s for failures, and data analysis was performed on them. This also deems this problem as a data classification problem.

Machine learning algorithms have shown great performance for classification problems. The popular algorithms that are used in this domain fall into two main categories: Decision Trees and Support Vector Machines (SVMs). Decision tree models are a type of supervised learning approach constructed in the form of "trees". Each tree consists of leaves which represent a class label and branches that represent logical conjunctions of features that lead into the class labels. Trees are built by splitting the feature set into subsets based on classification features by maximizing a target number such as reducing error percentage, entropy/loss reduction etc. Decision trees are generally good at dealing with non-linear relationships. They generally lead to high levels of interpretability of the model as well suggest outputs such as feature importance measures which highlights key parameters influencing shift. SVMs on the other hand are an unsupervised learning approach by using kernels to identify and separate clusters of data. It aims to draw lines to divide a high dimensional feature space by using a kernel trick classifying them

based on the relative distance from the kernel function. SVMs have known to perform well in high dimensional datasets however, they have far lower interpretability and are regarded as “black-box” models. For these reasons, this study is using Tree based methodologies.

While simple decision tree algorithms (Decision tree classifier, Random Forest classifier etc.) were tried, the high number of features led to complex leaf nodes or very low accuracy outputs. Several tree-based methods exist in literature, however, the following three major methods are chosen: Gradient Boosting Classifier (GBC), Extreme Gradient Boosting (XGBoost), and Light Gradient-Boosting Machine (LGBM).

Gradient Boosting classifier works based on building trees and reducing residual error. Multiple trees are constituted with each tree being built based on the amount of error from the previous tree as a label which is defined as a loss function. This process continues for all trees by minimizing the error of the loss function till the difference in errors is small enough. Figure 6 shows the modeling structure of each tree. While there may be several trees being built, due to the relatively smaller size of the dataset this process is not time consuming

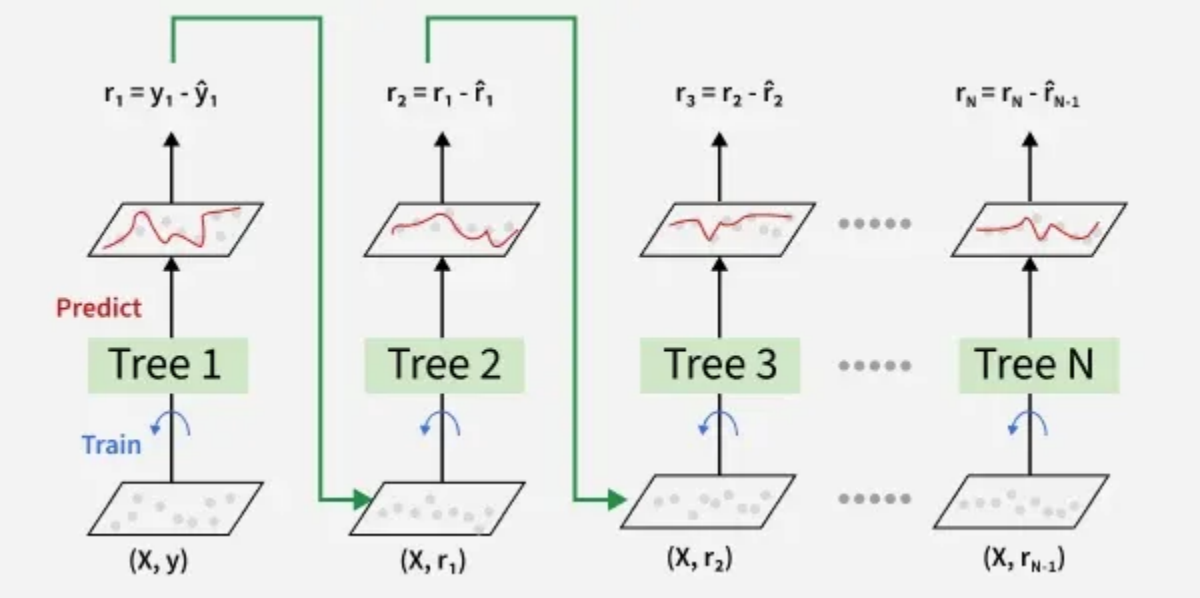


Figure 6: Development of Boosted Decision Trees [37]

XGBoost [37] builds on Gradient Boosting Classifiers by introducing regularization of the data. There are two methodologies used, Lasso regularization (L1) and Ridge regularization (L2). A linear regression is first applied to the dataset, and it is followed by calculating the L1 value as shown in Eq. 7.2:

$$L1 = \lambda(|\beta_1| + |\beta_2| + \cdots |\beta_n|) \tag{7.8}$$

Where λ is the regularization parameter and β_1, β_2 etc. are the linear regression coefficients. The objective function is therefore a twofold function that is similar to GBC, which minimizes the residual error function in addition to minimizing the L1 parameter. The L1 regularization term can shrink the absolute value of the coefficients. If λ is sufficiently large, the coefficients can be shrunk down to zero. This property makes it very useful for feature importance measurements. The optimization algorithm to minimize the loss function is done with coordinate descent.

Ridge regularization is calculated with the sum of squares of the regression coefficients as shown in equation

$$L2 = \beta_1^2 + \beta_2^2 + \cdots \beta_n^2 \tag{7.9}$$

L2 is an additional function that is minimized so as to not allow for overweighting of regression parameters in the loss function. The final loss function in XGBoost is represented by Eq. 7.4, where RSS is the residual sum of squares.

$$Total\ Loss\ Function = RSS + L1 + L2 \tag{7.10}$$

LGBM [38] uses Gradient One Side Sampling (GOSS) to select a subset of data based on the gradient of the loss function. The subset is divided into the largest K gradients and the remaining smaller N-K gradients. For the large gradients, all data points are considered whereas for the smaller gradients, only a few randomly sampled points are considered (this is known as bagging fraction). This lays a higher emphasis on more telling data points and can be used to speed up the model fitting. LGBM also uses Exclusive Feature Bundling which allows for grouping of features that are more mutually exclusive. First features are represented as graph nodes and nodes which are not mutually exclusive, are connected with an edge. Graph coloring is then done such that no two connected nodes are represented by the same color with a goal of minimizing the total number of colors (this is attained by using a greedy graph coloring algorithm). This allows for better feature reduction as well as feature importance calculation. Since all of the discussed algorithms have merits and varying degrees of performance they will all be integrated in the modeling experiment.

8 Parametric-Shift Yield: Scope of model

8.1 Introduction

This section defines the scope, inputs, outputs, and architectural boundaries of the yield estimation framework developed in this dissertation, establishing what the framework is designed to do, what data it requires, and where the limits of its applicability lie before the methodology chapters introduce each component in detail. The framework is organized around two parallel modeling methodologies that converge into a single yield output: A parametric-shift model that estimates yield loss from process-driven parameter variations using a physics-informed, data-driven process that combines BSIM models (Section 9), Gaussian copulas (Section 10), tree-based machine learning classifiers (Section 11), and design & process generalization factors (Section 12). The two models operate on different input data and capture different failure mechanisms, but share a common output format and are combined through a probabilistic OR-gate formulation into a unified Total Yield estimate suitable for use in an engineer-facing ATE workflow, with all components integrated into a single deployable calculator described in Section 13.

8.2 Model Inputs

The parametric shift stream requires WAT and WS datasets, together with a small number of process-level inputs needed for the theoretical WAT estimation stage:

- **Process node parameters** for theoretical WAT estimation — oxide thickness, oxide permittivity, gate work function, junction depth, mobility values, saturation velocity, and operating voltages, calibrated against BSIM Level 2.
- **Real WAT data (optional):** Process node data can estimate WAT parameters theoretically, however, WAT parameters such as threshold voltage, saturation current,

sheet resistance, via resistance, and other electrical parameters can be used to as a direct inputs.

- **Generalization factors (optional):** transistor area density (for Pelgrom's Law) and foundry-reported Cp/CpK values (for process maturity)

8.3 Model Outputs

The parametric shift stream produces a per-process-node yield estimate by running synthetic WAT vectors (generated by the Gaussian copula, optionally modified by Pelgrom's Law and Cp/CpK through trained tree-based ML classifiers and counting predicted passes. Each model (GBC, XGBoost, LGBM) produces its own estimate, and SHAP-based feature importance rankings accompany the yield output to support root-cause analysis.

8.4 Scope Boundaries and Assumptions

To clearly delineate what the framework does and does not address, the following scope boundaries are noted:

- **In scope:** Parametric shift yield from WAT-observable process variations, process-node sensitivity, and a unified Total Yield metric.
- **Out of scope:** Systematic yield (repeatable, non-random defects affecting every die in a wafer region due to CMP, photoresist patterning, etc.), Design Complexity effects beyond transistor density (which would require full circuit simulation), and post-packaging Final Test failures (the parametric pipeline is trained on WS data only to avoid packaging-induced confounders).

- **Independence between the two types of yield:** Defect-limited and parametric shift failures are treated as independent events in the Total Yield formulation. While there can be some common cause failures, the frequency of occurrence is relatively low, and capturing such a failure is impractical due to the difficulty in tracking and obtaining data on such faults and in formulating the probability of a joint occurrence. Thus, this assumption serves as a lower bound to the total yield estimate.
- **Generalization assumption:** The Gaussian copula assumes that the interdependencies between WAT parameters is reasonably similar across the devices within a given process node. Empirical evidence from Section 10 supports this for the devices studied from the correlation tables.

9 Parametric-Shift Yield: Theoretical Estimation of WAT parameter

9.1 Introduction

This section discusses the output of estimation of I_{sat} and V_t parameters used in the theoretical estimation of WAT parameters. This resulted in the usage of BSIM level 2 model for estimating the parameters which will be used as an input to the Copula to generate synthetic dataset (discussed in section 10).

9.2 Error reduction outcome

Several levels of error reduction were applied and the progression of error in the theoretical estimates for 5 different I_{sat} and V_t parameters are summarized in Table 7 and Table 8 (with the help of color coding where more red results indicate higher error percentage) with the final best stopping at level 2. While further levels of modeling may result in more phenomena being accounted for, they require a high degree of knowledge of model dependent parameters to be used with which may not be feasible /convenient. For the first module of the model, BSIM level 2 is hence used to estimate WAT parameters V_t and I_{sat}.

Table 7: Table showing the percentage error of theoretical threshold voltage estimates. Each version incorporates a model for a phenomenon; v1 uses Basic Vt model using only tox as a factor, v2 added adjustment for Metal-Semiconductor work function difference, v3 added correction factor for parasitic capacitance, v4 Readjusted dependency on L and W, v5 added level 1 BSIM model and v6 added level 2 BSIM model, and v8 added estimated doping concentration

	v1	v2	v3	v4	v5	v6	v7
Vt_A	-47%	-40%	-46%	-4%	-40%	-46%	7%
Vt_B	12%	25%	17%	74%	25%	17%	4%
Vt_C	-20%	-11%	-16%	24%	-11%	-16%	-2%
Vt_D	25%	53%	-25%	28%	28%	18%	4%
Vt_E	-46%	-40%	-44%	-17%	-40%	-27%	2%

Table 8: Table showing the percentage error of theoretical saturation current estimates. Each version incorporates a model for a phenomenon; v1 uses the basic equation for transistor drain current, v2 added real Vt for Isat and calculator for mobility value, v3 updated mobility values using mobility models, v4 Introduce max drift velocity factor, v5 updated Vgs values from Teradyne manuals, v6 added level 1 BSIM model, v7 added level 2 BSIM model, and v8 added estimated doping concentration

	v1	v2	v3	v4	v5	v6	v7	v8
Isat_A	-67%	-9%	-68%	-77%	-42%	-79%	-47%	-3%
Isat_B	107%	464%	98%	40%	136%	22%	-13%	3%
Isat_C	101%	448%	92%	35%	168%	22%	26%	-3%
Isat_D	308%	772%	290%	110%	256%	-24%	-46%	-1%
Isat_E	-7%	152%	-12%	-38%	94%	-42%	56%	1%

To accurately estimate transistor behavior and WAT features at the 16nm technology node, the baseline device physics parameters were rigorously calibrated against industry-standard predictive models and commercial foundry disclosures. Because traditional bulk planar CMOS approximations break down at deeply scaled sub-20nm geometries, the physical inputs were updated to reflect High-K / Metal Gate non-planar (FinFET) architectures.

The transition to the 16nm regime necessitates the modeling of advanced gate stacks. The dielectric constant ϵ_{ox} was updated to 22.0 to represent the integration of Hafnium Oxide (HfO_2) replacing traditional SiO_2, yielding the C_{ox}required for electrostatic channel control. Simultaneously, the gate work function $\phi_m$was adjusted to 4.5 eV, reflecting the use of Titanium/Tantalum Nitride (TiN/TaN) metal gates rather than heavily doped polysilicon. Operating voltages ($V_{ds} and\ V_{gs}$were scaled down to 0.70 V, directly aligning with low-power 16nm logic specifications.

To mitigate aggressive short-channel effects such as Drain-Induced Barrier Lowering (DIBL) and punch-through leakage, the source / drain extension junction depth was constrained to 15 nm. Furthermore, carrier transport models were modified to account for the physical realities of nanometer-scale conduction. Due to extreme surface roughness, phonon scattering, and heavy channel doping, the effective low-field mobility was degraded from ideal bulk silicon values to realistic FinFET estimates of 400 $cm^2/V.s$ for electrons and 150 $cm^2/V.s$ for holes, while the saturation velocity remained capped near the physical limit of 10^7 cm/s.

The integration of these calibrated parameters yields theoretical threshold voltages $V_t \approx 0.35V - 0.4V$ and saturation currents that strictly adhere to the academic benchmarks established by the ASU Predictive Technology Model (PTM) and the Berkeley Short-channel IGFET Model (BSIM-

CMG). Consequently, the derived formulas and subsequent loop-simulated WAT features serve as a highly accurate approximation of commercial 16nm foundry profiles (e.g., TSMC 16FF/16FF+). ASU models also have several calculations for smaller technology nodes (14nm, 10nm, and 7nm) and can be used in the calculator for smaller nodes.

9.3 Summary

This section established a physics-informed framework for theoretically estimating WAT parameters directly from foundry-provided process node data, with the goal of generating reference values that can anchor the synthetic datasets, to be used later in the parametric shift modeling pipeline. The approach progresses through escalating levels of model fidelity — beginning with first-order analytical expressions and culminating in BSIM-level compact models — selected so that accuracy gains are balanced against the input-knowledge burden placed on the end user of the yield calculator.

For simpler WAT parameters such as sheet resistances and via resistances, straightforward closed-form geometric expressions were shown to produce estimates within acceptable error of the measured WAT means, requiring no further refinement. For threshold voltage and saturation current, however, basic textbook formulas produced unacceptably large errors, motivating an iterative augmentation strategy. Successive model versions incorporated metal–semiconductor work function differences, parasitic capacitance corrections, channel length / width adjustments, updated mobility models, maximum drift velocity limits, and Vgs corrections drawn from Teradyne manuals. While these phenomenological additions reduced the errors meaningfully, residual discrepancies remained too large for downstream use.

The decisive accuracy improvement came from adopting BSIM compact models, which were evaluated at increasing levels of complexity. Among the first four BSIM levels, Level 2 was selected as the optimal trade-off: it captures the dominant short-channel and narrow-channel phenomena — non-uniform depletion-layer charge distribution, sub-threshold conduction, field-dependent mobility, narrow-channel effects, and non-linear gate–source capacitance — while still relying on parameters that are tractable for a non-expert user. Higher BSIM levels were rejected not on accuracy grounds but on usability grounds, as they demand specialized model-fitting expertise that conflicts with the calculator's intended audience.

Collectively, these results provide the theoretical WAT baseline that feeds the Gaussian copula stage in Section 10 and the machine learning pipeline in Section 11. The ASU PTM library also covers 14 nm, 10 nm, and 7 nm nodes, providing a natural path for extending the calculator to more advanced technology nodes without redeveloping the underlying methodology.

10 Parametric-Shift Yield: Copula Training Results

10.1 Introduction

Having established theoretical baselines for individual WAT parameters in Section 9, the next step toward a usable parametric shift model is to capture the statistical structure that links these parameters together. WAT measurements such as threshold voltage, saturation current, and sheet resistance are not independent quantities — they share underlying process origins (oxide thickness variation, doping fluctuations, lithographic offsets) that induce systematic correlations across the parameter set.

10.2 Gaussian copula results and outputs

In this study, the Gaussian copula was used to represent the WAT parameters because WAT parameters closely follow a normal distribution which uses the Gaussian function. The “copulas” library was used in python to estimate the correlation value between all WAT parameters. The WAT datasets used comprised of two devices from the 40nm node and one device from the 16nm node. The results of the fit were used to generate the synthetic dataset of WAT parameters. The full copula correlation variables for the ASICs can be seen in Appendix A.

The correlation structures obtained from the Opus and Atlantis datasets exhibit a high degree of similarity across the majority of process and electrical parameters. Visual inspection of the heatmaps shows that most matrix elements remain in the green-to-yellow region, indicating strong agreement in both correlation magnitude and direction between the two datasets. The sign-match heatmap further confirms that the dominant positive and negative relationships are preserved across the two populations, demonstrating that the underlying dependency structure is largely unchanged.

Quantitatively, several parameters show very high trend similarity between ASIC A and ASIC C. These parameters maintain nearly identical correlation patterns with the remainder of the technology monitor structures, indicating that the major process interactions are consistently represented in both datasets.

Table 9: Heatmap showing correlation similarities between ASIC A and ASIC C for several WAT parameters. Greener colors indicate a higher level of agreement, whereas yellower colors indicate lower levels of agreement

ilarity (1	Vtl_N4H	Isat_N4H	Vtl_P3H	Isat_P3H	Rs_NW	Vtl_1N3	Isat_1N3	Vtl_1P3	Isat_1P3	Vbd_1PW	Vbd_1NW	Vtl_N4	Isat_N4	Vbd_PW	Vtl_P3	Isat_P3	Vbd_NW
Rc_V3	0.897	0.937	0.978	0.987	0.883	0.884	0.757	0.883	0.862	0.983	0.980	0.901	0.859	0.903	0.933	0.972	0.956
Rc_V4	0.939	0.996	0.940	0.996	0.979	0.941	0.925	0.958	0.995	0.971	0.828	0.928	0.937	0.972	0.868	0.892	0.942
Rc_V5	0.928	0.998	0.950	0.861	0.946	0.745	0.927	0.751	0.977	0.941	0.825	0.914	0.989	0.966	0.858	0.790	0.946
Rc_V6	0.977	0.990	0.982	0.920	0.948	0.961	0.989	0.977	0.862	0.944	0.992	0.992	0.939	0.957	0.920	0.862	0.986
Vtl_N4L	0.987	0.901	0.981	0.917	0.963	0.965	0.932	0.969	0.959	0.898	0.922	0.993	0.995	0.978	0.986	0.948	0.959
Isat_N4L	0.992	0.936	0.972	0.969	0.924	0.967	0.959	0.936	0.958	0.882	0.918	0.969	0.941	0.975	0.910	0.877	0.993
Vtl_P3L	0.978	0.957	0.900	0.950	0.943	0.905	0.969	0.975	0.979	0.984	0.931	0.976	0.930	0.948	0.951	0.989	0.965
Isat_P3L	0.981	0.972	0.912	0.982	0.939	0.928	0.992	0.973	0.980	0.998	0.921	0.965	0.969	0.919	0.978	0.995	0.988
Vtl_N4H	1.000	0.987	0.980	0.946	0.980	0.904	0.993	0.944	0.955	0.931	0.964	0.980	0.953	0.925	0.997	0.978	0.997
Isat_N4H	0.987	1.000	0.889	0.884	0.935	0.927	0.940	0.982	0.985	0.943	0.910	0.992	0.971	0.960	0.971	0.960	0.923
Vtl_P3H	0.980	0.889	1.000	0.978	0.985	0.972	0.986	0.962	0.961	0.990	0.996	0.885	0.960	0.990	0.961	0.916	0.923
Isat_P3H	0.946	0.884	0.978	1.000	0.956	0.957	0.982	0.969	0.886	0.983	0.967	0.884	0.979	0.969	0.993	0.989	0.938
Rs_NW	0.980	0.935	0.985	0.956	1.000	0.981	0.983	0.922	0.887	0.949	0.995	0.950	0.921	0.952	0.995	0.973	0.982
Vtl_1N3	0.904	0.927	0.972	0.957	0.981	1.000	0.963	0.940	0.839	0.997	0.885	0.945	0.946	0.963	0.972	0.917	0.976
Isat_1N3	0.993	0.940	0.986	0.982	0.983	0.963	1.000	0.975	0.974	0.998	0.981	0.943	0.944	0.921	0.983	0.964	0.918
Vtl_1P3	0.944	0.982	0.962	0.969	0.922	0.940	0.975	1.000	0.760	0.936	0.888	0.951	0.947	0.958	0.953	0.960	0.914
Isat_1P3	0.955	0.985	0.961	0.886	0.887	0.839	0.974	0.760	1.000	0.894	0.851	0.990	0.996	0.988	0.989	0.952	0.971
Vbd_1PW	0.931	0.943	0.990	0.983	0.949	0.997	0.998	0.936	0.894	1.000	0.968	0.971	0.968	0.912	0.987	0.984	0.953
Vbd_1NW	0.964	0.910	0.996	0.967	0.995	0.885	0.981	0.888	0.851	0.968	1.000	0.981	0.979	0.984	0.965	0.952	0.942
Vtl_N4	0.980	0.992	0.885	0.884	0.950	0.945	0.943	0.951	0.990	0.971	0.981	1.000	0.964	0.978	0.930	0.985	0.979
Isat_N4	0.953	0.971	0.960	0.979	0.921	0.946	0.944	0.947	0.996	0.968	0.979	0.964	1.000	0.970	0.971	0.987	0.988
Vbd_PW	0.925	0.960	0.990	0.969	0.952	0.963	0.921	0.958	0.988	0.912	0.984	0.978	0.970	1.000	0.973	0.990	0.952
Vtl_P3	0.997	0.971	0.961	0.993	0.995	0.972	0.983	0.953	0.989	0.987	0.965	0.930	0.971	0.973	1.000	0.990	0.989
Isat_P3	0.978	0.960	0.916	0.989	0.973	0.917	0.964	0.960	0.952	0.984	0.952	0.985	0.987	0.990	0.990	1.000	0.963
Vbd_NW	0.997	0.923	0.923	0.938	0.982	0.976	0.918	0.914	0.971	0.953	0.942	0.979	0.988	0.952	0.989	0.963	1.000

.

The heatmap of absolute differences reveals that discrepancies are generally localized rather than systematic. Most differences appear as isolated cells rather than large correlated regions. If the copula models were fundamentally different between the two datasets, broad blocks of red regions would be expected across the matrix. Instead, the observed pattern suggests that the datasets share the same dominant dependence structure while exhibiting only moderate variations in correlation strength. The overall average absolute difference between ASIC A and ASIC C correlations came out to be 0.157.

As mentioned in section 7.1, the copula outputs were used to generate Synthetic data which are used to mimic WAT data and provide the user a way to get data without relying on the foundry. The distributions of the generated synthetic data were compared to the WAT parameters to check

for closeness. The results were found to be closely matching to each other as can be visualized in the example Figure 7 for I_{sat_B} (WAT parameter denoting saturation current) and R_{s_A} (WAT parameter denoting Sheet resistance). More copulas results are available in Appendix

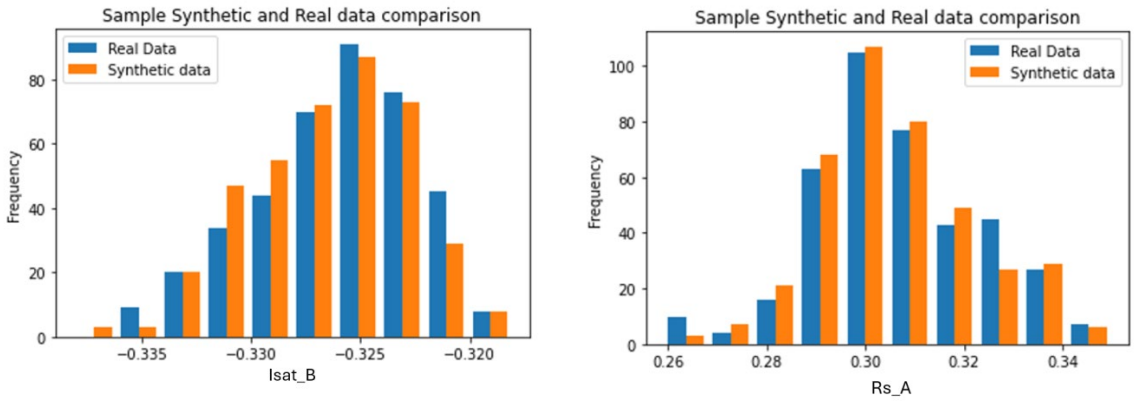


Figure 7: Example comparison of Real Data and generated Synthetic Data for two WAT parameters saturation current (left) and sheet resistance (right)

Kolmogorov-Smirnov testing was performed to compare the measured WAT distributions against copula-generated synthetic distributions. Of the 59 analyzed parameters, 54 (91.5%) exhibited p-values greater than 0.05, indicating no statistically significant difference between the real and synthetic distributions. Threshold voltage, source/drain resistance, contact resistance, and saturation current parameters demonstrated particularly strong agreement, with normalized distances frequently below 0.10. Wasserstein distance analysis also indicated strong agreement between the measured WAT parameters and the copula-generated synthetic distributions. Most parameters exhibited normalized Wasserstein distances below 0.20, corresponding to less than 20% of one standard deviation from the measured distributions. These results suggest that the Gaussian copula successfully reproduces the central characteristics of the majority of WAT distributions while exhibiting reduced accuracy for parameters with stronger tail behavior.

10.3 Summary

This section demonstrated that Gaussian copulas provide a faithful and computationally lightweight representation of the joint statistical structure of WAT parameters across multiple devices and process nodes. Fitting the Python copulas library to real WAT datasets from two 40 nm devices and one 16 nm device produced correlation matrices that remained largely consistent across devices, indicating that the inter-parameter dependencies captured by the copula reflect underlying process physics rather than device-specific artifacts and therefore generalize well across the technology nodes. Synthetic WAT vectors generated from the fitted copulas were validated against measured data and showed close agreement in both marginal distribution shape and joint behavior — illustrated for representative parameters such as saturation current (I_{sat_B}) and sheet resistance (R_{s_A}) — confirming that the copula-generated synthetic dataset is a viable stand-in for real WAT measurements. This validated synthetic dataset forms the statistically grounded input layer consumed by the machine learning models in Section 11.

11 Parametric-Shift Yield: Machine Learning Training Results

11.1 Introduction

With a validated synthetic WAT dataset now available from the copula framework in Section 10, the parametric shift estimation problem reduces to a supervised classification task: given a vector of WAT parameters for a die, predict whether that die will pass or fail WS testing due to parametric shift mechanisms. First, since the industry limits imposed on the viability of WAT parameters were very large, 3σ (3 times the standard deviation) limits were imposed on the WAT parameters to ensure they lie within reasonable limits. Next, this section applies tree-based machine learning models — Gradient Boosting Classifier (GBC), XGBoost, and LGBM — to the combined WAT and WS datasets from two 40 nm devices and one 16 nm device, with failure data pre-filtered to isolate parametric shift signatures from defect-limited and leakage-related failures. Two distinct training strategies are explored: A) Single-device models, which maximize accuracy on a specific product, and B) Mixed-device models spanning to multiple devices and process nodes. Beyond raw accuracy, this section evaluates yield estimates produced when synthetic WAT vectors are fed into the trained models — providing a direct, simulation-free pathway from process parameters to parametric shift yield — and uses Shapley Additive Explanations (SHAP) as a model-agnostic feature to compare which WAT parameters most strongly drive each model's predictions, supporting both model interpretability and downstream root-cause analysis.

11.2 Machine Learning results

Machine learning was performed on combined WAT and WS datasets for 2 devices from 40nm node and 1 device from the 16nm node. Unlike copulas however, ML was also done on mixed datasets. This involves mixing data from same process node but different devices (40 nm in this

case) and mixing data from different process nodes. The objective is to observe differences in how the ML model trains but also in how a more generalized model can be achieved for across process node estimation. Note that in order to estimate the variation in process nodes, two types of analysis were done since there are differences in device types present as well as number of devices (more transistor types for example). The first one is by using only common WAT parameters between the two devices, and the second is by using all WAT parameters but leaving the fields empty. The second methodology is a common methodology to use with ML models like XGBoost since it is good at isolating such cases.

Results are shown in Table 10 for the different ML models, mainly the accuracy in the testing set, as well as a comparison in the yield numbers when using a synthetic dataset. The results for GBC, XGBoost and LGBM are seen to be very good, boasting an over 90% accuracy result for all models. The corresponding estimate of yield using synthetic data is also seen to be close to the real estimate of yield, further solidifying the performance of these models. The results for mixed datasets also show high accuracy, albeit slightly lower for the 40nm and 16nm process node mixed dataset.

Table 10: Best yield estimates from different machine learning model, along with accuracy of each model during training.

Model	Accuracy	Predicted Yield estimate		
		ASIC A (40 nm)	ASIC B (40 nm)	ASIC C (16 nm)
Actual value	NA	72%	68%	71%
LGBM	91%	76%	70%	71%
GBC	95%	76%	60%	62%
XGB	93%	75%	71%	69%

Other statistics such as precision, recall and f-1 score are tabulated in Table 11 with all 3 metrics showing good agreement in terms of the prediction for each class.

Table 11: Prediction statistics for the different ML models

Model	Accuracy	Precision	Recall	F1
GBC	90.6%	0.90	0.89	0.89
LGBM	90.1%	0.94	0.897	0.915
XGB	90.06%	0.92	0.898	0.907

Since different process nodes have different WAT parameters due to the difference in components, mixing datasets from 40 nm and 16 nm nodes was done by merging the two datasets with their common features. This combining of datasets maintains the important WAT parameters such as Vt, Isat Rs, Con etc. and hence can be used to create a more holistic yield model. Unlike the individual dataset models, however, GBC ended up with a significantly worse accuracy compared to LGBM and XGB which maintained an accuracy of over 90%. While this combined model may help predict yield for multiple process nodes (and predict yields for unknown process nodes), a process node specific trained model should be used to obtain a higher degree of accuracy.

Table 12: Predictions statistics for ML models trained on combined datasets

Model	Accuracy	Precision	Recall	F1
GBC	39.6%	0.228	0.223	0.198
LGBM	92.5%	0.953	0.877	0.910
XGB	92.1%	0.942	0.874	0.903

Feature importance measurements were done for all ML models. However, due to differences in model structures and algorithms, comparing the feature importance results may be difficult. Shapley Additive Explanations (SHAP) which is a model agnostic feature importance algorithm was used to conduct feature importance analysis to compare the models. SHAP works by estimating the importance observed based on the sensitivity associated with inputting each individual data point into the model. An example of the output for SHAP is provided in Table 13

to show the results as well as a comparison between different models. As can be seen the average number for the result is high across the board for all of the models which allows for the feature to be considered highly important. The feature importance metric is also used as a justification for the usage of common WAT parameters between 40nm and 16nm node since the common WAT parameters also are some of the most important WAT parameters. This can explain the relatively high accuracy of the results even when fitting a mixed dataset.

Table 13: Samples (15 out of 42 WAT parameters) of Averaged SHAP values for XGB and LGBM for each WAT parameter. SHAP provides majorly consistent outputs determining feature importance. Higer values indicate higher importances

WAT parame	XGB	LGBM
Metal Res A	0.022034	0.0517653
Metal Res B	0.53667	0.467722
Metal Res C	0.010739	0.0165637
Metal Res D	0.252637	0.164093
Metal Res E	0.034808	0.0117627
Metal Res F	0.059144	0.055946
Metal Res G	0.034225	0
Via_res A	0.111382	0.089143
Via_res B	0.797429	0.660585
Via_res C	0	0
Via_res D	0.096584	0.0624017
Via_res E	0.183251	0.0612711
Via_res F	0.019485	0.0208452

11.3 Summary

This section demonstrated that tree-based ensemble models can accurately predict parametric shift-driven WS outcomes directly from WAT parameter vectors, providing a fast, data-driven alternative to circuit-level simulation. All three classifiers — GBC, XGBoost, and LGBM — exceeded 90% test accuracy on single-device datasets, and the yield estimates obtained by feeding copula-generated synthetic WAT vectors into the trained models closely tracked measured Sort

Yield, validating the end-to-end pipeline from process parameters to yield prediction. Mixed-device training across two 40 nm devices, and across the 40 nm and 16 nm nodes, retained strong accuracy (with a modest degradation in the cross-node case), indicating that a single generalized model is feasible when training is restricted to WAT parameters common across devices or when XGBoost's native handling of missing fields is exploited. SHAP analysis produced consistent feature-importance rankings across models, with the most influential WAT parameters being shared between the 40 nm and 16 nm common-feature set — an empirical justification both for the cross-node training strategy and for using the ranked parameters to guide future root-cause analysis of parametric shift failures. Of the three algorithms, XGBoost and LGBM emerged as the preferred choices for the calculator due to their higher accuracy, better robustness to input variability, and stronger interpretability through SHAP, with GBC retained as a baseline reference. These trained models, together with the synthetic dataset generator from Section 10, constitute the parametric shift yield engine that feeds the combined yield calculator in Section 13

12 Parametric-Shift Yield: Generalization Methodologies

12.1 Introduction

The machine learning models developed in Section 11 demonstrated strong predictive accuracy on the devices they were trained on, but a truly generalizable parametric shift estimator must also respond to the two factors that drive variation between devices and between fabs: design-side variation and process-side variation. Capturing these effects through full circuit simulation would reintroduce the very runtime cost that motivated the data-driven approach in the first place, so this section instead introduces two compact, physically grounded modifiers that act directly on the synthetic WAT dataset generated by the Gaussian copula. The first, Pelgrom's Law, provides a closed-form link between transistor area density and the variance of threshold voltage and saturation current, allowing the copula's standard deviations to be rescaled when transistor geometry changes. The second, the Cp/CpK process capability index, is a widely reported foundry metric that quantifies how tightly a process holds a parameter within its specification limits and can be used to compress or widen the copula-generated spread to reflect process maturity. Together, these two modifiers extend the trained machine learning pipeline from a device-specific estimator into a framework that can interrogate the expected yield impact of design choices and process drift without retraining the underlying models.

12.2 Pelgrom's Law

To improve model generalization, a few design and process node factors are considered. While circuit complexity is hard to quantify without actual circuit simulation (which defeats the purpose of the model), the transistor density has a known relationship with V_t and I_{sat} with the help of Pelgrom's law [39]. Pelgrom's Law states that the variation in transistor's threshold voltage, V_t is

inversely proportional to the square root area factor of the transistor, or the transistors area density as shown in Eq. 12.1.

$$\sigma_{VT}^2 = \frac{A_{VT}^2}{2WL}, A_{VT} = \frac{C\left(\sqrt[4]{q^3 \epsilon_{si} \phi_B}\right) T_{ox}}{\epsilon_{ox}} \tag{12.1}$$

q is the electron charge, $\epsilon_{si}$is silicon permittivity, ϕ_Bs is the work function of the substrate, $T_{ox}$is the oxide thickness, $\epsilon_{ox}$is the oxide layer permittivity, L and W are the transistor density parameters (can be replaced with area density if needed), and C is a correction constant (around 0.15).

The density factor of the transistor is also visualized in Figure 8 which shows the amount of variation for different transistor densities. This can be input along with the copula to modify the synthetic dataset by showing a higher variance in the corresponding WAT parameters.

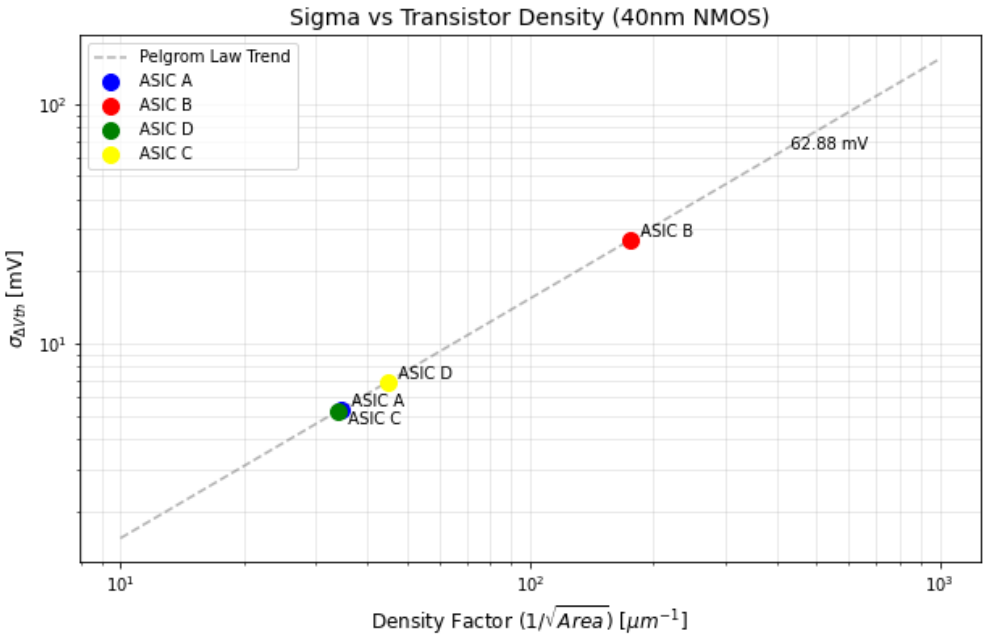


Figure 8: Variation of threshold voltage with transistor density based on Pelgrom's Law

12.3 Cp/CpK metrics for process parameter variability

A second factor being considered is the popular industrial estimate called Cp or CpK values. Cp/CpK values are statistical indices that measure the amount of spread observed in the process vs the estimated allowable limits for a parameter. Cp/CpK data is often something that is available from foundries as it is a metric which shows process maturity. Cp/CpK is measured as:

$$C_p = \frac{USL-LSL}{6\sigma} \; and \; C_{pk} = \min\left[\frac{USL-\mu}{3\sigma}, \frac{\mu-LSL}{3\sigma}\right] \tag{12.2}$$

Where μ and σ are the average and standard deviation of the samples respectively; USL and LSL stand for the Upper Specification Limit and Lower Specification Limit respectively. These limits are often specified based on the performance limits of an electrical parameter that is required for the device to run successfully. The higher the Cp value, the more mature is the process since there is lower variation (lower σ). Figure 9 shows this difference for a threshold voltage measurement for different Cp values.

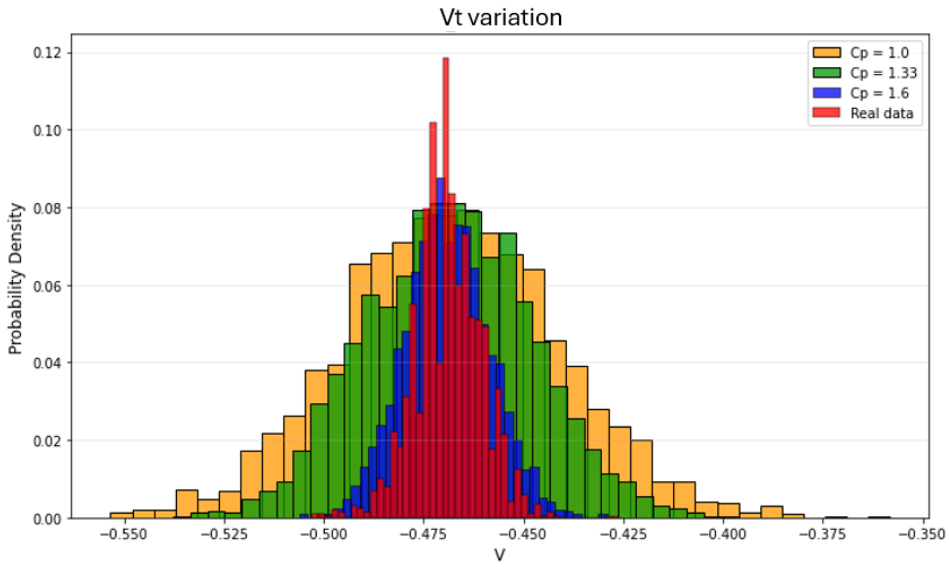


Figure 9: Variation of data spread based on Cp/CpK for threshold voltage

This updated synthetic dataset can now be used in conjunction with the ML models to reflect the difference in estimated yield. An example case for this is shown in Table 14 for the 16nm process node when using Cp values from June 2020 is compared to the current process node. This analysis highlights an issue in using GBC ML model where it shows higher sensitivity when compared to GBC and XGBoost. However, all three models still show the lower yield estimate which corresponds to the lower Cp values (or a less mature process).

Table 14: Estimated yield based on difference in process maturity (Cp/CpK index).

ML model	Yield Estimate from 2025 WAT+SORT data	Yield estimate from 2020 WAT Cp/CpK/CpU
XGB	67.0%	62.2%
GBC	67.6%	51.7%
LGBM	68.0%	62.2%

12.4 Summary

This section showed that two compact, physically motivated factors — Pelgrom's Law for design-side variation and Cp/CpK for process-side variation — are sufficient to extend the parametric shift pipeline beyond the specific devices and process conditions on which it was trained. Pelgrom's Law links the variance of threshold voltage (and, by extension, saturation current) to transistor area density through a closed-form relationship involving oxide thickness, oxide and silicon permittivities, substrate work function, and a small calibration constant, allowing the copula's standard deviations to be rescaled in a principled way whenever transistor geometry changes; the resulting variation profile (Figure 9) directly modifies the synthetic WAT dataset before it enters the trained ML models. Cp/CpK provides the complementary process-side lever: by tying the spread of WAT parameters to the foundry-reported ratio of specification width to process sigma, the framework can represent process maturity directly. The case study comparing 16 nm Cp values from June 2020 to the current node showed that all three ML models correctly predicted a lower yield for the less mature process, with XGBoost and LGBM responding smoothly while GBC showed disproportionate sensitivity to the Cp/CpK adjustment — reinforcing the model-selection guidance established in Section 11. Together, these generalization factors complete the parametric shift estimator: the copula supplies the joint structure, the ML models supply the WAT-to-yield mapping, and Pelgrom's Law and Cp/CpK supply the design- and process-aware modulation needed for the framework to estimate yield under conditions that differ from the training data, providing the final input to the combined yield calculator discussed in Section 13.

YIELD CALCULATOR AND COMBINED YIELD METRIC

13 Yield Calculator and Combined Yield Metric

13.1 Introduction

This research uses an Excel-based calculator, with 15 models to estimate defect-based yield for ASICs. It can estimate defect-based yield for 5 circuit types: IO; Logic; Memory; Pure Analog; and High-speed digital. While this calculator estimates defect-limited yield only, and not parametric shift-related yield; it allows for a percentage multiplier provided by the user. It also allows for the yield to be adjusted based on the Design complexity; and the Process node.

The parametric shift aspect of the calculator will be added to the calculator in the future. machine learning models and copulas are trained and estimated in Python and Excel integration for machine learning models. Copulas require complex coding. However, currently, the calculator allows for the usage of preset inputs to be used for the parametric shift module and in the future will allow the user to fine tune inputs based on process nodes.

13.2 Defect Yield module

To better understand the variation in yield, a calculator was developed with the following capabilities (as shown in Figure10):

- Estimate yield based on 15 different models.
- Estimate yield constants based on either WS yield data.
- Estimate yield based on different circuit types based on the area of the circuit type.
- Summarize and compare yields from different models.
- Adjust yield estimation with PAF multipliers as shown in Eqs. (5.1) and (5.2).

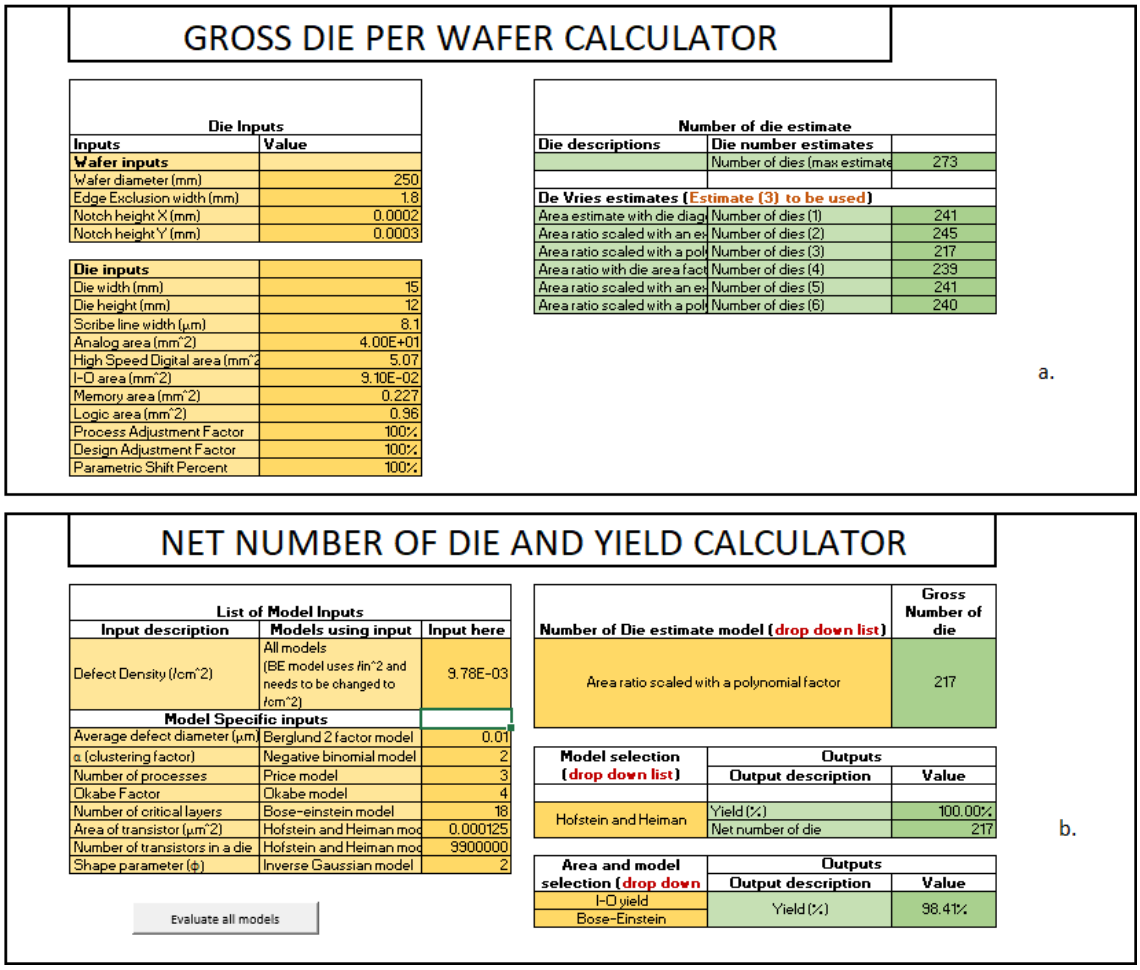


Figure10: Yield calculator with a. die and wafer-based inputs, b. model-based inputs and yield estimates with individual die area yield breakdowns.

De-Vries equations [40] are used to estimate gross number of dies on a wafer based on geometric factors including Wafer diameter (d), Edge Exclusion width (e_x), Notch height (H_n) and width (W_n), Die height (H_d) and width (W_d), scribe line width (W_{sl}), and ASIC area allocations based on circuit types. Eq. 10.1 calculates the maximum estimate of gross dies for the wafer (n_{max}) and Table 15 summarizes the De-Vries equations based on n_{max}value. These equations are derived from various geometrical estimates as well as analytical results based on different approximations described in the table. The area ratio scaled with polynomial factor (n_3) gave the best results.

$$n_{max} = \frac{\pi\left(\frac{d^2}{2} - e_x\right) - A_{io} - A_{mem}}{A_{die}} \tag{13.1}$$

Table 15: De-Vries Equations for Gross-die estimates

Estimate description	Equation
Area estimated with die diagonal length	$n_1 = \frac{n_{max}}{A_{die}} - \frac{\pi(d - 2e_x)}{\sqrt{\left(W_d + \frac{W_{sl}}{2}\right)^2 + \left(H_d + \frac{W_{sl}}{2}\right)^2}}$
Area ratio scaled with an exponential factor	$n_2 = n_{max} e^{-\frac{2\sqrt{A_{die}}}{d}}$
Area ratio scaled with a polynomial factor	$n_3 = n_{max}\left(1 - \frac{2\sqrt{A_{die}}}{d}\right)^2$
Area ratio with die area factor (Analytical)	$n_4 = n_{max} - \frac{0.58\pi d}{\sqrt{A_{die}}}$
Area ratio scaled with an exponential factor (Analytical)	$n_5 = n_{max} e^{-\frac{2.32\sqrt{A_{die}}}{d}}$
Area ratio scaled with a polynomial factor (Analytical)	$n_6 = n_{max}\left(1 - \frac{1.6\sqrt{A_{die}}}{d}\right)^2$

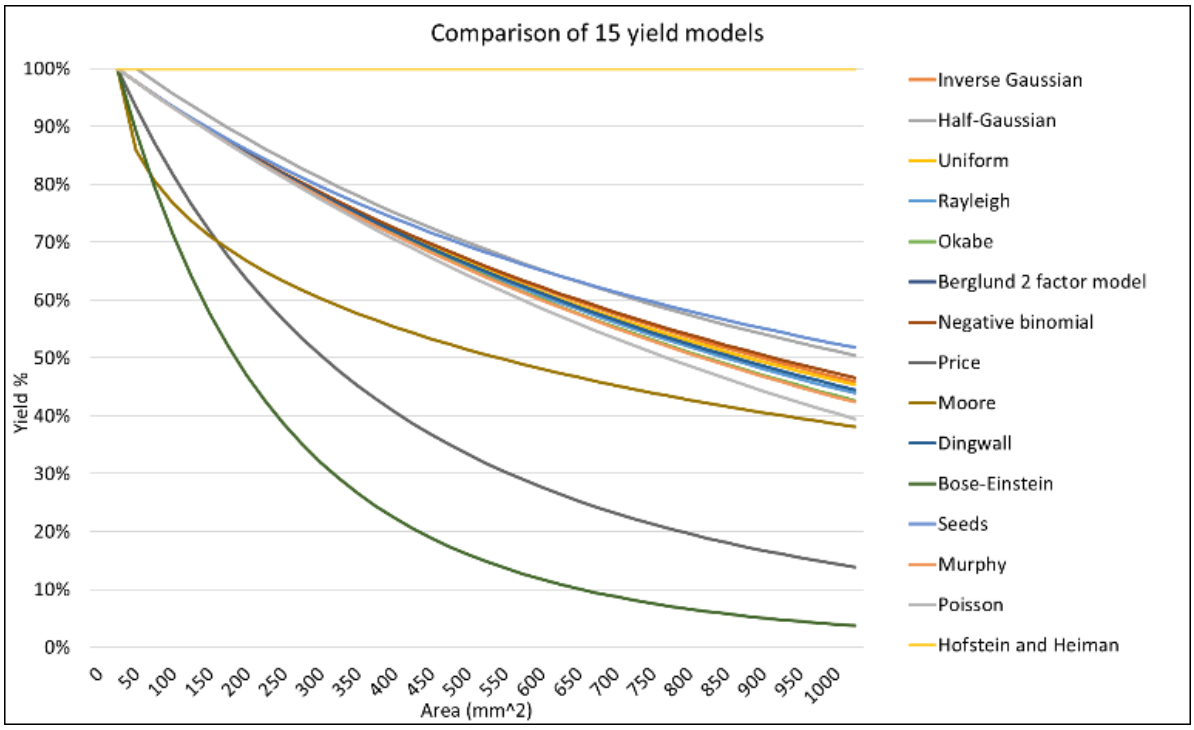
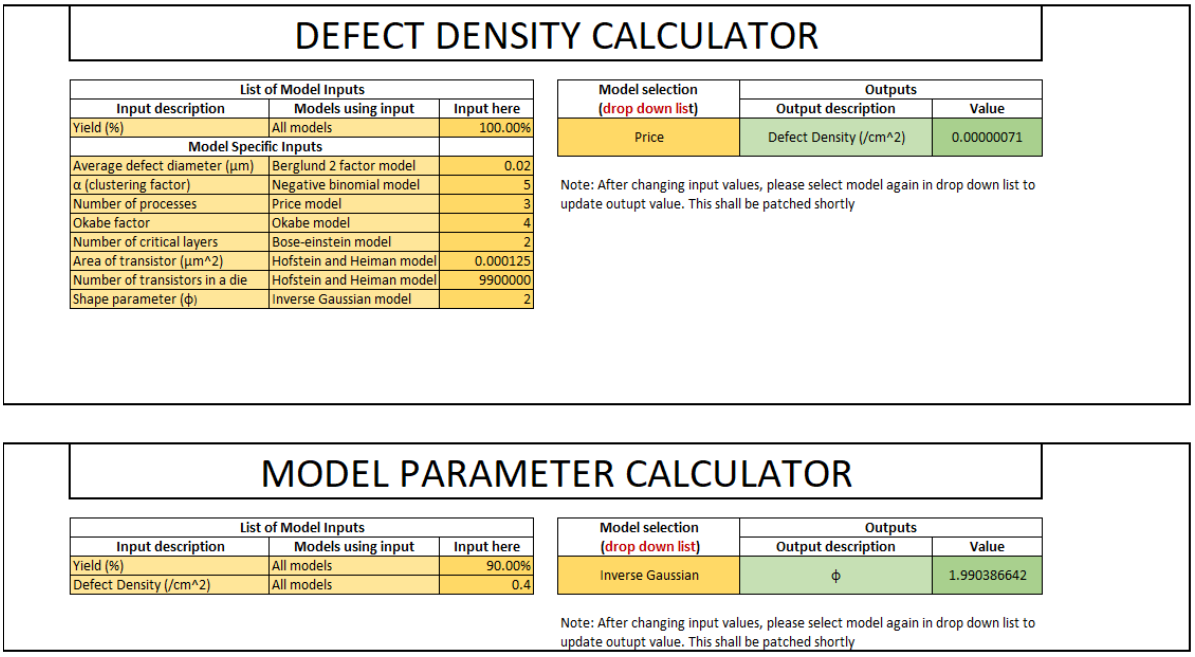


Figure11: Comparison of different models when varied by area and keeping defect density constant.

The calculator allows the user to adopt specific models for calculating yield based on the available parameters and constants. Figure11 provides a comparison of different models at constant defect density for different die sizes. Given that yield estimates from various models converge with an acceptable level of variability, it becomes possible to back calculate other process-related parameters that are not readily available (such as average defect diameter, clustering factor and number of critical layers). This can be particularly useful for predicting the behavior of dies with different areas or comparing the fab data with the user's own findings.

The calculator also provides the options to estimate the D_0 value, based on product yield. This allows users to compare estimated D_0 against fab's reported D_0. Both back calculation modules (as shown in Figure 12) use two main techniques for estimation based on the complexity of the model. A simple model like Poisson has a straightforward equation to estimate D_0. However, other models

like inverse-gaussian require the use of numerical solver tools (such as the one found in excel) for estimation of the model parameter.



DEFECT DENSITY CALCULATOR

List of Model Inputs		
Input description	Models using input	Input here
Yield (%)	All models	100.00%
Model Specific Inputs		
Average defect diameter (μm)	Berglund 2 factor model	0.02
α (clustering factor)	Negative binomial model	5
Number of processes	Price model	3
Okabe factor	Okabe model	4
Number of critical layers	Bose-einstein model	2
Area of transistor (μm^2)	Hofstein and Heiman model	0.000125
Number of transistors in a die	Hofstein and Heiman model	9900000
Shape parameter (ϕ)	Inverse Gaussian model	2

Model selection (drop down list)	Outputs	
	Output description	Value
Price	Defect Density (/cm^2)	0.00000071

Note: After changing input values, please select model again in drop down list to update outupt value. This shall be patched shortly

MODEL PARAMETER CALCULATOR

List of Model Inputs		
Input description	Models using input	Input here
Yield (%)	All models	90.00%
Defect Density (/cm^2)	All models	0.4

Model selection (drop down list)	Outputs	
	Output description	Value
Inverse Gaussian	ϕ	1.990386642

Note: After changing input values, please select model again in drop down list to update outupt value. This shall be patched shortly

Figure 12 Reverse parameter estimations in the calculator for Defect Density and Model Parameter calculation.

Lastly, the calculator provides a graph (Figure 13) that highlights yield estimates for different models based on the fixed die area and defect density, which is useful for engineers to see which model fits best with their data and use that for further estimations.

As shown in Figure 13, there is a wide variability of the yield estimates depending on the model used. The uncertainty distribution illustrated by these sets of points can inform the user of the possible distribution of yield estimates that can be obtained with a fixed set of inputs.

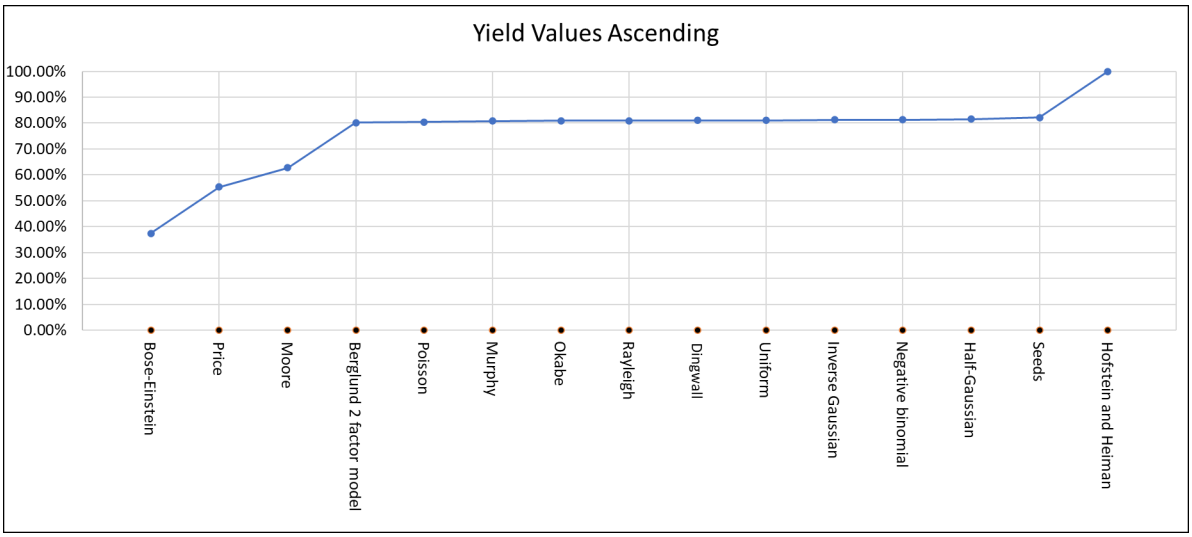


Figure 13: Comparison of yield estimates for various models for a given die area and defect density

Bayesian analysis approach is used to develop this uncertainty distribution of yield, to account for the variability of the yield as predicted by different models. The mathematical foundation of Bayesian analysis is based on Bayes Theorem (Eq. 13.2)

$$\pi(x|E) = \frac{L(E|x) * \pi_0(x)}{\int L(E|x) * \pi_0(x)\, dx} \tag{13.2}$$

where x the unknown of interest to be estimated (e.g., yield); $\pi_0(x)$ is an initial (prior) uncertainty distribution of x; $\pi(x|E)$ is the update (posterior) uncertainty distribution of x, give a certain amount of evidence E; and $L(E|x)$ is the likelihood of observing evidence E (likelihood function). The posterior distribution therefore is a fusion of the initial estimate and the new data or evidence.

In the present context E is the set of yield estimates from N models:

$$E = [y_1, y_2, \dots, y_n] \tag{13.3}$$

This analysis, namely the aggregation of yield estimates across different models, aims to preserve the model-to-model variability of estimates, meaning that the evidence E is treated as a non-homogeneous data set. Mathematical theory behind non-homogenous Bayesian estimation is more intense because the unknown of interest is not a single value but a distribution *f*(*y*) that can explain the variability of quantity of interest (*y*). Therefore, the Bayes theorem will be in form of:

$$\pi(f(y)|E) = \frac{L\big(E\big|f(y)\big) * \pi_0\big(f(y)\big)}{\int L\big(E\big|f(y)\big) * \pi_0\big(f(y)\big)\, df(y)} \tag{13.4}$$

To proceed, we can simplify the problem by postulating that the unknown distribution *f*(*y*) is a member of a parametric family of distributions $f(y|\boldsymbol{\theta})$with parameters θ_1 through θ_m. Therefore, the problem will be simplified to that of finding the distribution of $\boldsymbol{\theta}$ = [θ_1 through θ_m]:

$$\pi(\boldsymbol{\theta}|E) = \frac{L(E|\boldsymbol{\theta}) * \pi_0(\boldsymbol{\theta})}{\int L(E|\boldsymbol{\theta}) * \pi_0(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \tag{13.5}$$

The above distribution produces nearly an infinite set of candidates for $f(y|\boldsymbol{\theta})$. Normally an average over the entire set is used as the answer: That is

$$f_e(y) = \int_{\theta} f(y|\boldsymbol{\theta}).\ \pi(\boldsymbol{\theta}|E)\, d\boldsymbol{\theta} \tag{13.6}$$

The likelihood function $L(E|\boldsymbol{\theta})$ is structured based on the nature of the evidence $E = [y_1, y_2, \dots, y_n]$and the assumed form for $f(y|\boldsymbol{\theta})$. In this case, $0<y_i<1$, accordingly the suitable distributions for $f(y|\boldsymbol{\theta})$ are Beta distribution, or truncated normal or lognormal distribution. The truncated lognormal function is chosen which has several desirable properties. More specifically:

$$f(y|\theta_1,\theta_2) = \frac{1}{A\sqrt{2\pi}\theta_1 y}\exp\left(-\frac{1}{2}\left(\frac{\ln(y)-\ln(\theta_2)}{\theta_1}\right)^2\right) \quad (13.7)$$

where $0<y<1$, and θ_1, θ_2 are the two distribution parameters.

The likelihood function $L(E|\boldsymbol{\theta})$ is built based on the so-called multiplicative error model:

$$L(\mathrm{y_i}|y) = \frac{1}{A\sqrt{2\pi}\sigma \mathrm{y_i}}\exp\left(-\frac{1}{2}\left(\frac{\ln(y_i)-\ln(y)}{\sigma}\right)^2\right) \quad (13.8)$$

representing the accuracy/credibility of y_i, i.e., the yield estimate produced by model "*i*". The parameter σ of this truncated lognormal distribution is a measure of deviation of y_i from the true but unknown yield value *y*.

This methodology is used (proposed by Mosleh et. al [41] and Groen et. al. [42]) to develop the uncertainty distribution of yield estimates based on the results shown in Figure 13. In this case, the prior range for *y* is taken as the general industry standards for yield based on minimum acceptable yield thresholds for production and the expected yield during production and the evidence, *E* is the set of yield estimates from different models.

Results for the uncertainty analysis are performed with the help of R-DAT [43] and are displayed in Figure 14. Since the output of the uncertainty distributions comprise of a range of possible parameters, the distributions based on the mean, median, 5th and 95th percentile of the lognormal parameters are shown in Figure 14. However, referring to the mean distribution is sufficient to visualize the variability in yield estimates by plotting the 5th and 95th percentile points.

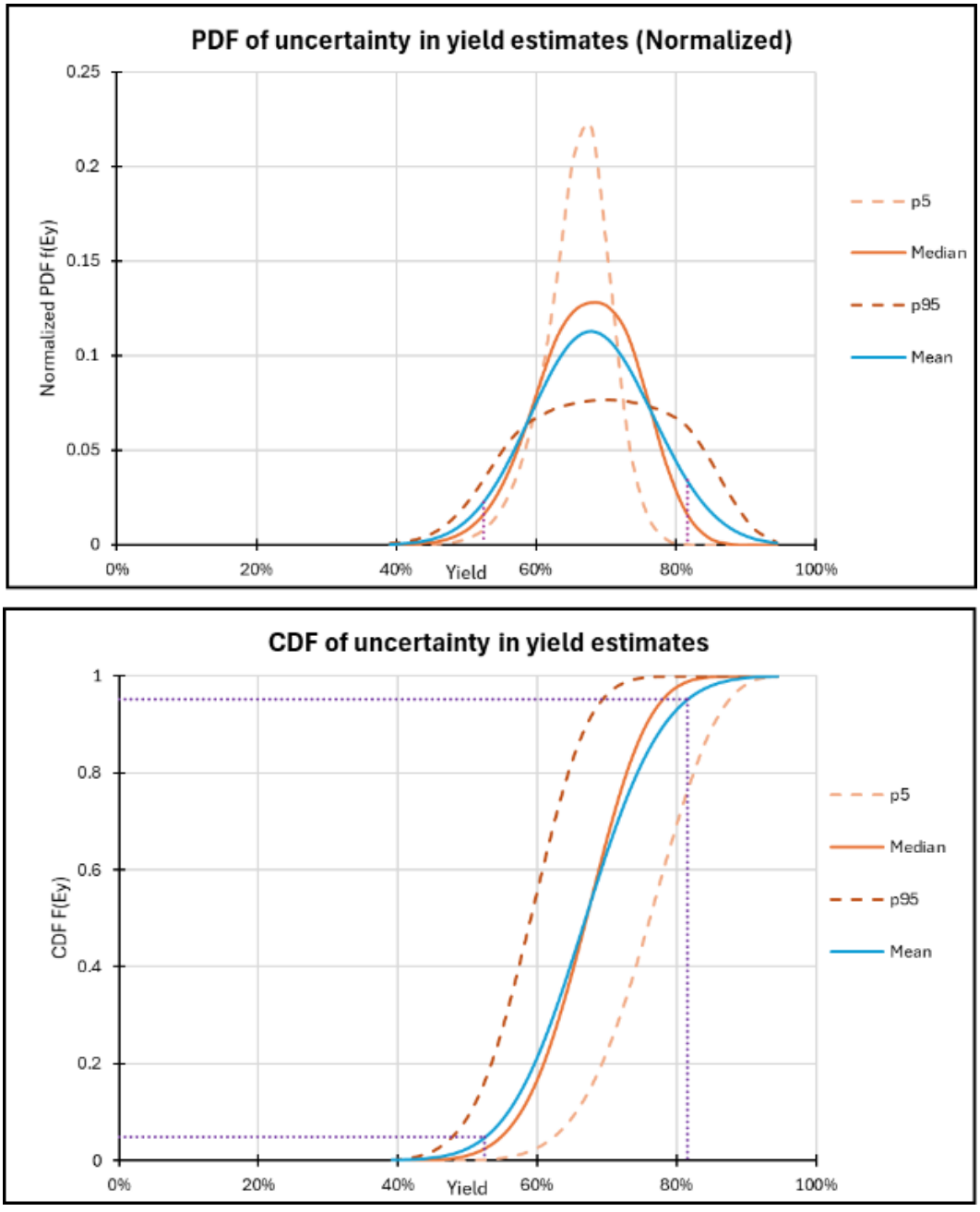


Figure 14: Truncated Lognormal Posterior PDF and CDF of Yield estimates using outputs of all models from Yield Calculator. The dotted lines indicate p5 and p95 uncertainty in the mean yield distributions.

13.3 Parametric-Shift Yield module

The Parametric-shift Yield model estimates the parametric shift based on the selected model and Process node. The estimation of yield is handled by a python script that starts a server to run the model and output based on the selected process node.

(drop down lists)	Output description	Value
I-O yield	Defect Yield (%)	71.56%
Bose-Einstein		
lgbm_Opus_Matrix_common	Parametric Yield (%)	70.42%
Total Yield		**50.39%**

Figure 15: Parametric Shift output based on input model and process node.

13.4 Combined Yield metric

This study considers parametric-shift limited and Defect-limited yield as independent events constituting the yield in semiconductor devices. This independence is attributed to two main concepts: the definition of the two types of yield and the causes of the two types of yield. Defect-limited yield is considered to be a form of catastrophic failure which causes the device to not perform its functionality. Parametric-shift limited yield on the other hand is considered to be a form of failure which causes the device to not reach its intended performance metrics for a given function. Defect-limited yield is also influenced by the interference of external particles or abrasive faults caused during the production process. Parametric yield is primarily attributable to the inherent variations in the production process which can cause the overall device output to vary. In a probabilistic perspective of yield, these concepts can be represented with an OR gate which defines the output yield as being brought about by either a Defect ($P(D)$) OR Parametric($P(PS)$). The mathematical formulation for the same is hence defined as

$$Yield = 1 - P(D) \cap P(PS) = 1 - (1 - (1 - P(D))(1 - P(PS)) \quad (13.9)$$

$$Yield = 1 - (1 - Y_D Y_{PS}) = Y_D Y_{PS} \quad (13.10)$$

Where defect yield (Y_D) can be estimated with the help of several existing defect distributions related models as discussed in section 3, and parametric-shift yield (Y_{PS}) is estimated as detailed in this research. For comparison, for a mature process, the defect limited yield can reach upto 90%. Using the yield estimate from the 40nm process of 72%

$$Yield = 90\% \times 72\% = 68\%$$

we get an estimated total yield of 68% using the combined metric.

13.5 Summary

This section integrated the dissertation's two parallel modeling streams — defect-limited and parametric-shift yield — into a single Excel-based calculator that engineers can use to estimate total semiconductor yield without running circuit-level simulations. The Defect Yield calculator implements all 15 defect-limited models reviewed in Section 3, supports five circuit types (IO, Logic, Memory, Pure Analog, and High-speed Digital), and applies De-Vries equations with a polynomial area-ratio scaling (n_3) to estimate gross die counts from wafer geometry. Beyond forward yield prediction, it also supports reverse parameter estimation — by back-calculating defect density D_0 or model-specific constants from measured WS data using closed-form solutions (where available) and numerical solvers (via Excel) for more complex models — and integrates the Process Adjustment Factor from Section 5 as a selectable multiplier keyed to process node and ISO fab class. A Bayesian uncertainty analysis built on a truncated lognormal posterior and a multiplicative-error likelihood function, implemented in R-DAT, aggregates the spread of estimates across all 15 models into a single posterior PDF and CDF of yield, giving the users a defensible uncertainty band (5th–95th percentile) rather than a single point estimate. The parametric yield calculator runs the trained Gaussian copula and ML models (XGBoost, LGBM,

GBC) from Sections 9–12 through a Python-backed server, currently using preset inputs per process node with planned support for user-tunable parameters. The two streams are unified through a probabilistic OR-gate formulation that treats defect-limited and parametric-shift failures as independent events. This combined metric is validated by applying to a mature 40 nm process, which produced an estimated total yield of 68% (from a 90% defect-limited and 72% parametric-shift baseline), illustrating how the framework consolidates the methodological contributions of the entire dissertation into a single, deployable Software Application suitable for inclusion in an ATE toolbox, as originally proposed in the abstract.

14 Conclusions and Future Work

This research focuses on modeling Defect limited yield and Parametric-shift limited yield.

This research started with a brief review and characterization of several defect-limited yield models available in literature. A yield calculator was developed to display the variations among the different yield models as well as to estimate defect density and model parameters based on user inputs. An uncertainty analysis was also done to show uncertainty in the estimates of yield based on the outputs of different models. This research also proposed a Process Adjustment Factor (PAF) that enhanced the existing yield models by incorporating dependencies on key process factors likely to influence defect yield. The PAF integrated data on these process factors alongside information about particle distribution within the fabrication plant by proposing a geometrical probability of failure. PAF was evaluated for current yield estimates on 40 nm devices with primarily favorable outputs and was integrated into the yield calculator to help estimate yield. Our comparison of model yields revealed an improvement in yield estimation for three out of four different ASICs. However, further enhancements can be achieved by redefining the PAF. Instead of being applied as a multiplier at the end of the defect estimation process, future studies should consider modifying the die area to a “Critical Area” based on circuitry..

A Physics-informed machine learning approach for estimating parametric shift was presented by using WAT and WS datasets. First, WAT parameters were theoretically estimated using process node inputs and Level 2 BSIM models. Gaussian copulas were trained on 40nm devices and 16nm devices to generate a correlation table for comparing the different WAT parameters. These correlation tables were found to have several similarities across different devices thereby suggesting a high degree of generalizability. Using the physics-based estimates and copulas, a synthetic dataset can be generated which closely mimics WAT parameters for a process node.

Three types of machine learning models were trained on a combined WAT and WS dataset to predict the outcome from a given set of WAT parameters. XGBoost and LGBM displayed the highest degree of accuracy for all of the models whereas GBC while showing high accuracy, had a higher amount of sensitivity to variability in the WAT parameters as was seen in the Cp/CpK dataset analysis. Different results were obtained by combining data from multiple devices to help generate a more generalizable model however at the cost of lower accuracy and fewer WAT parameters being accounted for. Furthermore, Pelgrom's Law and Cp/CpK parameters were introduced to help account for transistor design density and process node capability respectively. The machine learning models utilized the synthetic database to estimate the parametric shift yield.

Feature importance metrics were generated with the help of SHAP. Results from copulas and feature importance metrics can also be further used to develop an understanding of parametric shift phenomena as a tool to aid root cause analysis since it highlights certain WAT parameters as particularly involved with the help of feature importance metrics or the interlinking of WAT parameters with the help of the gaussian copula-based correlation table. Further improvements to the model will require more data points, not only from the same device but also from other devices from different process nodes. Larger datasets may also warrant the use of more complex data analysis methodologies such as neural networks albeit at the cost of lower interpretability.

15 References

[1] D. Petersen, "Global Semiconductor Sales Increase 25% from Q4 2025 to Q1 2026," *Semiconductor Industry Association,* 2026.

[2] D. S. K. R. J. B. a. D. B. J. Kusters, "Global Semiconductor Outlook," *Deloitte Center for Technology, Media & Telecommunications,* 2026.

[3] P. G. a. E. Papadopoulou, "Yield Analysis and Optimization," *Auerbach Publications,* 2009.

[4] G. O. D. A. W. D. a. S. C. G. A. I. Mirza, "Spatial Yield Modeling for Semiconductor Wafers," *IEEE/SEMI Advanced Semiconductor Manufacturing Conference,* 1995.

[5] A. A. S. Y. a. S. B. A. H. Sakr, "Building Discrete-Event Simulation for Digital Twin Applications in Production Systems," *26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA),* 2021.

[6] T. P. K. L. F. A. A. R. K. F. J. S. a. G. F. L. K. W. Tobin, "Content-Based Image Retrieval for Semiconductor Process Characterization," *EURASIP Journal on Applied Signal Processing,* pp. 704-713, 2002.

[7] S. G. R. P. David, "Using Multivariate Nested Distributions to Model Semiconductor Manufacturing Processes," *IEEE Transactions on Semiconductor Manufacturing,* vol. 12, 1999.

[8] B. P. B. a. D. A. Paszkiewicz, "Influence of External Plane Stress on Electric Parameters of AlGaN/GaN HEMT Heterostructures," *Phys. Status Solidi A,* 2023.

[9] G. E. Moore, "What level of LSI is best of you," *Electron Devices Meeting, Washington D. C.,* vol. 43, pp. 126-130, 1967.

[10] B. T. Murphy, "Cost-size optima of monolithic integrated circuits," *Proc. IEEE,* vol. 52, pp. 1537-1545, 1964.

[11] C. H. Stapper, "On Murphy's Yield Integral," *IEEE TRANSACTIONS ON SEMICONDUCTOR MANUFACTURING,* vol. 4, no. 4, 1991.

[12] R. B. Seeds, "Yield, economic, and logistic models for complex digital arrays," *IEEE Int. Convention Rec.,* vol. 6, pp. 61-66, 1967.

[13] C. H. Stapper, "Fact and fiction in yield modeling," *Microelectronics Journal,* vol. 20, no. 1-2, pp. 129-151, 1989.

[14] R. B. Seeds, "Yield and cost analysis of bipolar LS," *IEEE International Electron Devices Meeting (IEDM) Technical Digest,* vol. 13, p. 12, 1967.

[15] B. Meister, "On Murphy's yield formula," *IBM Journal of Research and Development,* vol. 27, no. 6, pp. 545-548, 1983.

[16] R. M. Warner, "A note on IC-Yield statistics," *Solid State Electronics ,* vol. 24, no. 11, pp. 1045-1047, 1981.

[17] J. E. Price, "A new look at yield of integrated circuits," *Proc. IEEE,* vol. 58, no. 8, pp. 1290-1291, 1970.

[18] T. N. M. a. S. S. Okabe, "Analysis on yield of integrated circuits and a new expression for the yield," *Electron. Eng. Japan,* vol. 92, pp. 135-141, 1972.

[19] Z. K. a. C. H. S. I. Koren, "A unified negative-binomial distribution for yield analysis of defect-tolerant circuits," *IEEE Trans. Comput,* vol. 42, pp. 724-734, 1993.

[20] A. S. a. P. S. M. Raghavachari, "Poisson mixture yield models for integrated circuits: A critical review," *Microelectronics Reliability,* vol. 37, pp. 565-580, 1997.

[21] A. V. Ferris-Prabhu, "On the assumptions contained in semiconductor yield models," *IEEE Trans. Computer Aided Design Integrated Circuits System,* vol. 11, pp. 986-975, 1992.

[22] S. H. a. F. P. Heiman, "The silicon insulated-gate field-effect trainsistor," *Proc. IEEE,* vol. 51, pp. 1190-1202, 1963.

[23] C. N. Berglund, "A Unified Yield Model Incorporating Both Defect and Parametric Effects," *IEEE Transactions on Semiconductor Manufacturing,* vol. 9, pp. 447-454, 1996.

[24] M. G. B. H. R. Sayah, "Comb/serpentine/cross-bridge test structure for fabrication process evaluation," *NASA Technical Reports Server,* 1988.

[25] "J750Ex-HD," Teradyne, [Online]. Available: https://www.teradyne.com/products/j750/.

[26] D. B. M. Klassen, "A unified mobility model for device simulation—I. Model equations and concentration dependence," *Solid-State Electronics,* vol. 35, pp. 953-959, 1992.

[27] D. Klaassen, "A unified mobility model for device simulation—II. Temperature dependence of carrier mobility and lifetime," *Solid-State Electronics,* vol. 35, pp. 961-967, 1992.

[28] X. X. J. H. W. L. K. M. C. X. J. J. J. O. M. C. A. M. N. a. C. H. M. V. Dunga, BSIM 4.6.0 MOSFET Model User's Manual, Berkeley, CA, USA: University of California, Berkeley, 2006.

[29] A. Sklar, "Fonctions de répartition à n dimensions et leurs marges," *Publications de l'Institut de Statistique de l'Université de Paris,* vol. 8, pp. 229-231, 1959.

[30] P. arbenz, "Bayesian copulae distributions, with application to operational risk management – some comments," *Methodology and Computing in Applied Probability,* pp. 105-108, 2013.

[31] S. D. a. A. J. McNeil, "The t Copula and Related Copulas," *International Statistical Review,* vol. 73, no. 1, p. 111–129, 2005.

[32] D. G. Clayton, "A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence," *Biometrika,* vol. 65, pp. 141-151, 1978.

[33] M. J. Frank, "On the simultaneuous associativity of F(x, y) and x+y-F(x,y)," *Aequationes Mathematicae,* vol. 19, pp. 194-226, 1979.

[34] E. J. Gumbel, "Bivariate Exponential Distributions," *Journal of the American Statistical Association,* vol. 55, no. 292, p. 698–707, 1960.

[35] R. B. Nielsen, An Introduction to Copulas 2nd ed., New York, USA: Springer, 2006.

[36] N. N. M. a. M. S. H. M. M. Ali, "A class of bivariate distributions including the bivariate logistic," *Journal of Multivariate Analysis,* vol. 8, no. 3, pp. 405-412, 1978.

[37] C. G. T. Chen, "XGBoost: A scalable tree bossting system," *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining,* pp. 785-794, 2016.

[38] Q. M. T. F. T. W. W. C. W. M. Q. Y. a. T.-Y. L. G. Ke, "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems,* vol. 30, 2017.

[39] A. C. J. D. a. A. P. G. W. M. J. M. Pelgrom, "Matching properties of MOS transistors," *IEEE Journal of Solid-State Circuits,* vol. 24, no. 5, pp. 1433-1439, 1989.

[40] D. K. d. Vries, "Investigation of gross die per wafer formulas," *IEEE Transactions on Semiconductor Manufacturing,* vol. 18, pp. 136-139, 2005.

[41] A. M. a. G.Apostolakis, "The Development of a Generic Data Base for Failure Rates," in *Proceedings of the International ANS/ENS Topical Meeting on Probabilistic Safety Methods and Applications*, San Francisco, CA, USA, 1985.

[42] E. L. D. S. J. a. A. M. F.J. Groen, "A Reliability Data Collection and Analysis System for Products Under Development," *Brazilian Journal of Operations & Production Management,* vol. 1, no. 1, p. 93–106, 2004.

[43] "R-DAT software," Prediction Technologies, [Online]. Available: https://www.prediction-technologies.com/rdat.

[44] P. V. Zant, Microchip Fabrication, 6th ed., McGraw-Hill, 2014.

Appendices

A: Copula Correlation tables

Figure A1: Copula Correlation Table for ASIC A

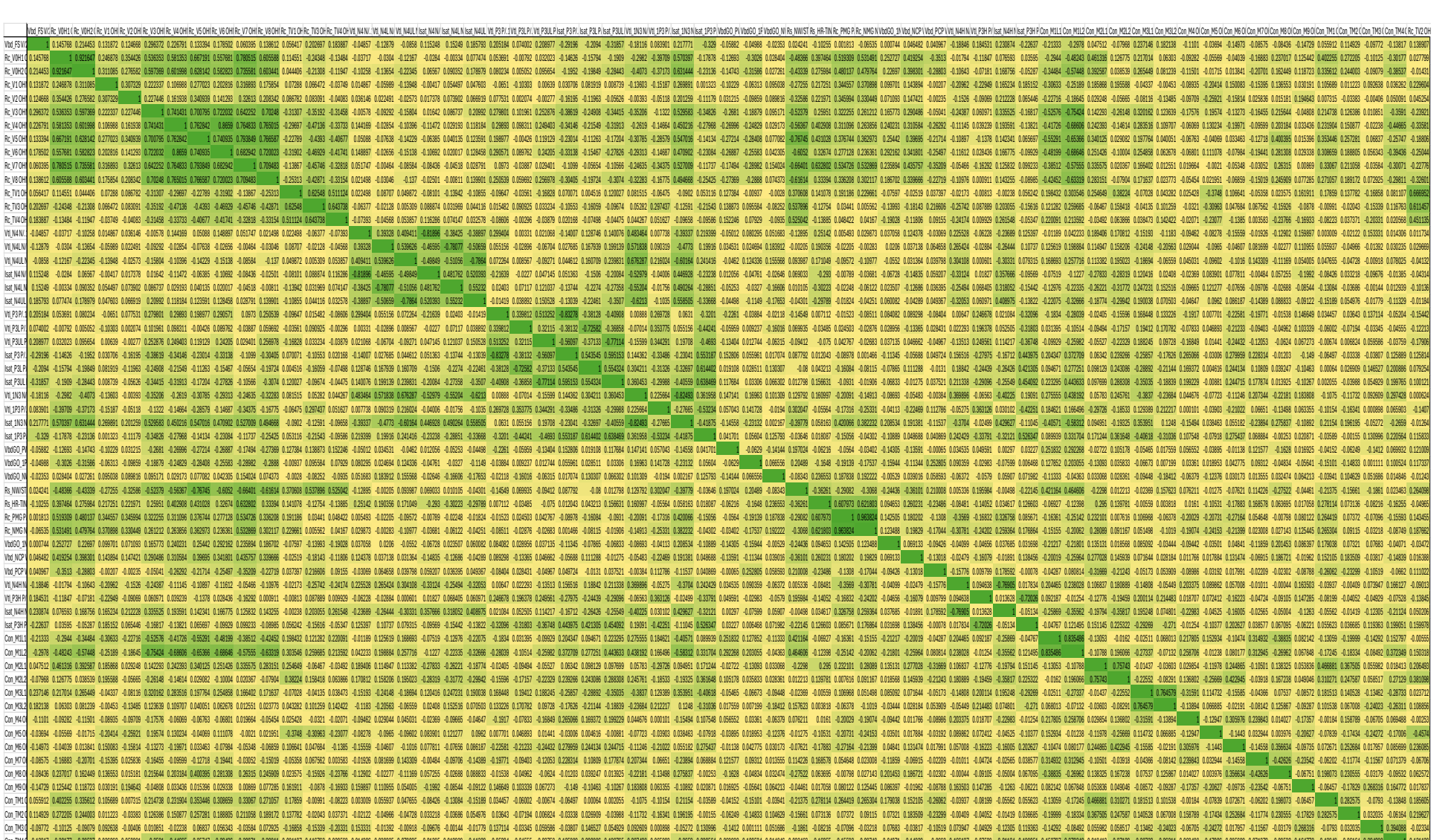


Figure A2: Copula Correlation Table for ASIC C

B: Copula outputs for all WAT parameters

This section shows the synthetic data generated and compared with real data for each WAT parameter for the 40nm process data.

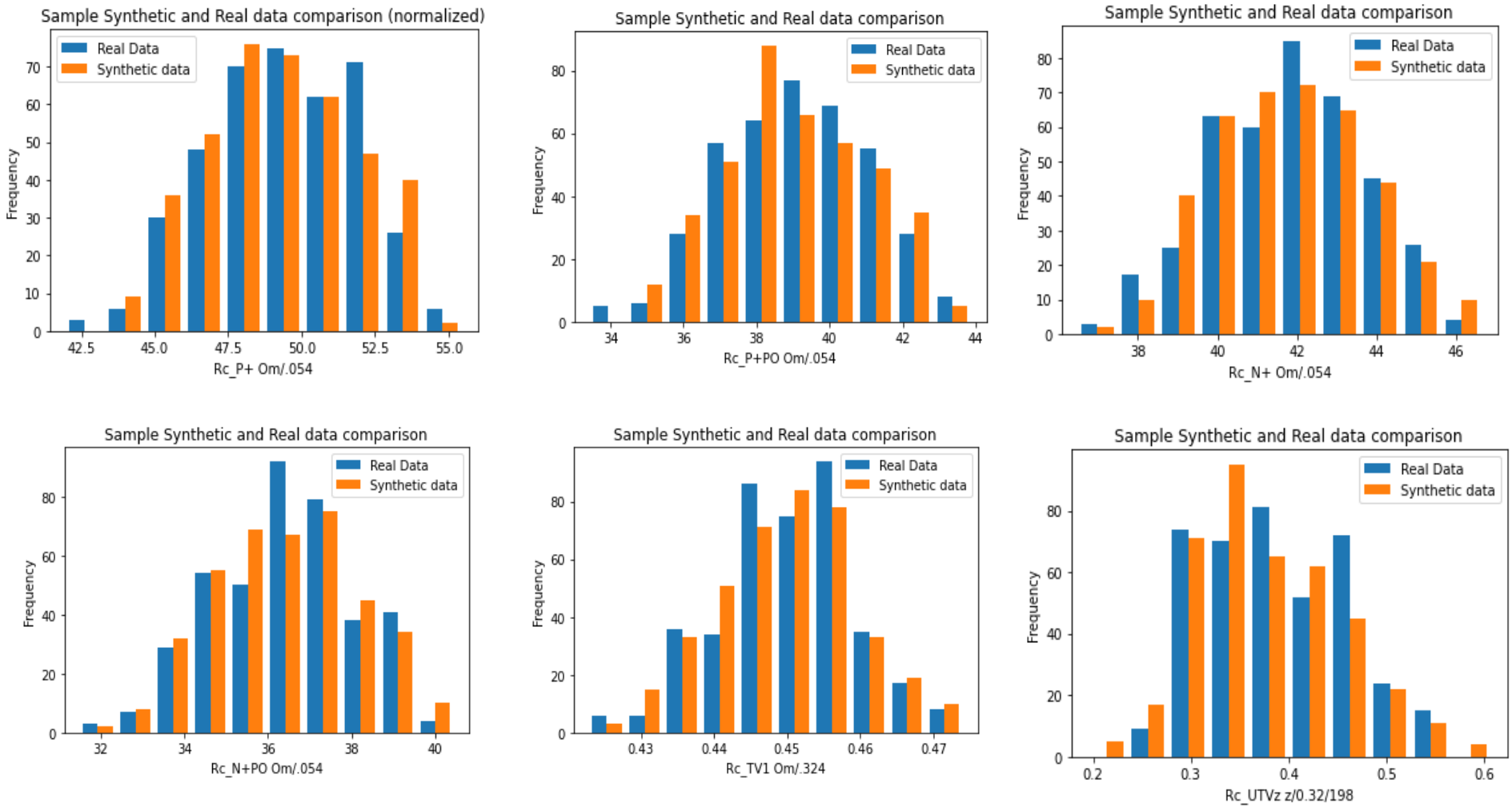


Figure A3: Synthetic Data for Rcs (contact resistances)

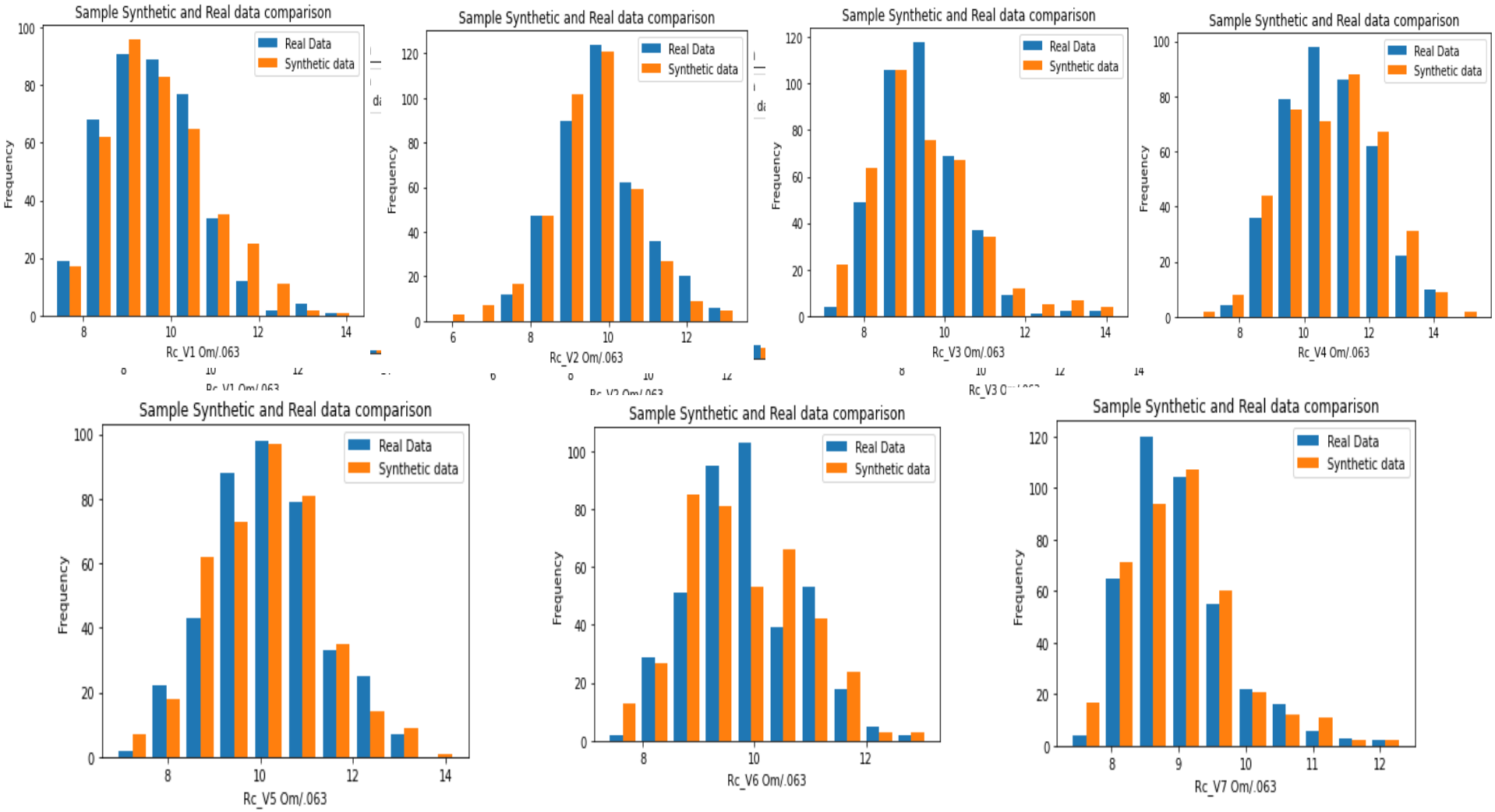


Figure A4: Synthetic Datasets for Via resistances (Rc)

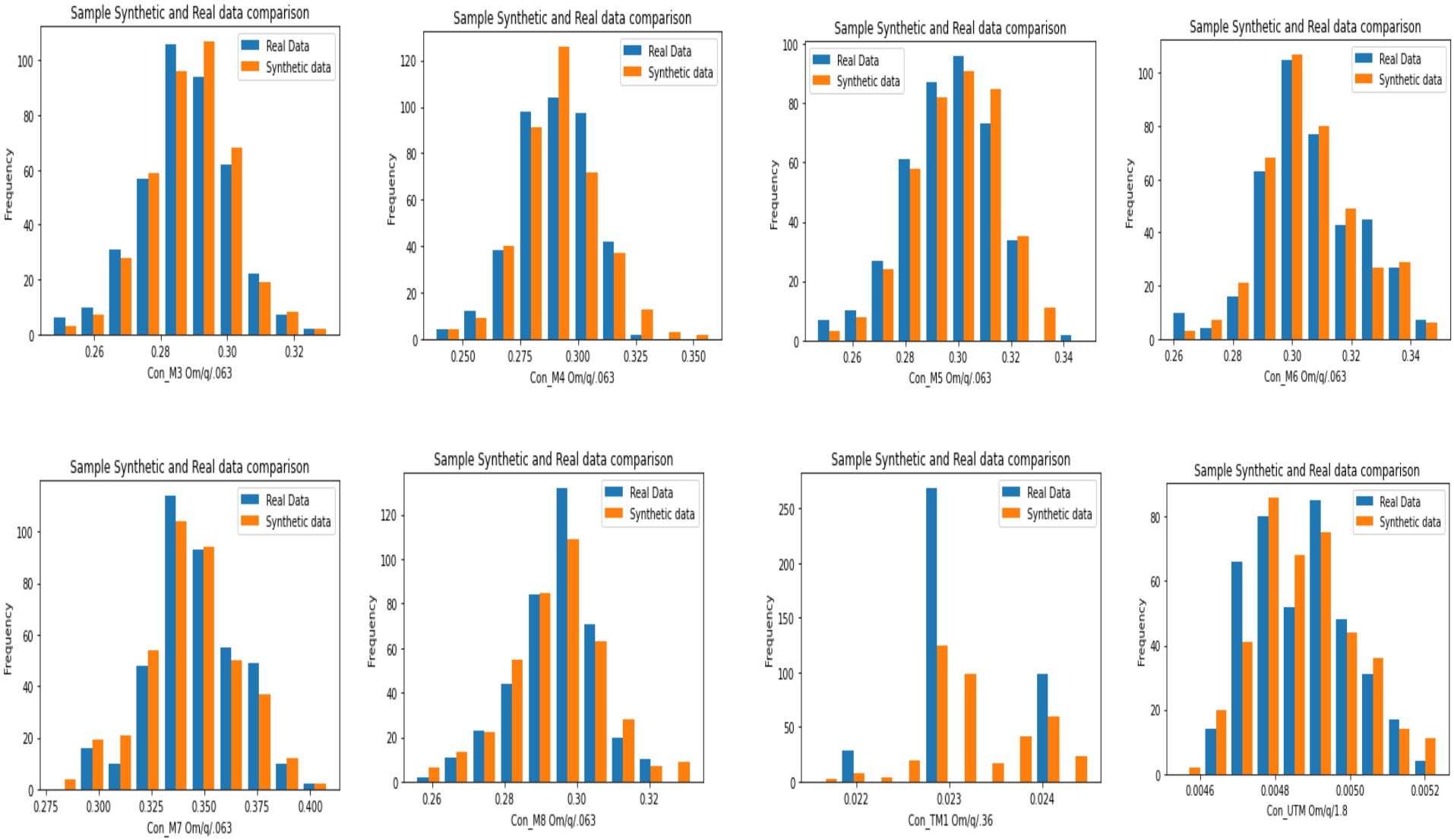


Figure A5: Synthetic Datasets for metal sheet resistances (Con)

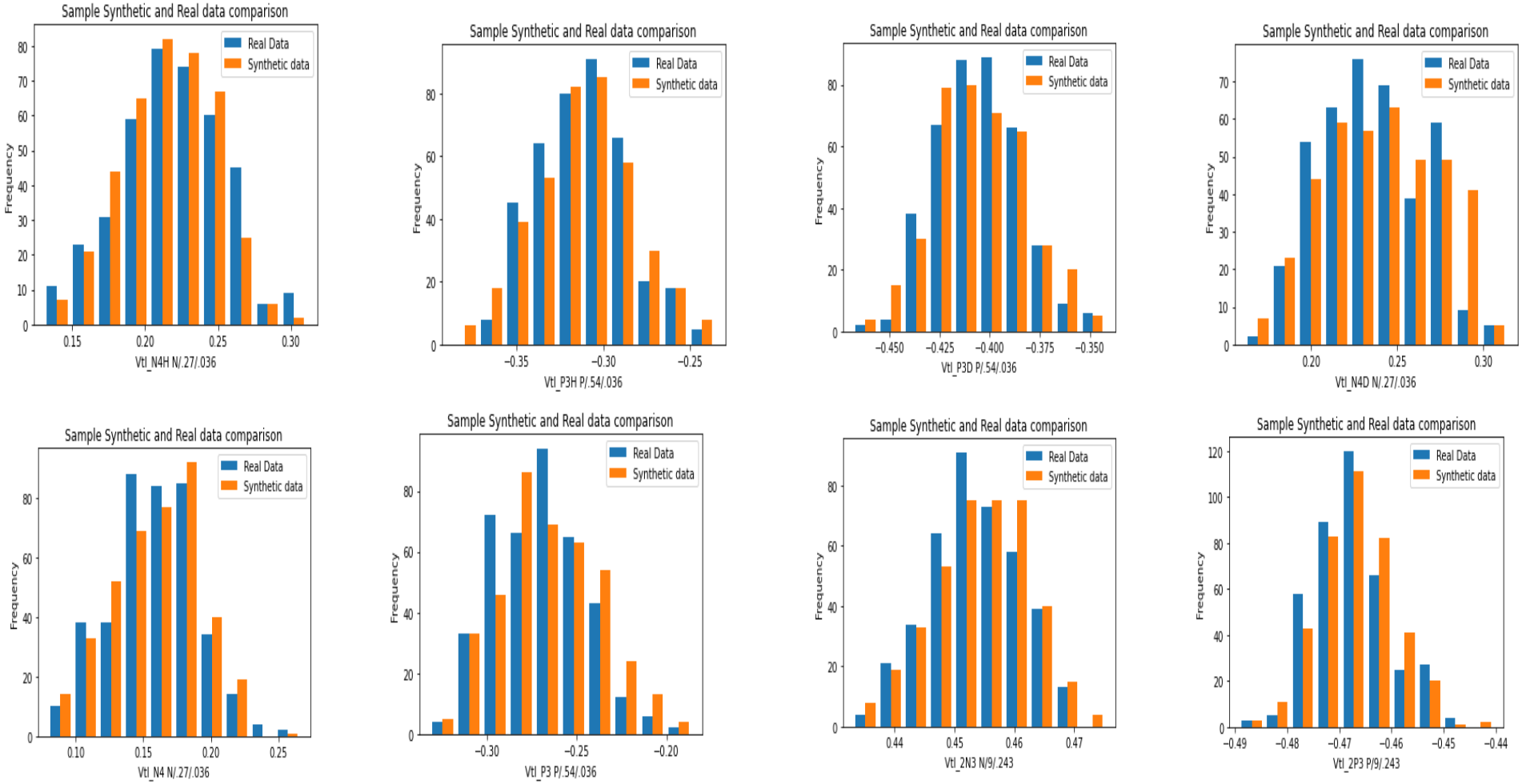


Figure A6: Synthetic data for threshold voltage (Vt)

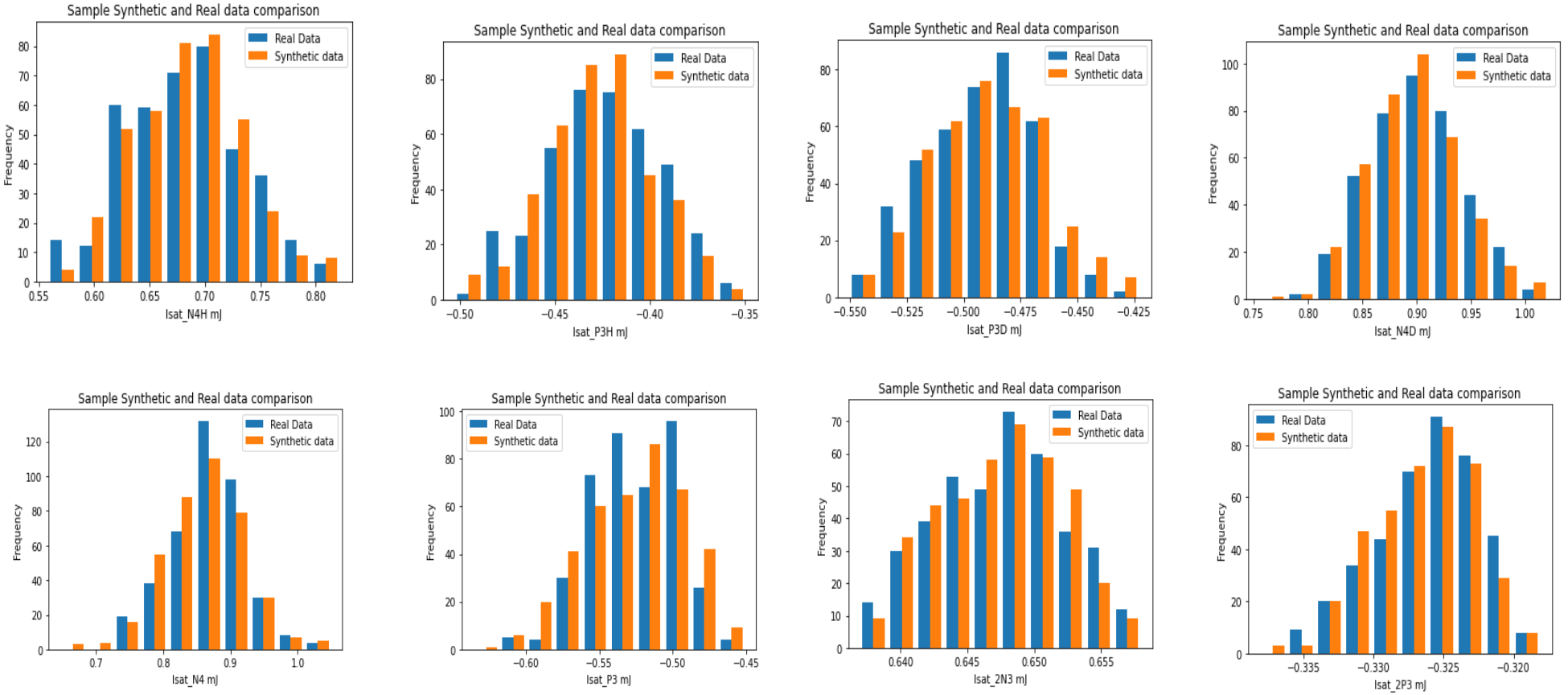


Figure A7: Synthetic Data for Saturation Current (Isat)

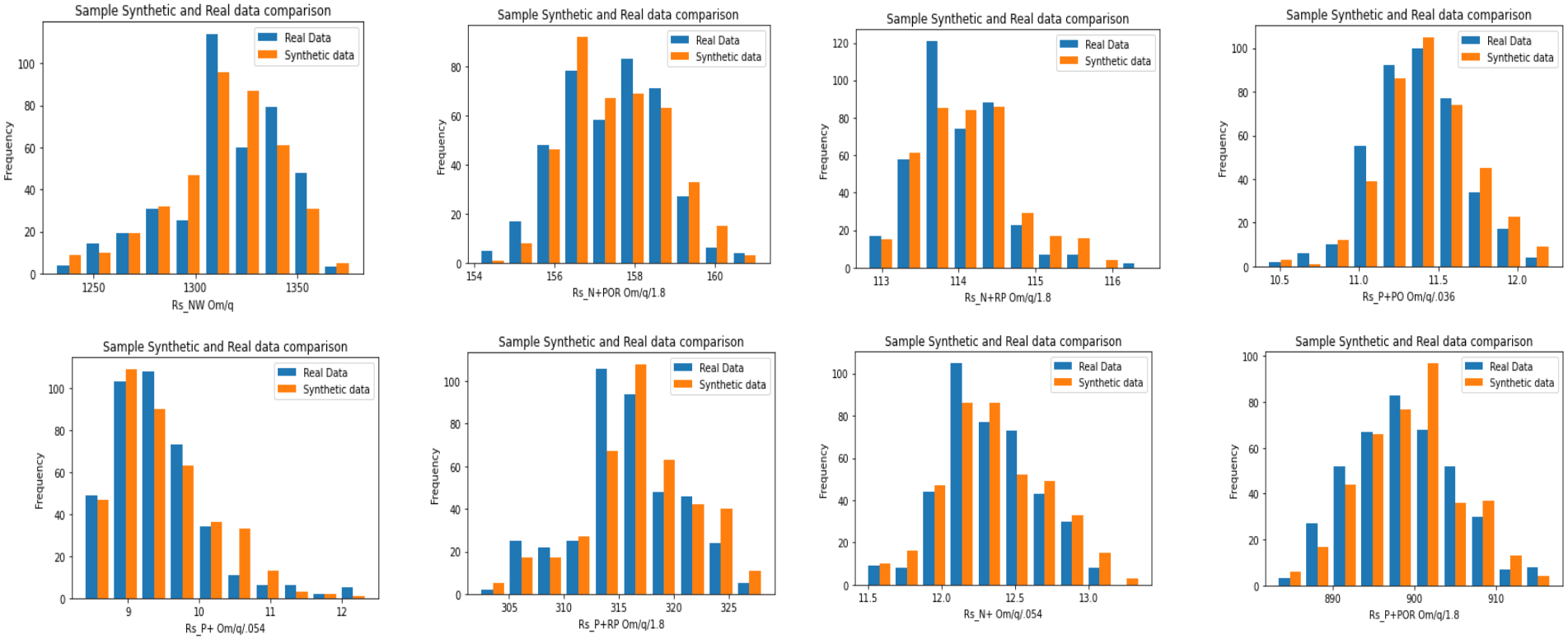


Figure A8: Sheet resistances for poly/oxide layers

C: Example Decision Tree

The decision tree here is trained on the 40nm device data and split based on freidman_mse parameter.

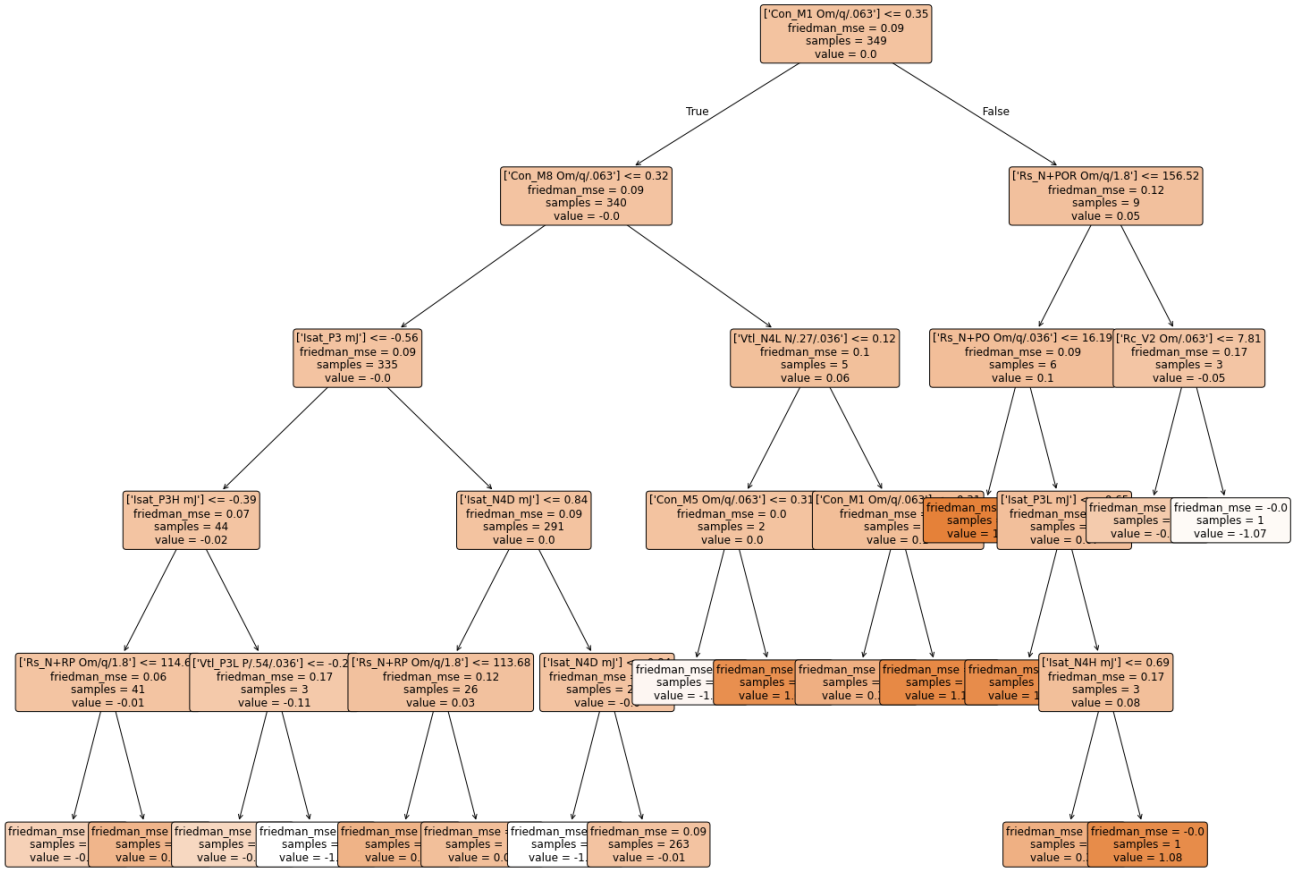


Figure A9: Example tree construction for 40nm GBC algorithm