The Thesis committee for Tyler Jackson Darwin certifies that this is the approved version of the following thesis:

Gaussian Process Regression for Virtual Metrology of Microchip Quality and the Resulting Strategic Sampling Scheme

APPROVED BY

SUPERVISING COMMITTEE:

____

Dragan Djurdjanovic, Supervisor

____

Roman Garnett

Gaussian Process Regression for Virtual Metrology of Microchip Quality and the Resulting Strategic Sampling Scheme

by

Tyler Jackson Darwin

Thesis

Presented to the Faculty of the Graduate School
of the University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science in Engineering

The University of Texas at Austin
August 2017

Acknowledgment

Thanks to Dragan Djurdjanovic for challenging me to contribute to this field to which he introduced me two years ago. I appreciate the opportunity.

Special thanks to Roman Garnett for his timely advice throughout the research process. This work would not have come to fruition without your support.

Additional thanks to my parents, Pam and Barnes, who never receive enough credit.

Gaussian Process Regression for Virtual Metrology of Microchip Quality and the Resulting Strategic Sampling Scheme

by

Tyler Jackson Darwin, M.S.E.
The University of Texas at Austin, 2017
Supervisor: Dragan Djurdjanovic

Manufacturing of integrated circuits involves many sequential processes, often executed to nanoscale tolerances, and the yield depends on the often unmeasured quality of intermediate steps. In the high-throughput industry of fabricating microelectronics on semi-conducting wafers, scheduling measurements of product quality before the electrical test of the complete IC can be expensive. We therefore seek to predict metrics of product quality based on sensor readings describing the environment within the relevant tool during the processing of each wafer, or to apply the concept of virtual metrology (VM) to monitor these intermediate steps. We model the data using Gaussian process regression (GPR), adapted to simultaneously learn the nonlinear dynamics that govern the quality characteristic, as well as their operating space, expressed by a linear embedding of the sensor traces' features. Such Bayesian models predict a distribution for the target metric, such as a critical dimension, so one may assess the model's credibility through its predictive uncertainty. Assuming measurements of the quality characteristic of interest are budgeted, we seek to hasten convergence of the GPR model to a credible form through an active sampling scheme, whereby the predictive uncertainty informs which wafer's quality to measure next. We evaluate this convergence when predicting and updating online, as if in a factory, using a large dataset for plasma-enhanced chemical vapor deposition

(PECVD), with measured thicknesses for ~32,000 wafers. By approximately optimizing the information extracted from this seemingly repetitive data describing a tightly controlled process, GPR achieves ~10% greater accuracy on average than a baseline linear model based on partial least squares (PLS). In a derivative study, we seek to discern the degree of drift in the process over the several months the data spans. We express this drift by how unusual the relevant features, as embedded by the GPR model, appear as the inputs compensate for degrading conditions. This method detects the onset of consistently unusual behavior that extends to a bimodal thickness fault, anticipating its flagging by as much as two days.

Table of Contents

Chapter 1 Introduction **1**

Chapter 2 Literature review **4**

Chapter 3 Bayesian modeling approach **7**

3.1 Conceptual background 7

3.2 Mathematical foundation 8

3.2.1 Incorporation of prior knowledge 9

3.2.2 Tuning of parameters 9

3.2.3 Prediction under uncertain embedding 10

3.2.4 Active learning of embedding 11

Chapter 4 Background of data from representative plasma process **13**

4.1 Plasma-enhanced chemical vapor deposition 13

4.2 Context of data from PECVD tool 15

Chapter 5 Results **17**

5.1 Description of underlying experiment 17

5.2 Learning curves 19

5.3 Investigating process drift 22

Chapter 6 Concluding discussion **26**

Appendices **28**

Bibliography **29**

Chapter 1

Introduction

To manufacture modern integrated circuits (IC), hundreds of sequential processes pattern, stack, and connect layers of electrical components that demand nanoscale tolerances across hundreds of millimeters of a wafer substrate [1]. The dimensions necessary to pack more processing power or memory onto a single chip have shrunk for decades, pushing tolerances to limits of what may be measured and challenging engineers to further understand each step to maintain acceptable yield. By the nature of a high-throughput industry, semiconductor manufacturers cannot afford to measure all critical quality characteristics of these intermediate steps for every chip, much less every 300mm wafer often gridded with over 100 chips. Should the chip's final electrical test reveal a short circuit, the root cause may not be apparent, especially if lacking knowledge of the quality characteristics from key steps, such as etching of transistor gates or deposition of thin films [1].

Given limited physical measurements of the product, we seek to predict such metrics of product quality through the concept of virtual metrology (VM) [2]. Note that mathematical models based on first principles, particularly those of plasma dynamics [3], typically struggle to explain the nanoscale deviations in these dimensions, mostly due to the sheer complexity of the physical phenomena involved [2]. Instead, models in the context of VM approximate the dynamics that govern the quality characteristic of interest using sensor traces that describe the environment within the relevant tool during processing of each wafer.

In this thesis, we propose a probabilistic approach to regress the metric of quality that uses Gaussian processes (GPs), which can model the nonlinear behavior typical of processes in semiconductor manufacturing. Such Gaussian process regression (GPR) predicts a Gaussian distribution for the target metric of quality given the features extracted from the corresponding wafer's sensor traces, so process engineers may assess its credibility based on predictive accuracy and confidence. Note that rarely does production halt to excite the system about its set points, so this regression depends on the seemingly repetitive data from a tightly controlled process. To approximately optimize the information extracted from the stream of manufacturing data, we let the GPR model's predictive uncertainty inform where to measure next [11], as constrained by the measurement budget.

In theory, GPR can approximate arbitrarily complex functions, or exhibit universal consistency, assuming *stationary*, Gaussian noise corrupts the observed outputs [12]. In practice, irrelevant features contribute conceptual noise, hampering inference of the underlying input–output relationship [10]. We therefore integrate feature selection into the GP prior, effectively projecting the inputs to the lower-dimensional space where the dynamics operate. To hasten convergence to a credible form when starting from no initial data, we let the GPR model select which wafer's quality to measure based on the expected information gain.

To thoroughly evaluate this modeling approach, we rely on an extensive PECVD dataset for which the quality characteristic of interest, film thickness averaged across the wafer, has been measured for each of the roughly 32,000 wafers produced over several months [5]. In chapter 2, we review related works in the VM literature. Chapter 3 outlines the Bayesian framework to learn a low-dimensional embedding to explain the data, following the techniques of Garnett et al. [10] (see [10] and references therein for context in the GP literature). Chapter 4 sketches how PECVD creates conformal thin films, and explains the features derived from sensor traces representing the given tool's environment. In chapter 5, we first present the experiment designed to mimic online prediction in a factory, then compare convergence to that of partial least squares, a baseline linear model, and finally investigate whether the process drifts before faults. Chapter 6 concludes by assessing the viability of this fully Bayesian approach for real-time prediction.

Chapter 2

Literature review

Methods in the VM literature share the basic goal of predicting product quality based on soft-sensor readings of the underlying environment during processing of that unit. Many have theorized of VM's potential to fine tune the control of processes in semiconductor manufacturing to limit yield loss [2–4]. However, none of these modeling frameworks have yet established sufficiently credibility in an industrial setting, at least for plasma processes such as etch or thin film deposition, to be thus configured.

Given a block of data with some measurements of a quality metric, one may default to a linear model like partial least squares (PLS) [2] for its computational efficiency and ability to handle highly correlated features, such as those extracted from sensor traces of a tightly controlled process. Furthermore, by projecting the inputs to a lower-dimensional space, PLS can project out conceptual noise from features irrelevant to the output metric of quality. However, such a linear approximation may not be appropriate for the nonlinear dynamics that govern this quality characteristic, which has prompted others to localize models in order to conduct piecewise regression. Notions of similarity between or locality of examples typically derive from distance in the input space.[1] For example, to refine the notion of distance before weighting examples based on proximity to the query point, Hirai & Kano [7] use a global PLS model to scale the raw input features from a plasma etch process. However, the notion of "local," expressed by a scaling parameter must be set by cross-validation.

Bleakie & Djurdjanovic [6] modeled data from a PECVD process by partitioning the input space through unsupervised clustering, then fit a local linear model within each

[1] Inputs may include a dimension for time, or order of the processed wafers, though rarely is this seen in the VM literature.

cluster. As the model updates, various heuristics inform how to move the centers of the assigned clusters to regions of high modeling error and if necessary, create new clusters. The authors also touched on the issue of when to trust a given prediction and the advantage of measuring at unscheduled times, when the raw inputs are "unusually" distant from the known data. This sampling strategy may be too myopic, as it neglects the relevance to the target metric of quality when discerning informational value.

Others localize models exclusively in terms of time, by regressing on the previous L examples before any given query point (i.e. applying a moving time window). For example, Lynn et al. [8] compare PLS, artificial neural networks (ANN), and Gaussian process regression (GPR)[2] for windows of various lengths, finding GPR to be about ~10% more accurate for predicting plasma etch depth. However, windowing marginally improved the GPR's accuracy (by <2%) compared to a global model, while the constant refitting greatly slowed execution.[3]

Artificial neural networks (ANN) offer a global, rather than piecewise regression of nonlinear functions. Feed-forward ANNs comprise an input layer, one or more "hidden" layers of weighted nonlinear transformations of the previous layer's activity, and an output layer [12]. Although appealing for their ability to discern patterns amidst conceptual noise from irrelevant features, the trained model's parameters tend to be uninterpretable, so which input features are relevant remains unclear. The weights exhibit many local optima during fitting, which prompted research in the '90s on how to regularize, or stabilize the weights by assigning Gaussian priors to them. Neal found that such Bayesian neural networks, in the limit of infinite hidden nodes, approached a Gaussian process prior.

[2] Lynn's GPR incorporates a squared exponential kernel with "automatic relevance determination," which restricts the linear embedding to scale along the axes of the input dimensions [12].

[3] Updating a GP involves inverting a covariance matrix, an operation with time complexity $\mathcal{O}(n^3)$, or cubic in the number training examples, but the delay here probably arose from re-initializing the model parameters for every query point. In the proposed, active sampling scheme, each update triggers re-tuning of the parameters, but the model tends to stay lean by approximately optimizing the informational value of the selected examples.

With infinite nodes in a hidden layer, an ANN can approximate arbitrarily complex functions, which may explain their popularity in the VM literature (see survey in [2]). Unless the GP prior incorporates feature selection, perhaps by structuring the covariance function to project inputs to a lower-dimensional space [12], conceptual noise from irrelevant features can render such inferential capacity to be merely theoretical.

Bayesian models naturally offer an estimate of predictive uncertainty, as the priors propagate to form predictive distributions of the output, or in the context of VM, the metric of product quality. Lee et al. [9] suggest that such estimates would facilitate decisions in an industrial setting, such as when to disregard the model's predictions. Note that their model regresses via k-Nearest Neighbor (k-NN), with Gaussian priors on the weights of the neighbors. Each query point triggers a linear reconstruction of its inputs using the known examples' inputs. The weights derive from how heavily an example factors in the query's reconstruction, but this approach ignores the relevance to the output, leading to degraded performance in practical situations.

We propose using the predictive uncertainty from a GPR not only to gauge its credibility, but also to assess the expected information gain associated with measuring a processed wafer's quality. Instead of complicating methods with heuristics, such as those to localize VM models, we confront a prevalent assumption in the VM literature: a fixed (perhaps random) sampling scheme for measuring product quality. Such an assumption limits the VM model to passively learning from a given block of data. Recall that recent work by Bleakie & Djurdjanovic [6] recommends measuring at unscheduled moments to update the trained model, if the raw inputs are "unusually" distant from the known data. The proposed, more principled approach lets the model decide which wafer's quality to measure based on this expected information gain, in order to hasten convergence when starting from no initial data.

Chapter 3

Bayesian modeling approach

3.1 Conceptual background

Bayesian inference offers a consistent way to update prior beliefs by conditioning them on the data [12]. By explicitly acknowledging uncertainty, especially when starting with no initial observations of the output and little to no prior knowledge of the input–output relationship, the process of "learning" from the data becomes more principled. In the context of the VM task, a Bayesian model would naturally express uncertainty in the predicted metric of product quality. Not only does this clarify whether the model is credible, but can also inform which output, given the inputs for a candidate set of examples, is expected to yield the greatest information gain by being observed or measured.

To match the complexity of the dynamics which govern the target metric of quality, we assumed a flexible, Gaussian process (GP) prior and a Gaussian likelihood for the measurement noise, following the technique for regression in [12]. This special case yields exact inference of the predictive, or posterior distribution through Bayes' rule. Since a GP evaluated at any finite set of (input) points take the form of a multivariate Gaussian distribution, GPs exhibit the same elegant properties as the Gaussian distribution itself, such as closure under affine transformations or under convolution with another Gaussian distribution [12]. In effect, GPR models predict a Gaussian distribution for each product's metric of quality, as expressed by the predictive mean and variance. The mean corresponds to the Bayesian estimate that is optimal in the mean squared sense, while two standard deviations above and below the mean bound the 95% credibility interval. The algebraic operations to update the GP prior take a similar form to how one conditions a

Gaussian distribution on observed outputs given the associated inputs.

Algorithm 1 Simultaneous active learning of function and linear embedding, or dynamics and operating space of PECVD data (pseudocode adapted from [10])

```
Require: $m$, $M$; kernel $\kappa$, mean function $\mu$; prior $p(R)$;$X \leftarrow \varnothing$; $Y \leftarrow \varnothing$; $\hat{y} \leftarrow \varnothing$
//Partition $32k$ wafers into 1280 lots of 25 wafers
$lot \leftarrow$ random integer from $[1, 1280]$
$lotsPerRun \leftarrow 100$; $finalLot \leftarrow lotsPerRun + lot$
$initTrain \leftarrow 10$; $initTrainEnd \leftarrow initTrainEnd + lot$
while $lot < finalLot$
    if $lot \geq initTrainEnd$
        $\hat{y} \leftarrow [\hat{y}; q(f(X_{lot}))]$        ▷ predict output for upcoming lot
    repeat on candidate wafers from given lot
        $q(R) \leftarrow$ LaplaceApprox $(p(R \mid X, Y, \kappa, \mu))$
            //approximate posterior on embedding R
        $q(f) \leftarrow$ ApproxMarginal $(p(f \mid R), q(R))$
            //approximate marginal on function f
        $x_* \leftarrow$ OptimizeUtility$(q(f), q(R))$
            //find approximate optimal evaluation point $x_*$
        $y_* \leftarrow$ Observe $(f(x_*))$
        $X \leftarrow [X; x_*]$ ; $Y \leftarrow [Y, y_*]$
    until budget depleted
    $lot \leftarrow lot + 1$
return $\hat{y}$
//Concludes one run of experiment
```

3.2 Mathematical foundation

Recall that the quality characteristic of interest may be insensitive to some of the features extracted from sensor readings, which then act as conceptual noise to the model and hamper standard GP inference [10]. The GP model proposed here learns the lower-dimensional space in which the dynamics operate, approximated by a linear embedding of the original features, while simultaneously learning the latent function mapping equipment signatures (inputs) to product quality (outputs).[1] During the learning process described in Algorithm 1, the embedding may be highly uncertain, which we seek to express in the predictive uncertainty [10].

[1] MATLAB implementation of the algorithmic core, a GP that actively learns a low-dimensional embedding to explain the data, is available from Garnett's repository: `https://github.com/rmgarnett/active_gp_hyperlearning`

3.2.1 Incorporation of prior knowledge

Assuming the target metric of quality tends to a set point (i.e. constant mean value, μ), the GP's covariance function encapsulates our prior knowledge of the structure of the latent function of the underlying dynamics that maps input features to the outputs. Let x and x' refer to inputs of dimensionality M for two examples, and assume a linear embedding, $R \in \mathbb{R}^{m \times M}$ with $m \ll M$, such that the covariance between the examples may be expressed as

$$K(x, x') = \gamma^2 exp(-\frac{1}{2}(x - x')R^T R(x - x')^T). \tag{3.1}$$

In other words, we modify the well-known squared-exponential kernel [12] to project the model inputs into a lower dimensional space through a linear mapping, R. Note that the resulting GP can still regress nonlinear interactions between these linearly embedded features that constitute a given input. The rows of R correspond to the directions of the most rapid change in the function, and the Euclidean length of each projection corresponds to its relevance, or inverse length scale [12]. For the purposes of inference, examples more than a few input length scales apart along a given direction in R are perceived as uncorrelated, or their observed outputs seen as irrelevant to each other.

3.2.2 Tuning of parameters

As the model actively selects which wafers to physically measure so as to inform R and learn the latent input–output relationship, we seek to maximize the likelihood of the model with respect to the entries of R and other parameters of the GP prior. Let all these parameters[2] be denoted by a vector, θ. Then let $y \in \mathbb{R}^{N \times 1}$ be the noisy measurements of the metric of product quality, and let the associated inputs be $X \in \mathbb{R}^{N \times M}$, such that the

[2] In deriving a GP, the underlying parametric model's weights are integrated out, so the tunable parameters of the GP prior are referred to as *hyperparameters* to emphasize the non-parametric nature of this approach.

marginal likelihood of the model for the given parameter values may be expressed by

$$\log p(y \mid X, \theta) = -\frac{(y-\mu)^T V^{-1}(y-\mu)}{2} - \frac{\log \det V}{2} - \frac{N \log 2\pi}{2}, \tag{3.2}$$

in which $V = K(X,X) + \sigma^2 I$ corresponds to the covariance of y for inputs X, and σ^2 is the variance of the measurement noise [12]. The first term of equation (3.2) penalizes poor fit to the data, whereas the second term penalizes the complexity of the model. In effect, this process of tuning the model's parameters by maximizing (3.2) balances its accuracy and simplicity.

3.2.3 Prediction under uncertain embedding

To acknowledge uncertainty in R, let us assume a Laplace approximation on $p(R)$:

$$p(R) = \mathcal{N}\Big(R; \hat{R}, \Sigma_R\Big) \tag{3.3}$$

$$\Sigma_R = \left(-\nabla\nabla p(R)|_{R=\hat{R}}\right)^{-1}, \tag{3.4}$$

in which $\hat{R}$ refers to the maximum likelihood estimator (MLE) and Σ_R represents the inverse Hessian about the MLE. The GP then regresses the target metric of quality on $u = xR^T$, or a linearly transformed Gaussian expressed as $p(u) = \mathcal{N}\Big(xR^T; x\hat{R}^T, x\Sigma_R x^T\Big)$. The perceived inputs, u are uncertain, but correlated and Gaussian-distributed through R [10].

When estimating the predictive distribution for a test point, or $f_* = f(x_*)$, we wish to integrate out this uncertainty in the parameters θ, of which entries of R constitute a majority:

$$p(f_* \mid D) = \int p(f_* \mid D, \theta)\, p(\theta \mid D)\, d\theta, \tag{3.5}$$

where D denotes the data. However, such an integral cannot be directly evaluated due to its intractability arising from the nonlinear dependence of f on R. In order to approximate (3.5), we rely on the method devised by Garnett et al. [10]: first linearly approximate the dependence of $p\left(f_* \mid D, \theta\right)$ on θ, then match moments to those of the exact integral.[3] This technique inflates the predictive variance based on the uncertainty in θ, as expressed by

$$\tilde{V}_{f|D}\left(x_*\right) = \frac{4}{3}\hat{V}\left(x_*\right) + \left(\frac{\partial \hat{m}\left(x_*\right)}{\partial \theta}\right)^T \Sigma \left(\frac{\partial \hat{m}\left(x_*\right)}{\partial \theta}\right) + \\ + \frac{1}{3\hat{V}\left(x_*\right)}\left(\frac{\partial \hat{V}\left(x_*\right)}{\partial \theta}\right)^T \Sigma \left(\frac{\partial \hat{V}\left(x_*\right)}{\partial \theta}\right), \quad (3.6)$$

in which $p\left(\theta \mid D\right) = \mathcal{N}\left(\theta; \hat{\theta}, \Sigma\right)$ and $p\left(f_* \mid D, \theta\right) = \mathcal{N}\left(f_*; \hat{m}\left(x_*\right), \hat{V}\left(x_*\right)\right)$. Note that though the predictive variance yielded by this marginal GP (mGP) has increased, the mean remains the same (i.e. the value at $\hat{\theta}$, the MLE).

3.2.4 Active learning of embedding

Let us leverage this predictive uncertainty to hasten the convergence of the model to a credible form. In particular, we want to take physical measurements of product quality for cases that are expected to be the most informative, not only about the dynamics' latent function but also of its linear embedding, which facilitates this GP framework to correlate examples and generalize to unseen ones [12]. To assess the perceived information gain, or utility of observing a product's quality, we employ *Bayesian active learning by disagreement* (BALD) [11], which assigns utility based on the mutual information, $I\left(F; \Theta\right)$ between the latent function, F and the hyper–parameters, Θ. Mutual information, or relative entropy, may be interpreted through Kullback-Leibler divergence, as seen in the

[3] Since the dependence of the latent function, f on the parameters, θ tends to be well-concentrated [15], approximating this dependence as a line through the MLE, $\hat{\theta}$ seems reasonable.

utility function:

$$\begin{aligned} \nu(x) &= I\left(F;\Theta\right) = D_{KL}(p(F,\Theta) \parallel p(F)p(\Theta)) \\ &= \mathbb{E}_{\Theta}\left(D_{KL}(p(F \mid \Theta) \parallel p(F))\right) = H(F \mid \Theta) - H(F), \end{aligned} \tag{3.7}$$

in which H refers to the entropy, or log-variance for Gaussian distributions. Essentially, the greatest information gain would be expected at the candidate point for which the standard predictive variance and inflated version, from $p(F \mid \Theta)$ and $p(F)$ respectively, diverge or "disagree" the most.[4]

Since the GP framework yields predictions in the form of Gaussian distributions, the perceived utility of candidate points to measure may be readily approximated by a ratio of predictive variances using the mGP [10]:

$$\nu'(x) = \frac{V_{f|D}(x)}{V_{f|D,\hat{\theta}}(x)}. \tag{3.8}$$

Garnett et al. [10] contrast this active sampling scheme with a more myopic one called uncertainty sampling, which arises from a utility function that uses only the standard predictive variance: $\tilde{\nu}(x) = V_{f|D,\hat{\theta}}$. Basically, by observing outputs that are expected to inform about an embedding, the predictive variance tends to shrink for all points, not just at the observed one.

[4] The invariance of mutual information to homeomorphic transformations [16] make this learning process robust to scaling or rotating of the current embedding, or to centering and scaling of the inputs and outputs.

Chapter 4

Background of data from representative plasma process

4.1 Plasma-enhanced chemical vapor deposition

We apply the framework for GPR detailed in the previous chapter to data collected from a major semiconductor fabrication plant, in order to predict the thickness of films created by a plasma-enhanced chemical vapor deposition (PECVD) process. In chemical vapor deposition, carrier gases containing volatile precursors decompose on the wafer substrate to yield a thin film of insulating or conducting material. To exclude contaminating particles, the process is enclosed in a vacuum chamber, which can be pumped down to low pressure, typically 0.1–10 Torr, in order to enhance the film's density and purity [1]. Energy for the reaction can be provided by heat or electrons from a plasma (see Fig. 4.1). Unlike directional deposition techniques, which may sputter or condense pure materials onto the substrate, the probabilistic reactions of chemical vapor deposition tend to apply a coat that conforms to the sharp corners of trenches and vias that comprise the etched topology [1]. Based on the gas temperature and pressure, the volatile precursors can bounce from 10 to 10,000 times before sticking and decomposing [1]. Such uniform coatings prove practical for insulating microelectronics, such as transistors, from each other and those on adjacent layers.

The concentrations of reactive species and their average kinetic energy dictate the rate of film growth on the wafer [1]. In the plasma-enhanced version of this process, high energy electrons collide and excite the reactive species, so measuring the environment within the plasma tends to be difficult. Invasive probes may perturb the plasma, or degrade due to deposition on (or etch of) the probes' surfaces [26]. Therefore measuring con-

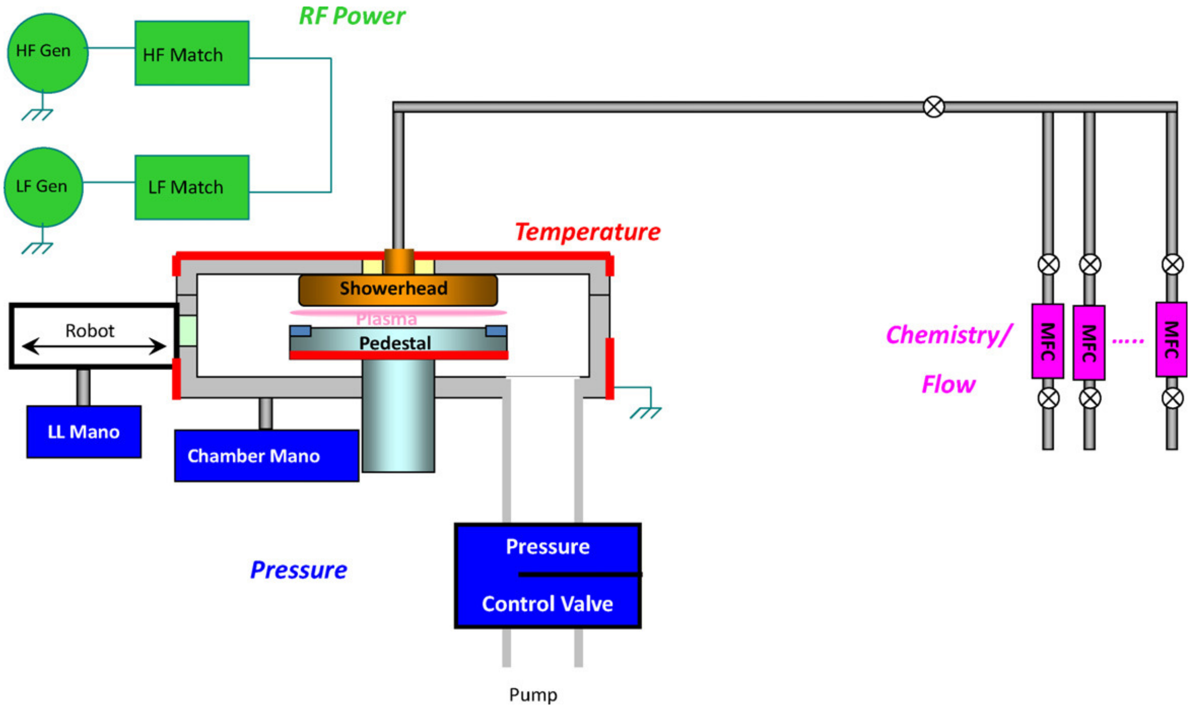

Figure 4.1: Schematic diagram of a PECVD system (reprinted from [5] with permission).

ditions at the most important junction, the wafer's surface where components form, has not yet been proven feasible [26]. However, through historical experience, manufacturers have largely stabilized these processes by strictly adhering to task-specific "recipes," or a series of set-points for the controllable variables [2].

The particular process analyzed here employs tetraethoxysilane (TEOS) that, when sufficiently excited, decomposes to form an insulating film of SiO_2 on the silicon substrate. The data gathered from the PECVD tool indirectly describe these dynamics through 11 sensors sampled at 10Hz. The associated traces describe temperatures at the wafer pedestal and chamber walls, pressure, gas flows into the chamber, and the power and voltage associated with the radio-frequency (RF) system that energizes the plasma [5]. As illustrated in Fig. 4.1, the showerhead delivers the reactive gases above the substrate, at a rate governed by the valve in the mass flow controller (MFC). Although the plasma supplies sufficient energy for the volatile precursors to decompose at room temperature, typically the wafer and walls of the chamber are heated up to 300–400°C to minimize the number of defects [1]. Note that without the plasma, the wafer would require heating to 700–900°C to make the decomposition energetically favorable [1]. Operating at lower temperature decreases the risk of heating the (300mm) wafer too quickly, in which case it temporarily bows, and the deposited film may stretch and crack upon relaxation.

Capacitive plates above and below the wafer supply a radio frequency (RF) signal to excite the gas, leading to an electron cascade that stabilizes into a plasma. In this case, a high and low frequency signal [21] was applied in order to decouple adjusting the plasma density and the ion bombardment energy; bombarding ions tend to alleviate tensile stresses associated with rapid film growth [1]. To maximize power delivered to the plasma and minimize that reflected to the generator, the RF matching system tweaks the voltages of the load and tune capacitors, respectively [25].

The pressure within the chamber contributes to the film's uniformity over the wafer's surface, or how the film conforms to sharp corners. Volatile by-products and other gases evacuate from the chamber at a rate mediated by the exhaust valve angle. Although the wafers may be tagged by batch or lot, as they arrive to load-lock in cassettes, each one undergoes individual processing upon delivery to the chamber.

4.2 Context of data from PECVD tool

As film thickness may be measured in-line by an ellipsometer [20], the data comprises measurements of mean wafer thickness (MWT) for each of the ~32,000 wafers processed over the span of several months. The TEOS-based recipe required a few minutes to execute, during which the 11 sensors sampled the given chamber's environment at 10Hz. We therefore condense the raw sensor traces into a set of features, describing each sensed parameter's steady-state and transient behavior, as well as transition times [5]. This recipe comprises two main deposition steps, and yielded 49 features (see list in Appendix). We denote the data, $D = (X, y)$ as the set of input–output pairs in which a wafer's input, x, comprises this set of 49 features. We do not assume to know which features are relevant prior to modeling. As mentioned earlier, the possibility that the film thickness may be insensitive to some of these features, yet they interact nonlinearly to dictate growth rate, motivated the integration of feature selection into the Bayesian model.

Over the time-span of the data, the tool failed four diagnostic tests, signaling faulty behavior that required the tool to shut down for repair. These faults correspond to bimodal thickness across the wafer, films exhibiting unacceptable range in thickness, particle contamination of the chamber, and Coulomb crystal formation in the chamber [5]. We let the model cross preventative maintenance events, such as the in-situ cleans. Though not as thorough as wet cleans, in which an operator opens the chamber and wet wipes the exposed surfaces, these in-situ cleans conducted every 25-100 wafers tend to remove residue that accumulates on the hot walls of the chamber [5], which can flake or otherwise disturb the plasma [1].

Chapter 5

Results

5.1 Description of underlying experiment

As if in a factory with no initial data, the following experiment simulates predicting and updating online under the constraints of a measurement budget. Let us assume only one one wafer's dimensions can be "measured" or observed per process lot of 25 wafers, a typical constraint for expensive-to-measure metrics of quality, like critical dimensions from plasma etch [4,7]. By conducting this study on a large PECVD dataset, we can thoroughly evaluate the convergence of the predictive models to credible forms.

As outlined in Algorithm 1, we first partition the data into sets of 25, yielding approximately 1280 lots, then from a random initial wafer/lot: predict one-lot-ahead, select the approximately optimal wafer from the next lot for which to measure the output MWT, update the model, and repeat for 100 consecutive lots. We conducted at least 50 of these 2500-wafer runs (see two examples of such runs in Fig. 5.1), excluding those that would cross a fault or a 150-wafer buffer before the fault was flagged.

Prior to modeling, we transformed the inputs, x to a box-bounded region of $[-1, 1]^D$, and normalized the outputs, y to zero mean and unit variance. The given framework for GPR approximately marginalizes the uncertainty in the parameters as part of the active sampling scheme, which tends to be more effective if corresponding priors are specified. For each element of the linear embedding, we assigned a diffuse prior, following [10]. This initial belief of a zero-mean i.i.d. Gaussian distribution with a standard deviation of $\frac{5}{4D}$ merely favors low magnitude values. Furthermore, it transforms the box-bounded inputs to $[-2.5, 2.5]^d$, which is a relatively broad domain about five (input) length scales

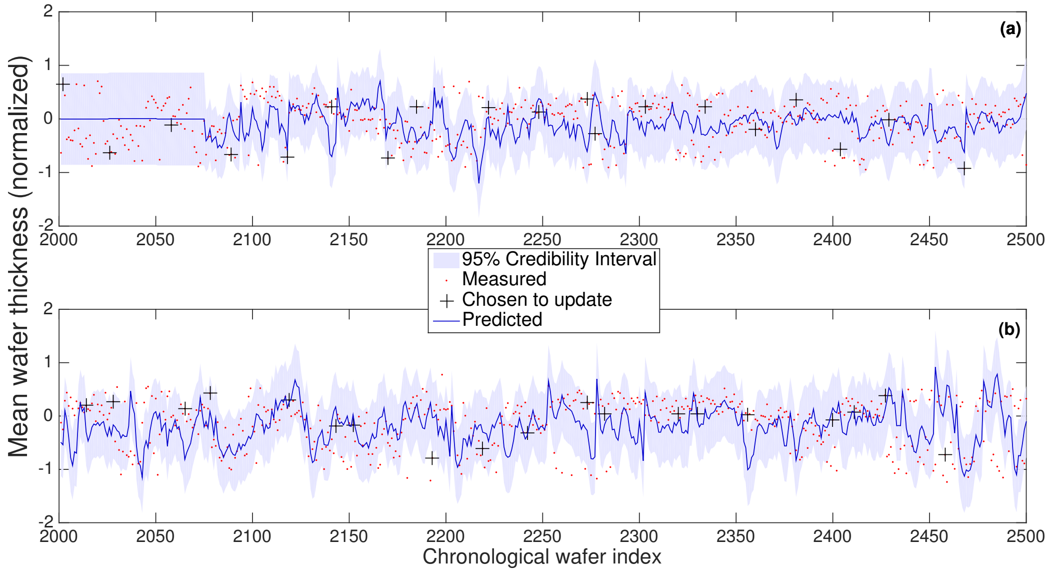

Figure 5.1: Measured mean wafer thickness compared to that predicted by GPR using an active sampling scheme, during last 20% of two runs. To simulate online setting of a factory, GPR predicts one lot of 25 wafers ahead, selects the approximately optimal wafer to measure, updates, and repeats for 100 consecutive lots. Note that predictions in subplot (a) tend to be conservative before the embedding converges to a plausible form.

long [10].[1] The remaining priors reflect assumptions of zero-mean measurement noise with relatively small variance, a constant-mean output, and an output scale, γ (from eqn. 3.1) that tended to be slightly below unity, the standard deviation of the normalized output.

When updating the GPR model, we estimated the parameters of the covariance and mean functions by minimizing the negative marginal log-likelihood via limited-memory BFGS [27]. Each search begins at the most likely (i.e. *maximum a posteriori*) values yielded by the previous update. In order to escape local minima, particularly those related to parameters in the embedding matrix R, we alloted at least one random restart from the prior, permitting two restarts for the GPR model using an active sampling scheme.

[1] Although the inputs, x were scaled to $[0,1]^D$ by omitting the final center-and-scale step, the GP regresses on xR^T, so the embedding can compensate. The diffuse prior on R helps by roughly mapping the inputs to $[-2,2]^d$, a domain four length scales long.

We compared the accuracy of GPR to that of partial least squares (PLS), the "baseline" linear model often used for VM purposes in industry [2]. Note that such a (non-Bayesian) linear model only predicts a point estimate, so it cannot use its predictive uncertainty to inform where to measure next. Therefore, we let PLS update using a random wafer from each lot. We also maintained an auxiliary GPR model that updates on the same set of random wafers measured thicknesses.

In order to evaluate each model's convergence, we considered how the accuracy or in the case of GPR, the marginal likelihood improves over time. Given these metrics represent predictive performance in an online scenario, we constructed each *learning curve* not by summarizing the improvement on a designated test set, but on all subsequent lots of wafers from a given point in the run. Note that averaging the metrics over these upcoming lots tends to smooth the curves backwards in time, but renders them more comparable across trials of the experiment.

Recall that the process may have drifted as discrete shifts in behavior accumulated between the known faults [5, 7, 8], so we checked for runs with unusually poor convergence. The threshold for "far outlying" examples, as suggested by Tukey [17], lies three interquartile ranges ($Q_3 - Q_1$) below the first quartile, Q_1 or above the third quartile, Q_3. For example, if predictive errors were Gaussian distributed, then these far outlying results would lie ~4.7 standard deviations beyond the mean.

5.2 Learning curves

Process drift and other non-stationary behavior may hamper modeling the dynamics during certain time periods, so 57 rather than the 50 nominal runs were executed. In order to check for runs with unusually poor convergence, we examined learning curves based on predictive accuracy, which both the Bayesian (GP) and non-Bayesian (PLS) mod-

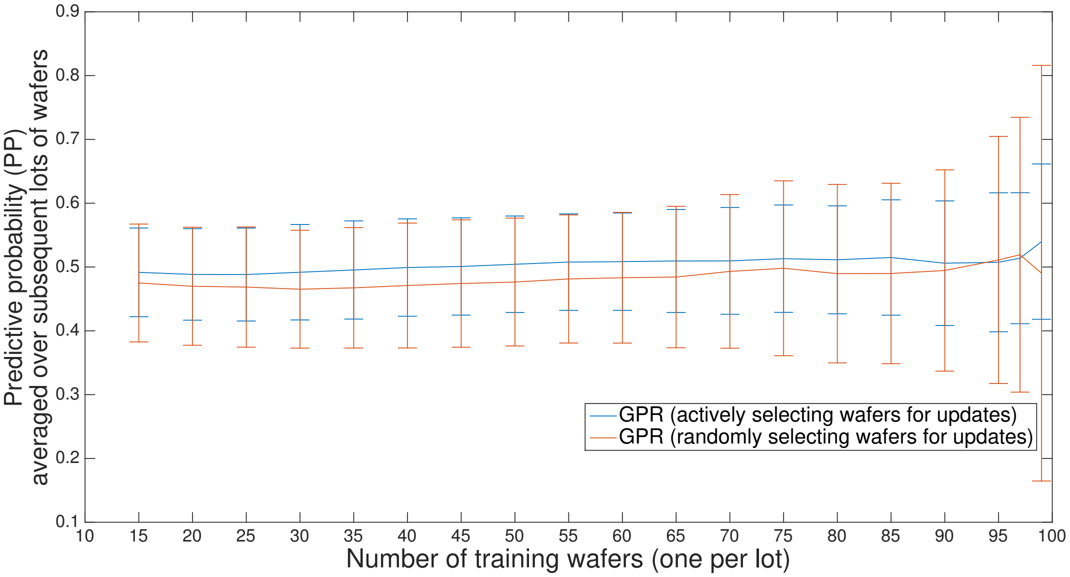

Figure 5.2: Learning curve for predictive probability, averaged over 52 runs of online scenario sketched in Alg. 1. The tighter standard deviation bars for the GP using an active sampling strategy suggest a more robust strategy for learning from the tightly controlled process' data.

els produce. Furthermore, we assumed outliers may be discerned based on root-mean-squared error (RMSE) during the last 15% of the run, or after the model has received 85 observations of the mean wafer thickness. Note that RMSE more heavily penalizes large predictive errors than mean absolute error (MAE), and typically yields clearer outliers. PLS flagged five runs as exhibiting "far outlying" RMSE, and subsets of those five were flagged by GPR given random wafers (four) and GPR using the active sampling scheme (three). Curiously, the four runs on which the GPR models and PLS agreed were temporally clustered before a fault, suggesting the process may have been destabilizing during that period.

Let us first consider learning curves exclusive to the Bayesian models. GPR predicts a Gaussian distribution for each wafer's MWT, so we may evaluate the predictions' accuracy and confidence in terms of the predictive probability (or marginal likelihood for

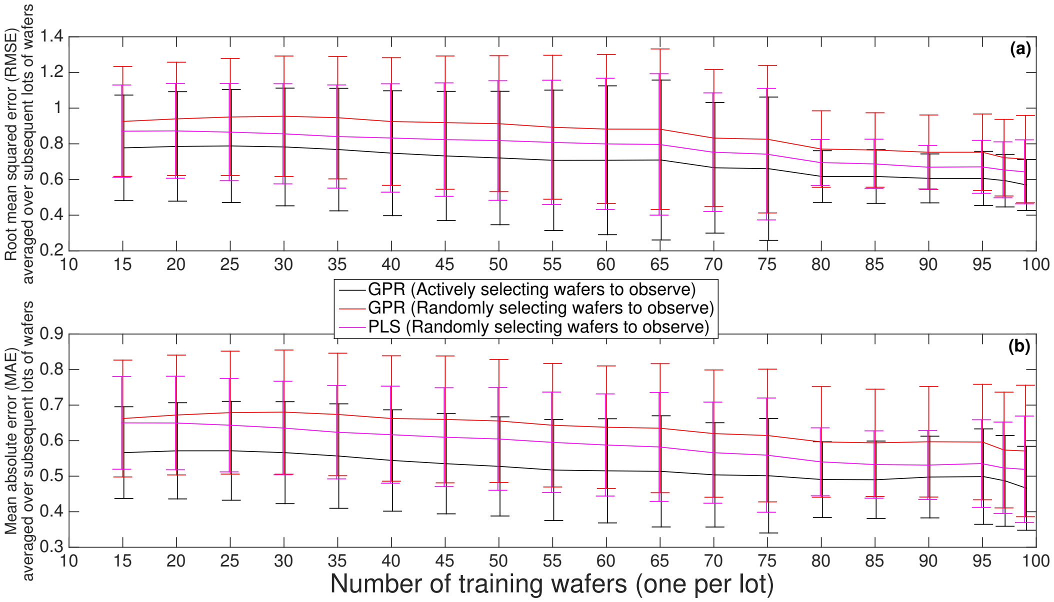

Figure 5.3: Learning curve for predictive inaccuracy, averaged over 52 runs of online scenario sketched in Alg. 1. Note that MAE $\leq$ RMSE, though these metrics share units of output standard deviations. By letting the GP inform which wafer to measure in the upcoming lot, it tends to obtain sufficient information to outperform PLS' linear approximation to the nonlinear dynamics.

test cases); its logarithm takes a simple form:

$$\log p\left(y_{*} \mid X\right) = -\frac{\left(y-\mu\right)^{2}}{2\sigma^{2}} - \frac{\log 2\pi\sigma^{2}}{2}, \tag{5.1}$$

in which μ and σ^2 correspond to the predictive mean and variance, respectively [12]. Fig. 5.2 summarizes the predictive probability during the remaining 52 trials for the two GPR models employing different sampling schemes. Note the greater standard deviation bars for the GPR model that updates on MWTs from random wafers, illustrating its struggle to extract sufficient information, as manifested in several runs with very low predictive probability. The tighter bars associated with GPR using BALD indicates a more robust strategy to learn from the seemingly repetitive dataset.

As PLS only offers a point estimate, let us compare its performance to that of the

GPR models on the basis of inaccuracy alone. Fig. 5.3 summarizes their convergence in terms of RMSE and MAE over the 52 remaining trials of the experiment. Since dimensions beyond the tolerances incur disproportionately greater cost, RMSE would be a more appropriate metric for this task, as it more heavily penalizes large predictive errors than metrics of absolute error do. The MAE and RMSE in Fig. 5.3 share units of standard deviations, σ of the SiO_2 thickness. As for the relative trends, when updating on MWTs from the same random wafers, GPR proves less accurate than the linear model. However, by letting the predictive uncertainty inform which wafer to measure next, GPR becomes about 10% more accurate on average than PLS, probably due to the nonlinear nature of the underlying dynamics.

5.3 Investigating process drift

To understand why some runs failed to converge, let us consider the abrupt shifts in process behavior that may factor in long-term drift [7]. These shifts may arise in various ways: a human operator could tune the recipe by adjusting set-points for temperature, pressure, or gas flows, the RF system could compensate for degrading chamber conditions (e.g. residue build-up on walls [1]), or the tool could undergo preventative maintenance, which may involve opening the vacuum chamber. As a result, the operating space, or relevance of the features extracted from sensor readings may drift as well [6–8]. However, each GPR model's linear embedding reflects the assumption of a consistent operating space over its 2500 wafer span.

In order to assess the degree of degradation or instability in the process, we examined a metric based on how "unusual" the relevant input features appear over time. During periods of persistent drift, a single linear embedding may be inadequate, so we seek to track how the operating space evolves by leveraging the embeddings of all of the converged trials. We leverage the parameters estimated by the end of each run to construct a T^2 statistic, which summarizes the embedded features for each of the 2500 wafers from which the associated GPR model actively sampled to train. Let us express this statistic as

$$T^2 = \sum_{i=1}^{m=7} \frac{s_i^2}{\mathrm{Var}\left(s_i\right)}, \tag{5.2}$$

where s_i refers to the ith "score" [18], or linearly embedded feature of the GPR model. The magnitude of the metric defined by (5.2) expresses how atypical the embedded features appear for a given wafer. Where runs overlap, we weight the mean T^2 based on how well each GPR explains the data during this period, or by its marginal likelihood (see eqn. 3.2). Given that the data has been partitioned into lots, we assigned each one a separate weight. Furthermore, we used the measured dimensions from all 25 of those wafers to evaluate the how well a given GPR explained that period's data.

We weight the variance of each score based on the marginal likelihood as well. Let us restrict the weights across the runs for a given wafer to sum to unity, and let the ith score for wafer *t* derived from the final model of run *j* be denoted $s_{i,j,t}$. Then, summing over all wafers from fault to fault, the weighted variance may be expressed by

$$\mathrm{Var}\left(s_i\right) \approx \tilde{\sigma}_i^2 = \frac{\sum_t \sum_j w_{j,t} \left(s_{i,j,t} - \hat{s}_i\right)^2}{\sum_t \sum_j w_{j,t}} \tag{5.3}$$

$$\hat{s}_i = \sum_t \sum_j w_{j,t} s_{i,j,t}, \tag{5.4}$$

where $\hat{s}_i$ refers to the weighted mean of the i^th score. Although the number of runs that fall between faults may vary (e.g. only two occur between the first two faults, of bimodal thickness and unacceptable range in thickness), all wafers assigned a T^2 during that period use the same set of weighted variances to approximately normalize their squared scores.

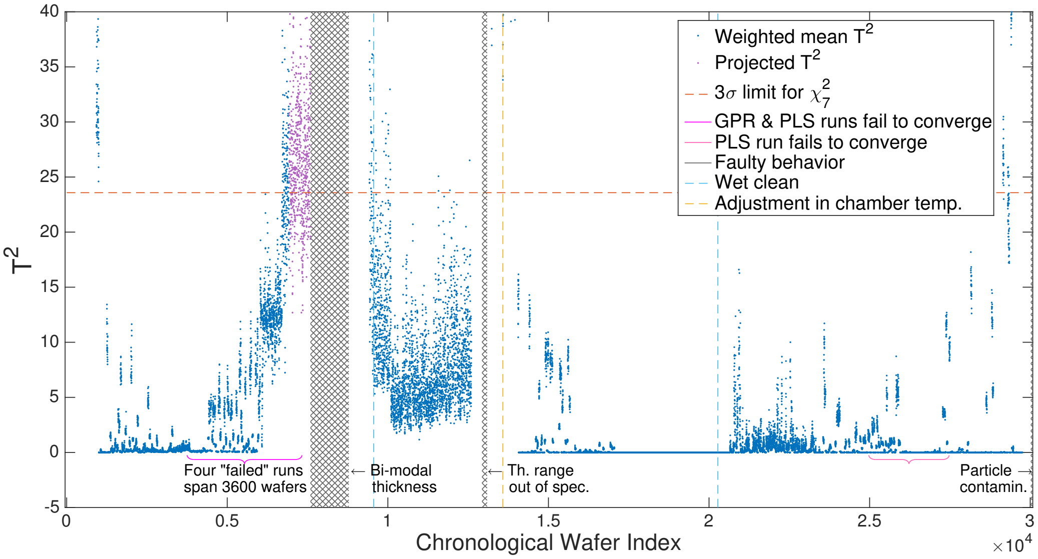

Figure 5.4: Statistic describing how "unusual" the relevant, embedded inputs of each wafer appear during the 52 trials randomly spread over the several months the data spans. This metric approximates the degree of degradation or instability in the process, as the inputs compensate. Although trials were restricted to not cross faults, they can overlap, at which points the mean T^2 is weighted by how well the associated models' explain the data, or by the marginal likelihood (eqn. 3.2).

Note that this method for approximating the variance of each score assumes the rows of the embedding matrix, R are ordered and therefore the scores are comparable across runs. We find this assumption to be plausible not merely because the trends in the resulting T^2 are interpretable in the historical context of the data, but also because the given implementation [28] permits symmetry to break between the rows of R. In particular, R was restricted to an upper triangular form, so provided the features on the diagonal were sufficiently relevant and distinct, symmetry should break. As listed in the Appendix, these first $m - 1$ features correspond to the temperature of the separately heated pedestal and walls of the chamber.

Figure 5.4 shows this proxy metric for how degraded or unstable the process appears over the span of the data's history. Note that the four runs for which the GPR models and PLS exhibited “far outlying” RMSE temporally cluster within a 3600 wafer span before the bimodal thickness fault. We assign a three-sigma (i.e. 99.7%) limit [19] for significantly unusual behavior based on a χ^2_m distribution, which the T^2 theoretically follows. Note that for the bimodal thickness fault, this metric detects the onset of consistently unusual behavior as much as two days before fault was flagged.

In the wake of major repairs associated with faults, the process may gradually recover, or return to more typical behavior as human operators clean the chamber and tweak the recipe. For example, the process seems to have stabilized soon after the wet clean that followed the final repairs of the bimodal thickness fault. A similar phenomenon may be seen soon after a technician corrected the shift in the chamber’s temperature following the repairs for an unacceptable range in wafer thickness.

Looking towards the end of the data’s history, we see sporadic periods of unusual behavior that precede the particle contamination fault. These moments of seeming instability may correspond to initial stages of particle formation. Further investigating the root cause of the faulty behavior, perhaps by decomposing the T^2 into percent by feature, lies beyond our scope.

Chapter 6

Concluding discussion

We find GPR can approximate the nonlinear dynamics that govern metrics of product quality, such as mean film thickness in a PECVD process, by approximately optimizing the information extracted from a tightly controlled manufacturing process. This strategy for predicting microchip quality leverages the framework devised and implemented by Garnett et al. [10], in which a marginal GP actively learns the underlying dynamics as well as the embedded space in which the dynamics operate. This Bayesian method yields interpretable parameters, as well as clear ways to assess convergence and uncertainty in the prediction, thereby estimating the probability that a wafer's quality characteristics are within tolerances rather than just offering a "best guess" of their value. In addition to providing greater predictive accuracy than a comparable linear model, the GP offers a integrated estimate of how relevant the features are to the target metric of quality, as conveyed by the linear embedding. We found that by observing the process inputs through the lens of this embedding, alarming trends can be perceived before conventional diagnostics declare a fault.

In order to integrate this approach into a factory setting, this strategy should also predict and decide which wafer's quality to physically measure in real-time. On a 2.3GHz Intel i7 processor with 8GB of RAM, predicting 25 wafers ahead and selecting which one to measure next required about 20–25s, then updating the ~350 parameters required another 20–35s. Deciding which wafer to measure next may be made linear rather than quadratic in the number of features, M, as algorithmically detailed in the "Computational Cost" section of [10]. Though not yet implemented, such code could select the approximately optimal wafer to measure in less than a second. Depending on the process and its given time constraints for each lot, one could also allot a larger initial measurement budget to hasten convergence.

It remains unclear whether the proposed model achieves sufficient accuracy or credibility to fine-tune control of the process, thereby preventing yield loss [2–4]. To choose how to adjust the inputs in a GP framework, one could select those that optimize the expected improvement (EI) [14] towards the desired thickness, constrained by the control limits for the inputs. When comparing the predicted values to those actually measured (see e.g. those prediction traces in Fig. 5.1), we find that values often lie within the 95% credibility interval. However, when considering accuracy alone, the GP achieves a MAE of ~0.5σ after training on 100 wafers, and the model roughly estimates the measurement noise as 0.15–0.2σ, so perhaps 30% of the dynamics that govern the average film thickness remain unexplained.

Recall that the rate of deposition for the given process depends on the concentration of TEOS with sufficient energy to decompose on contact with Si to form SiO_2. None of the sensors in this study directly describe the *active* species' concentrations within the plasma. However, these concentrations can spike to unpredictable levels when igniting the plasma over a given wafer, which may partly explain the nano-scale deviations in these critical dimensions [24]. Perhaps by incorporating spectroscopy to estimate the concentrations of the excited and radical species [22, 23], these plasma-based dynamics may be further understood [2].

Despite the difficulties of maintaining acceptable yield, this industry will continue to try fitting more processing power or memory onto a chip, by shrinking critical IC dimensions. Whether the width of a gate created by plasma etch, or the thickness of a film deposited in a gate stack [20], these dimensions are approaching the limits of what may be currently measured. In order to handle this uncertainty and ensure quality, applying such Bayesian methods to the task of virtual metrology may become more and more necessary.

Appendix

Table A1: List of features derived from PECVD sensor traces. Dep1 & 2 refer to the two steps of the process.

Chamber temperature 1	1. Mean
	2. Range
Chamber temperature 2	3. Mean
	4. Range
Pedestal temperature 1	5. Mean
	6. Range
Pedestal temperature 2	7. Mean
	8. Range
HF reflected power	9. Mean Dep1
	10. Mean Dep2
	11. Range Dep1
	12. Range Dep2
	13. Trigger-time Dep1
	14. Trigger-time Dep2
LF reflected power	15. Mean Dep1
	16. Mean Dep2
	17. Range Dep1
	18. Range Dep2
	19. Trigger-time Dep1
	20. Trigger-time Dep2
Flow rate of TEOS	21. Mean
	22. Range
	23. Over-shoot
	24. Rise-time
Load capacitor voltage	25. Mean Dep1
	26. Range Dep1
	27. Max Dep2
	28. Range Dep2
Chamber pressure	29. Rise-time of pump-up
	30. Fall-time of pump-down
	31. Peak
	32. Mean Dep1
	33. Range Dep1
	34. Rise-time Dep2
	35. Mean Dep2
	36. Range Dep2
	37. Min
Pendulum valve angle	38. Rise-time of pump-up
	39. Max of pump-up
	40. Mean of pump-up
	41. Range of pump-up
	42. Mean Dep2
	43. Range Dep2
	44. Max of pump-down
	45. Mean post-process
Tune capacitor voltage	46. Mean Dep1
	47. Range Dep1
	48. Max Dep2
	49. Range Dep2

Bibliography

[1] McInerney, Edward J. "Chemical Vapor Deposition." Semiconductor manufacturing handbook. Ed. Hwaiyu Geng. McGraw-Hill, Inc., 2005.

[2] Ringwood, John V., et al. "Estimation and control in semiconductor etch: Practice and possibilities." IEEE Transactions on Semiconductor Manufacturing 23.1 (2010): 87-98.

[3] Yang, Yang, Mingmei Wang, and Mark J. Kushner. "Progress, opportunities and challenges in modeling of plasma etching." Interconnect Technology Conference, 2008. IITC 2008. International. IEEE, 2008.

[4] Su, An-Jhih, et al. "Control relevant issues in semiconductor manufacturing: Overview with some new results." Control Engineering Practice 15.10 (2007): 1268-1279.

[5] Bleakie, Alexander, and Dragan Djurdjanovic. "Feature extraction, condition monitoring, and fault modeling in semiconductor manufacturing systems." Computers in Industry 64.3 (2013): 203-213.

[6] Bleakie, Alexander, and Dragan Djurdjanovic. "Growing Structure Multiple Model System for Quality Estimation in Manufacturing Processes." IEEE Transactions on Semiconductor Manufacturing 29.2 (2016): 79-97.

[7] Hirai, Toshiya, and Manabu Kano. "Adaptive virtual metrology design for semiconductor dry etching process through locally weighted partial least squares." IEEE Transactions on Semiconductor Manufacturing 28.2 (2015): 137-144.

[8] Lynn, Shane A., John Ringwood, and Niall MacGearailt. "Global and local virtual metrology models for a plasma etch process." IEEE Transactions on Semiconductor Manufacturing 25.1 (2012): 94-103.

[9] Lee, Seung-kyung, Pilsung Kang, and Sungzoon Cho. "Probabilistic local reconstruction for k-NN regression and its application to virtual metrology in semiconductor manufacturing." Neurocomputing 131 (2014): 427-439.

[10] Garnett, Roman, Michael A. Osborne, and Philipp Hennig. "Active learning of linear embeddings for Gaussian processes." arXiv preprint arXiv:1310.6740(2013).

[11] Houlsby, Neil, et al. "Bayesian active learning for classification and preference learning." arXiv preprint arXiv:1112.5745 (2011).

[12] Rasmussen, Carl Edward. "Gaussian processes for machine learning." (2006).

[13] Neal, Radford M. Bayesian learning for neural networks. Vol. 118. Springer Science and Business Media, 2012.

[14] Osborne, Michael A., Roman Garnett, and Stephen J. Roberts. "Gaussian processes for global optimization." 3rd international conference on learning and intelligent optimization (LION3). 2009.

[15] MacKay, David JC. "Comparison of approximate methods for handling hyperparameters." Neural computation 11.5 (1999): 1035-1068.

[16] Kraskov, Alexander, Harald Stögbauer, and Peter Grassberger. "Estimating mutual information." Physical review E 69.6 (2004): 066138.

[17] Tukey, John W. "Exploratory data analysis." (1977): 2.

[18] Wold, Svante, Michael Sjöström, and Lennart Eriksson. "PLS-regression: a basic tool of chemometrics." Chemometrics and intelligent laboratory systems 58.2 (2001): 109-130.

[19] Klein, Morton. "Two alternatives to the Shewhart X control chart." Journal of Quality Technology 32.4 (2000): 427.

[20] Hilfiker, James N., et al. "Survey of methods to characterize thin absorbing films with spectroscopic ellipsometry." Thin Solid Films 516.22 (2008): 7979-7989.

[21] Van de Ven, Evert P., I-W. Connick, and Alain S. Harrus. "Advantages of dual frequency PECVD for deposition of ILD and passivation films." VLSI Multilevel Interconnection Conference, 1990. Proceedings., Seventh International IEEE. IEEE, 1990.

[22] Granier, A., et al. "Optical emission spectra of TEOS and HMDSO derived plasmas used for thin film deposition." Plasma Sources Science and Technology 12.1 (2003): 89.

[23] Harvey, Kenneth, et al. "Method for Real Time Monitoring of Gas Composition with High Sensitivity Using Optical Emission Spectroscopy." AEC/APC Symposium XIX, Indian Wells, CA, USA. 2007.

[24] Nomura, Kazuhiro, et al. "Virtual metrology of Dry Etching Process Characteristics Using EES and OES." Proceedings of the AEC/APC Symposium Asia. 2011.

[25] Chen, Francis F. "Capacitor tuning circuits for inductive loads." UCLA Report(1992).

[26] Sobolewski, Mark A. "Real-time, noninvasive monitoring of ion energy and ion current at a wafer surface during plasma etching." Journal of Vacuum Science and Technology A: Vacuum, Surfaces, and Films 24.5 (2006): 1892-1905.

[27] Schmidt, Mark. "minFunc: unconstrained differentiable multivariate optimization in Matlab." Software available at http://www.cs.ubc.ca/ schmidtm/Software/minFunc. html (2005).

[28] Rasmussen, Carl Edward, and Hannes Nickisch. "Gaussian processes for machine learning (GPML) toolbox." Journal of Machine Learning Research 11.Nov (2010): 3011-3015.