NATIONAL UNIVERSITY OF IRELAND, MAYNOOTH

NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

# Virtual Metrology for Plasma Etch Processes

Shane Lynn

A thesis submitted in partial fulfillment for the degree of

Doctor of Philosophy

in the

Faculty of Science and Engineering

Electronic Engineering Department

Supervisor: Prof. John V. Ringwood

Head of Department: Dr. Seán McLoone

April 2011

# Declaration of Authorship

I, Shane Lynn, declare that this thesis titled, 'Virtual Metrology for Plasma Etch Processes' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

*"Twenty years from now you will be more disappointed by the things that you didn't do than by the ones you did do. So throw off the bowlines. Sail away from the safe harbor. Catch the trade winds in your sails. Explore. Dream. Discover."*

Mark Twain

# *Abstract*

Plasma processes can present difficult control challenges due to time-varying dynamics and a lack of relevant and/or regular measurements. Virtual metrology (VM) is the use of mathematical models with accessible measurements from an operating process to estimate variables of interest. This thesis addresses the challenge of virtual metrology for plasma processes, with a particular focus on semiconductor plasma etch.

Introductory material covering the essentials of plasma physics, plasma etching, plasma measurement techniques, and black-box modelling techniques is first presented for readers not familiar with these subjects. A comprehensive literature review is then completed to detail the state of the art in modelling and VM research for plasma etch processes.

To demonstrate the versatility of VM, a temperature monitoring system utilising a state-space model and Luenberger observer is designed for the variable specific impulse magnetoplasma rocket (VASIMR) engine, a plasma-based space propulsion system. The temperature monitoring system uses optical emission spectroscopy (OES) measurements from the VASIMR engine plasma to correct temperature estimates in the presence of modelling error and inaccurate initial conditions. Temperature estimates within 2% of the real values are achieved using this scheme.

An extensive examination of the implementation of a wafer-to-wafer VM scheme to estimate plasma etch rate for an industrial plasma etch process is presented. The VM models estimate etch rate using measurements from the processing tool and a plasma impedance monitor (PIM). A selection of modelling techniques are considered for VM modelling, and Gaussian process regression (GPR) is applied for the first time for VM of plasma etch rate. Models with global and local scope are compared, and modelling schemes that attempt to cater for the etch process dynamics are proposed. GPR-based windowed models produce the most accurate estimates, achieving mean absolute percentage errors (MAPEs) of approximately 1.15%. The consistency of the results presented suggests that this level of accuracy represents the best accuracy achievable for the plasma etch system at the current frequency of metrology.

Finally, a real-time VM and model predictive control (MPC) scheme for control of plasma electron density in an industrial etch chamber is designed and tested. The VM scheme uses PIM measurements to estimate electron density in real time. A predictive functional control (PFC) scheme is implemented to cater for a time delay in the VM system. The controller achieves time constants of less than one second, no overshoot, and excellent disturbance rejection properties. The PFC scheme is further expanded by adapting the internal model in the controller in real time in response to changes in the process operating point.

# *Acknowledgements*

It is a pleasure to thank those who made this thesis possible. First and foremost, I would like to express my sincerest gratitude to my PhD supervisor Professor John Ringwood for his help with the work contained within this thesis. From Skype calls discussing control theory in Costa Rica, to weekly meetings polishing responses to reviewers, to midnight discussions on the intricacies of red wine tasting in Chile, John has always provided great advice, encouragement, and stimulating conversation, continuously helping me through the good and bad times of my post-graduate work.

Secondly I would like to thank Niall MacGearailt for his unquenchable enthusiasm, his willingness to help, and his "unconventional" approach to getting things done throughout my research. Thanks to Niall for his vision and openness to the research community in Maynooth that provided the data sets required for this research, and access to the necessary tools and people when they were most required.

I would like to express my gratitude to Dr. Sean McLoone in the Electronic Engineering Department for not only his vast mathematical knowledge, but also his kindness and patience in distributing that knowledge. I would also like to thank Sean for his help in proof reading this document towards the end of my research.

Thanks to all of the staff of Ad-Astra Costa Rica, who welcomed me as one of their own during my time with them, and enriched my experience with all of the trimmings of Costa Rican life. Your enthusiasm for your work, and belief in your ultimate goal, will always inspire me. Thanks to the former members of the Semiconductor Research project in Maynooth, Emanuele Ragnoli and Bei Bei Ma. Thanks also to Conor Hilliard of Lam Research Ireland and David Kavanagh of Dublin City University.

Thanks to all of the staff in the Electronic Engineering Department in Maynooth, for making the department a friendly place. Thanks to the NUIM postgrads and my colleagues in the Engineering Department, Giorgio Bacelli, Shane Butler, Niall Cahill, Francesco Fusco, and Violeta Mangourova. Thanks to my parents, Liam and Marion, and my sisters, Helena and Thérèse, for their encouragement and belief throughout the years, and also their invaluable support during the "more challenging" times.

Thanks to the funding bodies that supported this research at different points, the Irish research council for science and engineering (IRCSET) and the FÁS Science Challenge scholarship.

And finally, a special thanks to my housemates and friends, Ciarán Pollard, Damian Kelly, Kevin Sweeney and Lorcan Walsh, for sharing countless laughs, great experiences, and valued friendship throughout the last four years. Long may it continue.

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| **ANN** | **A**rtificial **N**eural **N**etwork |
| **ANOVA** | **AN**alysis **O**f **V**ariance |
| **APC** | **A**dvanced **P**rocess **C**ontrol |
| **ARDE** | **A**spect **R**atio **D**ependant **E**tching |
| **BPR** | **B**eam **P**rofile **R**eflectometry |
| **BFGS** | **B**royden-**F**letcher-**G**oldfarb-**S**hannon optimisation method |
| **CD** | **C**ritical **D**imension |
| **CCD** | **C**harge-**C**oupled **D**etector |
| **CCP** | **C**apacitively **C**oupled **P**lasma |
| **CMP** | **C**hemical **M**echanical **P**lanarisation |
| **CLRT** | **C**losed **L**oop **R**esponse **T**ime |
| **CSTR** | **C**ontinous-flow, **W**ell-**S**tirred **T**ank **R**eactor |
| **CSV** | **C**omma **S**eparated **V**alues |
| **DC** | **D**irect **C**urrent |
| **DOE** | **D**esign **O**f **E**xperiment |
| **ECR** | **E**lectron **C**yclotron **R**esonance |
| **EP** | **E**tch **P**rocess |
| **EWMA** | **E**xponentially **W**eighted **M**oving **A**verage |
| **FCC** | **F**ederal **C**ommunications **C**ommission |
| **FDC** | **F**ault **D**etection and **C**lassification |
| **FSCA** | **F**orward **S**election **C**omponent **A**nalysis |
| **FSR** | **F**orward **S**tepwise **R**egression |
| **GRNN** | **G**eneralised **R**egression **N**eural **N**etwork |
| **ICA** | **I**ndependent **C**omponent **A**nalysis |
| **ICP** | **I**nductively **C**oupled **P**lasma |
| **ICRH** | **I**on **C**yclotron **R**esonance **H**eating |
| **kNN** | **K-N**earest **N**eighbour |
| **LARS** | **L**east **A**ngled **R**egression (**S**tagewise) |
| **LI** | **L**aser **I**nterferometry |
| **LIF** | **L**aser-**I**nduced **F**luorescence |

| | |
|---|---|
| **LRI** | **L**aser **R**eflectance **I**nterferometry |
| **LM** | **L**evenberg-**M**arquardt algorithm |
| **LQG** | **L**inear **Q**uadratic **G**aussian |
| **LSR** | **L**east **S**quares **R**egression |
| **MAPE** | **M**ean **A**bsolute **P**ercentage **E**rror |
| **MERIE** | **M**agnetically **E**nhanced **R**eactive **I**on **E**tch |
| **MFC** | **M**ass **F**low **C**ontroller |
| **MLP** | **M**ulti-**L**ayered **P**erceptron |
| **MLR** | **M**ultiple **L**inear **R**egression |
| **MPC** | **M**odel **P**redictive **C**ontrol |
| **MS** | **M**ass **S**pectrometry |
| **MSE** | **M**ean **S**quared **E**rror |
| **NI** | **N**ational **I**nstruments |
| **NIPALS** | **N**onlinear **I**terative **PA**rtial **L**east **S**quares |
| **OES** | **O**ptical **E**mission **S**pectroscopy |
| **OLRT** | **O**pen **L**oop **R**esponse **T**ime |
| **OMA** | **O**ptical **M**ultichannel **A**nalyser |
| **PCA** | **P**rincipal **C**omponent **A**nalysis |
| **PCR** | **P**rincipal **C**omponent **R**egression |
| **PI** | **P**roportional **I**ntegral |
| **PIC** | **P**article **I**n **C**ell |
| **PID** | **P**roportional **I**ntegral **D**ifferential |
| **PIM** | **P**lasma **I**mpedance **M**onitor |
| **PFC** | **P**redictive **F**unctional **C**ontrol |
| **PLS** | **P**artial **L**east **S**quares |
| **PM** | **P**reventative **M**aintenance |
| **RGA** | **R**esidual **G**as **A**nalysis |
| **RIE** | **R**eactive **I**on **E**tch |
| **RSM** | **R**esponse **S**urface **M**ethodology |
| **RF** | **R**adio **F**requency |
| **RLS** | **R**ecursive **L**east **S**quares |
| **RMSE** | **R**oot **M**ean **S**quared **E**rror |
| **SCCM** | **S**tandard **C**ubic **C**entimeter per **M**inute |
| **SE** | **S**quared **E**xponential |
| **SEERS** | **S**elf-**E**xcited **E**lectron **R**esonance **S**pectroscopy |
| **SEM** | **S**canning **E**lectron **M**icroscope |
| **SIMPLS** | **S**traightforward **impl**ementation of a statistically inspired **m**odification of the **PLS** method |
| **SNR** | **S**ignal to **N**oise **R**atio |

| | |
|---|---|
| **SPC** | **S**tatistical **P**rocess **C**ontrol |
| **SPE** | **S**quared **P**rediction **E**rror |
| **SSE** | **S**um **S**quared **E**rror |
| **STI** | **S**hallow **T**rench **I**solation |
| **STP** | **S**tandard **T**emperature and **P**ressure |
| **SVD** | **S**ingular **V**alued **D**ecomposition |
| **SVM** | **S**upport **V**ector **M**achine |
| **VASIMR** | **V**ariable **S**pecific **I**mpulse **M**agnetoplasma **R**ocket |
| **VM** | **V**irtual **M**etrology |
| **WAT** | **W**afer **A**cceptance **T**est |
| **WCSS** | **W**ithin **C**luster **S**um of **S**quares |

# Physical Constants

| | | | |
|---|---|---|---|
| Speed of Light | $c$ | = | $2.997\ 924\ 58 \times 10^{8}$ ms$^{-\mathrm{S}}$ |
| Boltzmann's constant | $k_B$ | = | $1.380\ 650\ 3 \times 10^{-23}$ m$^2$ kg s$^{-2}$ K$^{-1}$ |
| Pi | $\pi$ | = | 3.141 592 653 589 793 |
| Plank's constant | $h$ | = | $6.626\ 068\ 96 \times 10^{-34}$ J s |
| Electron charge | $e$ | = | $1.602\ 176\ 46 \times 10^{-19}$ C |
| Electron mass | $m_e$ | = | $9.109\ 382\ 15 \times 10^{-31}$ kg |
| Electric constant | $\epsilon_0$ | = | $8.854\ 187\ 82 \times 10^{-12}$ F m$^{-1}$ |

# Symbols

*Matrix Notation*

| | |
|---|---|
| $\mathbf{X}$ | Two-dimensional matrix |
| $\mathbf{X}^T$ | Transpose of $\mathbf{X}$ |
| $x_{ij}$ | Element at row $i$, column $j$, of matrix $\mathbf{X}$ |
| $\mathbf{x}_i$ | Column vector made up of $i^{th}$ column of $\mathbf{X}$ |
| $\vec{\mathbf{x}}_i$ | Row vector made up of $i^{th}$ row of $\mathbf{X}$ |
| $\lvert\mathbf{X}\rvert$ | Determinant of matrix $\mathbf{X}$ |
| $\lVert\mathbf{x}\rVert$ | Magnitude of vector $\mathbf{x}$ |

*List of Nomenclature*

| | | |
|---|---|---|
| $n_e$ | Plasma electron density | $\mathrm{m}^{-3}$ |
| $n_i$ | Plasma ion density | $\mathrm{m}^{-3}$ |
| $n_n$ | Plasma neutral density | $\mathrm{m}^{-3}$ |
| $T_e$ | Electron temperature | K |
| $T_i$ | Ion temperature | K |
| $v$ | Speed of particle | $\mathrm{ms}^{-1}$ |
| $\overline{v^2}$ | Mean square speed of particle | $\mathrm{m}^2\mathrm{s}^{-2}$ |
| $q$ | Particle charge | C |
| $m$ | Mass of particle | kg |
| $T$ | Temperature | K |
| $E_{ph}$ | Photon energy | J |
| $f_{ph}$ | Photon frequency | Hz |
| $f_c$ | Collision frequency | $\mathrm{s}^{-1}$ |
| $\bar{v}_e$ | Electron mean speed | $\mathrm{m\ s}^{-1}$ |
| $\bar{v}_e$ | Ion mean speed | $\mathrm{m\ s}^{-1}$ |
| $j_e$ | Electron current density | $\mathrm{A\ m}^{-2}$ |
| $j_i$ | Ion current density | $\mathrm{A\ m}^{-2}$ |
| $V_p$ | Plasma potential | V |
| $V_f$ | Floating potential | V |

| | | |
|---|---|---|
| $m_i$ | Ion mass | kg |
| $B$ | Magnetic flux density | G |
| $P$ | Power | W |
| $V$ | Voltage | V |
| $I$ | Current | A |
| $Z$ | Impedance | Ω |
| $R$ | Resistance | Ω |
| $X$ | Reactance | Ω |
| $f_r$ | Resonant frequency of microwave probe with plasma | Hz |
| $f_0$ | Resonant frequency of microwave probe without plasma | Hz |
| $L$ | Length of resonator on microwave probe | m |
| $T_s$ | Sampling period | s |
| $V_{bias}$ | Wafer self-bias voltage | V |
| $V_p$ | Plasma potential | V |
| $V_f$ | Floating potential | V |
| | | |
| $\lambda_{ph}$ | Photon wavelength | m |
| $\alpha_{is}$ | Plasma ionisation degree | % |
| $\Phi$ | Particle flux | $m^{-2}$ |
| $\rho$ | Space charge density | C $m^{-3}$ |
| $\phi_w$ | Work function of metal | J |
| $\omega_p$ | Plasma frequency | rad $s^{-1}$ |
| $\xi_0$ | Electric field amplitude | V |
| $\nu_e$ | Electron elastic collision frequency | $s^{-1}$ |
| $\lambda_d$ | Debye length | m |
| $\omega_{ce}$ | Electron cyclotron frequency | rad $s^{-1}$ |
| $\omega_{ci}$ | Ion cyclotron frequency | rad $s^{-1}$ |
| $\omega_p$ | Plasma frequency | rad $s^{-1}$ |
| $\phi$ | Phase angle between $V$ and $I$ | degrees |

*Dedicated to my parents, Liam and Marion; without their unwavering love and encouragement, this document would not exist.*

# Chapter 1

# Introduction

## 1.1 Background and motivation

Plasma, often termed the fourth state of matter, is an ionised gas consisting of positively and negatively charged particles with approximately equal charge densities [1]. Interestingly, plasma makes up as much as 99% of the mass in the universe, both by mass and by volume. Since its first discovery in 1879 by Sir William Crookes [2], plasma has found applications in many aspects of modern life, with applications as far reaching as plasma-arc welding, waste disposal, visual displays, propulsion systems, medical sterilisation techniques, fluorescent lamps, and semiconductor manufacture.

Control of plasma-based processes is difficult in general because of its non-linear behaviour and sensitivity to disturbances. In many cases, measurements required for accurate control are difficult to obtain due to the harsh conditions within plasmas, and a requirement to avoid perturbation of plasmas used in production.

Virtual metrology involves estimation of variables from a process that are not measured directly using surrogate measurements taken from the process. Virtual metrology (VM) is achieved using mathematical models that relate the in-situ measurements from the process to the variables of interest that are inaccessible at the time of processing. While VM is most widely associated with the semiconductor manufacturing industry, the idea of VM has strong ties to state estimation and observer design in control theory. VM has the potential to greatly improve the performance of semiconductor plasma processes by increasing the availability and response-time of process feedback variables for control and monitoring purposes. VM could also be used to replace expensive sensor equipment.

In this thesis, the application of virtual metrology to plasma processes is examined, with a particular focus on plasma etch, a semiconductor manufacturing process. As a demonstrative case study, the application of VM to a plasma-based space propulsion system is also examined. In this section, a short discussion on the motivation for both applications is provided, with emphasis on plasma etch as the main subject area of this thesis.

### 1.1.1 VM for VASIMR

Propulsion for space travel is predominantly achieved through the use of chemical combustion rockets that burn a fuel in an oxidising agent to produce thrust in space with a controlled explosion. The fuel efficiency of such engines is relatively low, and for interplanetary flights, chemical rockets present limitations in terms of the costs of transporting vast quantities of fuel into orbit, and the maximum velocities achievable with restricted fuel supplies. As space exploration turns towards more ambitious plans for interplanetary human flight, alternative technologies for space propulsion are being developed to enable faster and more efficient space travel.

The variable specific impulse magnetoplasma rocket (VASIMR ®) engine is a space propulsion engine being developed by the Ad-Astra Rocket Company. The VASIMR engine produces thrust by accelerating a gas propellant (typically argon) in plasma form, using large magnetic fields produced by an array of electromagnets. The production of thrust in this manner uses much less fuel than conventional chemical combustion rockets, and has the potential to revolutionise space travel by greatly increasing the achievable velocities for spacecraft with smaller fuel demands. VASIMR engines also have applications in satellite repositioning and lunar cargo transport. The VASIMR engine is currently in a prototype phase, with the first flight-ready engine to be tested on the international space station in the coming decade.

The VASIMR engine produces a great deal of excess heat during plasma production. Internal engine temperatures can quickly reach levels beyond the allowable limits of the engine's components. Monitoring of the internal engine temperatures is made difficult by the extreme environment in the rocket plasma, which can reach over one million degrees centigrade [3]. In prototype systems, the temperatures are measured using thermocouples attached to the engine, but such systems are not feasible for final flight-ready engines.

Virtual metrology provides a viable option for non-invasive estimation of internal engine temperatures. The plasma optical emissions are easily measured during plasma

production, and the optical emission can be related to the temperatures of interest through the use of VM models. Successful implementation of a reliable VM scheme reduces the need for invasive temperature measurements, and can be used at later developmental stages for feed-back control of active cooling systems. In this thesis, the application of VM techniques to the VASIMR engine is examined as a case study.

### 1.1.2 VM for semiconductor etch

Semiconductor manufacturing is difficult. The technology and engineering behind the microelectronics that now permeate almost every aspect of modern life is astonishingly complex. The proliferation of microprocessors has been made possible by tremendous advancements in semiconductor manufacturing techniques in recent decades, resulting in the minimisation of both the cost and size of electronic components. The trend of innovation famously follows "Moore's Law", first articulated in 1965 [4] by the co-founder of Intel corporation, Gordon E. Moore, who predicted that "The number of transistors that can be placed inexpensively on an integrated circuit has doubled approximately every two years." The exponential increase in microprocessor capabilities predicted by Moore has become both a benchmark and a target for semiconductor manufacturers worldwide, with current industrial development working towards a 22 nm node [5] (the dimension size of a manufacturing "node" is defined as half the distance between cells in a dynamic random access memory (DRAM) chip).

The semiconductor manufacturing cycle typically comprises over 350 different process steps to build nanometer scale circuits on silicon wafers. In modern semiconductor fabrication facilities, or *fabs*, wafers are typically 300 mm in diameter and processed in batches, or *lots*, of up to 25 wafers. The main manufacturing steps are deposition, lithography, etch, ion implantation, and planarisation. In deposition, layers of material are deposited on the wafer surface, usually using thermal processes. In lithography, patterns of photoresistive mask are transferred to the wafer surface. During etch, reactant gases in plasma form remove surface material that is not covered by the photoresistive mask. During ion implantation, the electrical properties of areas of the wafer surface are changed through semiconductor doping with different elements. Finally, in planarisation, wafers surfaces are smoothed with a combination of chemical and mechanical forces. Through repetitive applications of these five procedures, along with some other processes, elements of logic and memory circuits are constructed on silicon wafers [6].

Interdependencies exist between each of the processes carried out. Processing errors at one manufacturing tool invariably have knock-on effects that can reduce device yield. Such errors in manufacturing can cost companies hundreds of thousands of euro per year

due to the high-value nature of the product material. As such, process monitoring and control for every process in the fabrication environment is of paramount importance.

While wafers that are processed incorrectly during some processes, for example, lithography, can sometimes be stripped and reworked, such rework is invariably not possible for incorrectly etched wafers [7]. As a result, ensuring the etch process is operating within specifications is important. Measurement of plasma etch performance during processing is difficult due to the harsh environment within plasma etch chambers and, typically, metrology tools measure etch performance downstream from the plasma production tools. Non-invasive metrology techniques are desirable during the etch process to avoid perturbing the etching plasma and affecting the final process outcome on the wafer surface.

Etch processing is conducted within specialised etch chambers. *Process input variables* to the chambers are typically well controlled variables such as chamber pressures, component temperatures, and gas flow rates, specified by *set points*. In general, the required etch process input variables for each product are developed through extensive experimentation during the product development stage, early in a product's life cycle. Once decided upon, the etch process input variables are compiled into *recipes* that are distributed to different fabs for production. The etch recipes remain relatively fixed and, historically, the recipes were applied to product wafers in each fab environment in an open-loop manner [8], where repeatable results are assumed for each wafer processed. This open-loop application of process recipes is still used for some etch processes.

The time-varying dynamics of plasma etch processes causes difficulties in maintaining consistent etch results for processes using constant process recipes. Etch processes exhibit process drift and unpredictable shifts in behaviour, causing variance in the etch results for each wafer. Manufacturing processes in the semiconductor industry are predominantly managed using statistical process control (SPC), where variables measured in-situ during each process (*process variables*), or variables concerning the result of each process (*process output variables*) are monitored for deviations that indicate erroneous operation. Multi-variate statistics are also employed to allow multiple process variables from each process to be monitored using SPC [9].

Advanced process control (APC) is the next step in factory automation, which, as of yet, is not fully adopted by semiconductor factories worldwide. The ultimate aim of APC is to improve device yield, that is the number of "good" chips or *die* per wafer processed. APC is considered [6] to include four components,

- fault detection,

- fault classification,
- fault prognosis, and
- process control.

These tasks are achieved through the use of information about the material to be processed, measured data, and the desired results. APC includes lot-to-lot, wafer-to-wafer, and within wafer real-time control and has the capability to improve performance, yield, and throughput within manufacturing environments [10]. *Ideally*, measurements of important process variables and process output variables during every wafer processed are available to implement APC.

The implementation of APC for plasma etch in industry has broadly been restricted to lot-to-lot control [11], rather than wafer-to-wafer or real-time control as a result of two main difficulties. Firstly, measurements of important process output variables typically involve a prohibitively large time overhead, meaning that every wafer processed cannot be measured. Secondly, there is typically a considerable delay (several hours or even days) between the etch of a product wafer and the availability of metrology on the etch process output variables, i.e. a *metrology delay*. As a result, corrections to the etch process recipe, if required, cannot be implemented in real-time during the etch process or even immediately after each wafer is processed. If an etch system operates out of control, several wafers or lots can be processed erroneously before the problem is detected, potentially leading to multiple wafer scraps. With shrinking device dimensions, process control limits are becoming more stringent, and deviations in the etch process performance can more easily destroy valuable product wafers. Stricter control of the etch process is required for the continued advancement of the products being processed.

One potential solution to the problem of infrequent measurement is that of integrated metrology, that is the addition of metrology tools into each processing tool in the fab, that allow frequent measurements of product material to be taken during or after a processing step [12, 13]. However, due to a prohibitive set up cost, a lack of cooperation between semiconductor companies and tool manufacturers, and increased cycle time, large-scale adoption of integrated metrology has not yet occurred in the industry.

The second potential solution is *virtual metrology* (VM) which, as mentioned earlier in this section, is the use of in-situ measurements of process variables along with mathematical models of the process to estimate or predict process output variables of interest. A typical VM implementation for plasma etch is depicted in Figure 1.1. There are many advantages to VM in semiconductor manufacturing [14], including

- Reduction in wafer scraps by faster process monitoring: with increased numbers of immediately available "virtual" measurements of process output variables, errors in processing can be detected in a timely fashion, preventing further wafers from being processed incorrectly.
- Improved process control: VM estimates, available during or after each processed wafer, overcome the low frequency of measurements taken in semiconductor processes, enabling the adjustment of process input variables on a real-time or per wafer basis. Different control schemes are possible, as depicted in Figure 1.2: Plasma variables such as species concentrations and temperatures (see Chapter 2) can be controlled in real-time, etch process variables such as etch rate (the rate of material removal from the wafer surface) can be controlled in real-time or on a wafer-to-wafer basis, and process output variables such as etch depth can be controlled on a wafer-to-wafer or lot-to-lot basis.
- Increased throughput: When dependable VM schemes are implemented, the frequency of actual metrology operations could be reduced, simultaneously reducing the production cycle time and metrology costs, increasing fab efficiency.

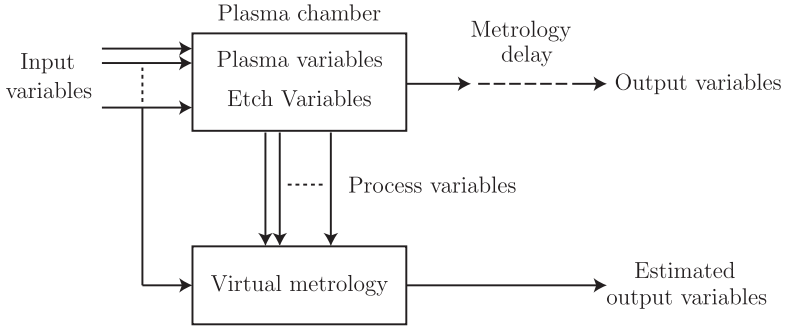



FIGURE 1.1: Virtual metrology principle. Estimates of process output variables of interest are made using process variables and mathematical models, or *virtual metrology models*. Similarly, plasma or etch variables can be estimated using virtual metrology models.

The model input data required for VM is, in some cases, already being collected from processes in many fabs for off-line analysis of faults and SPC. However, the successful implementation of VM for any process depends on the construction of a reliable process model. In the case of plasma etch, such a model is difficult to create and maintain [15]. The inherent complexities of the etch process means that modelling from first-principals is extremely complicated, and typically, such models cannot be computed in real time. Hence, many researchers rely on empirical black-box modelling techniques using data sets collected from either specially designed experiments or production wafers. A considerable amount of research has been completed in the area of plasma etch modelling for VM [8].

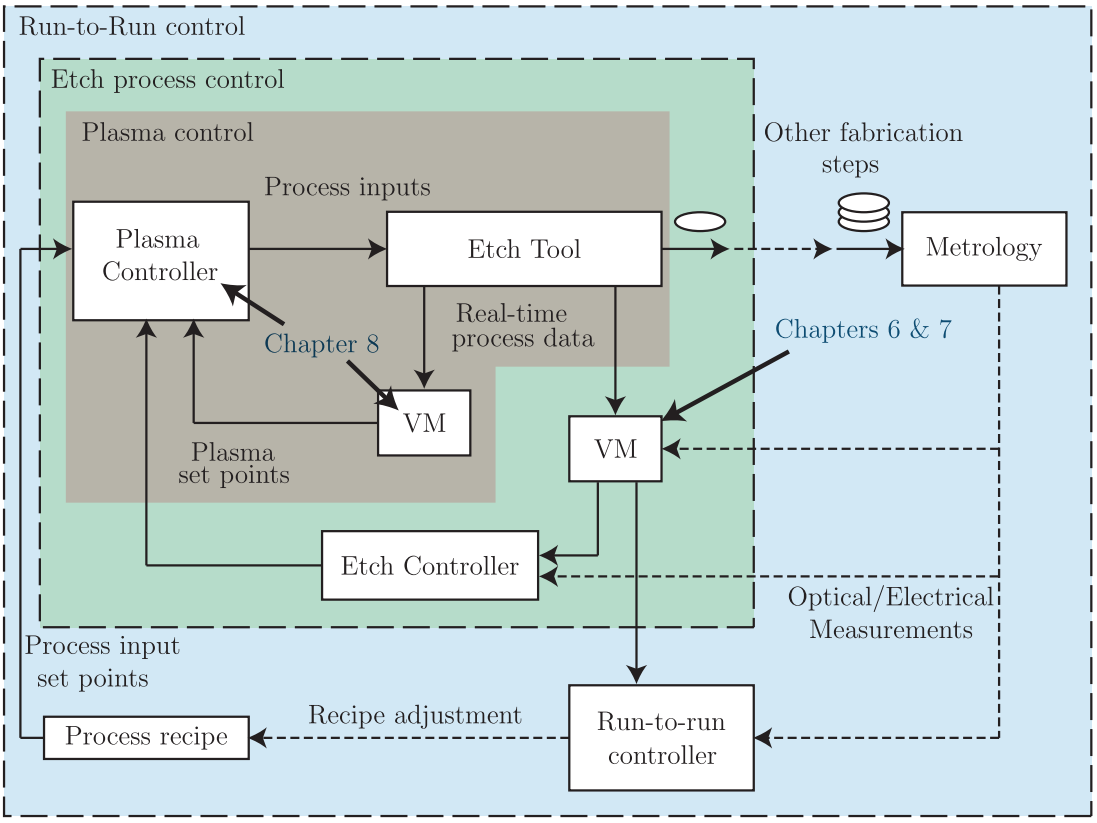



FIGURE 1.2: Virtual metrology applications for plasma etch.

Successful implementation of VM for all semiconductor manufacturing production processes has the potential to revolutionise the semiconductor production process, enabling APC implementation on a fab-wide basis. Such implementation is seen as essential as the semiconductor industry moves simultaneously towards smaller critical dimensions and larger diameter wafers [16, 17]; the international technology roadmap for semiconductors (ITRS) points towards production at the 16 nm node and the introduction of 450 mm diameter wafers in the coming decade [5]. As a result, VM has been highlighted as a topic of interest by a number of research consortia, symposia, and funded research collaborations as a key area for development in the semiconductor industry. Examples of research initiatives with VM component streams include the global semiconductor manufacturing technology (SEMATECH) consortium, the IMPROVE research project in Europe, the KAP (knowledge, awareness, prediction) research project also in Europe, and the Irish centre for manufacturing research (ICMR) competence center in Ireland. VM is now a mainstream topic at annual advanced equipment control / advanced process control (AEC/APC) symposia in Europe, the U.S., and Asia. Research on VM topics is also published often at the annual advanced semiconductor manufacturing conference (ASMC).

## 1.2 Objectives

The main objective of this thesis is to examine the application of VM techniques to industrial plasma processes, with a particular focus on semiconductor plasma etch. This objective is achieved in four main sections.

Firstly, the thesis aims to comprehensively describe the present state of the art in virtual metrology in plasma etch through the examination of published literature.

Secondly, the thesis aims to demonstrate the versatility of VM techniques for plasma applications through the development of a VM system for temperature monitoring purposes on the space propulsion engine, the VASIMR engine. This particular case study aims to demonstrate how VM is not limited to the semiconductor industry, for which it is most commonly known. The VM system aims to provide accurate real-time estimates of engine temperatures to operators.

Thirdly, for semiconductor etch, the thesis aims to develop VM models for a production plasma etch process. The research aims to develop VM models that achieve the maximum etch rate estimation accuracy possible. The data set used is representative of typical industrial data, consisting of measurements of process and output variables from an industrial etch process. The aim of the VM system is to estimate etch rate on a wafer-by-wafer basis for the sake of process monitoring, and potentially for implementation of a wafer-to-wafer control system. This research implements the VM block in the etch process control loop of Figure 1.2.

Finally, the thesis also aims to investigate the application of real-time VM techniques for control of the plasma electron density in an industrial etch chamber, implementing the real-time plasma variable control loop depicted in Figure 1.2. The aim of this section of the research is to investigate the feasibility of using VM to eventually facilitate specification of process recipes in terms of plasma variables, rather than process input set points. During real-time control, changes to the process input variables are made in real time to maintain consistent plasma electron density even the presence of disturbances representative of maintenance events, hence producing more predictable etch performances.

## 1.3 Contributions of this thesis

This thesis claims the following original contributions:

1. A comprehensive literature review of the state of the art in virtual metrology applied to semiconductor plasma etch is conducted. The literature review examines the modelling techniques and applications of VM used by different researchers in academia and industry. The extensive literature is divided into logical subsections for ease of reference.

2. In a case study contained in this research, the novel use of optical emission spectroscopy (OES) data to accurately estimate a spatial temperature distribution in a plasma rocket engine is detailed. A linear state space model and Luenberger estimator is used to achieve this goal.

3. A comprehensive investigation into the maximum achievable accuracy of global and local VM models for a modestly sampled plasma etch system is carried out. The performance of a global VM modelling scheme is compared to three different local VM modelling schemes to determine the best modelling approach to cater for the peculiarities of plasma etch process dynamics.

4. A novel weighting system based on the maintenance history of the plasma etch chamber is proposed for windowed partial least squares (PLS) regression VM models. The suggested weighting scheme is found to increase the accuracy of the etch rate estimates compared to non-weighted models.

5. This thesis details the first application of Gaussian process regression (GPR) models to semiconductor etch data, and finds GPR models to generate more accurate estimates of plasma etch rate for unseen data compared to the other modelling techniques investigated.

6. A novel real-time VM and model-based predictive control scheme is implemented to achieve non-invasive real-time control of electron density in a production plasma etch chamber. To the best of the authors knowledge, this research details the first application of virtual metrology for real-time control of plasma electron density in an industrial etch system.

7. The research reports the first application of predictive functional control (PFC), to an industrial plasma etch chamber.

## 1.4 List of publications

### 1.4.1 Peer-reviewed publications

1. Lynn, S., Ringwood, J., and MacGearailt, N., "Global and local virtual metrology models of a plasma etch process", *IEEE Transactions on Semiconductor Manufacturing*, submitted for publication Nov. 2010.

2. Lynn, S., Ringwood, J., and MacGearailt, N., "Real-time virtual metrology and control of electron density in an industrial plasma etch chamber", *IFAC World Congress 2011*, accepted for publication and presentation Aug. 2011.

3. Lynn, S., Ringwood, J., and MacGearailt, N., "Gaussian process regression for virtual metrology of plasma etch", *Irish Signals and Systems Conference*, oral presentation, Cork, Ireland, Jun. 2010, pp. 42–47 (received best student paper award).

4. Lynn, S., Ringwood, J., and MacGearailt, N., "Weighted windowed PLS models for virtual metrology of an industrial plasma etch process", *2010 IEEE International Conference Industrial Technology (ICIT)*, oral presentation, Vina del Mar, Chile, Mar. 2010, pp. 309–314.

5. Ringwood, J.V., Lynn, S., Bacelli, G., Ma, B., Ragnoli, E., and McLoone, S., "Estimation and control in semiconductor etch: Practice and possibilities, *IEEE Transactions on Semiconductor Manufacturing*, vol. 23, no. 1, pp. 87–98, Feb. 2010.

6. Lynn, S., Ringwood, J.V., and Del Valle Gamboa, J.I., *"Temperature Estimation for a Plasma-Propelled Rocket Engine - Inferential Measurement using Optical Emission Spectrometer Data"*, IEEE Control Systems Magazine, Vol. 29, No. 6, pp. 15–25, Dec. 2009.

7. Ragnoli, E., McLoone, S., Lynn, S., Ringwood, J.V., and MacGearailt, N., "Identifying key process characteristics and predicting etch rate from High-Dimension Datasets", in *Proceedings of the Advanced Semiconductor Manufacturing Conference (ASMC)*, Berlin, Germany, May. 2009, pp. 106–111.

8. Lynn, S., Ringwood, J.V., Ragnoli, E., McLoone, S., and MacGearailt, N., "Virtual Metrology for Plasma Etch using Tool Variables", in *Proceedings of the Advanced Semiconductor Manufacturing Conference (ASMC)*, Berlin, Germany, May. 2009, pp. 143–148.

9. Lynn, S., Ringwood, J.V., and Del Valle Gamboa, J.I., "State Estimation for the VASIMR Plasma Engine", in *Proceeding of the 16$^{th}$ Irish Signals and Systems Conference*, oral presentation, Galway, Ireland, 2008, pp. 24–29.

### 1.4.2 Internal technical reports

1. Lynn, S., *"Local modelling of a plasma etch data set."* Technical Report, EE/JVR/1/2010, Dept. of Dept. of Electronic Engineering, National University of Ireland, Maynooth, February 2010.

2. Lynn, S., *"Global modelling of a plasma-etch data set."* Technical Report, EE/JVR/3/2009, Dept. of Electronic Engineering, National University of Ireland, Maynooth, 2009.

3. Lynn, S., *"Virtual metrology for plasma etch - A literature review."* Technical Report, EE/JVR/3/2007, Dept. of Electronic Engineering, National University of Ireland, Maynooth, Dec. 2007.

4. Lynn, S., *"An introduction to plasma and plasma etching."* Technical Report, EE/JVR/2/2007, Dept. of Electronic Engineering, National University of Ireland, Maynooth, Apr. 2007.

## 1.5 Thesis layout

The thesis begins in Chapter 2 by providing background information on plasma physics and plasma etch processing so that readers unfamiliar with these core topics can familiarise themselves with key principals and terminology that are encountered in each chapter thereafter. Explanations are provided at a relatively basic level for an audience with a general scientific or engineering background.

Chapter 3 introduces the mathematical modelling techniques that are employed throughout this thesis to perform VM. Explanations of the workings of each technique are provided along with some discussion on the advantages and disadvantages of each. The techniques described in Chapter 3 are used in Chapters 6 – 8 to perform VM, and are referred to regularly when discussing related research in Chapter 4.

Chapter 4 contains a comprehensive literature review of the state of the art in VM and modelling for plasma etch. Chapter 4 is included to provide background information on existing work in the literature so that the reader can understand the context and relevance of the research described in the thesis.

Chapter 6 details the application of global models to an industrial plasma etch data set. An in-depth discussion on the etch process studied is included to describe the peculiarities of the data and to describe the main sources of variance in the data. The VM techniques described in Chapter 3 are used to create the global models. Chapter 7 then details the development and application of three different local modelling schemes to the same plasma etch data set. Each local modelling scheme is discussed in turn, and the motivation for each is clearly provided.

Chapter 8 examines the development of a real-time VM and model predictive control system for plasma electron density in an industrial plasma processing chamber. Because this chapter is the only chapter primarily concerned with the application of control algorithms, a literature review on control research in plasma etch and an introductory discussion on the methods used for model predictive control are provided before presentation of the results of the experiments.

Finally, Chapter 9 presents the general conclusions that can be drawn from the body of research presented in the thesis and discusses potential future work arising from the research.

# Chapter 2

# Plasma and plasma etch fundamentals

In this chapter, the basic principles of the plasma etch process are examined. An introduction to the basic features and phenomena of plasma physics is provided along with a broad overview of plasma processing technology and common diagnostic tools encountered in the area of plasma etching. Plasma physics is a vast and complex area of study, the complete details of which are beyond the scope of this thesis. For a more complete examination, the interested reader is directed to the work by Lieberman and Lichtenberg [18].

## 2.1 Basic plasma physics

### 2.1.1 What is a plasma?

Plasma, or "radiant matter" as it was first dubbed, was discovered by Sir William Crookes in 1879 in a Crooke's tube [2]. Joseph J. Thompson identified the nature of the fluorescent "cathode rays" identified by Crookes in 1897 with his discovery of the electron [19]. The ionised gases were first named "plasmas" by Irving Langmuir in 1928 [20], choosing the name as they reminded him of blood plasmas. Langmuir went on to introduce the concepts of electron temperature and invent the Langmuir probe, a plasma measurement tool still in use today (see Section 2.5.6).

Referred to as the fourth state of matter, plasma is an ionised gas consisting of positively and negatively charged particles with approximately equal charge densities

[1]. Plasma can be produced by heating a gas to such a temperature that the random kinetic energy of the gaseous molecules exceeds the ionisation energy of the constituent gases. At such temperatures, the atoms and molecules of the gas become ionised due to collisions with other molecules leaving a large number of ions (atoms with at least one of their electrons taken away, or with one extra electron) in a sea of freely moving electrons.

Natural plasmas make up as much as 99% of the mass in the universe, both by mass and by volume. Most stars, including the Sun, and a significant fraction of the interstellar medium are made up of plasma. In the Earth's atmosphere, examples of natural plasmas include lightening and the aurora borealis. The ionosphere and the magnetosphere are layers of plasma that surround the earth at altitudes above 80km. Since plasmas contain many freely moving charged particles, they are highly responsive to magnetic and electric fields. Man-made plasmas have become commonplace in society, with applications as far reaching as plasma-arc welding, waste disposal, visual displays, and fluorescent lamps, as mentioned in Chapter 1.

### 2.1.2 Degree of ionisation

The degree of ionisation of a given plasma describes the proportion of the gaseous molecules that have been ionised, that is the proportion of molecules which have lost or gained one or more electrons via energetic collisions with other particles. The degree of ionisation is expressed in terms of the charged particle densities, $n_e \approx n_i$ particles / $\mathrm{m}^3$, where $n_e$ is the number of electrons per cubed meter, or the *electron density*, and $n_i$ is the *ion density* of the plasma. The degree of ionisation of a plasma is

$$\alpha_{is} = \frac{n_i}{n_i + n_n} \times 100\% \tag{2.1}$$

where $n_n$ is the density of neutral molecules. A gas may begin to exhibit plasma behaviours with a degree of ionisation $\alpha_{is}$ as little as 0.01 %. Note that the relationship $n_e \approx n_i$ only holds for plasmas where the average charge state $\nu$ (an integer) of the ions is one, otherwise $n_e = \nu n_i$.

A source of energy is required to maintain the plasma, usually in the form of an electric or a magnetic field from which charged particles gain energy. This energy appears as kinetic energy for each particle, given by $\frac{1}{2}mv^2$, where $m$ is the mass of the particle, and $v$ is it's speed. Because electrons have a much smaller mass than ions, but carry the same magnitude of electric charge, electrons move at much faster speeds through

the plasma since they absorb the same amount of energy from the applied electric or magnetic field.

### 2.1.3 Ion and electron temperatures

Plasmas can be further classified into "hot" or "cold" plasmas, depending on the amount of energy supplied to the constituent molecules. The degree of heating will be related to the degree of ionisation observed. The relationship between the kinetic energy of a particle in a gas, and its temperature, is described [21] by

$$\frac{1}{2}m\overline{v^2} = \frac{3}{2}k_B T \tag{2.2}$$

where $\overline{v^2}$ is the mean square speed of the particle, $k_B$ is Boltzmann's constant, and $T$ is the temperature in Kelvin. It follows from Equation (2.2) that the mean square speed is given by $3k_BT/m$. A more useful parameter is the mean speed $\bar{v}$, which is not simply equal to the square root of $\overline{v^2}$ (as 'mean' and 'root mean square' are defined differently), but can be shown to have a value:

$$\bar{v} = \sqrt{\frac{8k_BT}{\pi m}} \tag{2.3}$$

The definition of a mean speed implies that some molecules travel slower and some molecules travel faster than $\bar{v}$. In a plasma, each of the species that exists within it can have their own temperatures, $T_i$, $T_e$ and $T_n$, for ion, electron and neutral temperatures respectively. Temperature can essentially be viewed as a measure of the speed at which the particles move through the plasma. Typically, a fluid or gas is in thermal equilibrium such that $T_i \approx T_e$, and the atoms and molecules of the fluid have a Maxwellian (Gaussian) velocity distribution $f(v)$ as they move randomly [21]. $f(v)$ is described by

$$f(v) = Ae^{-(\frac{1}{2}\frac{mv^2}{k_BT})}, \tag{2.4}$$

where A is a normalisation factor and the temperature $T$ determines the 'width' of the velocity distribution.

The plasmas found in stars are in thermal equilibrium. However, since the low-pressure discharges used in plasma processing are electrically excited and relatively

weakly ionised, the applied power preferentially energises the relatively low-mass electrons, while the relatively heavy ions exchange energy through collisions with the neutral gas particles. Hence $T_e \gg T_i$ for these plasmas. When expressed in Kelvin (K), the electron temperature $T_e$ of low-pressure discharges can reach extremely high values, e.g. 23000 K. However, since the heat capacity of the electrons is very small, the extreme temperatures do not mean that the vessel containing the plasma will melt. Merely the electrons will be moving at high speeds in the plasma. Ions, since they have lower speeds, will have temperatures $T_i$ only slightly above the ambient temperature, e.g. 500 K, in accordance with Equation (2.2).

### 2.1.4 Gas phase collisions

The processes that dominate the behaviour of a plasma are the random collisions that occur between the constituent gas particles. It is through collisions with one another that energy is transferred between ions, electrons and neutrals to maintain the Gaussian distribution of energies described in Equation (2.4). Collisions in a gas or plasma can be classified under two main headings:

- Collisions in which there is an interchange of kinetic energy only, similar to colliding billiard balls. These are *elastic collisions*.
- Collisions in which the internal energies of the colliding particles are changed. These are *inelastic collisions*.

In this context, internal energy changes refer to electronic excitation, ionisations, dissociations, etc. Elastic collisions are the simpler of the two collision types, where kinetic energy is conserved, and no new particles are formed. In contrast, inelastic collisions are capable of creating and annihilating particles in the plasma.

#### Elastic collisions

For two masses of mass $m_i$ and $m_t$, assuming that $m_t$ is initially stationary, and that $m_i$ collides with velocity $v_i$ at angle $\theta$ to the line joining the centres of $m_i$ and $m_t$ at the moment of collision (as depicted in Figure 2.1), an expression for the energy lost by the moving particle $m_i$ to mass $m_t$, $E_L$ [18] can be found

$$E_L = \frac{4m_i m_t}{(m_i + m_t)^2} \cos^2 \theta. \tag{2.5}$$

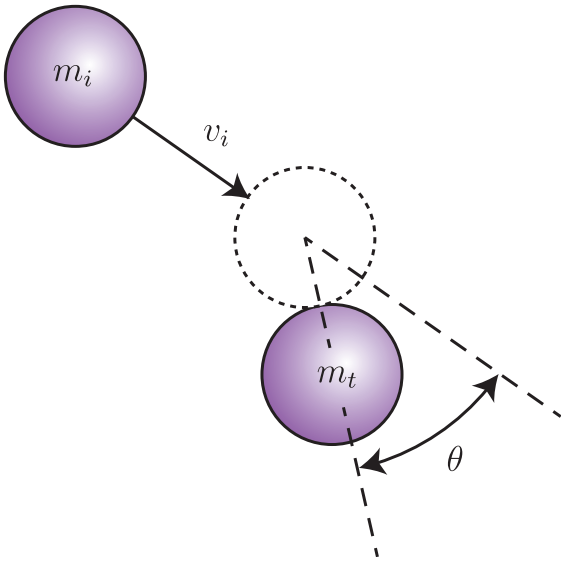



FIGURE 2.1: Two-mass elastic collision. Mass $m_i$ moving at speed $v_i$ collides with $m_t$ at angle $\theta$, assuming that $m_t$ is initially stationary.

This proportion of energy transferred has a maximum of $\cos^2\theta$ when both masses are equal ($m_i = m_t$). Energy is transferred evenly between two colliding particles of equal mass. In cases where electrons strike molecules or atoms, the difference in mass between the particles causes the speed of the electron to not be changed by much, but its direction is. Molecules are largely unaffected by kinetic collisions with electrons. In the case of a molecule with large mass striking a particle with much lower mass in a head on collision, it can be shown [21] that the particle with low mass will travel away from the collision at approximately twice the impact velocity.

The mean group velocity of electrons moving in the plasma under the influence of an applied electric field are limited by elastic collisions with other particles. While the elastic collisions described in Figure 2.1 apply to collisions involving neutral particals, the forces acting between charged particles for elastic collisions are the strong electrostatic Coulombic forces, determined from Coulomb's Law,

$$F = \frac{q_1 q_2}{4\pi\epsilon_o d^2}, \tag{2.6}$$

where $q_1$ and $q_2$ are the charges on each particle, $\epsilon_o$ is the permittivity of the medium, and $d$ is the distance between the particles. The Coulombic forces extend to relatively large distances around charged particles and typically, Coulombic collisions result in small-angle ($< 10°$) scattering of particles where the charged particles effectively "swing" around each other due to attractive or repulsive electrostatic forces acting at a distance

(depending on the relative charges of the particles). However, the cumulative effect of many small-angle collisions scatters particles by large angles ($> 90°$) [18, 22].

### Inelastic collisions

**Ionisation** Ionisation is a collision where one electron or more is knocked from or attached to a stable atom or molecule to create an ion. This process is necessary to maintain a plasma discharge. The main type of ionisation in a plasma discharge is *electron impact ionisation*, where an energetic electron strikes an atom and removes an electron from the atom.

$$\mathrm{e} + \mathrm{A} \rightarrow 2\mathrm{e} + \mathrm{A}^{+}$$

The two free electrons can now gain energy and cause further ionisation, maintaining the discharge. An electron that causes ionisation must have enough energy to overcome the *ionisation energy* of the atom or molecule with which it collides. The ionisation energy is the energy required to remove the outermost electron in an atom or molecule in its ground electronic state. Ionisation can occur as a result of a variety of energy sources, including photo-ionisation (where molecules or atoms are ionised using the energy from incident photons of light) or thermal activation (where molecules or atoms become ionised after gaining energy from heating).

### Excitation

During excitation, an electron in an atom gains enough energy to jump to a higher energy level in the atom as a result of the collision. The existence of discrete energy levels for electrons orbiting the atoms nucleus is described in the Bohr model of the atom introduced in 1913 [23]. In general, excitation is a less energetic collision compared to ionisation. Excitation occurs when the atom absorbs energy, and can be brought about by particle collisions, photo excitation or thermal excitation. An excited atom is usually indicated using an asterisk superscript ($^*$). The energy below which excitation will not occur is known as the *excitation energy* which, as expected, is typically much less than the ionisation energy.

$$\mathrm{e} + \mathrm{A} \rightarrow \mathrm{e} + \mathrm{A}^{*}$$

### Relaxation

Relaxation is the opposite process to excitation. Relaxation is the movement of an electron in an excited atom from a higher energy level to a lower one.

$$A^* \rightarrow A + hf_{ph} \text{ (photon)}$$

In this movement, the excess energy of the electron is generally released in the form of a photon as depicted in Figure 2.2. The energy of the photon corresponds to the difference in energy between the two atomic levels between which the electron moves. The frequency of the released photon is directly proportional to its energy, and is given by

$$E_{ph} = hf_{ph} = \frac{hc}{\lambda_{ph}} \tag{2.7}$$

where $E_p$ is the energy of the released photon, $h$ is Plank's constant, $c$ is the speed of light, and $f_{ph}$ and $\lambda_{ph}$ are the frequency and wavelength of the released photon respectively. This phenomenon leads to the characteristic glow of plasma discharges, and is used by optical emission spectroscopy (OES) (discussed later in Section 2.5.1) to deduce the gaseous species that exist in a plasma.

### Recombination

Recombination is the opposing process to ionisation whereby an electron and a positive ion combine to form a neutral atom. However, to conserve energy and momentum, a third body is often required for this collision to occur. This third body can be the wall of the chamber, or a third, neutral, molecule.

$$e + A^+ + A \rightarrow A^* + A$$

This is known as a *three body collision*. The probability of a gas atom being used as a third party for recombination, over a wall, increases with pressure, considering that the chamber walls are always present, but the number of atoms per $cm^3$ increases with pressure. Other, more unlikely, recombination processes that can occur are *two-stage* recombination processes,

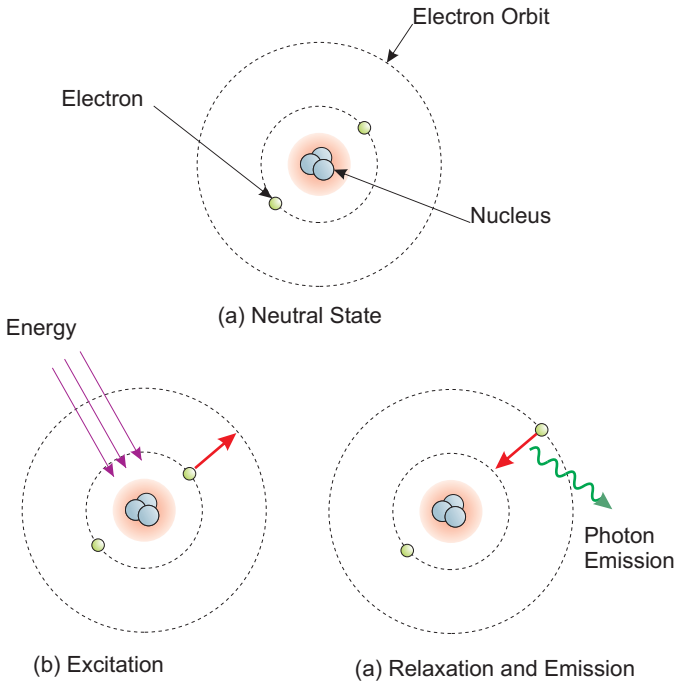



FIGURE 2.2: Excitation and relaxation with photon emission. Photons are released from atoms when electrons drop from higher to lower energy levels. The frequency of the emitted photon is proportional to the energy difference between the original and final electron energy levels. This diagram shows: (a) The atom in a neutral state where all electrons are in the lowest orbits available. (b) The excitation process. Energy is introduced to the atom from an outside source to excite electrons to higher energy orbits. (c) Relaxation. Excited electrons fall from their unstable outer orbits and release energy in the form of photons in the process.

$$\mathrm{e} + \mathrm{A} \rightarrow \mathrm{A}^-$$
$$\mathrm{A}^- + \mathrm{A}^+ \rightarrow \mathrm{A} + \mathrm{A}$$

and *radiative recombination*, whereby the excess energy from the collision process is carried away with a photon.

$$\mathrm{e} + \mathrm{A}^+ \rightarrow \mathrm{A} + hf_{ph}$$

### Dissociation

Dissociation is a process whereby collisions of sufficient energy break apart a molecule into its constituent atomic species. For example, an oxygen molecule could be broken into two atoms of oxygen through electron impact dissociation. The energy needed to achieve this, the *dissociation threshold*, depends on the strength of the chemical

bond between the atoms. Dissociation may also be accompanied by ionisation. This is the main process responsible for the creation of chemically active radicals in typical production plasmas.

$$e + A_2 \rightarrow e + A + A$$

**Electron attachment**

During electron attachment, an electron attaches to an electronegative atom to form a negative ion. The likelihood of atoms forming negative ions is determined by their electron configuration. Atoms with spaces left in their outer shell will have a strong affinity for electrons, whereas the noble gases, with full outer shells will not form negative ions with electrons. A common and important electron attachment reaction used in many industrial applications [22] is the dissociative attachment of $SF_6$ to form negative $SF_5^-$.

$$e + SF_6 \rightarrow F + SF_5^-$$

**Ion-neutral collisions**

Ions and neutrals often interact with collisions exchanging both kinetic and internal energy in the process. Ionisation can occur by fast ion or atom bombardment of a neutral, provided the incident particle has at least enough energy to overcome the ionisation threshold of the bombarded neutral. Charges can also sometimes be transferred between ions and atoms in these collisions:

$$A + A^+ \rightarrow A^+ + A$$

**Metastable collisions**

A metastable atom is one which has been previously energised above its ground state, and has a long lifetime (will exist for some time before relaxation occurs). These atoms can collide with all of the other particles in the plasma as normal, with slightly different results due to their relatively high energies. *Penning ionisation* is the ionisation of a neutral by a metastable atom. Two colliding metastables can have enough energy to overcome both of their ionisation thresholds, ionising each other in a collision. Similar

to ground state atoms, metastables can be ionised by simple electron impact ionisation. Because metastable atoms are energised above ground state, there are many more electrons in the discharge that have enough energy to ionise them.

### 2.1.5 Mean free path and collision cross section

The mean free path refers to the average distance travelled by a particle between two successive collisions. At a given gas temperature, the mean free path $l_f$ is inversely proportional to the gas density $n_n$ and the gas pressure $p$. In kinetic theory, the mean free path of particles with a Maxwellian distribution of velocities is given by:

$$l_f = \frac{k_B T}{\sqrt{2}\pi d_p^2 p} \tag{2.8}$$

where $d_p$ is the diameter of the gas particle and $p$ is the gas pressure. Using the mean free path, and the average speed of particles that is described in Equation (2.3), an expression for collision frequency $f_c$ is given by

$$f_c = \frac{\bar{v}}{l_f}. \tag{2.9}$$

To all but the slowest moving electrons, ions can be seen as relatively stationary during the approach of an electron to a collision. An effective collision cross-section of a gaseous atom/molecule can be defined which is used to express the likelihood of interaction between particles. As an electron approaches an atom, the Coulombic interaction between the electron and the nucleus and orbiting electrons of the atom is governed by the relative speeds and trajectories of the approach. There is an element of probability to the collision results and this probability is implicit in the definition of a collision cross-section. The collision cross section is dependant on the approach velocity as the particle interaction time will be dependant on the velocities at which they are travelling.

The idea of a collision cross-section is an alternative view to the mean free path. While the simpler model of a mean free path is usually reserved for elastic collision processes, the collision cross-section is often employed during the analysis of inelastic collisions. Every collision process documented in Section 2.1.4 has a collision cross-section that varies with electron velocity or energy. Since each collision is defined by a probability, it is usual to culminate all of the individual probabilities into one *total collision cross-section*, which is a measure of the probability of a particle being scattered during a collision.

### 2.1.6 Floating substrates and sheath formation

As defined so far, plasma is made up of an equal number of electrons and positive ions moving randomly in a sea of electrically neutral atoms and molecules. The plasma density is defined as the electron or ion density which is usually much less than the density of the neutrals present. It can be shown that the flux of particles (ions, electrons, or neutrals) per unit area, $\Phi$, is

$$\Phi = \frac{n\bar{v}}{4}, \tag{2.10}$$

where $n$ is the density of the particles of interest and $\bar{v}$ is their mean speed, calculated using (2.3). Following this, the current densities to a small electrically isolated substrate suspended in the plasma will [21] be

$$j_e = \frac{en_e\bar{v}_e}{4} \tag{2.11}$$

$$j_i = \frac{en_i\bar{v}_i}{4} \tag{2.12}$$

where $j_e$ and $j_i$ are the electron and ion current densities respectively. As $\bar{v}_e$ is much greater than $\bar{v}_i$, the electron current is much greater than the ion current, leading to a build up of negative charge on the isolated substrate.

The accumulated negative charge repels further incoming electrons to the substrate and attracts slower moving positive ions. The charge continues to build until eventually the electron flux to the substrate is retarded such that it evenly balances the ion flux. Apart from disturbances in potential such as this accumulated charge, the plasma is equipotential at the *plasma potential*, $V_p$. The isolated substrate in the plasma is charged to the *floating potential*, $V_f$. Because $V_f$ repels electrons, it is at a lower potential than $V_p$. Since electrons are repelled from the substrate, a region of net positive charge forms in the space around the surface of the substrate. This region of positive charge is known as a *sheath* and has an associated *space charge density* $\rho$ around it [21]. Poisson's equation relates the variation in potential $V$ with the distance $x$ across the sheath.

$$\frac{d^2V}{dx^2} = -\frac{\rho}{\epsilon_o}, \tag{2.13}$$

where $\epsilon_o$ is the permittivity of free space. The electric field across a distance $x$ changes in non-equipotential regions such as the sheath. The sheath has a net positive charge that causes a reduction in electron density close to the substrate and also a reduction in the amount of luminescence from this region (because of less electron impact collision processes). Hence, the sheath appears as a dark space around an object in the plasma.

The only electrons that can penetrate the sheath by overcoming the sheath voltage $V_p - V_f$, are those who strike the sheath with energy greater than $e(V_p - V_f)$. The fraction of electrons able to achieve this, $n'_e$, is found using the Maxwell-Boltzmann equation (Equation (2.4)) such that [21]

$$\frac{n'_e}{n_e} = exp\Big[ - \frac{e(V_p - V_f)}{k_B T_e} \Big] \tag{2.14}$$

The only electrons that make it through the sheath begin with a very high energy, and after traversing the sheath this is reduced to approximately $\bar{v}_e$ by the sheath voltage. The sheath has the effect of accelerating the ions that enter it, where they eventually strike the substrate surface with the energy of the sheath voltage, assuming no inter-particle collisions within the sheath itself.

The effects of the sheath, however, do not cease at the line where the ion and electron densities become equal again. There exists a quasi-neutral transition region of low electric field that extends into the plasma [21], first discovered by Bohm (1949). This region is known as the *pre-sheath*, and its effect is to increase the speed of the ions approaching the sheath.

Using the principles of conservation of momentum and energy, Bohm showed that the ions entering the sheath must have an initial velocity of

$$v_i = \left( \frac{k_B T_e}{m_i} \right)^{\frac{1}{2}} . \tag{2.15}$$

where $m_i$ is the mass of the ions. The ions acquire this energy in the pre-sheath region, which gives ions a directed velocity by the time they strike the substrate [22], hence increasing the ion flux to objects in the plasma. The separation of plasmas into quasi-neutral bulk regions and positive space charge sheaths is important in all plasma discharges because the directionality of ions striking substrates is important for many plasma processes. In the bulk plasma region, the instantaneous and time averaged electric fields are low, whereas in the sheath regions, high electric fields are present.

### 2.1.7 Debye Length

The Debye length $\lambda_d$ is a measure of the distance over which mobile charge carriers (e.g. electrons) screen out electric fields in plasmas. The Debye length is a measure of how rapidly a perturbation in potential is attenuated in the plasma, and its value is approximately calculated [21] as

$$\lambda_d = \left( \frac{k_B T_e \epsilon_o}{n_e e^2} \right)^{\frac{1}{2}}. \tag{2.16}$$

The shielding effect occurs because if a charge appears in a plasma, opposing mobile charge carriers cloud around it and shield its effect from the surrounding plasma. $\lambda_d$ increases with increased $T_e$, and would collapse to an infinitely thin layer in the absence of thermal movement [24].

Such charge screening maintains the quasi-neutrality of the plasma as a whole. After one Debye length, perturbations in charge are reduced to 0.37 of their initial value. The edge of the cloud that surrounds points of charge could be seen as the point where the electrostatic potential reduces to the thermal energy of the electrons and ions in the rest of the plasma, $\sim k_B T_e$. Effectively, a charge at a point in a plasma will be affected by interactions with other particles that fall within a sphere with a radius of one or two Debye lengths. For particles outside of this sphere, the effect of the interactions will be negligible.

### 2.1.8 Secondary electron emission

Secondary electron emission is the emission of an electron that occurs when a particle strikes a surface. Secondary electron emission can occur for ion, electron, photon and neutral bombardment. The *secondary electron emission coefficient* is defined as the average number of electrons emitted per incident particle [21]. The secondary electron emission process is an important source of electrons for plasma discharges, contributing electrons to the discharge to counteract electron loss mechanisms.

A secondary electron emission coefficient can be defined for every type of particle in the plasma that comes in contact with the surface in question. Both the secondary electron emission yield from electron bombardment and from ion bombardment depends heavily on the surface chemistry of the bombarded surface. This dependence is important in manufacturing processes where the target surface can be changed in time.

The emission coefficient also varies with the incident ion or electron energy. In general, insulators have a much larger secondary electron emission yield for ions than conductors.

Secondary electron emission as a result of photon bombardment of a surface is well understood and is known as *photoemission*. Photoemission occurs if the incoming photon has enough energy to remove one of the outermost electrons from the atoms of the surface. This energy is known as the *work function*, $\phi_w$, of the metal. Photoelectric yield is generally a low value, but rises with increasing photon energy, $E_{ph} = hf_{ph}$.

### 2.1.9 Plasma oscillations

When the quasi-neutrality of a plasma is perturbed, the plasma reacts to restore its neutrality, causing waves and oscillations to move through the plasma particles. These waves can be electromagnetic or acoustic (longitudinal) waves, and are known as plasma oscillations.

In the case where a set of the electrons in the plasma are moved by a small amount in one direction, a group of positive ions will be left behind. The electrons will immediately be attracted back to the ions by the excess positive charge, overshoot their original positions, and then return again. This movement repeats until the electrons settle, resulting in very fast, small amplitude oscillations that occur at the *plasma frequency* given by [21]

$$f_p = \sqrt{\frac{e^2 n_e}{\epsilon_0 m_e}} \frac{1}{2\pi} \text{ Hz}, \tag{2.17}$$

assuming ions of infinite mass. The plasma frequency is generally in the gigahertz range.

The waves that move through the ions in the plasma act in a different manner, and behave as ion acoustic or sound waves, which are longitudinal oscillations of the ions much like acoustic waves travelling in neutral gas. Similar to electron waves, ion waves are caused when the ions move from their equilibrium positions. The surrounding electrons can move quickly enough to shield out the electric field caused by this ion movement. However, since a portion of the electron motion is random thermal motion, the shielding effect is not perfect, allowing electric fields to leak out and create the ion acoustic waves. The speed of the waves $c_s$ is given [22] by

$$c_s = \sqrt{\frac{k_B T_e}{m_i}}. \tag{2.18}$$

Because the ions have a much larger mass than the electrons, the ion oscillations are much slower than the electron oscillations, typically having a frequency between zero and the megahertz ranges. Particularly slow ion acoustic waves can be detected with the naked eye as variations in the luminous intensity of the plasma .

## 2.2 Basic plasma discharges

In this section, a description of two basic plasma discharges are provided, direct current (DC) and radio frequency (RF) discharges. Although most plasma processing applications use a RF excited plasma, DC discharges are simpler to analyse and many of the principals transfer to RF plasmas.

### 2.2.1 DC discharges

A DC discharge is created by applying a DC potential between an anode and a cathode inside of a chamber filled with a low pressure gas. As free electrons and ions that exist in the gas from random thermal processes accelerate under the influence of the electric field between the electrodes, ionisation of other molecules begins as described in Section 2.1.4. The gas will take on the familiar glow of a plasma as the process continues due to the excitation and relaxation processes occurring between the excited particles.

DC discharges consist of several distinctive regions of glowing and dark spaces. Figure 2.3 shows the regions that appear, most of which are visible with the naked eye. The positive column varies with the length of the discharge tube, all other parts remaining a relatively constant size until the tube is made too small to have a positive column. The smallest size of discharge tube within which a plasma can be maintained is approximately twice the dark space thickness; at any smaller sizes the discharge is extinguished [21].

To maintain a continuous current in the system, the currents to each electrode of the discharge seen in Figure 2.3 must be equal. Typically the cathode may be at a potential of $-2000V$, and have a current density of 0.3 mA/cm$^2$, which is much less than the random electron flux expected at the electrodes when calculated from Equations (2.11) and (2.12). Hence, it is assumed that electric fields exist to retard the electron flux at

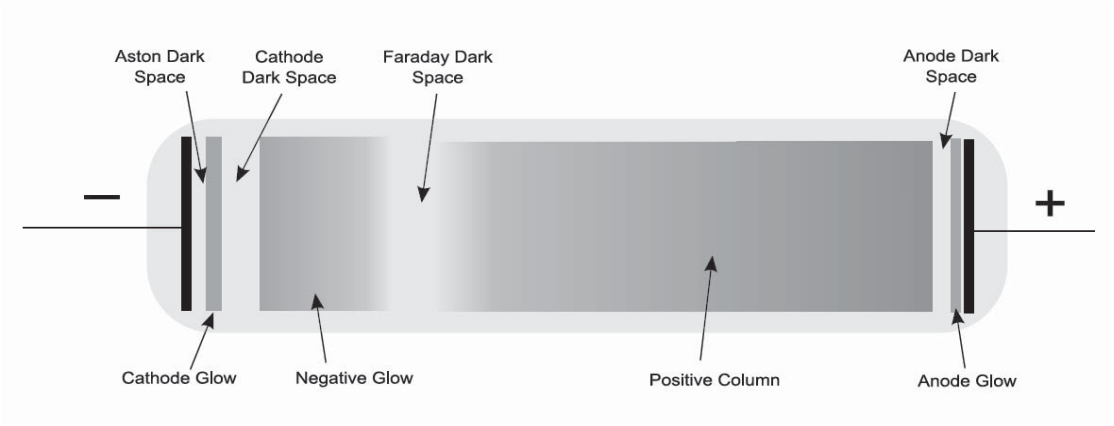



FIGURE 2.3: Structure of a DC excited plasma discharge. Several distinguishable regions form as a result of the collision processes occurring under the influence of the DC potential between the anode and cathode.

both the anode and the cathode, i.e. the plasma is at a higher potential $V_p$ than both of the electrodes. However, some current still flows, so the anode is more positive than $V_f$, the floating potential (Section 2.1.6). Figure 2.4 shows a simplified model of the potential variations across the discharge.

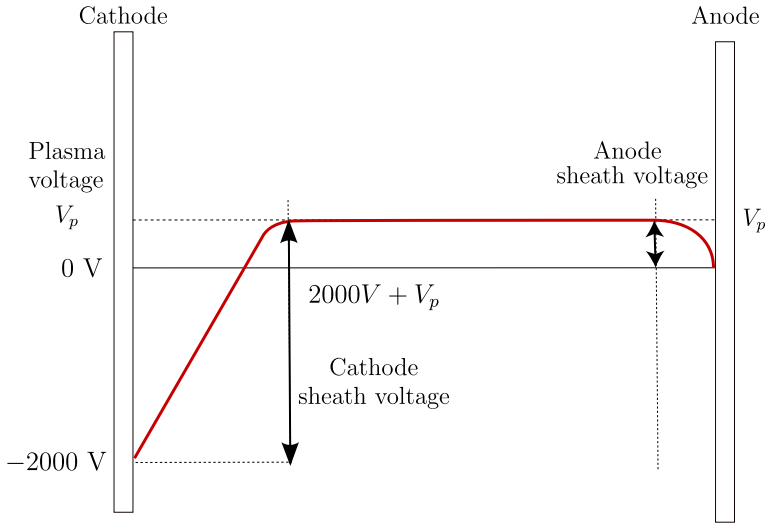



FIGURE 2.4: Voltage distribution in an example DC glow discharge process [21]. The plasma does not take a potential intermediate between those of the electrodes, but rather it is the most positive body in the system. The electric fields at the sheaths are such as to repel electrons from the electrodes.

For the discharge to be continuous, the energy losses from the discharge must be balanced by the energy input, and all recombination and relaxation must be balanced by ionisation and excitation.

In the cathode region of the discharge, secondary electron emission from the cathode plays a considerable part in electron impact ionisation in the cathode sheath region. Each

secondary electron can produce an avalanche effect of ionisation in the sheath area due to the strong electric field, helping maintain ion flux to the cathode. *Fast electrons* are also generated in the cathode sheath. These fast electrons are electrons that manage to travel through the sheath without collision with other particles, gaining all of the energy from the sheath potential. Such electrons have a small collision cross-section due to their high speeds and are likely to traverse the full discharge to impact the anode with considerable energies.

At the anode, the sheath is typically one order of magnitude smaller than that at the cathode. The fast electrons that are generated at the cathode strike the anode and create further secondary electrons that are accelerated by the anode sheath back into the glow region. These secondary electrons are a major source of electrons and energy to maintain the glow region of DC discharges. The anode sheath is of sufficient magnitude to repel some of the random electron flux from the plasma bulk. However, this repulsion is not as strong as for floating substrates since the net current to the anode is maintained to be an electron current.

The negative glow region of the discharge consists of an ionised gas of approximately neutral charge overall. However, due to the fast electrons produced at the cathode, this plasma is unlike the simple plasmas described in Section 2.1.1. Collisions between neutrals, metastable ions, and thermally excited electrons are the main sources of ionisation in the glow region. The electrons remain trapped in the glow region due to the potentials of the cathode and anode sheath regions.

A full description, and a more detailed physical examination of the constituent parts of the discharge, is provided in [18], [21] and [24]. A more detailed summary has also been completed in [25].

### 2.2.2 RF discharges

For most industrial applications, glow discharge processes are powered using oscillating electric fields, with high frequency power supplies. RF discharges are different from DC discharges since there are no dedicated anodes or cathodes, and no defined floating potential.

#### Why use RF?

The main reason for using RF discharges in plasma processes is that a DC bias cannot be applied to a semiconductor wafer in a processing chamber, as some of the layers

of the wafer may be made up of insulating materials. A DC bias is useful in plasma processing applications to ensure that reactive species are attracted to the wafer surface in a directional manner. Insulating layers ensure that upper layers on the wafer surface behave as electrically isolated conductors that assume the floating potential and are unusable for processing applications where directional ion fluxes and energies are required at the wafer surface [21]. It is, however, possible to impose a time-averaged DC bias to insulators using RF frequencies [22].

If the cathode/target in an RF discharge is an insulator, it gathers positive charge by losing electrons to incoming ions attracted during the negative parts of the RF cycle, and this positive charge is neutralised by electron bombardment during the rest of the cycle. With low RF frequencies ($\sim 50$ Hz), an on/off effect is observed where a series of short duration DC discharges are created. To create a continuous discharge, high frequencies are necessary, and in practice, a discharge can be maintained with frequencies above approximately 100 kHz [21].

Another advantage of RF frequency excited discharges is that they are more efficient at promoting ionisation and sustaining the discharge [21] than their DC counterparts. The enhanced ionisation arises from the fact that the movement of the electrons in the discharge are fast enough to be modulated by the applied RF frequency. Let us take for example an electric field of $\xi$ of amplitude $\xi_o$ and angular frequency $\omega$ along the $x$ direction.

$$\xi = \xi_0 \cos \omega t \tag{2.19}$$

The electron position and motion can be derived as

$$\begin{aligned} m_e \ddot{x} &= -e\xi_0 \cos \omega t \\ \dot{x} &= -\frac{e\xi_0}{m_e \omega} \sin \omega t \\ x &= \frac{e\xi_0}{m_e \omega^2} \cos \omega t \end{aligned} \tag{2.20}$$

$x$ gives the electron displacement from a point centered between the electrodes. The enhanced ionisation of RF discharges, when compared to DC discharges, arises when electrons make elastic collisions and reverse direction at the same time as the electric field reverses polarity. In this way, electrons can rapidly gain energy to reach the ionisation energy of the neutral atoms, for quite weak electric fields [21]. The electrode sheaths are also modulated by the RF frequencies and electrons can "collide" with these also, rebounding into the discharge with a greater momentum.

The RF discharge does not have as many clearly defined regions as the DC discharges shown in Section 2.2.1. Rather, RF discharges consist typically of only three parts: two electrode sheaths and the plasma bulk. The discharge is maintained by secondary electrons from ion bombardment of the electrodes, by exciting electrons in the plasma glow via the oscillating electric field, and finally by exciting electrons that "collide" with the modulated sheath.

### Discharge frequency

The frequency of the RF power applied to the chamber electrodes determines whether electrons are trapped in the inter-electrode space, or lost to the electrodes in each half cycle. The maximum displacement of electrons during each half cycle is given by $\frac{e\xi_o}{m_e\omega^2}$ which arises from Equation (2.20). Equation (2.20) was derived without taking into account the electron collisions in the plasma. Introducing a term $\nu_e$ as the collision frequency for momentum transfer, the maximum displacement $x_{max}$ can be shown [18] to be

$$x_{max} = \frac{e\xi_o}{m_e} \frac{1}{\omega(\nu_e^2 + \omega^2)^{\frac{1}{2}}}. \tag{2.21}$$

With an inter-electrode spacing of $d$, the cutoff frequency $f_{ce}$ for which electrons are trapped between the electrodes is found when $x_{max} = \frac{1}{2}d$, and given [24] by

$$f_{ce} = \frac{e\xi}{\pi m_e \nu_e d}, \tag{2.22}$$

assuming that $\nu_e \gg \omega$. If $f > f_{ce}$, where $f = \omega/2\pi$, the electrons become trapped and oscillate between the two electrodes. These electrons are now only lost from the inter-electrode region by lateral diffusion and other loss processes. Because ions have a much greater mass than electrons, the oscillatory motions of ions will be less by a factor of $\sim m_i/m_e \sim 10^3$, and so the ions can be considered relatively stationary at high frequencies. At frequencies above $f_{ce}$, a true continuous RF discharge is maintained between the electrodes.

### Self-bias of electrodes

As electrons in the discharge have a much smaller mass than that of the ions, their velocity is more affected by the applied RF field. Hence, for the same electric field, electrons carry more current than ions. In a discharge excited by a voltage with a square wave shape, more electrons strike the *target* electrode during positive sections

of the cycle than ions during the negative sections, resulting in the target assuming a negative bias (assuming the target is isolated from ground via a blocking capacitor or has underlying insulating layers). As a result, high energy ion bombardment is alternated with low energy electron bombardment at the target, ensuring that there is a fixed current conducted during each portion of the RF cycle.

When the applied voltage has a sine wave shape, the resulting voltage at the target electrode will be the same shape but offset in the negative direction by a fixed voltage known as the *dc offset voltage*. The target electrode has acquired a *self-bias* $V_{bias}$ as seen in Figure 2.5. While large ions may not move quickly enough under the influence of the applied RF power, they are accelerated by the persistent DC bias on the target electrode.

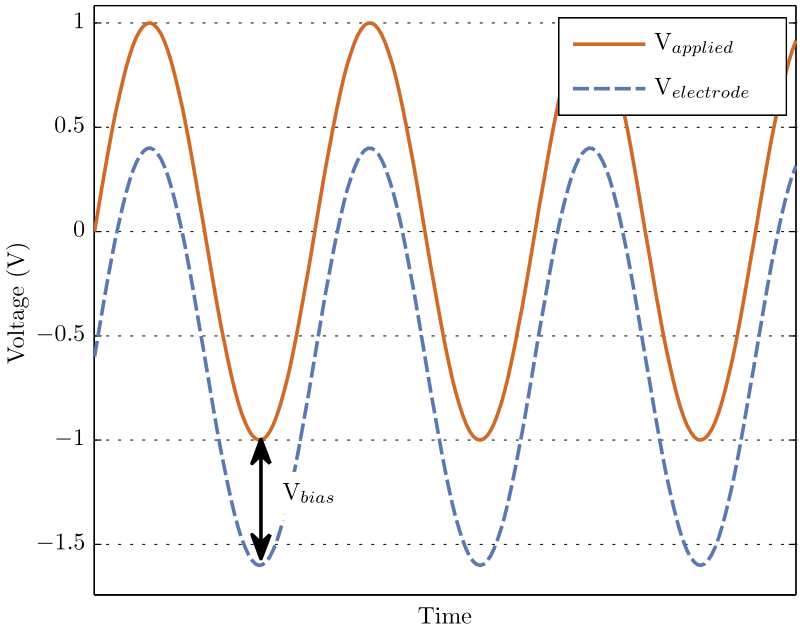



FIGURE 2.5: Voltage waveforms at the generator and target in an RF discharge. The electrode acquires a negative bias with respect to the applied potential from the generator. This effect is due to the relatively high mobility of electrons compared to the ions in the discharge.

### Voltage distribution in RF systems

For many applications, the energy of the ions that strike the surfaces of the electrodes, targets, and chamber walls is an important parameter. This energy is related to the voltage drop across the sheath at the plasma-surface interface. For the parallel-plate type RF reactor of unequal electrode areas shown in Figure 2.6, the relationship between the electrode voltage drops, $V_1$ and $V_2$, and the surface areas of the electrodes $A_1$ and $A_2$ is given [24] by

$$\frac{V_1}{V_2} = \left(\frac{A_2}{A_1}\right)^q. \tag{2.23}$$

$V_1$ and $A_1$ refer to the powered electrode, where the wafers are placed, and $V_2$ and $A_2$ refer to the grounded electrode. In some processing tools, the wafers are placed on a grounded electrode, and power is supplied to the other. A theoretical work by Koenig and Maissel in 1970 gave the exponent $q$ a value of 4, using a simplified plasma model. Their scaling law is restricted to low pressure situations and assumes also that the current density is equal at both electrodes and therefore, cannot be applied to all circumstances. The actual value of $q$ is typically found to be $< 2.5$ at higher pressures where collisions become more influential. A detailed discussion on the factors affecting $q$ and the $q$ for a number of chamber configurations is given in [18].

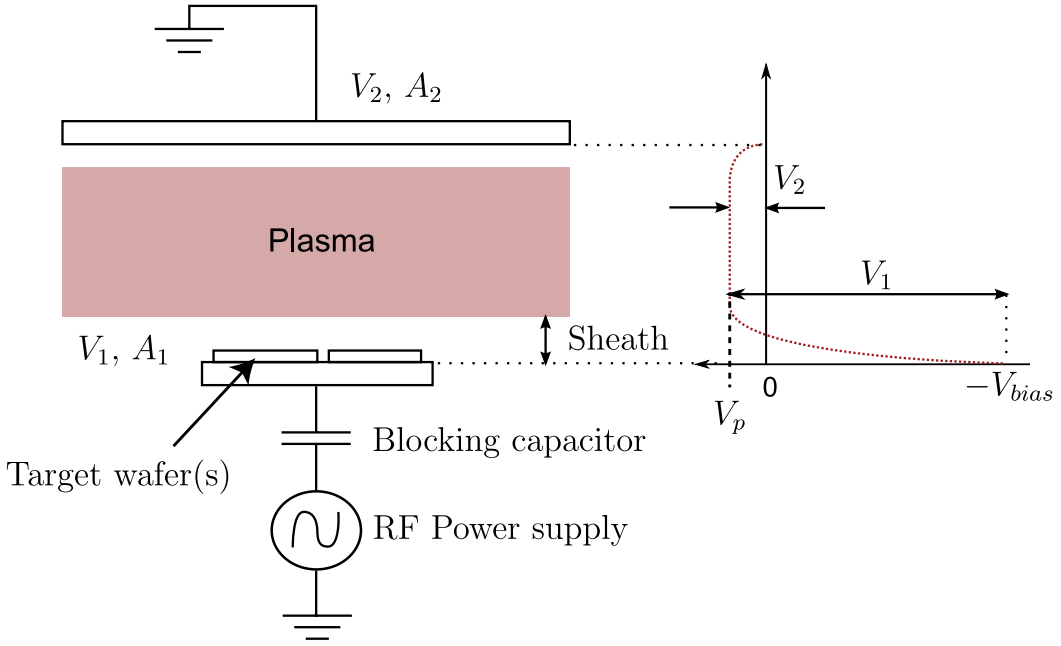



FIGURE 2.6: Voltage distribution in an parallel plate RF discharge with unequal electrode sizes [24]. In this diagram, $A_1 \ll A_2$. A larger sheath voltage is formed at the electrode with smaller surface area.

Hence, wall sheaths can be reduced and ion energy to the target can be increased by reducing the target surface area. However, current trends in microprocessing techniques require target areas/wafers that are becoming larger, which reduces the wall to target area ratio. This tends to increase the wall sheaths in the processing chamber, and encourage high energy ions to participate in wall reactions.

### Matching networks

To maximise the power transferred from the power supply to the plasma, it is common practice to include a matching network circuit on the powered electrode supply circuit. Power is not transferred efficiently if the RF power source is connected directly to the plasma discharge.

To understand this, the discharge is modelled as a complex load $Z_l$ with resistive $R_l$ and reactive $X_l$ components such that $Z_l = R_l + jX_l$. The RF power source is modelled as an ideal voltage source $V_s$ with complex amplitude $\|V_s\|$ in series with a source impedance $Z_s = R_s + jX_s$ (a Thevenin equivalent representation). The current in the circuit has a complex magitude $\|I\|$, which is given by

$$\|I\| = \frac{\|V_s\|}{\|Z_s + Z_l\|}. \tag{2.24}$$

The average power dissipated in the plasma is given by the square of the mean current multiplied by the resistive portion of the plasma impedance

$$P_l = I_{rms}^2 R_l = \frac{1}{2}\|I\|^2 R_l = \frac{1}{2}\left(\frac{\|V_s\|}{\|Z_s + Z_l\|}\right)^2 R_l \tag{2.25}$$

$$P_l = \frac{1}{2}\frac{\|V_s\|^2 R_l}{(R_s + R_l)^2 + (X_s + X_l)^2} \tag{2.26}$$

The maximum power transfer from the source to the load is obtained when $X_l = -X_s$ and $R_l = R_s$ such that

$$P_{max} = \frac{1}{8}\frac{\|V_s\|^2}{R_s}, \tag{2.27}$$

and the source and the load are said to be *matched* [18].

Typically, $R_l \ll R_s$ and $X_l \neq -X_s$ and so maximum power transfer is not achieved with a direct connection between the power source and plasma electrodes. A matching network circuit between the RF source and the plasma discharge is used to achieve maximum power transfer as shown in Figure 2.7. The matching network is designed to present a purely resistive load to the generator equal to the resistance of $R_s$. Typical RF generators have an output resistance of 50 Ω.

Matching networks contain variable elements that allow the impedance presented to the RF generator to be changed in order to match the changing plasma parameters. For parallel-plate plasma reactors, the plasma discharge is typically found to be

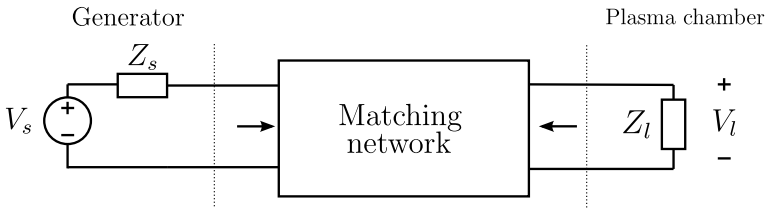



FIGURE 2.7: Matching Network between plasma and generator. The matching network is automatically controlled to ensure that maximum power transfer between the generator and the discharge is maintained.

capacitive, with the capacitance changing in response to changes in the process (temperature, pressure, gas flows etc.). A typical three element matching network is shown in Figure 2.8. This configuration is known as a "pi" network, due to the configuration of the three impedances. Three element networks typically consist of a fixed inductance value in combination with two variable capacitive elements, termed the *load* and *tune* capacitors. Two element "L-type" networks are also used extensively for low resistance loads.

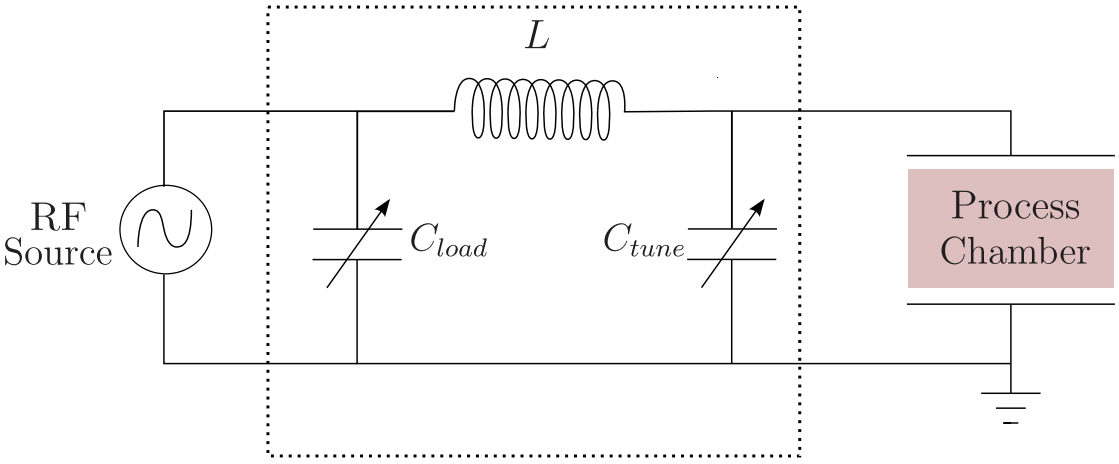



FIGURE 2.8: RF Matching Network in pi configuration. Two variable capacitors are used to allow the matching network to adjust to counteract for the variable complex impedance presented by the process chamber. The pi-network configuration gets its name from the shape of the inductance and dual capacitor circuit.

The matching network maximises power transfer to the plasma by ensuring that the load that the RF power source sees is purely resistive, and protects the generator from any reflections that may be induced in the power circuitry. The tunable vacuum capacitors tend to be large and expensive [22], and considerable RF expertise is required to set up the matching network, as every wire has an appreciable inductance and capacitance at the high frequencies in use.

## 2.3 Plasma etching

With a basic understanding of plasma discharges, plasma etch is now examined. Plasma etching is the use of a plasma to remove material from a surface in a controlled manner. The main process is the reaction between the ions created in the plasma with material on the target surface. This reaction is engineered to form a volatile etch product that is removed by means of a gas flow through the entire plasma etch chamber. Ions from the plasma are attracted to the wafer surface by a DC bias voltage on the wafer. The ions are accelerated towards the wafer in a highly directional manner by means of the bias voltage and sheath voltage at the wafer surface.

### 2.3.1 Why plasma etch?

Before the advent of plasma etching (pre 1960), the first etching processes used liquid-phase ("wet") etchants. Although cost effective and often providing infinite *selectivity* (the ability to etch one material and not another), wet etching produces an *isotropic* etch, that is etching that proceeds in all directions simultaneously. The minimum feature size of such techniques is hence limited, and plasma etching (sometimes called *dry* etching) is required to obtain a more directional or *anisotropic* etch to cater for the dense packing of today's microchips.

Let us assume a surface where the lithographic pattern is in the x-y plane and the z direction is normal to this plane. An etch process is described as isotropic if the etch rate in the x, y and z directions are equal [26]. Anisotropic etch processes usually have etch rates which are faster in the z direction as shown in Figure 2.9.

Since the ions of a plasma discharge bombard the wafer surface in a downward direction governed by the sheath voltage and the DC bias that is placed on the wafer, a highly directional etch can be achieved using plasma etch for two main reasons [27]. Firstly, ion bombardment damages the wafer surface so that it becomes more reactive, and secondly, the bombarding ions help remove etch-inhibiting species from the surface. Ion collisions in the sheath can reduce the directionality of the approaching ions, resulting in sidewall bombardment and some lateral etch. This effect is counteracted with reduced chamber pressure to lower the number of neutrals present in the sheath, reducing the number of collisions that occur.

Plasma etching is typically used to remove thin layers from wafer surfaces. The etch is complete when the layer is fully etched and the etched trenches have attained a desirable profile. Accurate detection of this *end point* is crucial to prevent unwanted

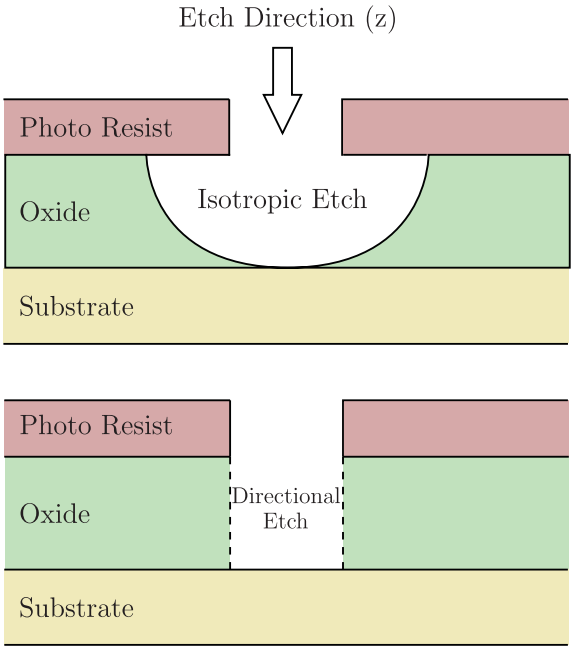



FIGURE 2.9: Isotropic and directional (anisotropic) etch profiles. There is a limit to the smallest achievable feature size that can be etched using isotropic etch methods.

etching of the next layer. Manufacturing processes typically include an *over-etch* step that uses less aggressive chemistry than the main etch step to clear all material from the bottom of etched trenches once end point has been detected.

### 2.3.2 Plasma etch mechanisms

There are a number of different mechanisms that occur during the etch of a wafer surface. The complete etching process can be reasonably approximated by the following steps [22]:

1. Reactive species are created by electron collisions in the plasma.
2. Reactive species are transported to the wafer surface by means of the DC bias on the wafer and the sheath at the plasma-surface interface.
3. The species are adsorbed on the surface (physisorption or chemisorption)
4. The etch product is formed on the wafer surface by dissociation of the reactant, formation of bonds to the wafer surface or diffusion into the surface.
5. The volatile etch product desorbs from the wafer surface.
6. The etch product is transported back into the plasma.

Figure 2.10 shows a schematic of the processes that are occurring in a simplified plasma etch chamber. With this overall view of the etch process, details of the reactions and mechanisms that occur are now examined.

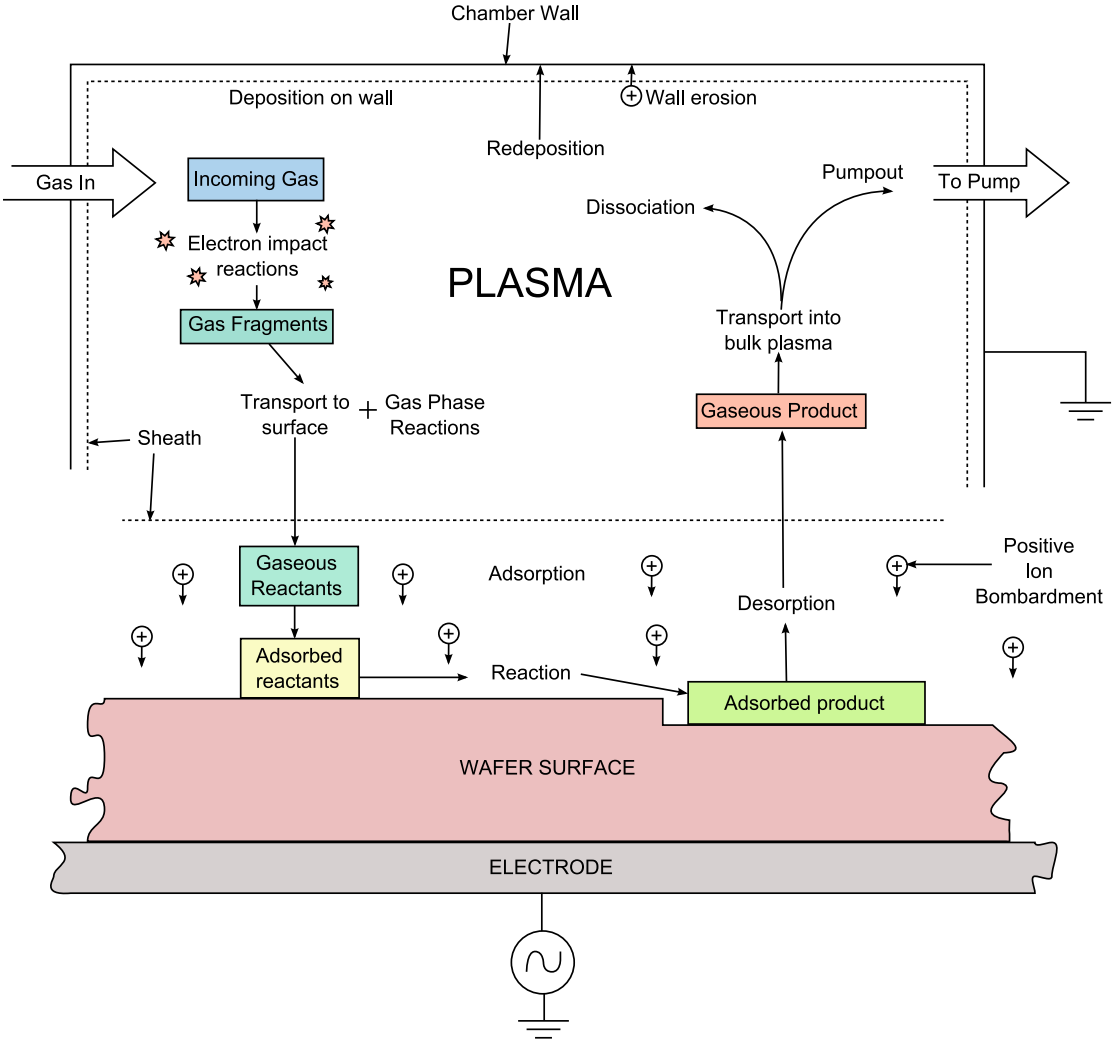



FIGURE 2.10: Schematic of Etch Process [27]. Reactive etchant species are created in the plasma which react with the wafer surface to form volatile etch products. A continuous gas flow is maintained to replenish the supply of etchant species.

### Spontaneous surface etching

Spontaneous surface etching, or *chemical etching*, is a chemical method in which neutral reactive species generated by the plasma interact with the materials surface to form volatile products [27]. Examples of such processes are the reactions between F with Si, or $Cl_2$ with Al.

Chemical etching usually provides good selectivity, isotropic etch profiles and a low ion-bombardment-induced damage to the wafer. The etch rate of chemical methods

depend on a number of factors such as the chemicals in use, the temperature of the surface and the silicon doping level. More details on these effects are discussed in [22].

### Mechanical etching

This type of etching is a more physical process than chemical etching and is also known *sputter etching*. During sputter etching, the target surface is bombarded by high-speed positive ions. With the energy of each impact, particles are physically removed, or *sputtered* from the target surface. Most of the particles are ejected by momentum transfer from the incoming ions. The bombarding ions are accelerated by the sheath potential as they approach the wafer surface and their impact energy can be controlled by adjusting the bias voltage on the wafer. Sputtering yield is affected by the angle of collision, ion bombardment energy and the mass and energy of the incoming ions [22].

Sputter etching yields very directional etch performance, but results in surface damage from ion bombardment of the wafer. Selectivity is difficult to achieve during sputter etch [27].

### Ion-enhanced chemical etching

This is a hybrid technique, that uses a physical technique to enhance the chemical etching of a surface. First discovered in 1979, early experiments found that when a surface is exposed simultaneously to both chemical etching neutrals and ion bombardment, the resulting etch rate was greater than the sum of each method individually. This famous experiment was carried out by Coburn and Winters [26], using $XeF_2$ and $Ar^+$ to etch Silicon. Their findings are shown in Figure 2.11.

During ion-enhanced chemical etching, the bombarding ions give energy to particles on the surface and encourage the etching reaction. Studies have demonstrated that ions with a greater mass contribute to faster etching since they dissipate more energy on the target surface [22].

The combination of etching mechanisms takes the advantages of each technique individually, and is associated with anisotropic etch profiles, good selectivity and relatively little ion-bombardment damage to the wafer surface [27].

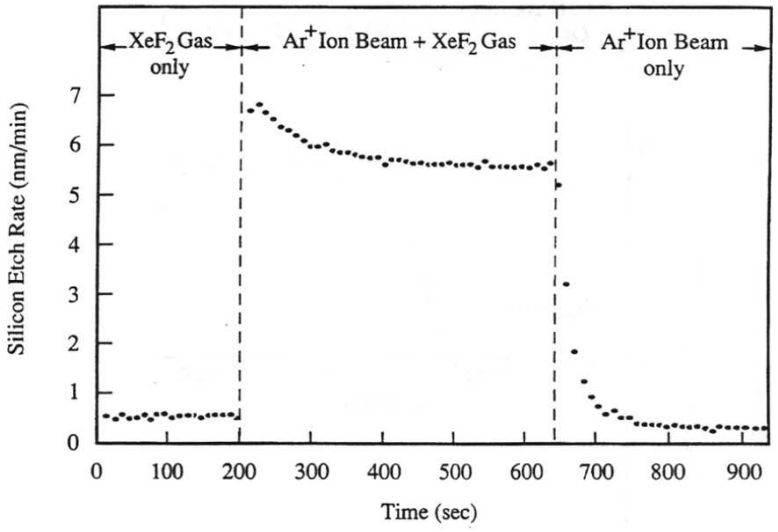



FIGURE 2.11: Enhancement of chemical etching etching via ion bombardment. These results were first obtained by Coburn and Winters in 1979 [26] using $XeF_2$ and $Ar^+$ to etch silicon surfaces.

### 2.3.3 Selectivity and Polymerisation

Selectivity in an etch process is the ability to alter the etch rate of one material compared to another. Selectivity allows finer control over the etch process and ensures that only the required material is etched from the wafer surface. As an example, the etching of silicon (Si) and silicon dioxide ($SiO_2$) using a $CF_4$ discharge is examined. In this process, through electron impact ionisation, $CF_4$ molecules are dissociated (in 90% of cases) [21] as

$$e + CF_4 \rightarrow CF_3^+ + F + e.$$

Less often, $CF_2^+$ and $CF^+$ are produced. The $CF_3^+$ molecules bombard the silicon wafer surface under the influence of the sheath voltage and applied DC bias. The impact energy at the surface cause the molecules to dissociate further into carbon and fluorine atoms. The fluorine atoms F react readily with the silicon to form $SiF_4$, a volatile product that desorbs from the wafer surface and is transported to the exhaust of the etch chamber.

$$Si + 4F \rightarrow SiF_4$$

Silicon Dioxide ($SiO_2$) is etched in a similar process such that

$$SiO_2 + 4F \rightarrow SiF_4 + O_2.$$

The carbon that remains on the surface after the etching reactions is removed by forming carbon oxides using oxygen from the discharge. Without the use of oxygen, the carbon forms $C_2F_6$ and desorbs back into the plasma. This formation of $C_2F_6$ consumes fluorine, hindering the silicon etch process. On $SiO_2$ surfaces, oxygen is more readily available as a product of the etch reaction and so the carbon atoms are removed easily.

Introducing variations in the gaseous make up of the plasma allows the selectivity of the etch process to be adjusted as required. It is possible to increase the etch rate of silicon by adding oxygen to the discharge. This oxygen removes the carbon produced during the etch reaction readily from the wafer surface, exposing it for further etching and freeing up more fluorine atoms so that etching can proceed faster. The addition of hydrogen gas to the discharge causes the etch rate of Si to decrease while the $SiO_2$ rate remains fairly constant [24]. Hydrogen has a high affinity for fluorine and readily forms HF [21], reducing the number of F atoms available for etch, hence reducing the etch rate. In $SiO_2$, this effect is less dramatic as it is offset by the "built-in" oxygen supplied by the etch reaction. Effectively, the F:C ratio in the plasma is adjusted in this selectivity tuning process. Solid materials can also absorb fluorine from the discharge to reduce the F:C ratio, or the input gas could be changed to one that contains fewer fluorine atoms per carbon atom, e.g. $C_5F_{12}$.

Hence, high selectivity of $SiO_2$ etching over Si etching is achieved by limiting the F:C ratio. However, a coating of carbon in the form of a polymer $(CF_2)_n$, begins to form on all surfaces in the chamber if the F:C ratio is sufficiently limited. The condensation of these polymers from the plasma onto surfaces is known as *plasma polymerisation* which stops the etching process everywhere. The achievement of high $SiO_2$ selectivity is therefore an exercise in trying to get as close to the onset of polymerisation as possible [24] while still etching $SiO_2$.

Ion bombardment of the wafer surface slows the onset of polymerisation on the wafer surface. Often, polymerisation can be occurring on the walls of the plasma chamber, but under ion bombardment, etching is simultaneously occurring at the wafer surface [21]. Polymerisation can be used to form protective coatings in some applications, and so is not always unwanted. A combination of polymerisation on the walls of etched trenches with directional etching at the bottom of the trenches can be used in some processes to produce very high aspect ratio trenches on wafer surfaces.

### 2.3.4 Uniformity

Uniformity in plasma etch refers to two things: the evenness of etching across a single wafer and the degree to which etch rates are maintained from wafer to wafer in the same chamber [24].

Uniformity across a single wafer requires spatial uniformity of plasma properties such as plasma density, species concentrations, and electron temperature, across the entire surface of the wafer. A uniform supply of etchant chemicals to the wafer area is important to maintain a consistent etch rate, and is often achieved via a shower-head gas delivery system. Etch rates can also vary across the wafer depending on feature size and pattern density. These uniformity problems fall under two categories - aspect ratio dependant etching (ARDE) and pattern dependent etching, or *micro-loading* [27]. The *aspect ratio* of an etched trench is the ratio of the depth of the trench to its width. ARDE causes trenches in the wafer with a large aspect ratio ($> 5:1$) to etch more slowly than trenches with a smaller aspect ratio (mainly due to gas transport limitations). The *loading* effect in plasma etch is the effect that different exposed areas on the etch wafer surface have on the etch rate achieved. Differing exposed areas lead to different consumption rates of etchant species from the plasma. Micro-loading occurs when etchant species are depleted from localised areas of the wafer due to uneven distributions of exposed substrate across the wafer surface.

Uniformity between different wafers processed in the same chamber depend on a variety of effects that can be difficult to quantify. Etch chambers typically undergo scheduled preventative maintenance (PM) operations on a regular basis that can dramatically alter the operating characteristics of the etch process as components are replaced and/or cleaned. As wafers are processed between each PM event, etch chambers undergo a conditioning effect where material arising from each etch cycle is deposited on the chamber walls. This conditioning can alter the chamber behaviour, influencing the etch rate achieved for each wafer. Etch chamber performance is typically monitored in fabrication environments through the use of statistical process control (SPC) [9] and regular cleaning and maintenance cycles are used to keep process performance within specifications. Etch behaviour also differs between wafers processed in individual lots as etch chambers heat and condition due to repeated etching processes. The *first wafer effect* describes the typical phenomenon whereby the first 1-2 wafers etched in each lot yield considerably different results compared to the remaining wafers.

### 2.3.5 Plasma property effects

The etching plasma is highly sensitive to variations in the plasma chamber conditions. The main characterising properties are the etchant gas, the chamber pressure, the frequency of the discharge, the gas flow patterns, and the size and shape of the chamber itself [24]. In this section, the effects of pressure, frequency, and gas flow on the plasma are examined in some more detail.

#### Pressure effects

The pressure of the plasma chamber influences the properties, and hence the etching behaviour, of the plasma [24]. The plasma properties influenced include:

1. RF voltage amplitude, which affects sheath potential and ion bombardment energy,
2. Sheath thickness, in situations with mobility controlled ion motion (collisional plasmas),
3. Electron temperature, which controls ion-to-radical abundance ratios, and
4. Relative rates of different chemical processes in the plasma.

The most dramatic effect of pressure variation is on the sheath potential. As pressure decreases below approximately 1 Torr, the sheath voltage drops begin to increase sharply [24] and the increased potential causes ions to strike the target with much more energy. Since the mean free path of particles is inversely proportional to pressure, this also adds to the higher energy ion flux to the surface, shifting the main etch mechanism from chemical to physical etching. There is a threshold value for this shift to occur that depends on the ions used and the surface being bombarded. Too high of an ion bombardment energy can lead to damage of the wafer surface and a loss of selectivity.

#### Frequency effects

As well as DC and RF excited plasmas, there are also those that are created using microwave energy. RF discharges usually operated at 13.56 MHz, because this frequency has been set aside by the Federal Communications Commission (FCC) authority to avoid communications interference. Microwave discharge sources usually include waveguides, magnetrons and other speciality equipment, and operate at 2.45 GHz, again in accordance with FCC regulations.

The frequency used to excite the plasma discharge affects the plasma properties in a number of different ways [24]:

1. Frequency can affect the spatial distribution of species and electric fields across the discharge.
2. Frequency determines whether the energy and density of species in the plasma are constant or fluctuating in time.
3. The frequency of the RF power controls the electron energy distribution, thereby affecting ion-electron interactions.

Frequency also has an effect on etching behaviour in an etch chamber, affecting selectivity ratios, polymerisation activity, and etch rates. Considerable research has been done in this area, and is discussed in the text by Sugawara [24].

#### Gas flow rate effects

Etch rates vary significantly with the flow rate of the reaction gases. Flow rates are measured in "standard cubic centimetres per minute" or *sccm*. One sccm is a cubic centimetre of gas at standard temperature and pressure (STP), that is 273.15 K (0 °C) at 100 kPa (1 bar).

In general, etch rates increase rapidly as flow rates increase, reach a maximum, and then tend to fall off as the flow rate increases further. With very low flow rates, the etch rate is limited because etch products dominate the discharge for a considerable time before being replaced by the inflow of new etchant gas. The consumption rate of etchant species is related to the exposed area on the wafer being processed. To achieve maximum etch rate, the gas flow rate must be altered so that new etchant species are supplied to replace the etch products in a timely fashion. However, 100% utilisation of the etchant species is difficult to achieve since etchant species are lost unpredictably to the chamber walls and other surfaces.

As the gas flow rate is increased further, the etch rate is reduced because etchant species exit the chamber before they etch the target surface.

## 2.4 Plasma processing reactors

Plasma processing systems used in industry have four main subsystems [22]:

1. *The vacuum system:* Base pressures in plasma processing chambers, before gas introduction, are as low as $10^{-6}$ Torr. Such low pressures are typically attained using a combination of two pumps. Firstly, a turbo-molecular pump, or *turbo-pump*, uses a high speed fan to remove gas from the chamber. Secondly, a *fore-pump*, a large mechanical pump, is placed between the turbo-pump and atmospheric pressure to reduce the pressure gradient across the turbo-pump.

2. *The gas handling system:* Gases are introduced to processing chambers in controlled amounts using mass flow controllers (MFCs). MFCs consist of a flowmeter, a controller, and a valve, and they are located between the gas source and the chamber itself [28]. MFCs alter the amount of gas passing through them according to a provided set point. A large gate valve at the chamber exhaust controls the gas flow rate out of the chamber and regulates the chamber pressure.

3. *The cooling system:* The heat generated during plasma processing must be removed to avoid interference with the process and/or product. Chamber components are typically water cooled to control their temperature. The wafer is maintained at a constant temperature by feeding a coolant gas to the backside of the wafer. Electrostatic chucks are typically used to hold wafers in place in etch chambers, where a DC voltage is placed on the chucks to induce an opposite charge on the back of the wafer and fix it in place. The wafer coolant, typically helium, passes through grooves in the chuck and along the backside of the wafer, regulating the wafer temperature.

4. *The power system:* A steady supply of power is required to maintain a stable plasma discharge. For RF plasmas, power is supplied by solid state power amplifiers with built in oscillators that generate RF signals [22]. Directional couplers measure the power flowing to and back from the antenna / matching unit. In plasma excited using microwave energy, microwaves are produced at 2.45 GHz by magnetrons outside of the chamber. The microwaves travel down a waveguide, through a quartz window and into the chamber to excite the plasma.

In this section, three of the main types of processing chambers that are used for plasma etch are examined.

### 2.4.1 Capacitively coupled plasmas

Capacitively coupled plasmas (CCPs) are plasmas in which power is capacitively transferred to the input gas to form a plasma. Typically, parallel-plate reactive ion etch

(RIE) chambers using CCPs consist of two flat circular electrodes separated by a gap in which a plasma is generated [22]. Such chambers are arguably the simplest type of etch chambers and were used extensively in industry from their invention in 1970 until the mid-90s, when newer technology became more widespread. The target wafer is placed on the lower electrode and fixed in place using a chuck (either electrostatic or mechanical). Electrostatic chucks are preferable because they prevent the wafer surface bulging from the pressure of the coolant circulated on its backside. RF power may be applied to either one or both electrodes to produce the plasma. The sheaths between the plasma and the wafer control the ion flux to the wafer. A schematic of a parallel-plate CCP chamber is shown in Figure 2.12.

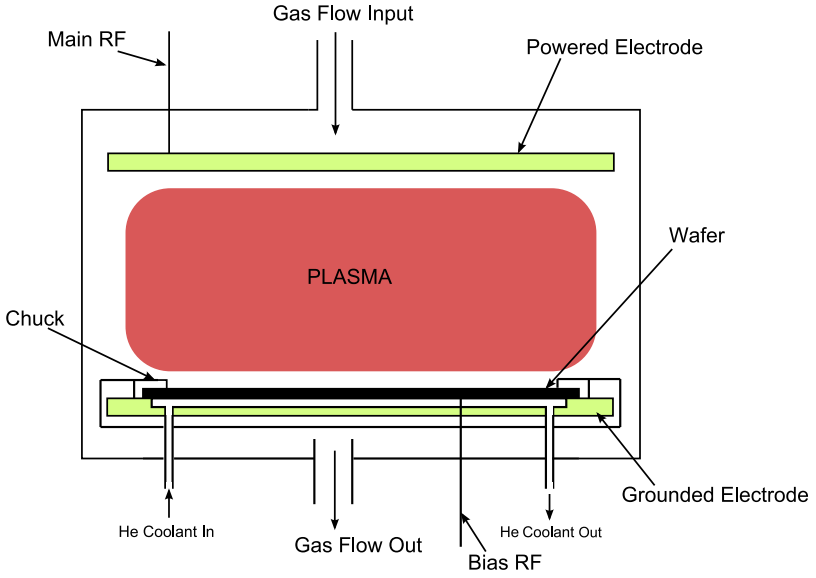



FIGURE 2.12: Parallel-plate reactive ion etch (RIE) chamber [22]. The etch chamber in this diagram is a top powered RIE chamber with a mechanical chuck system. The plasma is generated capacitively between the electrodes.

Although simple to maintain and understand, parallel-plate RIE chambers have a number of disadvantages. Firstly, since RF power controls both plasma density and the magnitude of the sheath voltage, ion flux to the wafer cannot be controlled independently of ion energy. Additionally, parallel-plate RIE discharges generally produce lower density plasmas and require higher operating pressures than newer technologies. Finally, the electron temperature tends to be higher in parallel-plate RIEs, and this can lead to excessive heating of the wafer. Regardless of these disadvantages, RIE chambers are still in use today in semiconductor manufacturing plants for some etch processes.

A further modification to parallel-plate RIE reactors is the magnetically-enhanced RIE (MERIE) chamber. MERIE chambers use permanent magnets behind the wafer or DC coils around the chamber to generate a magnetic field that is parallel to the

wafer surface [27]. The magnetic field has the effect of confining electrons to circular trajectories near the electrodes, reducing losses to walls and increasing collision frequency and hence plasma density. By confining electron movement, the electron flux into the sheath is also reduced, which reduces the Coulomb barrier at the sheath (the potential that normally retards electron flux). The reduction in the Coulomb barrier has the effect of reducing the RF fluctuation amplitude of the sheath potential. However, the magnetic fields used by MERIE chambers are known to increase damage to thin oxide layers on wafer surfaces and potentially cause non-uniformities in the plasma. To decrease non-uniformities, the direction of the magnetic field can be rotated slowly during processing [22].

### 2.4.2 Inductively coupled plasmas

Inductively coupled plasma (ICP) etch chambers differ from CCP (as shown in Figure 2.12) etch chambers since the RF field used to generate the plasma is inductively coupled to the plasma by an external antenna. ICP chambers are capable of generating high-density, low-pressure plasmas and allowing independent control of ion flux and ion energy at the wafer surface [27] through the use of a separate power supply to create a bias voltage at the wafer. No internal electrodes are used in ICP systems, and no DC magnetic field is required (as in MERIE chambers). The plasma can be generated close to the wafer surface by placement of the antenna, allowing high etch rates to be achieved.

The simplest form of ICP is a water cooled copper coil surrounding a cylindrical chamber in which the plasma is generated. The coil acts as an electromagnet, inducing an RF magnetic field (and hence an electric field), in the chamber to create a plasma.

Several variations on this design exist. A *helical resonator* uses a coil that is designed to naturally oscillate at the drive frequency, allowing the RF power to be supplied to one end of the antenna only, acting as a tank circuit. A *transformer coupled plasma* (TCP) chamber places an antenna in a spiral on top of the chamber so that as much energy as possible is coupled to the centre of the plasma. A *detached plasma source* (DPS) combines the windings of both the helical and transformer coupled designs to form a dome shaped antenna. The advantage of detached plasma sources is that the plasma is further removed from the wafer itself, allowing it to diffuse and become more uniform before etching the surface [22]. Figure 2.13 provides an overview of the main ICP sources.

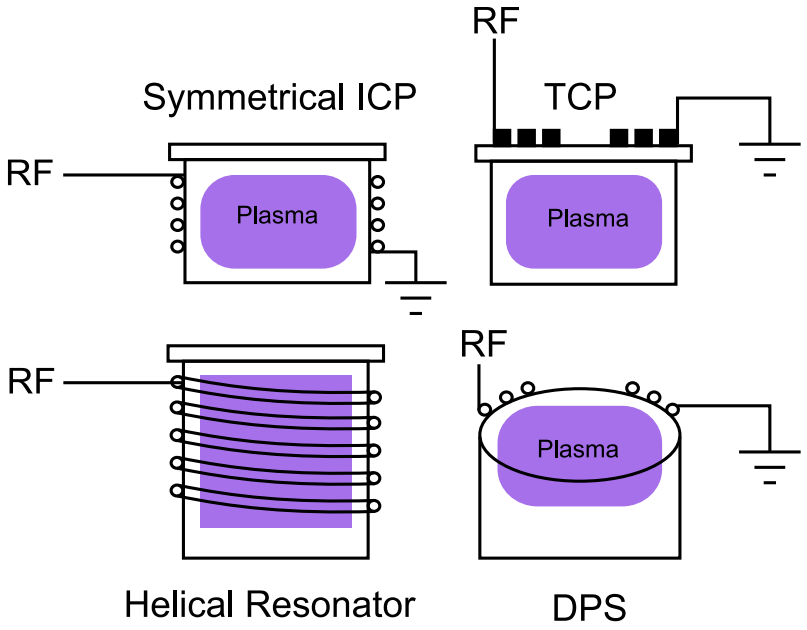



FIGURE 2.13: Inductively coupled plasma (ICP) sources [22]. In ICP sources, the plasma is excited by an RF field that is induced in the gas via an external antenna. A transformer coupled plasma (TCP) uses a spiral antenna at the top of the plasma chamber, and a detached plasma source (DPS) combines the windings of both the helical and transformer coupled designs to form a dome shaped antenna.

### 2.4.3 Electron cyclotron resonance sources

Electron cyclotron resonance (ECR) sources use electromagnetic radiation at microwave frequencies, along with with strong magnetic fields, to create a plasma. ECR sources are popular in semiconductor manufacturing since they can produce low-pressure, high-density plasmas with independent control of ion energy and plasma density. However, the strong magnetic field and complex microwave waveguide equipment makes these reactors more complicated and expensive than other chamber designs.

In a magnetic field, electrons rotate around the magnetic lines of force with angular frequency [24] $\omega_{ce}$ given by

$$\omega_{ce} = \frac{eB}{m_e}, \tag{2.28}$$

where $e$ is the electron charge and $B$ is the magnetic flux density. The frequency $\omega_{ce}$ is the *electron cyclotron frequency* or gyrofrequency. In ECR plasma sources, microwaves oscillating at the electron gyrofrequency $\omega_{ce}$ are used to rapidly energise the electrons through a resonance effect in the etchant gases. The energised electrons then proceed to ionise the surrounding molecules by high speed collisions [24] to create a high density

plasma. For 2.45 GHz microwaves, the resonance effect occurs at a magnetic flux density of 875 G.

As shown in Figure 2.14, in ECR chambers, the wafer and plasma are contained within a quartz bell jar separate to the microwave entrance. The microwaves enter the bell jar to reach the plasma gases and excite the plasma. Around the chamber, large DC coils create a magnetic field that runs through the chamber towards the wafer target. In some ECR chambers, the bell jar is replaced with a flat quartz window to separate the microwave entrance from the plasma.

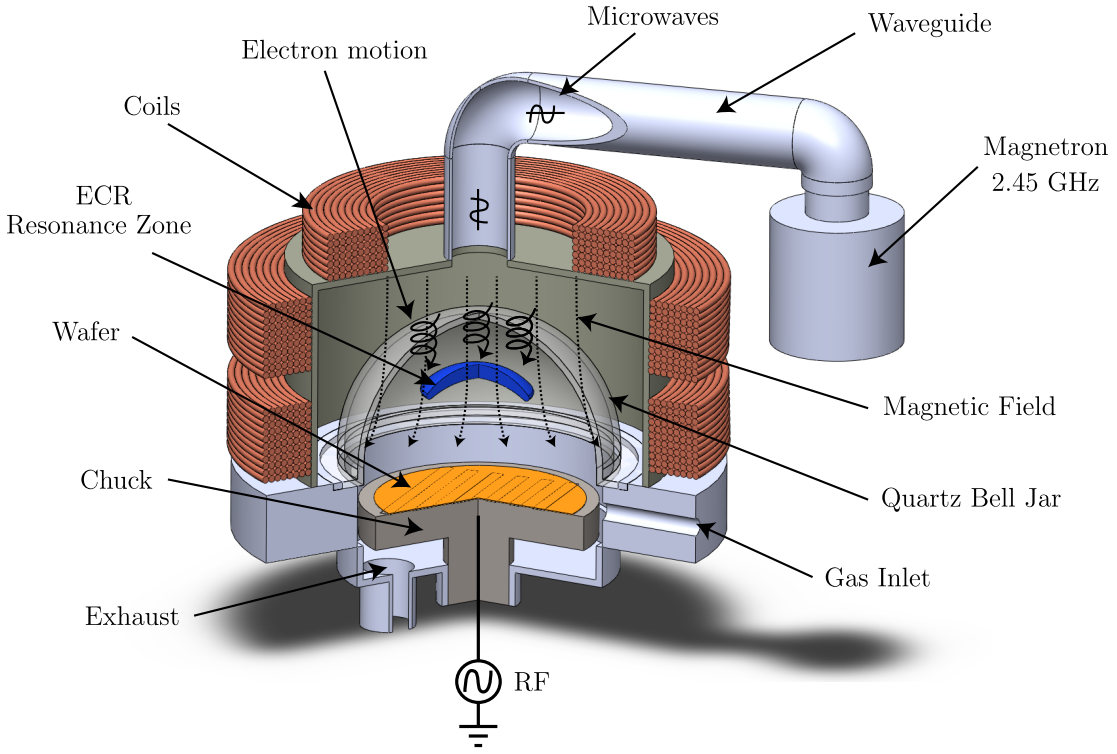



FIGURE 2.14: Electron cyclotron resonance (ECR) plasma etch chamber. In ECR-based chambers, microwaves at 2.45 GHz are used to accelerate electrons spiralling in a magnetic field. Resonance occurs at a magnetic flux of 875 G, when the electron gyrofrequency is equal to the microwave frequency.

A resonance zone is created in the plasma chamber where the magnetic flux density is constant at 875 G. The resonance zone, usually shaped like a shallow dish, is localised since the magnetic field in the chamber is non-uniform [22]. Collisions prevent the electrons from gaining excessive amounts of energy, along with the fact that the resonance zone is of fixed size.

To change the distribution of ions on the wafer surface, the position of the resonance zone can be changed by varying the currents in the magnetic coils that surround the chamber. This effort is assisted by the fact that microwaves exhibit a high degree

of spatial localisation. Hence, the resonance zone where plasma is generated can be separated from the wafer surface.

The plasma streams, or diffuses, along the magnetic field lines towards the wafer. An RF potential is placed on the wafer to control the wafer sheath, providing control over the ion energy to the wafer surface. In this way, the ion energy to the wafer can be controlled independently of the plasma generation, and this control is one of the main advantages of the ECR etch chambers.

An in-depth discussion of microwave propagation in plasmas, electron heating mechanisms, and further details on ECR etch operations can be found in [22], [18], and [24].

## 2.5 Measurement techniques

Complete information on the chemical and physical properties of plasma and the etch process performance is important in semiconductor processing for process monitoring, process control, fault detection, and process design. Measurement techniques can be divided into invasive and non-invasive subgroups. Invasive measurements are those measurements that require physical interference with the plasma, and affect the plasma physically. Non-invasive, or remote, measurements are measurements that can be taken without physical perturbation of the discharge. In this section, some of the more common techniques used to measure plasma properties and etch variables for both industrial and academic applications are examined. Particular focus is given to those techniques used in this thesis.

### 2.5.1 Optical emission spectroscopy

Optical emission spectroscopy (OES) measures light emitted from a plasma as a function of wavelength, time, and location, and is one of the most commonly used plasma measurement tools [22].

In an active plasma discharge, particles are continuously undergoing the processes of excitation from the sustaining external energy source, and relaxation, which is the loss of the previously gained energy. As described in Section 2.1.4, the energy of the photons released from a plasma a relaxation process is a function of the gaseous makeup of the plasma, and the energy levels between which the electrons move. Since each species in the plasma has an individual electron configuration, and since energy levels are quantised

to allow only certain transitions, the luminescence from plasma is a characteristic of its gaseous composition and the excitation levels of its molecules.

In OES, the emitted light from the plasma is examined to gather information about the internal condition of the discharge. The light is collected from the plasma chamber and focused by a lens onto a detector. The detector can be a photodiode, a photomultiplier, or an optical multichannel analyser (OMA) [22]. With photodiodes, filters can be used to isolate single wavelengths of interest. Photomultipliers can measure multiple wavelengths from a limited portion of the complete spectrum and are very sensitive. Using OMAs with a charge-coupled detector (CCD), the complete spectrum can be recorded at a regular interval to monitor plasma processes. While typical OES devices measure the full spectrum from the plasma emissions, it is not unusual to employ *monochromators*, which are devices that only measure the emission at one wavelength, for process monitoring applications such as end point detection. A series of spectra recorded at regular intervals during a plasma etch process are shown for example in Figure 2.15. Shifts in the characteristic peaks of the spectra occur when the gaseous makeup of the plasma changes, corresponding to different steps of the etch process.

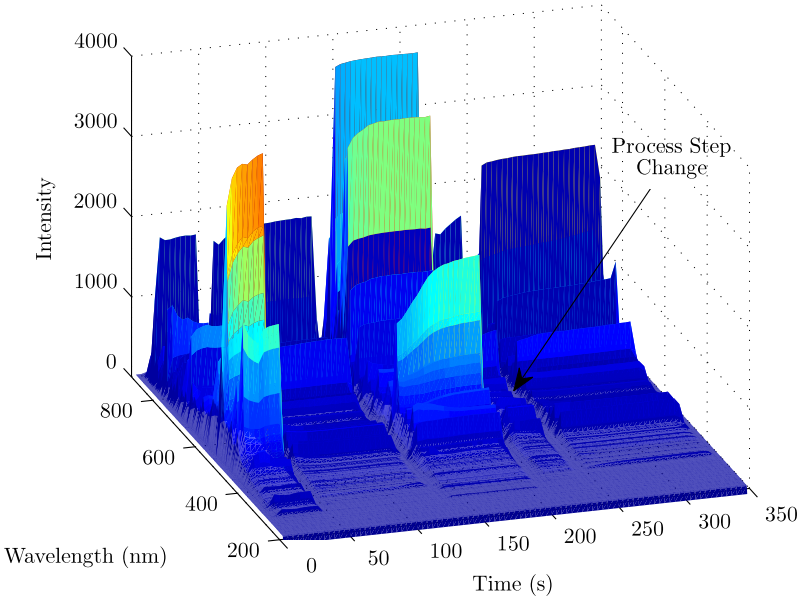



FIGURE 2.15: Optical emission spectra collected during a multi-step plasma etch process.

Information about the discharge is determined by comparing the wavelengths of the light emitted from the plasma to known emission spectra from atomic and molecular species. Typically, relative species concentrations are determined by analysing emission intensities. With correctly calibrated spectrometers, comparisons of the intensities of different wavelengths can be used to calculate electron temperature, density and ionisation fraction of plasmas [22].

Unfortunately, the intensity of the emitted light depends not only on the density of the species of interest but also on the properties of the discharge. The electron density and electron energy distribution in the plasma alter the effectiveness of a plasma in exciting given species. To overcome this variability and avoid complex calculations of the electron energy distribution etc., a baseline referencing technique known as *actinometry* [29] is sometimes used. Actinometry involves the addition of an inert gas such as argon (Ar) in known quantities to the plasma. By comparing the relative intensities of the emissions from the reference gas with the intensities of emissions from species of unknown concentration, variations in the electron distribution function can be overcome, and the species concentrations can be ascertained.

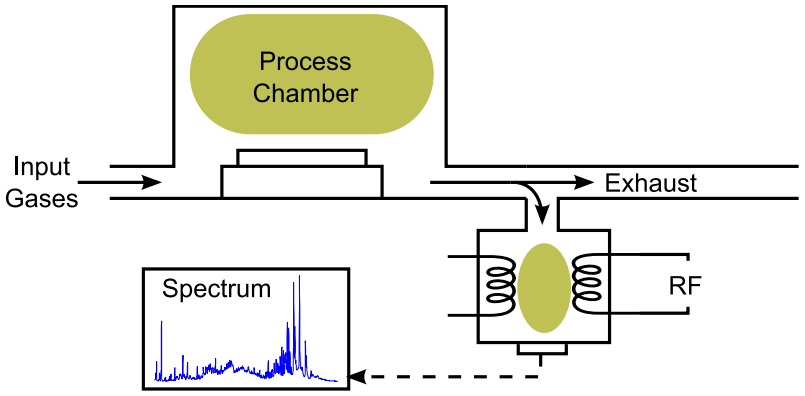



FIGURE 2.16: Downstream measurement of OES data. In some plasma etch processes, separate discharges are ignited using the exhaust gases from the etch process for analysis with OES techniques. This may be desirable in cases where there is no visual access to the primary plasma discharge.

The main advantage of OES is that it is a non-invasive technique that is easily implemented on any chamber with visual access to the plasma. In situations where visual access to the plasma is not available, the exhaust gases from the chamber can be excited separately to the main plasma, and OES data can be collected in the downstream manner shown in Figure 2.16. OES provides information in both the spatial and temporal dimensions [22]. Information is supplied in real time, allowing plasma processes to be monitored constantly.

OES techniques, however, are not without some disadvantages and complications. Deposition from the plasma etching process can cloud the viewing window and act as a filter to the light emissions, thus affecting the recorded spectra. While the spectra produced by the electron transitions in atomic species are typically made up of strongly defined peaks that may be relatively easy to interpret, for molecular gases, OES data can be quite complex and challenging to interpret correctly because of the many vibrational and rotational energy levels possible. Molecular spectra is typically more spread out and less predictable. In processing applications, chemical reactions in the plasma can

also cause chemiluminescence, further complicating the output. Optical spectrometers are delicate devices that can lose their calibration if physically knocked or interfered with. Many OES devices have an unevenly spaced wavelength scale, and can also have a non-linearity in wavelength sensitivities, requiring exact calibration curves for accurate analysis.

The duration of time for which photons of light are accumulated in the photodetector is the *integration time*. The integration time effectively low-pass filters the light intensity signals, where longer integration times correspond to lower bandwidths, while also affecting the signal-to-noise ratio (SNR) of the measured light intensity. Since the noise level is constant, increases in integration time produce a roughly proportional increase in SNR. The choice of integration time is therefore a trade-off between SNR and bandwidth of the OES signals. Since a single integration time for the photodetector must be specified and mean intensities of the spectral lines vary with wavelength, care must be taken to ensure that weaker spectral lines appear above the noise threshold, while stronger lines do not saturate the photodetector. Longer integration times are preferred for a more accurate spectral reading but as integration times increase, saturation becomes an issue and the technique loses its "real-time" essence as the sample rate decreases.

### 2.5.2 Laser induced fluorescence

Laser-induced fluorescence (LIF) is a non-invasive optical measurement technique used to determine the concentration of different species within a plasma. During LIF, optical emission from the plasma is stimulated through the introduction of laser light at specific wavelengths to the plasma. Certain electron transitions in the plasma molecules can be stimulated through photo-excitation by changing the wavelength of the laser light introduced to the plasma. When the excited electrons fall back to their original positions, the intensity of the light produced during the relaxation process can be analysed to allow the species to be identified.

An example of this technique is the use of light with a wavelength of 226 nm to induce a two-photon excited optical transition of atomic oxygen. The relaxation of the artificially excited electron releases light at 844 nm [22]. This particular technique is used to identify and measure the concentration of atomic oxygen in the plasma. As with OES, LIF is affected by the same problems with window contamination during plasma processing operations. LIF equipment is complex and expensive and as a result is not typically installed in manufacturing environments, but confined to research institutions.

### 2.5.3 Laser interferometry

Laser interferometry is a non-invasive technique that is used to measure the etch rate of plasma etch processes using laser light. During laser interferometry, laser light is directed at the wafer surface at an oblique angle, and the reflected light rays are analysed using a detector. Typically, the thin layers of material close to the surface of the wafer are transparent and, at each interface between the layers, the laser light is reflected, absorbed, or transmitted. Destructive and constructive interference occurs between the light beams reflected from the top and bottom of each layer. As the thickness of the layer being etched changes during the etch process, the interference oscillates between destructive and constructive interference, hence changing the intensity of the reflected light reaching the detector. The time between the maxima and minima of the intensity of the reflected light is related to the rate of change of the film thickness (the process etch rate) [27]. When the layer is completely etched (the process end point), the changes in intensity stop.

Figure 2.17 shows a schematic of laser interferometry being used to measure etch rate. The figure shows an incident ray of light at wavelength $\lambda$ travelling from air of refractive index $n_1$ into a layer of thickness $d$, with refractive index $n_2$ at an angle $\theta_1$. As the thickness of the layer $d$ is changed, the interference recorded between reflected rays 1, 2 and 3 will be altered. Rays 1 and 2 are caused by reflection between the interfaces of air-film and film-substrate ($n_3$) respectively, whereas ray 3 is caused by multiple reflections within the film.

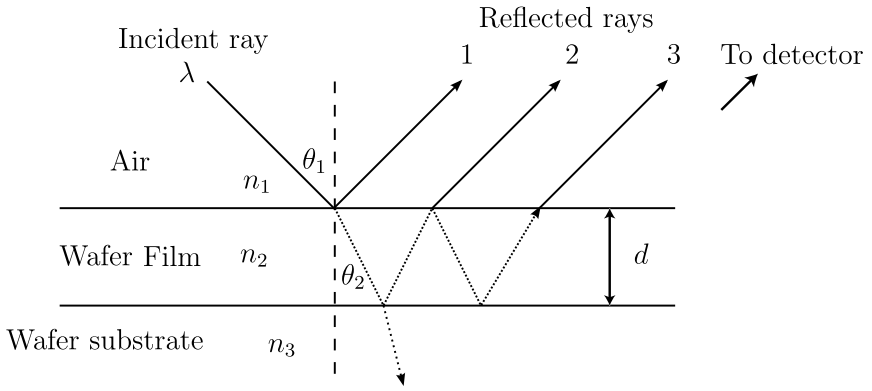



FIGURE 2.17: Laser interferometry [22]. Laser Interferometry is used to measure etch rate in real time. The measurement of etch rate is achieved by analysing the beat frequency created by the interference of reflected light from different layers of the wafer surface.

Laser interferometry requires an abrupt interface change between the individual layers of material on the wafer surface [22] and to ensure reflection, layers must be sufficiently thick and transparent. Laser interferometry equipment can be used to detect

the end point of a process as well as give in-situ etch rate measurements. On the other hand, the etch rate is only obtained from the point of light impact on the wafer surface, the reflected signal is less clear from rough surfaces, and as with OES, window clouding can become an issue as chambers become conditioned. Furthermore, as device dimensions continue to shrink, the trenches etched on wafer surfaces are becoming too small to be penetrated by the wavelengths of light used during laser interferometry, preventing easy measurement of etch rate.

*Full-wafer interferometry* is an extension to the interferometry process where a CCD camera is used to collect reflected light from the full wafer surface over time. Unlike laser interferometry, full-wafer interferometry uses the glow from the plasma as a light source. Full wafer interferometry is able to measure etching uniformity across a wafer, within etched patterns, and ARDE, if structural information within etch patterns is available [22].

### 2.5.4 Ellipsometry

Ellipsometry is a popular measurement technique that relies on the polarisation changes that occur when light is reflected from, or transmitted through, a medium. The changes in polarisation are a function of the optical properties of the medium, its thickness, and the wavelength and angle of incidence of the light beam relative to the surface normal [28].

Spectroscopic ellipsometry uses multiple light beams with different wavelengths, and is a fundamentally more accurate technique than interferometry for obtaining film thickness. Typically, linearly polarised light is incident on the surface being measured and the elliptical polarisation state of the reflected light is analysed. The wafer surface structures are modelled, and the model parameters, including the thickness of the surface layer of interest, are changed iteratively using an optimisation technique until the theoretical model outputs match the measured data.

Because ellipsometry relies on an intensity ratio instead of absolute intensity measurements during operation, it is relatively robust to intensity changes in the light source and contamination of optical windows in plasma chambers.

### 2.5.5 Mass spectrometry

Mass spectrometry (MS) is used to determine the composition of gases. A typical mass spectrometer setup is shown in Figure 2.18. During MS, gaseous molecules are ionised by

electron impact and the ions produced are separated according to their mass-to-charge ratios. Separation is achieved by accelerating the ions down a deflection pipe and using a magnetic field to deflect the ions trajectories. Heavy, lower-charged ions are deflected by lesser amounts than light, highly-charged ions under the force of the magnetic field. By varying the magnetic field using an electromagnet, it is possible to focus ions with different mass-to-charge ratios onto a detector at the end of the deflection pipe. Ions that reach the detector create a current that is recorded to form the mass spectrometer signal. Figure 2.19 shows an example mass spectrum resulting from an analysis of zirconium t-butoxide with oxygen.

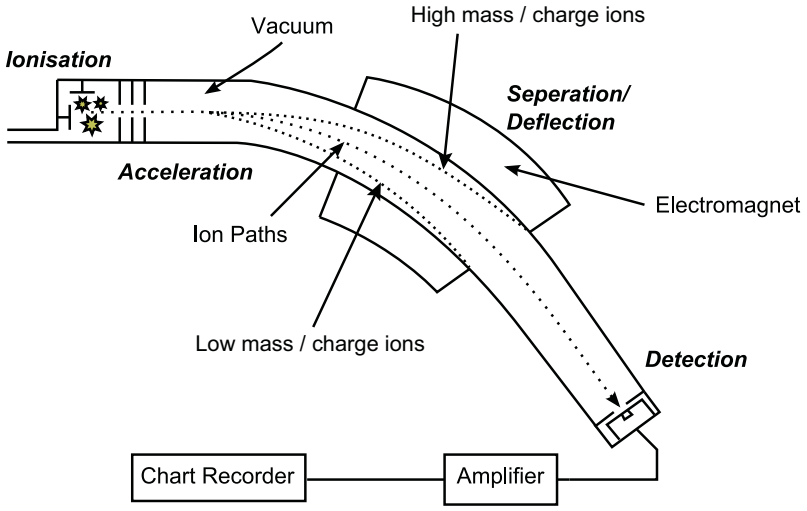



FIGURE 2.18: Mass spectrometer apparatus. Gaseous molecules are ionised and then separated according their mass-to-charge ratios through the use of a variable magnetic field.

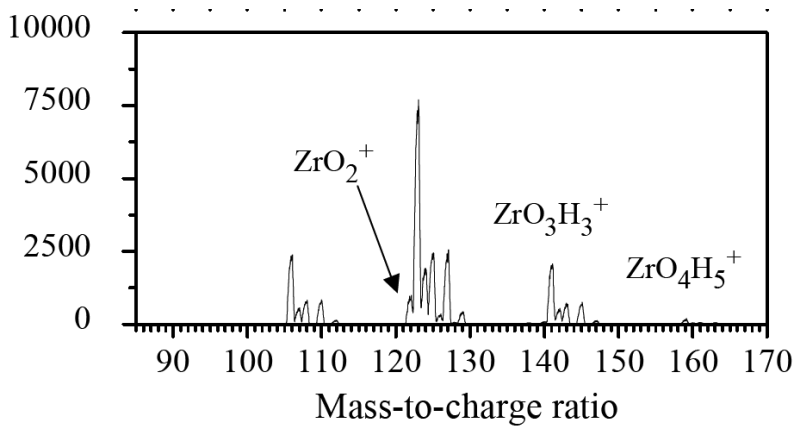



FIGURE 2.19: Typical Mass Spectrum of zirconium t-butoxide (ZTB) with $O_2$.

MS requires substantial changes to the plasma chamber to collect ions from the plasma. In plasma processing applications, the mass spectrometer is installed as close to the main discharge as possible to avoid collection of by-products from wall reactions.

Alternatively, the spectrometer can sample from the chamber exhaust to avoid disturbing the plasma. However, sampling the chamber exhaust does not provide as direct a measurement of the processing plasma as mass spectrometers that take samples from the main discharge, because the exhaust gases contain species from the chamber walls and there is a transport delay between the processing plasma and the point of measurement.

To fully analyse the neutrals and radicals in the plasma, they must first be separated from the ions, before being ionised themselves. A complete understanding of the ionisation products present in the plasma is necessary for the analysis of mass spectra. MS is often used to analyse plasmas consisting of large molecular species, for which OES data can be complicated as a result of a multitude of spectral lines that overlap. However, for atomic species, OES techniques are preferential to MS since the optical spectra are simpler and hence more easily analysed than the mass spectra.

### 2.5.6 Langmuir probes

The most direct measurements of a plasma are obtained from probes within the plasma chamber. A Langmuir probe is an invasive probe that can be used to determine electron temperature, electron density, electron energy distribution function, and plasma potential. The Langmuir probe consists of a small conductor that is placed within the plasma. The potential of the conductor is varied and the resulting current-voltage trace, or *I-V characteristic*, is used to determine the properties of the plasma [22].

Langmuir probes are designed to withstand the harsh conditions within the plasma without being destroyed. A 2–10 mm conducting tip at the end of a conducting rod collects the current for the I-V characteristic. The conducting tip is constructed of a material capable of withstanding high temperatures, typically tungsten or platinum. The conducting rod is threaded through a ceramic tube to insulate all but the current-collecting tip from the plasma. In low-temperature plasmas, Langmuir probes can operate without melting or excessive sputtering at the tip [22] but in dense, high-temperature plasmas, the probe is only exposed to the plasma for short intervals (less than one second) to prevent the tip from melting.

An ideal I-V characteristic for a Langmuir probe is shown in Figure 2.20. I-V characteristics are usually shown such that the electron current to the probe is in the $+I$ direction.

Ion current dominates to the left of the $V = 0$ point on Figure 2.20, because ions are attracted to the probe tip by the low potential. At the extreme left of the I-V

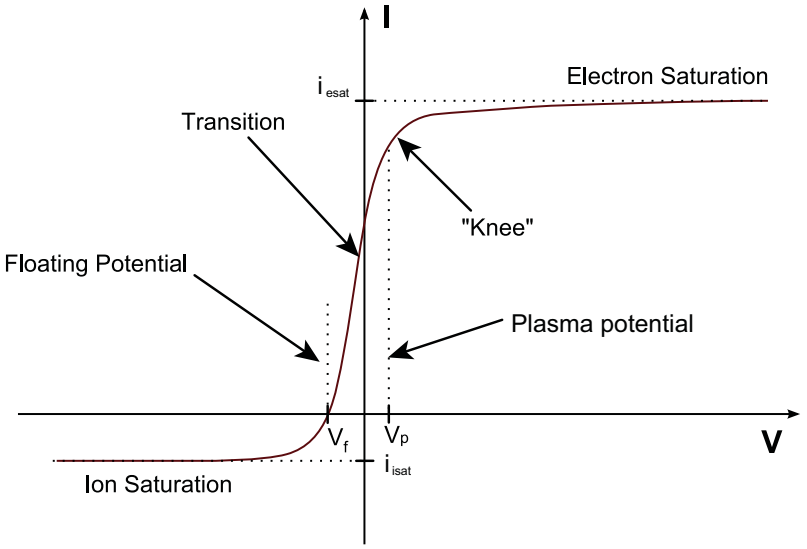



FIGURE 2.20: Ideal I-V characteristic from a Langmuir probe. Real I-V characteristic curves typically do not have such easily identifiable features.

characteristic, the trace levels out to the *ion saturation current* $i_{isat}$, where the electron current is zero, all electrons from the plasma being repelled by the strong negative charge on the probe. The ion current is determined by the rate of ion arrival at the sheath around the probe tip due to the random movement of ions in the plasma [21].

As the tip potential is increased, the number of electrons with enough energy to overcome the repulsive electric field at the tip, and hence contribute to the tip current, starts to increase. The point at which the trace crosses the V-axis is the floating potential $V_f$ because it is at this point that the ion and electron currents to the probe tip are equal.

To the right of $V_f$, electron current is dominating. The collected current rises throughout the *transition region* as the probe tip becomes more positive. The shape of the transition region can be used to derive information about the electron temperature and the electron energy distribution of the plasma. When the plasma potential $V_p$ is reached, the trace takes a sharp turn, known as the "knee" and finishes at a relatively constant *electron saturation current* $i_{esat}$. Further increases in the potential of the probe tip increase the energy of the electrons collected by the tip, but do not result in an increase in the collected current.

Langmuir probes are not ideal and a number of factors can complicate probe readings. The correct analysis of the I-V characteristic traces requires considerable expertise. For example, the effective current collecting area of the tip is a factor in the correct determination of the electron and ion currents. The current collecting area of the probe is not equal to the exposed surface area of the the tip itself, but rather the surface area of the sheath surrounding the tip. The sheath thickness at the probe tip, and hence the

surface area of this sheath, changes with the probe potential, and affects the collection of current from the plasma. Secondary electron emission at the tip and heating of the tip as a result of particle collisions both cause inconsistencies in the current recorded. Damage to the probe is possible from high energy electron and ion impact. Highly positively biased probes can draw relatively large electron currents from the plasma, perturbing the plasma properties and distorting the measurements. Along with these complications, in RF powered plasma processes, Langmuir probes are subject to RF interference that distorts the I-V characteristic, requiring extra processing to recover the correct signal. The I-V characteristic shown in Figure 2.20 is ideal in that it ignores such disturbing processes.

For a more in-depth analysis and discussion on Langmuir probes, the interested reader is directed to the "Plasma Diagnostics" chapter of work by Chen and Chang [22].

### 2.5.7 Hairpin resonator probe

The electron density in a plasma chamber can be measured using a microwave hairpin resonator, or *hairpin probe*. The hairpin probe is an invasive monitoring tool introduced by Stenzel [30] in the mid 1970's to perform spatially-resolved plasma density measurements in low-pressure plasmas.

A hairpin probe is an open-ended quarter wavelength transmission line whose resonant frequency is related the dielectric constant of the medium that surrounds it. The probe gets its name from the U-shaped resonator that resembles a hairpin, as shown in Figure 2.21. A microwave signal source drives a small amplitude time-varying current through a small diameter coupling loop located near the lower part of the U and this current is swept over a range of frequencies. Energy is coupled into the U-shaped structure, and at resonance, a standing wave occurs on the hairpin such that voltage is maximised at the open end and minimised at the shorted end of the transmission line. At resonance, the hairpin weakly radiates energy into the surrounding space, whereas off resonance almost all energy incident from the current source is reflected [31]. A well-pronounced maximum in the amplitude of the magnetic field around the hairpin resonator occurs at the resonant frequency. The frequency at which this resonance occurs is used to determine the relative permittivity, and thus the electron density, of the plasma in which the resonator is submersed. The hairpin resonator can be directly coupled to the input loop or electrically isolated. Electrically isolated floating probes are both DC and RF isolated from ground and hence, the potential of the resonator varies with that of the plasma. This reduces the number of corrections required to compensate for RF interference [31].

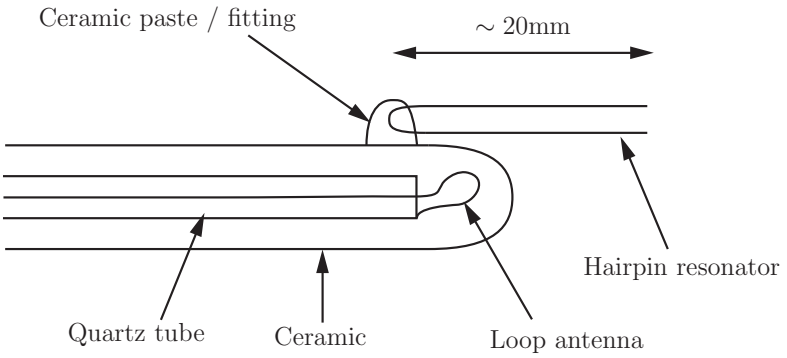



FIGURE 2.21: Microwave resonator "hairpin" probe. Hairpin probes are invasive monitoring tools immersed in the discharge to determine the electron density.

There are two main types of hairpin probes. Stenzel [30] describes the construction of a transmission probe which uses two loop antennae, one for coupling current in the hairpin, and a second for detection of the microwave signal at the resonant frequency. Two coaxial connections are required, one for the transmission of the microwave current, and one for the pick-up loop.

The second type of hairpin probe, described by Piejak *et al.* [32], is the reflection probe. The reflection probe differs from the transmission probe in that it requires only one coaxial feeder connection. Resonance is determined using the current reflected from the probe. As a result, the reflection probe is simpler to build and produces less plasma perturbation. A directional coupler between the microwave sweep source and the input loop to the hairpin probe allows the reflected current to be monitored using an oscilloscope. The electron density measurements from both probe types have been shown to be equally accurate [32], and due to its simpler design and smaller bulk, a reflection probe is used during the work described in Chapter 8 of this thesis.

As shown in Figure 2.21, typically the coaxial cable carrying the driving microwave current is threaded through a quartz tube to isolate it from the plasma. The quartz tube is in turn surrounded by ceramic piping to provide further protection from the temperatures and chemicals present during etching. Platinum or tungsten wire is used to form the hairpin resonator because of their resistance to the harsh environment inside the etch chamber.

### Operating principal

The operating principal of the hairpin probe is clearly described by Stenzel [30] and an abbreviated description of the probe operation is given here. The resonant frequency of

the hairpin is given by

$$f_r = \frac{c}{4L\sqrt{\epsilon}}, \tag{2.29}$$

where $c$ is the speed of light in m/s, $L$ is the length of the resonator (m) (from the bottom to the top of the U), and $\epsilon$ is the relative permittivity (dimensionless) of the medium surrounding the probe. In a vacuum, $\epsilon = 1$, and $f_0 = c/4L$ Hz. The relative permittivity of a non-magnetised plasma is

$$\epsilon = 1 - \frac{f_p^2}{f_r^2} \tag{2.30}$$

where $f_p$ is the plasma frequency given by Equation (2.17). By substituting Equation (2.30) into Equation (2.29), the resonant frequency $f_r$ in a plasma can be expressed as

$$f_r = \frac{f_0}{\sqrt{1 - f_p^2/f_r^2}}, \tag{2.31}$$

which rearranges to

$$f_r^2 = f_0^2 + f_p^2. \tag{2.32}$$

The electron density is related to the frequency difference between the probe resonant frequencies with and without the plasma, as summarised by

$$n_e = \frac{f_r^2 - f_0^2}{0.81} \times 10^{10}, \tag{2.33}$$

where $n_e$ is the electron density (per $m^3$), and $f_r$ and $f_0$ are given in gigahertz [32].

### 2.5.8 Plasma impedance monitor

In plasma equipment, harmonics of the fundamental frequency are generated in the power supply circuitry to the chamber because the plasma presents a non-linear load to the RF generator and matching network. The non-linearity of the plasma arises due to the modulation of the plasma sheaths with the applied RF power. The sheath width is sensitive to plasma electron density and changes to the wafer surface (e.g. etching through a layer) or chamber electrodes (e.g. coatings or erosion). Hence, the amplitudes of the harmonics of voltage and current signals are sensitive to changes in the bulk plasma conditions and the wafer state [33]. RF-based sensors have been used successfully in plasma etch for both fault detection [34, 35] in etch chambers and end-point detection for plasma processes [36, 37].

A plasma impedance monitor (PIM) is an electronic sensor that is added to the circuitry between the matching network and the plasma electrodes or antennae. The

PIM sensor samples the current and voltage waveforms from the RF power line and uses dedicated hardware to perform a Fourier decomposition of these waveforms. The sensor records the amplitudes of the current and voltage at a number of frequencies, usually harmonics of the supplied RF signal. The phase angle between current and voltage at each harmonic is also recorded. Typically, measurements are available at the fundamental frequency of 13.56 MHz and modern PIM sensors can also record data for over fifty harmonics of this frequency.

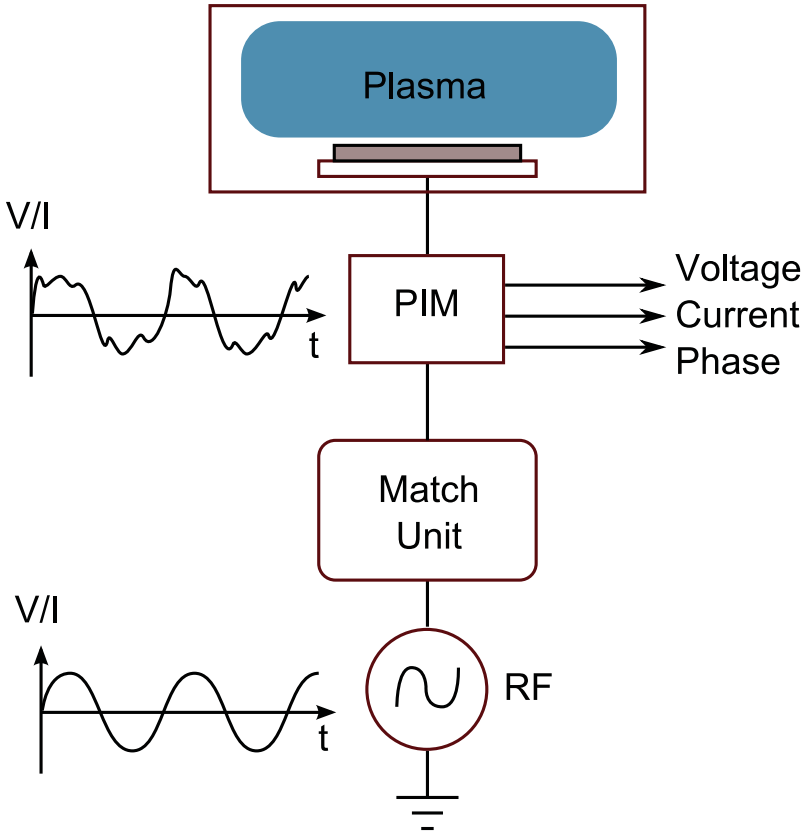



FIGURE 2.22: Schematic of PIM location and plasma effect on signals [33]. The non-linear impedance presented by the plasma generates harmonics in the supplied RF signals.

The first and fourth harmonic from voltage, current and phase supplied to the target electrode during a five-step plasma etch process are shown in Figure 2.23. The PIM signals vary over the course of each etch step, and vary between different steps as the wafer state and chamber set points are changed.

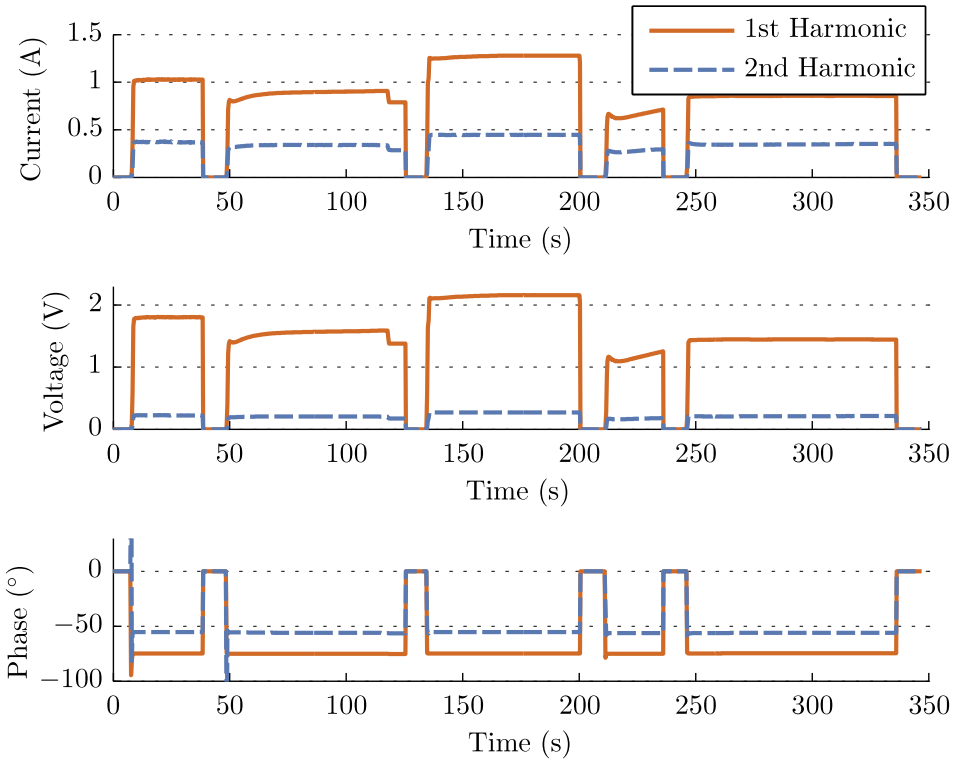



FIGURE 2.23: Sample PIM signals from a five-step ECR etch process.

Various electrical properties of the plasma can be calculated using the PIM measurements. Power, impedance, resistance and reactance can be calculated using

$$P = VI\cos(\phi), \tag{2.34}$$

$$Z = \frac{V}{I}, \tag{2.35}$$

$$R = \frac{V}{I}\cos(\phi), \tag{2.36}$$

$$X = \frac{V}{I}\sin(\phi), \tag{2.37}$$

respectively, where $P$ is power, $Z$ is impedance, $R$ is resistance, $X$ is reactance, $V$ is voltage, $I$ is current and $\phi$ is phase.

PIM sensors are sensitive to both physical (e.g. changes in layer) and chemical changes (e.g. species concentrations) in the chamber [38] and have the advantage over OES techniques that no optical access to the plasma is required. RF plasma impedance sensors placed between the matching network and the plasma electrode provide non-invasive information that is fast to respond to changes in the plasma for real-time monitoring. RF data are less noisy and do not suffer from problems that hinder the performance of OES-based techniques, such as cloudy windows, optical set ups, and detector drift [39].

It is important to ensure that measurements are made in high resolution because calculations of powers are highly sensitive to small errors in phase measurements. This sensitivity arises because the phase angle between the fundamental current and voltage signals are often close to $-90°$ for CCP sources. An expression for the sensitivity of power to changes in phase angle can be derived as

$$\frac{\Delta P}{P}/\frac{\Delta\phi}{\phi} = \frac{\Delta P}{\Delta\phi}\frac{\phi}{P} = \frac{\Delta(VI\cos\phi)}{\Delta\phi}\frac{\phi}{VI\cos\phi} = -VI\sin\phi\frac{\phi}{VI\cos\phi} = -\phi\tan\phi \quad (2.38)$$

As phase angles tend towards $-90°$, $\tan\phi \to \text{inf}$, and hence small errors in measurements of $\phi$ can lead to large errors in calculations of $P$.

One of the disadvantages of PIM sensor measurements is that variations detected in harmonic data as a result of phenomena in the etch chamber lack intuitive explanation. While a complex physical relationship exists between the harmonic data recorded and the properties of the plasma being measured, the relationship is difficult to predict, and may only be found using black-box modelling techniques. Often the mechanisms responsible for changes in measured phase and impedance during etching are not completely understood [40]. Important correlations between the sensor data and the etch process variables may go un-noticed without an exhaustive search of the data recorded from the PIM sensor.

# Chapter 3

# Virtual metrology techniques

Virtual metrology (VM) is performed by first collecting measurements from the process of interest and then modelling variations in the process output using these measurements. This chapter provides an overview of the modelling techniques that are used for the VM modelling described in this thesis. There are seven techniques discussed that can be broadly divided into three categories:

1. Linear regression techniques,
2. Latent variable-based techniques, and
3. Techniques with non-linear capabilities.

Least squares regression, stepwise regression, least angled regression and weighted least squares regression are included as linear regression techniques. Principal component regression and partial least squares regression are included as latent variable-based techniques. Finally, artificial neural networks and Gaussian process regression are discussed as modelling methods with non-linear capabilities. Due to the complexity of the processes explored in the thesis, all of the modelling techniques used are purely empirical, black-box methods, since, typically, the functional relationship between the VM model input and output variables is unknown or too complex to model analytically.

## 3.1 Least squares regression

A multiple linear regression (MLR) model is a model that describes a linear relationship between a number of *input* variables $x_j$ and a *response* or *output* variable, $y$, where the

subscript $j$ denotes the input variable number $j = 1, 2, \cdots p$. The term "regression" was first coined by Sir Francis Galton in 1885 [41] to describe a biological phenomenon. There is some dispute about the discovery of the method of least squares used to determine MLR model parameters which is described in this section. It appears that the method was discovered independently by Legendre [42] and Gauss [43] in the early 1800's where both mathematicians used the method to determine the movement of celestial bodies from astronomical observations.

A MLR model may take the form

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \tag{3.1}$$

where the parameters $\beta_j$, $j = 0, 1, \cdots p$ are the regression coefficients and $\epsilon$ represents an (ideally random, zero mean) modelling error term. MLR models are often used as empirical models, when the true functional relationship between $y$ and $x_1, x_2, \cdots, x_p$ is unknown. When used to approximate unknown non-linear functions, linear regression models can provide adequate approximations over specific, but usually small, ranges of the input variables [44].

The values of $\beta_j$ are determined most simply using the method of least squares, forming least squares regression (LSR) models. Suppose that $n > p$ observations of $x_1, x_2, \cdots, x_p$ and $y$ are available. Let $y_i$ denote the $i$th observed response and $x_{ij}$ denote the $i$th observation of input variable $x_j$. It is assumed that the resulting error terms $\epsilon_i$ after the model is formed have a mean value of zero, variance $\sigma^2$, and are uncorrelated. For clarity, the variance of a variable $x$ is universally defined as

$$var(x) = E[(x - \mu)^2], \tag{3.2}$$

where $x$ is a random variable with mean $\mu$ and $E$ is the expected value operator. Variance is the second central moment of a variable and is often denoted using $\sigma^2$, where $\sigma$ is the standard deviation of $x$.

For the $i^{th}$ observation, $y_i$ is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,\ i = 1, 2, \cdots, n. \tag{3.3}$$

$$= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i,\ i = 1, 2, \cdots, n. \tag{3.4}$$

$\beta_0$ acts as a bias term that can generally be excluded for notational simplicity if it is assumed that all variables are normalised to have zero mean prior to LSR modelling.

Equation (3.3) can be written in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.5}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

In general, $\mathbf{y} \in \mathbb{R}^{n\times 1}$ is a vector of $n$ observations of the output variable, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a matrix of $n$ observations of the $p$ input variables, $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ is a vector of the regression coefficients, and $\boldsymbol{\epsilon} \in \mathbb{R}^{n\times 1}$ a vector of random errors.

It is required to find the vector of least-squares regression coefficient estimates $\hat{\boldsymbol{\beta}}$ that minimises the least-squares cost function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{3.6}$$

$S(\boldsymbol{\beta})$ can be expanded as

$$\begin{aligned} S(\boldsymbol{\beta}) &= \mathbf{y}^T\mathbf{y} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \end{aligned} \tag{3.7}$$

since $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}$ is a scalar, as is its transpose $(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y})^T = \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}$. For the minimum solution to Equation (3.6), the regression parameters $\boldsymbol{\beta}$ must satisfy

$$\left.\frac{\delta S}{\delta \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 0, \tag{3.8}$$

which simplifies to

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}. \tag{3.9}$$

Hence, the *least-squares estimate* of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3.10}$$

Equation (3.10) yields a unique solution for $\hat{\boldsymbol{\beta}}$ provided that the inverse matrix of $\mathbf{X}^T\mathbf{X}$ exists [45, 46]. Models created using LSR are computationally efficient to develop as they require only relatively low-complexity matrix operations.

The system is described as *over-determined* when the number of samples $n$ outnumbers the number of input variables $p$. In this case, the rank of $\mathbf{X}$ is usually $p$ given uncorrelated input variables, and $\hat{\boldsymbol{\beta}}$ is unique. LSR experiences difficulties in situations where input variables are significantly correlated together or in *under-determined* systems where $n < p$. In the case of correlated input variables, $\mathbf{X}^T\mathbf{X}$ can be close to singular, causing difficulties in finding its inverse. If $n < p$, the rank of $\mathbf{X}$ is $n$ and there are an infinite number of solutions for $\hat{\boldsymbol{\beta}}$. A single solution can be chosen, but usually this solution has poor predictive capability for any new samples collected. Often the *minimum-norm* solution is chosen, that is the solution for which $\|\hat{\boldsymbol{\beta}}\|$ is minimised.

In cases where non-linear relationships exist between the input and output variables to be modelled, non-linear transformations can be carried out on the model input variables to form new input variables. LSR can then be employed to determine the model parameters, and, since the model is linear in the parameters $\hat{\boldsymbol{\beta}}$, the model is still referred to as a linear LSR model. It is typical to include squared $\mathbf{x}_j^2$ and interaction $\mathbf{x}_j\mathbf{x}_k$ terms as model inputs during the development of quadratic response surface models of processes using experimental data. The derivation for $\hat{\boldsymbol{\beta}}$ can also be extended for systems with more than one output variable.

## 3.2 Weighted least squares regression

In typical least squares applications, all observations in the analysis are given equal weighting, and the least squares algorithm minimises the sum of squared residuals, given by Equation (3.6). In some applications, it can be shown that particular observations used in a regression analysis are less reliable than others [47], for example if some of the observations are taken with less precision than others. When it is reasonable to assume that not all of the samples provide equally valuable information, *weighted least squares* can be used to give each data point different amounts of influence during determination of the model parameter estimates. A weighting scheme is introduced to the least squares algorithm so that the sum of squared residuals to be minimised becomes

$$S = \sum_{i=1}^{n} w_{ii}\epsilon_i^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{3.11}$$

where $\mathbf{W} \in \mathbb{R}^{n \times n}$ is a diagonal matrix of weights applied to each of the samples $i = 1, 2, \cdots, n$. Typically, these weights are chosen to be inversely proportional to the variance of the output sample values recorded for each combination of input variables involved in the regression.

The weights given to each sample operate relative to the weights given to the other samples, with the absolute values of the weights being relatively unimportant. Following the same calculations as detailed in Equations (3.6)–(3.10) for the cost function in Equation (3.11) yields

$$(\mathbf{X}^T\mathbf{W}\mathbf{X})\boldsymbol{\beta}_{\mathbf{W}} = \mathbf{X}^T\mathbf{W}\mathbf{y} \tag{3.12}$$

which can be solved for a unique solution of $\hat{\boldsymbol{\beta}}_W$, the parameter estimates that take the sample weights into account, as

$$\hat{\boldsymbol{\beta}}_W = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}. \tag{3.13}$$

## 3.3 Stepwise regression

In particular modelling applications, it may be desirable to include only a subset of the available input variables in the regression equation. Such situations arise where there is a pool of input variables from which to choose for regression and the input variables that are most influential on the system output are unknown. Ideally, the best subset of input variables that forms an accurate model is selected.

One of the disadvantages in using many or all of the input variables in the regression equation is that the variance of the output estimates increases as the number of input variables increases. Also, in cases with few samples and many input variables, the input data matrix $\mathbf{X}$ may be under-determined. Reduction of the number of input variables in a linear regression model is also useful to reduce the number of measurements required in situations where the measurement of input variables is expensive [47].

One possible approach to the problem of variable-subset selection is to examine all possible regression equations produced by all possible subsets of the input variables, resulting in $2^p$ total models for evaluation, where $p$ is the number of available input variables. While this approach may be feasible for small $p$, the number of models requiring evaluation increases rapidly with the number of candidate input variables. Efficient algorithms for the generation of all possible regressions have been developed, but these procedures are typically only useful for $p < 30$ (for example, see [48]).

*Stepwise selection* is a statistical variable selection procedure for selection of a suitable subset of input variables for use in a LSR model, in cases where it may not be desirable to use all of the input variables available. The combination of stepwise selection and linear regression is *stepwise regression*. Stepwise regression is computationally more efficient than examining all possible regression equations for $p > 30$ [45]. The algorithm is a combination of two variable selection techniques:

1. forward selection and
2. backward elimination.

As shown in Section 3.1, the standard linear regression model is given by

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \cdots + \beta_n \mathbf{x}_n + \boldsymbol{\epsilon}, \tag{3.14}$$

where $\mathbf{y}$ is the estimated output variable, $\mathbf{x}_1, \mathbf{x}_2..\mathbf{x}_n$ are the model input variables, $\boldsymbol{\epsilon}$ is a random error vector, and $\beta_0, \beta_1, \cdots, \beta_n$ are the model parameters.

*Forward selection* begins with no terms in the regression model apart from the intercept term, $\beta_0$. An iterative sequence is performed, where, at each iteration, each available input variable is added to the linear regression model separately and a statistical $F$-test for testing significance of regression is evaluated. The statistical test is structured such that the null hypothesis is that the regression parameter for the new input variable would have a zero coefficient if added to the model. The input variable having the largest partial $F$-statistic, given the other input variables already in the model, is added to the model if its partial $F$-statistic exceeds a preselected entry level $F_{toEnter}$. This process repeats, adding variables to the regression equation one at a time, and the algorithm ends when there are no input variables remaining with a partial $F$-statistic greater than $F_{toEnter}$, a user-defined variable, or when all of the input variables have been added to the model.

*Backward elimination* operates in the opposite manner, beginning with all of the input variables already in the model. Partial $F$-statistics are calculated for each input variable as if it were the last input variable added to the model. The smallest of these partial $F$-statistics is compared with a preselected value $F_{toRemove}$ and, if the partial $F$-statistic is less than this value, the corresponding input variable is removed from the model. This procedure continues until the smallest partial $F$-statistic of the remaining input variables is not less than the preselected cutoff $F_{toRemove}$.

*Stepwise regression* starts in the same manner as forward selection with no input variables included in the model and proceeds to add variables depending on their partial F-statistics. However, at each iteration in stepwise regression, input variables entered into the model during previous iterations are reassessed for removal via their partial $F$-statistics. Input variables can be removed, using the same criteria as the backward elimination algorithm, if they become redundant as a result of newer input variables more recently added to the model. Two cutoff values are specified for the stepwise regression algorithm, $F_{toEnter}$ and $F_{toRemove}$. Stepwise selection techniques are often used to select subsets of input variables for use in other modelling techniques that do not perform optimally with large numbers of input variables.

The choice of the cutoff values is a user preference. Larger $F_{toEnter}$ values are selected to reduce the chance of spurious input variables entering the model, whereas smaller $F_{toEnter}$ values will allow more input variables to enter the model that can then be assessed manually at a later stage for significance, using experience or further modelling. $F$-test values are often specified in terms of percentage points or significance levels, $\alpha$, which specify the probability of adding or removing an input variable from the regression equation in error. Draper and Smith [47] recommend settings of $\alpha = 0.05$ or $\alpha = 0.10$ for both entry and exit $F$-tests. Other analysts [45] frequently prefer to choose $F_{toEnter} > F_{toRemove}$, making it more difficult to add an input variable to the model than to delete one.

Although the stepwise regression technique is quick to compute, easy to implement, and readily available on popular mathematical computing packages, the technique has attracted substantial criticism from many authors. Montgomery *et al.* [45] highlight that the models produced during stepwise selection may not be the "best" models with respect to any standard criterion, and some users may be misled into thinking that the result is optimal. The regression equation and selected input variables should always be examined manually by the user. Experience, professional judgement in the subject matter field, and subjective considerations should be taken into account when analysing the results of such automatic variable selection techniques.

## 3.4 Least angle regression

Least angle regression is a model selection technique that is related to the forward selection regression technique described in Section 3.3. Least angle regression is abbreviated as LARS, where the 'S' arises from related techniques, *lasso* and *stagewise* regression. The forward selection algorithm described in Section 3.3 is sometimes criticised as an

aggressive fitting technique that can be overly hasty during selection of input variables, perhaps eliminating useful input variables that are correlated with model input variables variables previously chosen [49]. Such algorithms can become trapped in local optima, failing to exploit multivariate correlations in the data, instead exploiting only univariate correlations between one input variable and the model residual [50]. Before the presentation of the LARS procedure, stagewise regression, upon which the LARS algorithm is based, is first described.

### 3.4.1 Stagewise regression

*Stagewise regression* is a model creation technique that attempts to refine the forward selection algorithm by taking many small iterative steps when building up the regression function. Unlike forward selection, input variables are added *slowly* to the regression model by slowly increasing the corresponding regression parameter $\beta$ for each variable of importance.

Suppose an initial estimate of $\hat{\mathbf{y}} = 0$, and a vector of *current correlations* $\mathbf{c}(\hat{\mathbf{y}})$ is defined such that

$$\mathbf{c}(\hat{\mathbf{y}}) = \hat{\mathbf{c}} = \mathbf{X}^T(\mathbf{y} - \hat{\mathbf{y}}) \tag{3.15}$$

where each element $\hat{c}_i$ of $\hat{\mathbf{c}}$ is a measure of the correlation between input variable $\mathbf{x}_i$ and the residual vector $\mathbf{y} - \hat{\mathbf{y}}$. At each iteration in stagewise regression, $\hat{\mathbf{y}}$ is incremented in the direction of the greatest current correlation

$$\hat{k} = \underset{k}{\operatorname{argmax}} \|\hat{c_k}\| \quad \text{and} \quad \hat{\mathbf{y}} \to \hat{\mathbf{y}} + \eta \cdot sign(\hat{c}_{\hat{k}}) \cdot \mathbf{x}_{\hat{k}}, \tag{3.16}$$

where $k$ is used as an index for the model input variables, $\hat{k}$ is the index for the model input variable with the greatest correlation to the current residual vector, and $\eta$ is a small constant defining the step size. The step described by Equation (3.16) describes a small increment to the value of $\beta_{\hat{k}}$, the model parameter for input variable $\hat{k}$:

$$\beta_{\hat{k}} \to \beta_{\hat{k}} + \eta sign(\hat{c}_{\hat{k}}). \tag{3.17}$$

The stagewise algorithm differs from forward selection because instead of including input variables at each step, the estimated regression parameters are increased in a direction equiangular to each input variables correlations with the residual. For example, if an input variable $\mathbf{x}_i$ is being added to the regression model, and another input variable $\mathbf{x}_j$ becomes more correlated to the model residual after one particular iteration, the algorithm switches to adding $\mathbf{x}_j$ in the next iteration. As shown in Figure 3.1, in the space

spanned by the input variables, this step corresponds to a movement in the direction of $\mathbf{x}_j$ towards the solution $\mathbf{y}$ (where the solution has been projected orthogonally onto the input subspace). The drawback of stagewise regression is an increased computational load compared to forward selection, which varies depending on the step size $\eta$ taken at each iteration. If $\eta = \|\hat{c}_{\hat{k}}\|$, Equation (3.16) describes the forward selection routine. The LARS algorithm reduces the computational burden by taking large steps, but not as large as forward selection.

### 3.4.2 The LARS algorithm

The LARS algorithm uses a simple mathematical technique to reduce the number of iterative steps taken by the stagewise algorithm, requiring only $p$ steps for a full solution, where $p$ is the number of input variables. LARS starts in a similar manner to stagewise regression, but at each step, the $\hat{\mathbf{y}}$ vector is updated in the largest step possible in the direction of the current input variable $\mathbf{x}_j$ until some other input, say $\mathbf{x}_k$, has as much correlation with the current residual vector as $\mathbf{x}_j$. At this point, LARS continues in a direction equiangular between the two predictors $\mathbf{x}_j$ and $\mathbf{x}_k$ until a third variable $\mathbf{x}_i$ is as correlated with the residual as the current direction is. LARS then proceeds in the direction equiangular to $\mathbf{x}_i$, $\mathbf{x}_j$, and $\mathbf{x}_k$ [49].

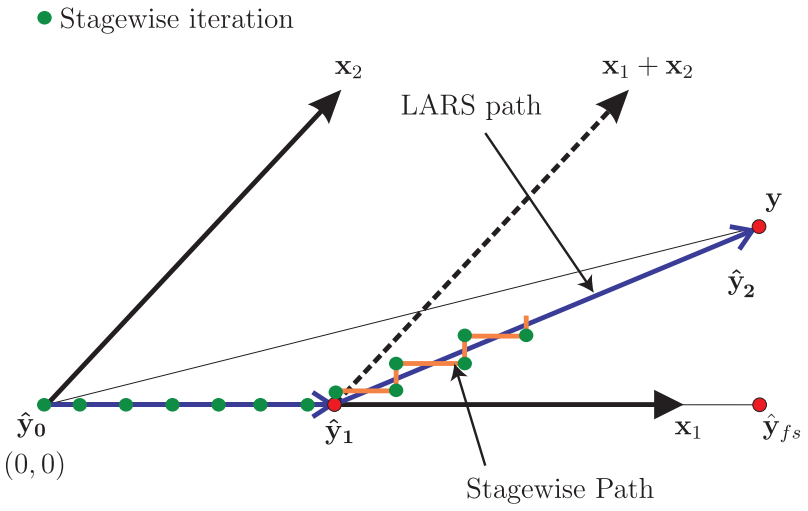



FIGURE 3.1: The LARS algorithm depicted for $p = 2$ input variables. The step-like path is a typical stagewise algorithm progression. $\mathbf{y}_{fs}$ depicts the point where forward selection algorithms proceed to upon the first iteration [49].

The progression of the LARS algorithm for a $p = 2$ example is shown in Figure 3.1. In Figure 3.1, $\mathbf{y}$ is the projection of the target vector into the linear space spanned by $\mathbf{x}_1$ and $\mathbf{x}_2$, $\mathcal{L}(\mathbf{X})$. The algorithm starts with $\hat{\mathbf{y}}_\mathbf{0} = 0$, and the correlation matrix is constructed as

$$\hat{C} = X^T(\mathbf{y} - \hat{\mathbf{y}}_\mathbf{0}). \tag{3.18}$$

From Figure 3.1, $\mathbf{y} - \hat{\mathbf{y}}_\mathbf{0}$ makes a smaller angle with $\mathbf{x}_1$ than it does with $\mathbf{x}_2$. Hence $\hat{\mathbf{y}}_\mathbf{0}$ is augmented in the direction of $\mathbf{x}_1$, to

$$\hat{\mathbf{y}}_\mathbf{1} = \hat{\mathbf{y}}_\mathbf{0} + \eta_1 \mathbf{x}_\mathbf{1}. \tag{3.19}$$

In Equation (3.19), the selection of $\eta_1$ differs between the forward selection, stagewise regression, and LARS algorithms:

- Forward selection chooses $\eta_1$ so that $\hat{\mathbf{y}}_1 = \mathbf{y}_{fs}$, the projection of $\mathbf{y}$ onto $\mathcal{L}(\mathbf{x}_\mathbf{1})$.
- Stagewise regression chooses $\eta_1$ to be a small constant, and iterates the calculation of $C$ and $\hat{\mathbf{y}}$ many times.
- LARS chooses $\eta_1$ so that $\mathbf{y} - \hat{\mathbf{y}}_\mathbf{1}$ is equally correlated with $\mathbf{x}_1$ and $\mathbf{x}_2$.

During LARS, after the first iteration, $\mathbf{y} - \hat{\mathbf{y}}_\mathbf{1}$ bisects the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$. The next LARS estimate will be

$$\hat{\mathbf{y}}_2 = \hat{\mathbf{y}}_1 + \eta_2 \mathbf{x}_\mathbf{2}, \tag{3.20}$$

where $\eta_2$ is chosen to make $\hat{\mathbf{y}}_\mathbf{2} = \mathbf{y}$. For examples with $p > 2$, $\eta_2$ would be smaller, leading to more changes in direction before a full solution is found. The LARS algorithm is quick to compute, has a small memory requirement, and multivariate relationships are captured in the developed model [50].

### 3.4.3 Training stop criteria

The stopping criteria for the LARS algorithm vary between users. The selection of input variables can be stopped when the correlations between new input variables and the residual vector level off, meaning that the model is not improving at each iteration. A similar restriction can be applied to the model error at each iteration. Some users employ metrics such as $C = (\epsilon_i)^2 - p + 2 \times$(number of variables in model), which incorporates the residual error at the current iteration $i$, along with the number of input variables in the

model. The stopping criteria for LARS is a user preference decided before commencing analysis.

For the analysis in this thesis, *cross validation* is used in an attempt to obtain the optimal model from the LARS training procedure. Cross validation involves the use of separate data set, the *validation data set*, to monitor the progress of the training procedure. The validation data set is a set of previously unseen input/output data points that are not used in the determination of the LARS model. As model training progresses (using the *training data set*), the model estimation error is minimised for the training data with successive iterations, and the model estimation error on the validation data set is monitored. Typically, model performance on the validation data set will deteriorate with extensive training because the model loses its generalisation capability, and becomes overly specialised to the specific variations in the training data. This phenomenon is known as *overtraining* or *overfitting*. In computationally intensive applications, when the validation data error starts to increase, *early stopping* is employed, where training is stopped prematurely, and the model is deemed optimal at the point of the lowest validation error. Otherwise, the error performance curves are analysed after the training procedure has completely finished, and the model state is returned to the optimal training point. Cross validation techniques are not specific to LARS modelling, and are employed in many of the different iterative model training procedures used in this thesis.

Figure 3.2 depicts the model mean squared error (MSE) (see Section 3.10.1) for the training and validation data sets for an example LARS training procedure. As more variables are added to the LARS model, the model becomes overtrained on the training data set, and this effect is detected by monitoring the errors produced on the validation data set. The optimal number of variables in the model is determined by finding the minimum value of the validation error, as exemplified in Figure 3.2.

## 3.5 Principal component regression

Principal component analysis (PCA) is an unsupervised data reduction technique that extracts explanatory or *latent* variables from a data set using a matrix decomposition. An unsupervised technique is one which acts without knowledge of desired outputs. The earliest descriptions of PCA-like algorithms were given by Pearson [51] in 1901 and Hotelling [52] in 1933. Both papers adopted different approaches. Pearson concentrated on finding lines and planes that best fit a set of points in $p$-dimensional space while Hotelling's motivation was to find a "fundamental set of independent variables

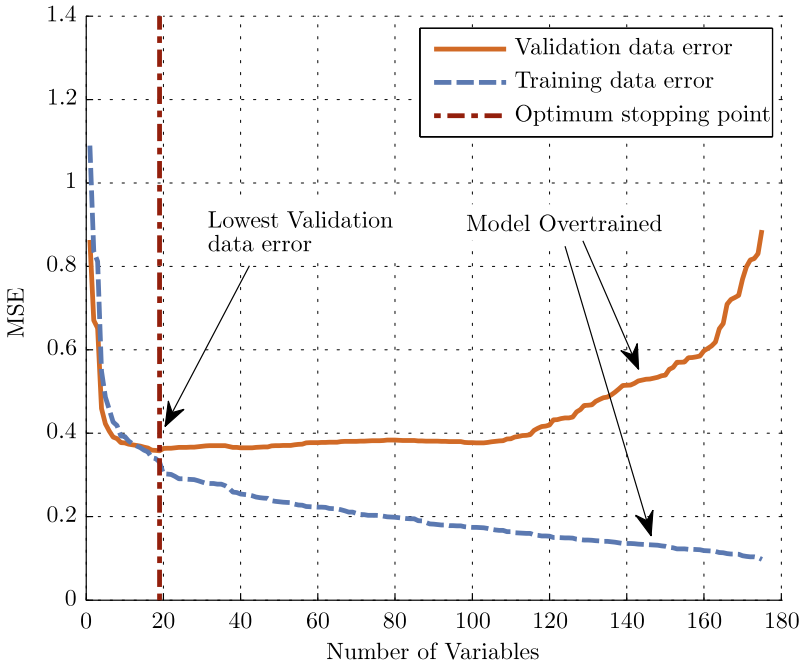



FIGURE 3.2: Typical training mean squared error (MSE) curves for LARS modelling. The validation error curve reaches a minimum and then begins to increase as the model becomes overtrained on the training data. The point of lowest validation error is highlighted in the figure.

... which determine the values" of the original $p$ variables [53]. The explanatory variables extracted using PCA are called principal components by Hotelling, and are a new set of uncorrelated variables extracted from the original data set that explain the main sources of *variance* in the data set. The principal components are arranged in order of the variance that each one explains [54].

Principal component regression (PCR) is the use of the principal components as input variables to a linear regression model. Pearson [51] stated that PCA could be calculated by hand for $p < 4$ but calculations quickly become cumbersome for systems with more dimensions, where PCA is most useful. As such, the full potential of PCA was not exploited until the advent of computers. Today, PCA is employed as a data analysis tool in a wide variety of application areas. The popularity of the technique was illustrated in the text by Jolliffe [53] by the fact that the *Web of Science* identified over 2000 articles published in the two years 1999–2000 that include phrases "principal component analysis" or "principle components analysis" in their titles or keywords. A repeat of this search for the most recent five years (2006 - 2010) yields almost $12,000$ articles spanning over 200 different subject areas.

Before PCA is performed on a data set, some preprocessing steps are normally taken. The mean is subtracted from each variable so that variables have zero mean

and, typically, normalisation to unit variance is also performed when the original data have multiple amplitude scales. Normalisation in this manner gives all variables equal importance during analysis, but such normalisation should be applied with care if it is known that some input variables contain significant amounts of noise.

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a data matrix made up of $n$ samples of $p$ variables. PCA performs an eigenvalue decomposition of the covariance matrix $\mathbf{X}^T\mathbf{X}$ to decompose $\mathbf{X}$ as the sum of the outer product of the *score* vectors $\mathbf{t}_i \in \mathbb{R}^{n \times 1}$ and the *loading* vectors $\mathbf{p}_i \in \mathbb{R}^{p \times 1}$, plus a residual matrix $\mathbf{E} \in \mathbb{R}^{n \times p}$ [55]:

$$\mathbf{X} = \mathbf{t_1}\mathbf{p_1}^T + \mathbf{t_2}\mathbf{p_2}^T + \cdots + \mathbf{t_l}\mathbf{p_l}^T + \mathbf{E} \tag{3.21}$$

$$= \mathbf{T}\mathbf{P}^T + \mathbf{E}, \tag{3.22}$$

where

$$\mathbf{T} = [\mathbf{t}_1\ \mathbf{t}_2 \cdots \mathbf{t}_l], \mathbf{P} = [\mathbf{p}_1\ \mathbf{p}_2 \cdots \mathbf{p}_l], \tag{3.23}$$

$l$ is the number of principal components, $\mathbf{T} \in \mathbb{R}^{n \times l}$ the principal component score matrix, and $\mathbf{P} \in \mathbb{R}^{p \times l}$ is the principal component loadings matrix. The principal components are arranged in descending order, consistent with the amount of variance explained in the original data set by each one. Each principal component loading vector $\mathbf{p}_i$ corresponds to an eigenvector of the covariance matrix of $\mathbf{X}$ and the amount of variance explained by each principal component is proportional to the size of the corresponding eigenvalue $\lambda_i$.

For PCA, the decomposition of $\mathbf{X}$ is such that the columns of the loading matrix $\mathbf{P}$ are orthonormal to each other and the columns of the principal component matrix $\mathbf{T}$ are orthogonal to each other. The first principal component is the linear combination of the $p$ original input variables that explains the greatest amount of variability in $\mathbf{X}$ ($\mathbf{t_1} = \mathbf{X}\mathbf{p_1}$). In the $p$-dimensional variable space, the loading vector $\mathbf{p_1}$ defines the direction of the greatest variance in the data matrix $\mathbf{X}$. Overall, the loadings represent how the original variables $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p$ are weighted to form the principal component scores $\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_l$, the principal component scores model $\mathbf{X}$, and, finally, the residual $\mathbf{E}$ represents the data that is left unrepresented by that model. The principal components of $\mathbf{X}$ are typically calculated using the singular valued decomposition (SVD) [53] or the non-linear iterative partial least squares (NIPALS) algorithm [56, 57].

An example of PCA applied to an artificially created, two-dimensional ($p = 2$) data set is depicted in Figure 3.3. The example data set is constructed with two variables, $x_1$ and $x_2$, as plotted with the grey circles on the diagram. The principal component loadings described in the $P$ matrix are overlayed on the diagram as two perpendicular

vectors. The first principal component loading $\mathbf{p}_1$ describes the direction of greatest variance in the data set. The diagram shows the projections of the original data points onto the first principal component. These points are the scores on the first principal component, represented by $\mathbf{t_1}$. PCR is carried out by using the score values as inputs to a LSR model.

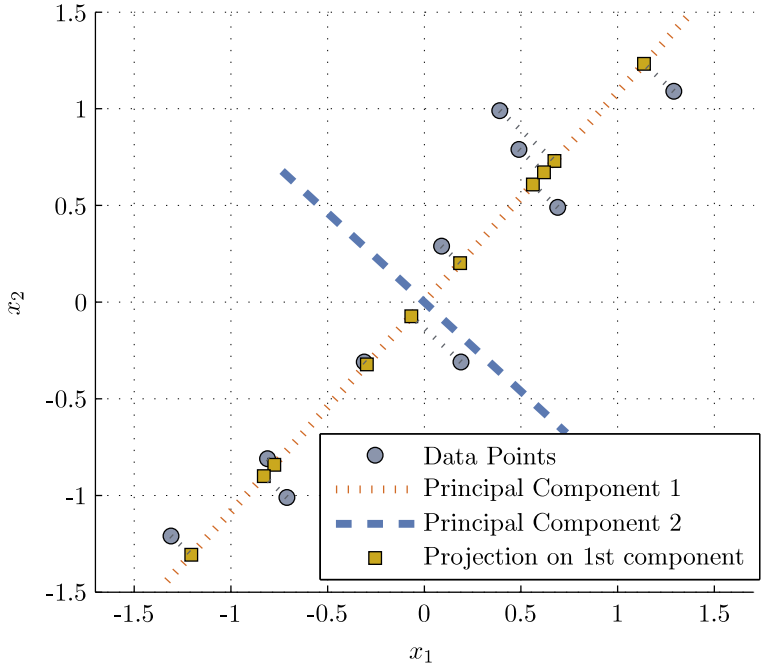



FIGURE 3.3: Principal component analysis depicted for an example data set with $p = 2$ variables. The original samples are shown as grey circles, with the two principal components overlayed for clarity. The first principal component lies along the direction of most variance in the data set. Projections of the original data points onto this principal component are shown as squares.

For a matrix $\mathbf{X}$ of rank $r$, $r$ principal components can be calculated. However, the first $l$ ($l < r$) of the principal components may be sufficient to explain the majority of the variance in the data. If $l = \text{rank}(\mathbf{X})$, then $\mathbf{E} = 0$, and the representation of the data is exact using the principal components. The total variance explained by a number of principal components for an example PCA analysis, using multivariate data from a semiconductor etch process, is shown in Figure 3.4. Figure 3.4 demonstrates that the total explained variance increases monotonically as more principal components are included in the model and that more variance is described by the first principal components compared to the later ones.

In the case of a data set with a number of highly correlated variables, PCA can be employed as a data compression or data reduction technique, extracting a small number of orthogonal principal components that almost completely describe the variations in the multitude of original variables. In this way, the problem of multicollinearity,

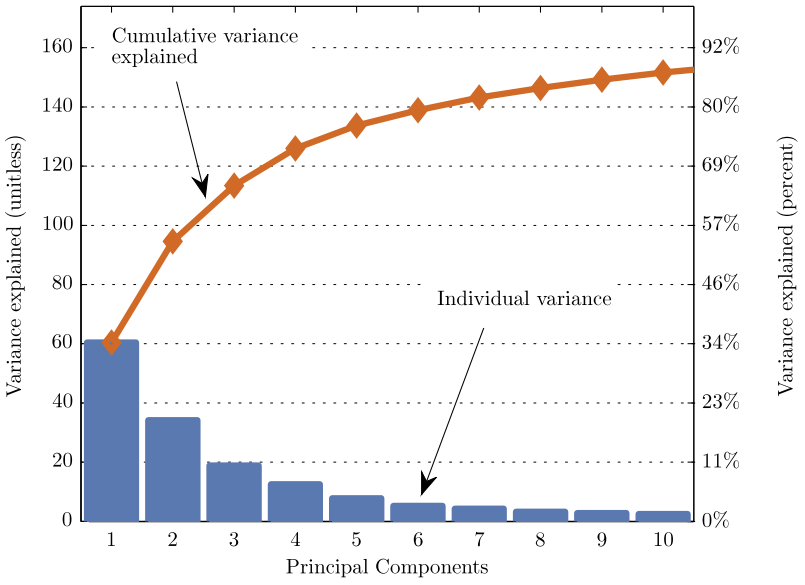



FIGURE 3.4: Variance explained by each component in a PCA analysis using plasma impedance monitoring data from a semiconductor etch process. Note that the the principal components are arranged in order of variance explained, and the variance explained increases monotonically with the number of principal components included in the model.

which presents mathematical difficulties for LSR (see Section 3.1), can be avoided during modelling exercises. However, because PCA is an unsupervised technique, where principal components are extracted from $\mathbf{X}$ without reference to the output $\mathbf{y}$, there is no guarantee that the components used in the PCR model contain information suitable for modelling. Important information can sometimes be disregarded unintentionally as noise in the lower principal components.

One of the disadvantages of PCA is that it takes no explicit account of the ordered nature of the data set [34]. Multi-way PCA (MWPCA) is an extension of PCA routine that is more suited to three dimensional data that is common with process monitoring applications. Data is normally arranged at first in three dimensions: wafers/units, variables and time. In MWPCA, this array is *unfolded* to form a large two-dimensional matrix upon which PCA is subsequently performed, as depicted in Figure 3.5. The raw data can be unfolded in different ways to analyse variability along the three dimensions. However, MWPCA is sometimes regarded as an inferior technique as it essentially relies upon a two-dimensional method (PCA) to analyse multi-dimensional data. For a comparison of MWPCA with some truly multi-dimensional methods in the area of fault detection, namely, trilinear decomposition and parallel factor analysis, the reader is guided to work by Wise *et al.* [34].

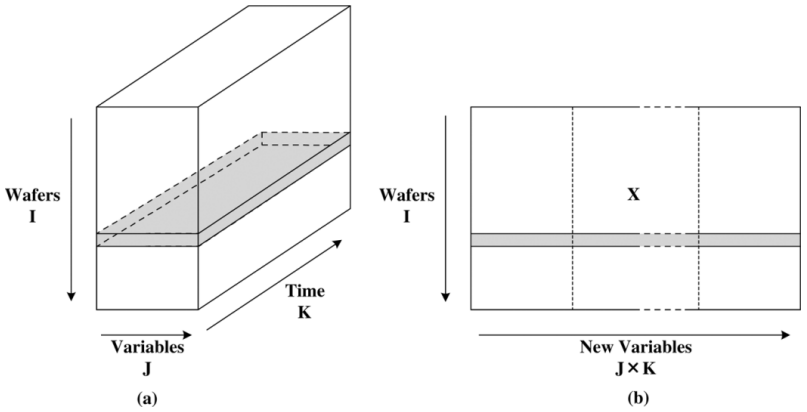



FIGURE 3.5: Example of data unfolding during multi-way PCA [58]. (a) Typical three dimensional data structure. (b) Data after unfolding prior to PCA. Data for one complete wafer is highlighted in gray in both (a) and (b).

### 3.5.1 Hotelling's $T^2$ statistic

At this point, it is worthwhile mentioning two statistics that are commonly employed in conjunction with PCA, Hotelling's $T^2$ statistic and the lack-of-fit statistic $Q$. Hotelling's $T^2$ statistic measures the variation of each sample $\vec{\mathbf{x}}_i \in \mathbb{R}^{1\times p}$ within the PCA model. The $T^2$ statistic is defined as the sum of normalised squared principal components and is calculated for the $i^{th}$ sample in $\mathbf{X}$ [59] as

$$T_i^2 = \vec{\mathbf{t}}_i \mathbf{\Lambda}^{-1} \vec{\mathbf{t}}_i^T = \vec{\mathbf{x}}_i \mathbf{P} \mathbf{\Lambda}^{-1} \mathbf{P}^T \vec{\mathbf{x}}_i^T \tag{3.24}$$

where $\vec{\mathbf{t}}_i \in \mathbb{R}^{1\times l}$ is the $i^{th}$ row of $\mathbf{T}$, $\mathbf{\Lambda} \in \mathbb{R}^{l\times l}$ is a diagonal matrix of the eigenvalues associated with the $l$ principal components retained in the principal component model, and $\mathbf{P}$ is the PCA loadings matrix truncated to the $l$ components of interest. To find the $T^2$ value for a sample not used to create the PCA model, $\vec{\mathbf{x}}_*$, the new sample should first be mean centered if this step was taken with the original data and then Equation (3.24) can be used, substituting $\vec{\mathbf{x}}_*$ for $\vec{\mathbf{x}}_i$.

The $T^2$-statistic can be used as an indicator of movement within the principal component space defined by the loadings $\mathbf{P}$ and is often employed in statistical process control for fault detection and process monitoring [60, 61].

The $T^2$ statistic is also calculable from the original data distribution [62] (taking all variables into account), and is also used often in this form for multivariate statistical process control with performing PCA, as

$$T_i^2 = (\vec{\mathbf{x}}_i - \bar{\vec{\mathbf{x}}})\mathbf{S}^{-1}(\vec{\mathbf{x}}_i - \bar{\vec{\mathbf{x}}})^T \tag{3.25}$$

where $\bar{\vec{\mathbf{x}}} \in \mathbb{R}^{1\times p}$ is a matrix made up of the column means of $\mathbf{X}$, and $\mathbf{S}$ is the covariance matrix of $\mathbf{X}$ such that the elements of $\mathbf{S}$, $s_{jk}$, are

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{\mathbf{x}}_j)(x_{ik} - \bar{\mathbf{x}}_k), \tag{3.26}$$

where $\bar{\mathbf{x}}_j$ denotes the mean of column $j$ of $\mathbf{X}$.

### 3.5.2 Lack-of-fit $Q$-statistic

The $Q$-statistic, also known as the squared prediction error (SPE), is a measure of the degree of variation exhibited by a sample of the input variables (a row of $\mathbf{X}$) that is unexplained by a PCA model. The $Q$-statistic is the sum of squares of each row of $\mathbf{E}$, and is calculated for the $i^{th}$ sample of the data matrix $\mathbf{X}$ [34] as

$$Q_i = \vec{\mathbf{e}}_i\vec{\mathbf{e}}_i^T = \vec{\mathbf{x}}_i(\mathbf{I} - \mathbf{P}\mathbf{P}^T)\vec{\mathbf{x}}_i^T, \tag{3.27}$$

where $\vec{\mathbf{e}}_i$ is the $i^{th}$ row of $\mathbf{E}$, $\vec{\mathbf{x}}_i$ is the $i^{th}$ row of $\mathbf{X}$, $\mathbf{I}$ is the $p \times p$ identity matrix and $\mathbf{P}$ is the PCA loadings matrix, truncated to the $l$ principal components of interest. The $Q$-statistic indicates how well each row of $\mathbf{X}$ conforms to the PCA model and is a measure of the amount of variation not captured by the model [63]. The $Q$-statistic can be calculated for a new observation $\vec{\mathbf{x}}_*$ by replacing $\vec{\mathbf{x}}_i$ with $\vec{\mathbf{x}}_*$ in Equation (3.27) after appropriate normalisation.

## 3.6 Partial least squares regression

Partial least squares (PLS) regression is a technique that combines features of principal component analysis and multiple linear regression [64]. Although PLS is similar to PCR (Section 3.5) in that latent variables describing the data set are extracted using eigenvalue decompositions of the data matrices to form a process model, PLS has the advantage of being a supervised technique, using information in the output variable(s) to create the prediction model. Unlike simpler linear regression techniques, PLS can

construct predictive models even in the presence of collinear input variables since it extracts independent explanatory components from the original data prior to modelling.

The development of PLS is attributed to Wold [56] in 1966 where he applied the technique to the social sciences. PLS regression gained popularity recently, partly due to Wold's son [65], who applied the technique to chemometrics (computational chemistry) for statistical process modelling. PLS is sometimes referred to as projection to latent structures, and has been shown to be a robust multivariate linear regression technique for the analysis and modelling of noisy and highly correlated data [11, 59].

PLS operates in a similar manner to PCA. Component scores $\mathbf{T}$ and loadings $\mathbf{P}$ are extracted from the input data matrix $\mathbf{X}$ such that $\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$ (Equation (3.22)). Simultaneously, a similar decomposition is carried out on the output matrix $\mathbf{Y}$ such that

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F}, \tag{3.28}$$

where $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2 \cdots \mathbf{y}_m]$, $\mathbf{Y} \in \mathbb{R}^{n \times m}$, is the output matrix, $\mathbf{U} \in \mathbb{R}^{n \times h}$ and $\mathbf{Q} \in \mathbb{R}^{m \times h}$ are the $\mathbf{Y}$-scores and $\mathbf{Y}$-loadings respectively, $\mathbf{F} \in \mathbb{R}^{n \times m}$ is the $\mathbf{Y}$-residual matrix, and $h$ is the number of principal components used in the output matrix decomposition.

In PCA, the input principal component scores are chosen to explain as much of the input variable variance as possible, but in PLS, the input principal component scores (in $\mathbf{T}$) and output principal component scores (in $\mathbf{U}$) are chosen so that the relationship between successive pairs of principal component scores is as strong as possible [66]. This choice of scores is the equivalent of rotating the input principal components so that they lie in directions in the input subspace associated with high levels of variation in the output variables.

Equations (3.22) and (3.28) are known as the outer relations of the $\mathbf{X}$ and $\mathbf{Y}$ matrices, respectively. The inner relation, which relates the two decompositions in Equations (3.22) and (3.28) is

$$\mathbf{U} = \mathbf{T}\mathbf{B}, \tag{3.29}$$

where $\mathbf{B}$ is a diagonal matrix of weights optimised to maximise the covariance between the components in $\mathbf{U}$ and $\mathbf{T}$. The two main algorithms used to calculate PLS models (consisting of the $\mathbf{P}, \mathbf{Q}$, and $\mathbf{U}$ matrices) are the non-iterative partial least squares (NIPALS) algorithm [57] and the straightforward implementation of a statistically inspired modification of the PLS method, or "SIMPLS", algorithm [67].

Estimates from PLS models are obtained using the multivariate regression formula $\hat{\mathbf{Y}} = \mathbf{T}\mathbf{B}\mathbf{Q}^T$ [64]. The components $\mathbf{T}$ are extracted from new $\mathbf{X}$ values using the

previously determined **P** loadings matrix. The optimal number of components to include in PLS models can be determined by analysing the amount of variance explained by the components and including enough components to reach a pre-specified percentage explained variance, or through the use of cross validation techniques as described in Section 3.4.3. For the purposes of this thesis, validation data sets are formed and cross validation is used to optimise the number of components retained in the PLS models used. Error performance curves for PLS model errors are similar to those shown in Figure 3.2. The model is deemed optimal at the minimum value of the validation data error.

## 3.7 Artificial neural networks

Artificial neural networks (ANNs) are networks of interconnected processing elements called nodes or neurons that can be used for information processing, and are an attempt to mimic the functionality and structure of the human brain. ANNs have been in existence for several decades, and the start of their development is credited to McCulloch and Pitts [68] who first theorised how knowledge is learned by brain neurons in 1943. Further development was achieved in work by Hebb [69], Rosemblatt [70], and Minsky and Papert [71] in the mid-twentieth century. With vast advances in computing power and a number of significant breakthroughs in ANN technology, more recent decades have seen the implementation of ANNs in classification and regression applications across a wide variety of industries. The popularity of ANNs is born from their ability to perform highly complex mappings on non-linear data, and the capability to generalise well enough to learn overall trends in functional relationships from training data [28].

A number of different ANN types have been developed, distinguished by the type of model input variables (continuous or binary), the network structure, and the type of learning procedures employed to learn trends from data. These ANN types include:

1. Multi-layer perceptrons (MLPs),
2. Radial basis function networks (RBFs), [72]
3. Hopfield networks, [73] and
4. Kohonen's self-organising feature maps, among others. [74]

MLP networks trained using the backpropagation (BP) algorithm are the most generally applicable and most popular networks employed in process modelling in the semiconductor manufacturing industry [75]. Hopfield networks and Kohonen's self-organising

feature maps are unsupervised methods *unsuitable* for input-output function approximation. Although RBF networks are suited for function approximation, they are known to suffer from the curse of dimensionality, where a large number of neurons is required to map high-dimensional input spaces [76]. Difficulties with high dimensional spaces arise because RBF networks construct "local" function approximations using exponentially decaying localised non-linearities, while MLP networks construct "global" function approximations. Although local approximations can allow RBF networks to learn some functions faster and give better generalisation at specific regions around the training data points than MLP networks, this is brought about at the expense of a prohibitively large number of radial basis functions to completely map the input space adequately [77]. For this reason, MLP networks are chosen as the network structure for the semiconductor-etch process modelling work described in this thesis (in Chapters 6 and 7).

### 3.7.1 Multi-layered perceptrons

A multi-layer perceptron (MLP) neural network is an ANN where neurons are arranged in several layers: an input layer, one or multiple hidden layers, and an output layer. The most common model for each neuron is based on the McCulloch-Pitts neuron, introduced in 1943 [68] and shown in Figure 3.6. Each neuron receives a number of inputs $in_i$, and associated with each input is a weight value $w_i$. The neuron calculates the weighted sum of the inputs and adds a bias value $b$. Inputs consist of information provided to the neural network from external sources, or information from other neurons in the network. The calculated sum is passed to an activation function $f(.)$ in the neuron to yield the neuron output. The McCulloch-Pitts neurons originally used a threshold activation function and the *perceptron* neuron introduced by Rosenblatt in 1957 used a linear activation function. Multiple layer MLPs using linear activation functions can be expressed as single layer networks, which are incapable of implementing particular classification functions (for example the XOR function) [71]. Hence, non-linear activation functions must be employed in the hidden layers of the network to allow a broad range of functions to be approximated.

In an MLP network, the neurons in each layer receive weighted inputs from all neurons in the preceding layer along with a constant bias value, calculate an output value using their internal activation function, and pass their outputs to the next layer. This *feedforward neural network* structure is depicted in Figure 3.7. The MLP neurons can use linear or non-linear activation functions such as step, log-sigmoid, or tan-sigmoid functions. The ANN outputs are limited to a small range if the output layer of the MLP uses sigmoid activation functions and as such, networks with sigmoid activation

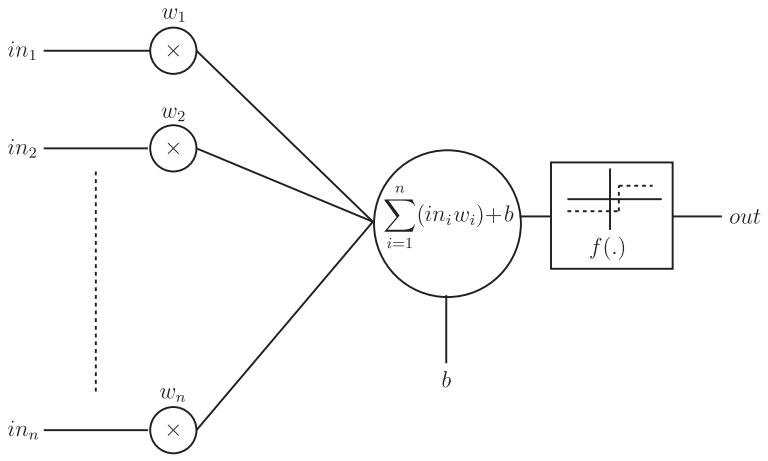



FIGURE 3.6: McCulloch-Pitts neuron. Neuron calculates weighted sum of its inputs along with a constant bias term $b$ and uses an activation function $f(.)$ to produce an output.

functions in the output layer are often applied to classification problems. Linear activation functions are typically used in the output layer neurons to avoid limiting the ANN output range for regression applications.

It has been shown that a feed-forward MLP network using an arbitrary number of neurons in a single hidden layer can approximate almost any continuous function [78, 79]. It is important to note, however, that this result is merely an existence proof and that approximation of complex functions may require a prohibitively large number of neurons in the hidden layer of the neural network. Cybenko [80] suspected that many approximation problems will require astronomical numbers of terms. Such networks would be difficult to implement and train. Zhang *et al.* [81] stressed in their work that neural networks can indeed approximate virtually any non-linear mapping, but assume that noiseless data are available in arbitrary amounts. In realistic situations with limited amounts of noisy data, the number of the parameters in the models may become excessively large.

Neural networks have been broadly applied in the literature to plasma processes due to their ability to approximate non-linear functions, and have been shown to outperform statistical techniques such as PCR, LSR, and PLS regression on some data sets [82, 83].

### 3.7.2 Backpropagation training

Information in an MLP is stored in the values of the weight and the bias values in the network. The optimal weight and bias values are determined through the use of a

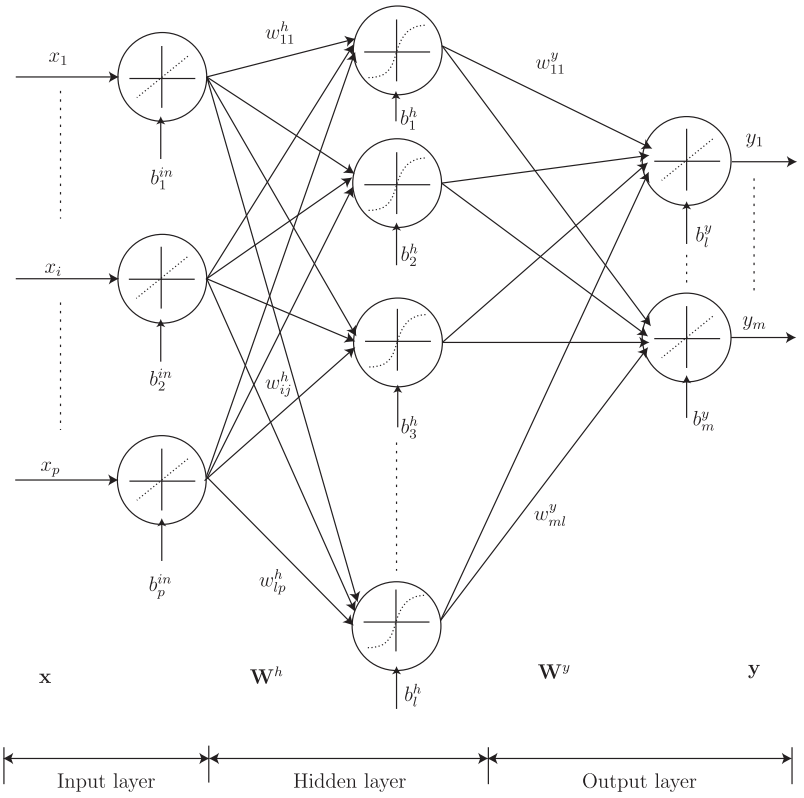



FIGURE 3.7: Multi-layer perceptron (MLP) feed-forward neural network. This diagram shows a neural network with one hidden layer consisting of $l$ neurons, showing the internal neurons, weights, and biases in the network. The input is supplied in the vector $\mathbf{x} \in \mathbb{R}^{p \times 1}$ and the network calculates an output vector $\mathbf{y} \in \mathbb{R}^{m \times 1}$. Weights are labelled as $w_{kj}^h$, which describes the weight between the $k^{th}$ neuron in layer $h$ and the $j^{th}$ neuron in the previous layer. The $in$ superscript denotes the input layer, $h$ the hidden layer, and $y$ the output layer. The weight vectors between each layer are denoted $W^h$ and $W^y$ (for the hidden and output layers, respectively). The neuron biases are denoted $b$, with a subscript for neuron number and superscript to indicate the relevant network layer.

training data set and a learning or training procedure, after which the network can be used to produce outputs from data not in the training data set (unseen data). For MLPs, where specific outputs are required, learning is supervised and the training data contain sample input-output pairs for the network to use to learn the underlying functional relationship of the data.

The *backpropagation* algorithm was introduced by Rumelhart *et al.* [84] as a learning method for MLPs. The backpropagation algorithm begins by first initialising the network weights, typically in a random fashion, but initialisation can also be achieved incorporating intelligent initialisation techniques [85]. The output of the network is

calculated and the sum squared error (SSE), defined as

$$\text{SSE} = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \hat{y}_{ij})^2, \tag{3.30}$$

is used as a training error measure or cost function, where $n$ is the number of samples in the training set, $m$ is the number of outputs from the network, $y_{ij}$ is the desired value for output $j$ at sample $i$, and $\hat{y}_{ij}$ is the estimated value for output $j$ at sample $i$. After the forward pass through the network, the error is propagated backward from the output layer and learning occurs by minimising the SSE through modification of the weights between each neuron [28]. This error minimisation is typically achieved via an iterative *gradient descent* algorithm, where the weights are adjusted at each iteration in the direction of decreasing SSE. A general gradient-based rule for MLP weight optimisation is

$$\mathbf{w}(m+1) = \mathbf{w}(m) - \upsilon\nabla(m) \tag{3.31}$$

where $\mathbf{w}$ is a vector containing the weights and biases for the network, $m$ and $\upsilon$ are the training iteration number and the *learning rate*, respectively [86]. $\nabla(m)$ is the gradient vector of the error function with respect to the weight vector $\mathbf{w}$,

$$\nabla(m) = \frac{\partial \text{SSE}}{\partial \mathbf{w}(m)}. \tag{3.32}$$

The gradient descent method proceeds iteratively until an error minimum is found. Two variants of the propagation algorithm are generally used. The *batch* backpropagation algorithm calculates the mean weight update based on the gradient calculations for all of the input-output sample in the training data. The weight update in Equation (3.31) is then applied based on this mean gradient vector. The *incremental* or instantaneous backpropagation algorithm applies the weight update after the presentation of each training sample to the network.

In practise, more sophisticated and computationally efficient error minimisation techniques are used, such as the 2nd order gradient descent Broyden-Fletcher-Goldfarb-Shannon (BFGS) method [87] or the Levenberg-Marquardt (LM) method [88, 89] that interpolates between gradient descent and the Gauss-Newton optimisation algorithm. Xia *et al.* [90] recently demonstrated that ANNs trained with the BFGS method converged faster and resulted in less overfitting than ANNs trained using gradient descent on a plasma etch data set.

While the learning rate $\upsilon$ is constant in Equation (3.31), more advanced learning rate schemes can also be adopted during ANN training. A common adaptation to the gradient descent update is the introduction of a *momentum* term $\alpha_{mom}$ to improve convergence and overcome local minima solutions,

$$\mathbf{w}(m+1) = \mathbf{w}(m) - \upsilon\nabla(m) + \alpha_{mom}\nabla(m-1). \tag{3.33}$$

It is common to train several networks with the same structure, initialising the network weights randomly each time to to remove the sensitivity of the solution to the ANN initial conditions. The network with the best performance is then retained for future use. Typically, early stopping techniques, as described in Section 3.4.3, are used for training ANN models, and typical ANN error curves for validation and training data sets are similar to those seen in Figure 3.2.

### 3.7.3 Advantages and disadvantages

Neural networks provide a black-box model with which analysts can approximate complex non-linear functions without a complete understanding of the underlying physical relationships in the data. Although advantageous in some cases, careful selection of model topology and inputs should be carried out to ensure the best modelling results and parsimony of the model. In general, neural networks can be computationally demanding to train due to the number of iterations required for the training method to converge and the repeated training required to avoid local minima solutions during the optimisation of the network weights. ANNs are also relatively data-hungry compared to other modelling techniques, typically requiring a large number of samples to develop a useful model [91].

## 3.8 Gaussian process regression

The use of Gaussian processes (GPs) for regression and classification is a relatively new concept. In 1996, Williams and Rasmussen [92] introduced the use of GPs to high dimensional problems that have been traditionally tackled using other modelling techniques such as neural networks and decision trees. Known in the past under the name of 'kriging' in the spatial statistics community, GPs have become increasingly popular over recent years as a modelling technique [93].

GP modelling does not impose a specific model structure on the underlying function, $f(x)$, being modelled [94]. Instead, a Gaussian prior is placed on the range of possible functions that could represent the mapping of inputs $\mathbf{x}$ to outputs $y$. The Gaussian prior incorporates the analyst's knowledge about the underlying function in the data where available, and is specified using the GP covariance function. As such, GP modelling is considered to be a non-parametric modelling technique, where the training data are used to discover the model *properties* in a supervised manner. However, some basic assumptions must be made about $f(x)$ and these are specified in the GP covariance function.

### 3.8.1 Gaussian process covariance function

For the purposes of this discussion, a one-dimensional input-output process is first assumed for simplicity. A Gaussian process is specified by a mean function $m(x)$ and covariance function $k(x_i, x_j)$ as

$$f(x) \sim GP(m(x), k(x_i, x_j)). \tag{3.34}$$

where

$$m(x) = E[f(x)] \tag{3.35}$$

$$k(x_i, x_j) = E[(f(x_i) - m(x_i))(f(x_j) - m(x_j))] \tag{3.36}$$

where the notation $x_i$ denotes any input sample from the one-dimensional input-output process and $x_j$ is any other input sample from the same process. The Gaussian process can be viewed as a set of random variables that have a joint multivariate Gaussian distribution and represent the value of the function $f(x)$ at location $x$. $f(x_i)$ is a random variable corresponding to the single input-output pair $\{x_i, y_i\}$, where $i$ here denotes sample $i$ in the data set available for modelling. For simplicity, a zero-mean process is assumed such that

$$f(x_1), f(x_2), \cdots, f(x_n) \sim N(0, \mathbf{\Sigma}), \tag{3.37}$$

where $\Sigma$ is the process covariance matrix such that $\Sigma_{ij}$ gives the value of the covariance between $f(x_i)$ and $f(x_j)$, and is a function of $x_i$ and $x_j$, $\Sigma_{ij} = k(x_i, x_j)$.

The covariance function $k(x_i, x_j)$ can be any function, provided that it generates a positive definite covariance matrix $\mathbf{\Sigma}$. One of the most commonly used covariance

functions is the *squared exponential* (SE) covariance function, which has the form:

$$k(x_i, x_j) = \nu^2 exp(-\frac{\|x_i - x_j\|^2}{2l^2}) \tag{3.38}$$

where $\nu$ and $l$ are *hyperparameters* that define the properties of the covariance function. The SE covariance function assumes that input points that are close together in the input space correspond to outputs that are more correlated than outputs corresponding to input points which are further apart (if $x_i \simeq x_j$, $k(x_i, x_j)$ tends towards its maximum, $\nu^2$). Variations in $l$ and $\nu$ control the smoothness of the covariance function. The parameter $\nu$ controls the scale of the variations between points $x_i$ and $x_j$ in the output space, while $l$, the *length scale*, determines the degree of variation possible over the input space. Examples of the effects of different length scales for a single-input single-output GP are shown in Figure 3.8. It can be shown that the use of a GP with a squared exponential covariance function is equivalent to modelling with a linear combination of an infinite number of Gaussian shaped basis functions in the input space [95].

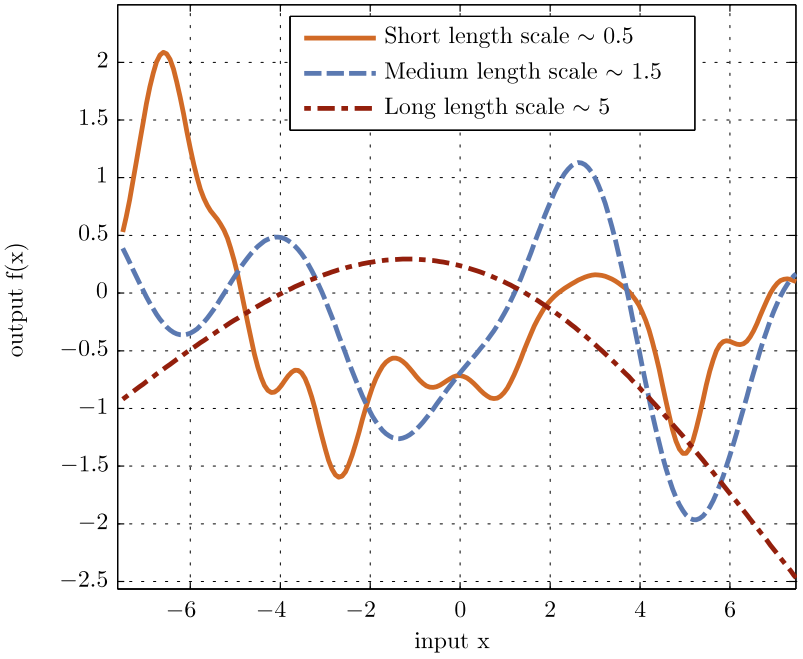



FIGURE 3.8: Three possible outputs from GP models with differing length scales. The length scale can be thought of as the required distance in the input space over which the output values can change significantly. Hence, the GP models with shorter length scales have more "flexibility" in the output space.

### 3.8.2 Optimising model hyperparameters

Because the observed data in a realistic system typically includes noise, it is assumed that the underlying function of our data is $y = f(x) + \epsilon$, where $\epsilon$ is a Gaussian white

noise term with variance $\sigma_n^2$ such that $\epsilon \sim N(0, \sigma_n^2)$. A Gaussian process prior is put on the range of possible underlying functions $f(x)$ with covariance function as exemplified in Equation (3.38) with unknown hyperparameters.

Hence, for this function,

$$y_1, y_2, \cdots, y_n \sim N(0, \mathbf{K}) \tag{3.39}$$

$$\mathbf{K} = \mathbf{\Sigma} + \sigma_n^2 I \tag{3.40}$$

where $\sigma_n^2 I$ represents the covariance between outputs due to white noise, where $I$ is the $n \times n$ identity matrix, and $y_i = f(x_i) + \epsilon_i$.

The aim is to use the set of training data points $\{x_i, y_i\}_{i=1}^n$ to find the posterior distribution of $y_*$, given input $x_*$, that is $p(y_*|x_*, \mathbf{x}_{tr}, \mathbf{y}_{tr})$, where $\{x_*, y_*\}$ denotes an unseen test data point and $\mathbf{x}_{tr} \in \mathbb{R}^{n\times 1}$ and $\mathbf{y}_{tr} \in \mathbb{R}^{n\times 1}$ denote the input and output training data. Before the posterior distribution of $y_*$ is found, the unknown hyperparameters of the covariance function in Equation (3.38), $l$, $\nu$, and the noise variance $\sigma_n^2$, must be optimised to suit the training data. This can be performed via a Monte Carlo method or, more typically, via maximisation of the log marginal likelihood given by

$$log(p(\mathbf{y}_{tr}|\mathbf{x}_{tr})) = -\frac{1}{2}\mathbf{y}_{tr}^T\mathbf{K}^{-1}\mathbf{y}_{tr} - \frac{1}{2}log(|\mathbf{K}|) - \frac{n}{2}log(2\pi). \tag{3.41}$$

The log marginal likelihood can be used to choose between different models. Equation (3.41) is made up of a combination of a *data fit* term, $\frac{1}{2}\mathbf{y}_{tr}^T\mathbf{K}^{-1}\mathbf{y}_{tr}$ that determines the success of the model at fitting the output data, a *model complexity* penalty $\frac{1}{2}log(|\mathbf{K}|))$, and a constant term that depends on the training data set size $\frac{n}{2}log(2\pi)$. Ideally, by maximising the log marginal likelihood the covariance function hyperparameters to fit the training data with the least complexity that fits the input-output training data set most accurately is found. The optimisation problem is non-convex and requires the computation of the derivative of Equation (3.41) with respect to each of the hyperparameters in the covariance function. Since typical gradient descent optimisation routines are sensitive to the initial choice of hyperparameters, multiple random initialisations of the routine, are used in an attempt to ensure that the global optimal solution is found [96]. Initial values for SE hyperparameter initialisation, $v^2 = 1$, $\frac{1}{l^2} = \frac{1}{p}$ and $\sigma_n^2 = 1$, where $p$ is the number of input dimensions, are suggested by Rasmussen [97]. However, through experimentation on the data sets in this thesis, these initial values are not found to always lead to a global optimal solution.

The above arguments can be expanded to the multi-dimensional input case by including the extra input dimensions in $x_i$ and $x_j$. Although $x_i$ and $x_j$ become vectors with multiple dimensions $\vec{\mathbf{x}}_i \in \mathbb{R}^{1\times p}, \vec{\mathbf{x}}_j \in \mathbb{R}^{1\times p}$, $k(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j)$ remains a scalar value and the remainder of the calculations remain the same. The GP covariance function can be extended to many input dimensions by introducing individual hyperparameters for each dimension. For example, in a multi-dimensional application of the SE covariance function, a separate length scale is employed for each input dimension. During optimisation of the covariance function hyperparameters, dimensions which do not influence the process being modelled are automatically assigned longer length scales than variables of greater influence. The analysis of length scales to determine variable importance in GPR is a form of variable selection, or automatic relevance determination (ARD).

### 3.8.3 Prediction with Gaussian process models

When the hyperparameters are optimised, the GP model can be used to predict the distribution of $y_*$ for the input $x_*$ (for a single input dimension). The predictive distribution of $y_*$, $p(y_*|x_*, \mathbf{x}_{tr}, \mathbf{y}_{tr})$, can be shown to be Gaussian [95], with mean and variance

$$\mu(y_*) = \mathbf{k}_*^T \mathbf{K}^{-1} \mathbf{y}_{tr} \tag{3.42}$$

$$\sigma^2(y_*) = k_{**} - \mathbf{k}_*^T \mathbf{K}^{-1} \mathbf{k}_* + \sigma_n^2 \tag{3.43}$$

respectively, where $\mathbf{k}_* = [k(x_*, x_1)k(x_*, x_2) \cdots k(x_*, x_n)]^T$ is a column vector of covariances between the test and training data points and $k_{**} = k(x_*, x_*)$ is the autocovariance of the test input.

In Equation (3.42), the mean prediction $\mu(y_*)$ is a linear combination of the observed outputs $\mathbf{y}_{tr}$, where the linear weights are given by the vector $\mathbf{k}_*^T \mathbf{K}^{-1}$. The variance of the predicted value $\sigma^2(y_*)$, defined in Equation 3.43, is given by the prior variance $k_{**}$, which is a positive term, minus the posterior variance $\mathbf{k}_*^T \mathbf{K}^{-1} \mathbf{k}_*$ which is also positive. The posterior variance will be inversely proportional to the distance between the test point and the training points in the input space, since it depends on $\mathbf{k}_*$. Figure 3.9 shows an example set of input-output training data that are approximated using a GPR model. The prediction mean and confidence intervals for test data between the training points are shown to demonstrate that large variances are produced for test data far from training data in the input space. The prediction variance can be interpreted as a level of confidence in the prediction [96].

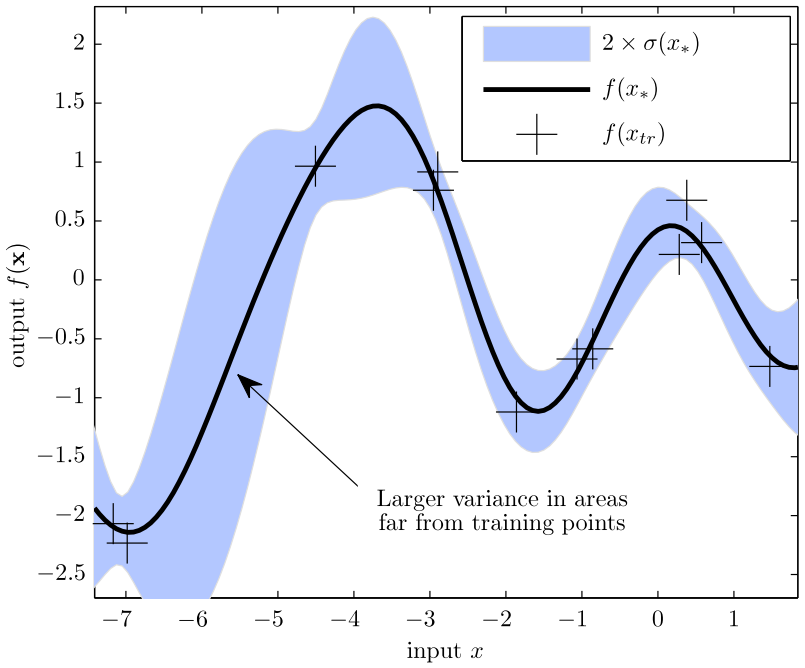



FIGURE 3.9: Example prediction and 95% confidence intervals (2 × standard deviation distance) for one dimensional input output process. The variance on predictions is increased for areas of the input space that are further from observed training data. Relatively simple calculation of confidence intervals on predicted outputs is possible because predictions are given in the form of a distribution.

### 3.8.4 Other covariance functions

The prior beliefs of the analyst about the target data are expressed when specifying the covariance function for a Gaussian process model. To allow a wide variety of underlying functional behaviours to be approximated, new covariance functions can be devised or existing ones combined to cater for different distributions of data. Useful covariance functions include:

1. The *squared exponential* (SE) covariance function (as above) detailed in Equation (3.38) is one of the most commonly used covariance functions in GP modelling applications and has the general form (for multiple input dimensions)

$$k_{SE}(\vec{\mathbf{r}}) = \nu^2 exp\left(-\frac{\vec{\mathbf{r}}\mathbf{Q}^{-1}\vec{\mathbf{r}}^T}{2}\right) \tag{3.44}$$

where $\vec{\mathbf{r}} = \|\vec{\mathbf{x}}_i - \vec{\mathbf{x}}_j\|$, $\nu$ determines the degree of variability in the output variable space, and $\mathbf{Q} = \text{diag}(q_{ii}), \in \mathbb{R}^{p\times p}$ is a diagonal matrix consisting of an ARD length scale parameter for each input dimension $(q_{11}, q_{22}, \cdots q_{pp} = l_1^2, l_2^2, \cdots l_p^2)$. The SE covariance function is infinitely differentiable and, therefore, is very smooth.

2. The *linear* covariance function has the general form

$$k_{lin}(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = \vec{\mathbf{x}}_i \mathbf{Q}^{-1} \vec{\mathbf{x}}_j^T \tag{3.45}$$

where again $\mathbf{Q} = \text{diag}(q_{ii}) \in \mathbb{R}^{p \times p}$ is a diagonal matrix of ARD parameters. The linear covariance function arises from an analysis of linear regression using GPs [95].

3. The *Matérn* class of covariance functions is given by

$$k_{Matern}(\vec{\mathbf{r}}) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{\sqrt{2\nu}\vec{\mathbf{r}}}{l} \right)^{\nu} J_{\nu} \left( \frac{\sqrt{2\nu}\vec{\mathbf{r}}}{l} \right) \tag{3.46}$$

with positive parameters $\nu$ and $l$, where $J_\nu$ is a modified Bessel function of order $\nu$ and $\Gamma(\nu)$ is the Gamma function. The Matérn covariance functions are $\nu - 1$ times differentiable. Hence the parameter $\nu$ can be used to allow very jagged outputs. As $\nu \to \infty$, $k_{Matern}(\vec{\mathbf{r}}) \to k_{SE}(\vec{\mathbf{r}})$. During regression, $\nu$ is typically set to $3/2$ or $5/2$ [95].

4. The *rational quadratic* (RQ) covariance function has the form

$$k_{RQ}(\vec{\mathbf{r}}) = \nu \left( 1 + \frac{\vec{\mathbf{r}} \mathbf{Q}^{-1} \vec{\mathbf{r}}^T}{2\varrho} \right)^{-\varrho} \tag{3.47}$$

with $\varrho > 0$ and can be visualised as an infinite sum of squared exponential covariance functions of differing length scales. The RQ covariance function hence allows the GP to vary the length scale over the range of each input dimension. The limit of the RQ covariance for $\varrho \to \infty$ is the SE covariance function.

5. A *periodic* covariance function can be introduced by first using a nonlinear mapping $u(\vec{\mathbf{x}})$ of the input $\vec{\mathbf{x}}$ and then using a different covariance function in $u$-space. In the case of a one-dimensional input variable $x$, a mapping function $u(x) = (\cos(x), \sin(x))$ is chosen to give a periodic random function of $x$. The periodic covariance function has the general form

$$k(x_i, x_j) = exp \left( -\frac{2sin^2(\pi \frac{x_i - x_j}{2})}{l^2} \right) \tag{3.48}$$

which arises from the application of the SE covariance function in $u$-space.

Depending on the requirements of the data, covariance functions can be summed or multiplied to form suitable GP models depending on the prior knowledge of the physical system being modelled. For example, in [98], a covariance function consisting of a

SE component, a linear component, a periodic component, and a noise component is employed by the authors to model electricity loadings. Different models can be compared using the marginal likelihood or prediction performance of each. A more complete explanation of GP modelling and a discussion of additional covariance functions can be found in [95].

### 3.8.5 Advantages and disadvantages of GP models

Gaussian process models have several advantages over other modelling techniques:

- Useful models can be created from training data sets with a relatively small number of training points.
- The analyst's prior beliefs about the data distribution can be encapsulated in the covariance function, which is modular, and can be made up of several different components to accommodate as much information as possible.
- The model form is not necessarily specified before training and both non-linear and linear functions can be approximated.
- Confidence intervals on predictions can be evaluated easily as each prediction is given in the form of a distribution. Estimates from areas in the input space with a large amount of training data will have a small variance whereas estimates from scarcely populated areas will be highlighted through a high variance.

Although GPs are a useful modelling tool, they are not without their disadvantages. During training and estimation, GP models require the inversion of large covariance matrices, the size of which are determined by the amount of training data points. Although modern computers can invert matrices of up to $1000 \times 1000$ relatively quickly, larger training sets may cause unacceptable computation delays. While this may be a disadvantage in some applications, for the purposes of this thesis, smaller training sets will be prevalent, and GP covariance matrix calculations will not pose a problem.

## 3.9 Model Robustness

As phrased by Jaynes [99], "One seeks data analysis methods that are *robust*, which means insensitive to the exact sampling distribution of errors, ...or are, *resistant*, meaning that large errors in a small proportion of the data do not greatly affect the conclusions." Hence, the *robustness* of a modelling technique is the ability of the technique to

produce useful models in the presence of data that does not strictly obey the mathematical assumptions made by the modelling technique. A range of assumptions are made during the development of each of each of the modelling techniques described in Sections 3.1 – 3.8. While model robustness is difficult to quantify, and often, model assumptions can be reasonably stretched, it is important for analysts to be aware of the limitations of the modelling techniques being used, to check how well their data conforms to the mathematical assumptions of the modelling technique, and to know how violations of the assumptions can affect the modelling results. *Robust* versions of modelling techniques can sometimes be employed to construct useful models in the presence of data that violate modelling assumptions.

For LSR, the following assumptions are made:

- There exists a linear relationship between the input and output variables.
- The errors are independent.
- The errors are normally distributed with zero mean and constant variance (homoscedasticity).
- The input variables are linearly independent, that is, the input data matrix is non-singular.

While data fulfilling these assumptions lead to unbiased and consistent model parameter estimates, all of the assumptions are rarely satisfied with real-world data. Violations of the assumptions will not completely invalidate the results of models produced, but may weaken them. For example, violations of heteroscedastic errors in the data make it difficult to gauge the standard deviation of the forecast errors, resulting in inaccurate confidence intervals. The error metrics and model fit measures described in Section 3.10.1 are typically used to give a measure of the usefulness of the developed models.

Non-normally distributed variables, and variables that violate the linearity assumption are typically tackled by applying transformations to form normally distributed variables that are linearly related to the output variable(s) (for example, the Box-Cox transform [109] is a commonly used normalising transform). However, such transformations must be completed with care as the variable units can be lost, adding difficulty to the analysis of the regression results. Non-normality is typically identified visually (see Section 3.10.2) or using skewness and kurtosis statistics. Because parameter estimation is based on the minimisation of squared error, LSR techniques are particularly sensitive to the presence of outliers in the input data. *Robust regression* techniques are sometimes used to create linear models in the presence of outliers and non-normal data. A number

of variants of robust regression exist, see M-estimation [100] or least trimmed squares regression [101] for examples. The effectiveness of robust regression is demonstrated in Figure 3.10.

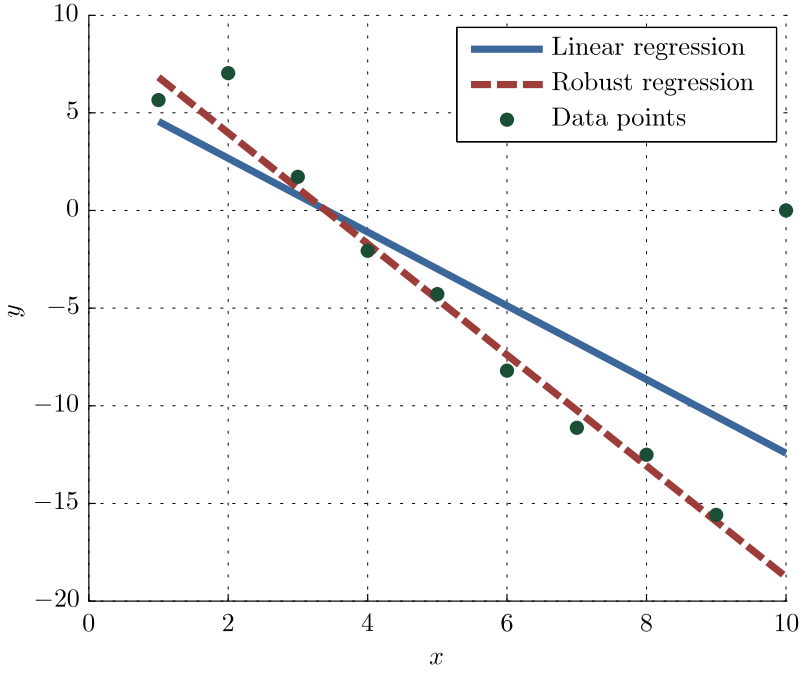



FIGURE 3.10: Robust regression example. The regression line created using the robust regression technique is much less susceptible to errors introduced by the single outlying point at $x = 10$.

PCA uses the data covariance matrix to extract the principal components, and assumes that data is linear and obeys a unimodal normal distribution. The calculations of the mean and covariance matrices are sensitive to outliers, and hence PCA is affected by anomalous observations. PCR suffers the same sensitivities as LSR and PCA because it relies on both techniques. He and Wang [58] demonstrate the effect of non-linearities and multi-modal distributions on PCA.

Although PCA assumes unimodal data when determining the directions of greatest variance in the input data, it should be noted that the use of multi-modal data does not invalidate the results, depending on the application. For high dimensional multi-modal data sets, PCA is a useful tool to visualise and detect systematic variations in covariance structures. However, the use of PCA for outlier detection in multi-modal or non-Gaussian data will be invalid, and the use of metrics such as the Mahalanhobis distance or the $T^2$ statistic will be invalid.

Robust PCR techniques typically involve a robust calculation of the sample covariance matrix, followed by the use of a robust regression technique. For example, the robust covariance matrix technique developed by Walczak and Massart [102] is based

on ellipsoidal multivariate trimming, a technique that iteratively converges on a robust estimate of the sample covariance matrix by systematically removing outlying points. Filzmoser [103] implements robust PCA through the use of a robust measure of variability, the median absolute deviation, to quantify variance, followed by least median of squares regression. Hubert and Engelen [104] propose a robust PCA method based on *projection pursuit*.

Similarly, PLS algorithms such as NIPALS and SIMPLS are known to be sensitive to outliers in the data set, and a range of techniques have been developed to robustify the calculation of the PLS regression coefficients to the presence of outliers. González *et al.* [105] provide a review of robust PLS techniques, and demonstrate that, if the sample covariance matrix is properly robustified, further robustification of the linear regression steps of the PLS algorithm is unnecessary.

Typically, neural network users pay less attention to assumptions than users of statistical techniques. However, ANNs are subject to similar distributional assumptions as statistical techniques about the data being modelled. ANN training procedures typically use least square error metrics during the determination of the network weights, and hence are highly sensitive to outliers in the error terms [106]. Outliers in the data set can lead to a bias in the model outputs, reducing the accuracy of predictions on unseen data. It is generally recommended to preprocess data prior to ANN modelling to remove outliers and test for normality [107], although some research [108] has shown that, in some applications, ANN performance is not highly sensitive to the distribution of the data. Should non-normal data be found, it is typically advised to use a transformation to obtain a normal distribution where possible.

A number of different approaches can be taken to develop robust ANN modelling techniques. For example, a robust error suppressor function is proposed by Walczak [106] to make the backpropagation learning algorithm more robust against outliers. Briegel and Tresp [110] propose to replace the Gaussian noise model in neural network models in favour of a more flexible noise model based on the Student-t distribution, demonstrating increased robustness to outliers.

GPR assumes that the data being modelled is distributed as a multivariate Gaussian distribution, with Gaussian noise on the individual points. Typically, the covariance functions used assume homoscedastic Gaussian noise. Significant outliers in the output variables to be modelled can cause the Gaussian process to assume artificially small length scales, invalidating the confidence intervals produced on predictions. However, GPR models using local covariance functions such as the SE function are relatively robust with respect to outliers in the input data [111]. The accuracy and usefulness of

the output predictions and confidence intervals produced by a GPR model can also be reduced through inaccurate choice of covariance function form.

A number of approaches are taken by researchers to increase the robustness of GPR modelling techniques. For instance, Le *et al.* [112] present an algorithm that assumes heteroscedastic noise terms during GPR. Noise is treated as a random variable that is adapted to suit different local areas of the input variable space. Snelson *et al.* [113] examine non-linear transformations of the Gaussian process outputs to allow for non-Gaussian processes and non-Gaussian noise. In work with a similar aim, Kuss *et al.* [111] use a two-model approach to incorporate the existence of outliers explicitly in the GPR framework. Sollich [114, 115] investigates improvements to the learning procedure of GPR models to make GPR more robust to model mismatch.

The area of robust statistics and robust modelling encompasses a vast array of different modelling approaches that are outside the scope of this thesis. While robust methods are not discussed further in this research, care is taken to remove gross outliers from data prior to modelling in Chapters 6 and 7. Where applicable, model residuals are examined for normal distributions as a metric of model suitability.

## 3.10 Model comparisons

A number of different metrics are employed to compare the performance of different VM models in this thesis. These metrics focus on the error performance of the models. The errors, or residuals, from a predictive model are defined as

$$\epsilon_i = y_i - \hat{y}_i,\ i = 1, 2, 3, ...n \tag{3.49}$$

where $y_i$ is the actual output at observation $i$, and $\hat{y}_i$ is the model estimate for observation $i$. Since the residual may be viewed as the deviation between the observed value and the predicted value, it is also a measure of the variation in the output variable not explained by the prediction model [45].

A number of assumptions are made about the residuals for linear models fitted to data [44]:

1. The residuals $\epsilon$ have zero mean.
2. The residuals $\epsilon$ have constant variance $\sigma^2$.
3. The residuals are uncorrelated in time.

4. The residuals are normally distributed.

Although these assumptions are generally applied to the LSR modelling case, the analysis of residuals for model evaluation applies to any situation where a model is fitted to observed data. The numerical values of the errors along with the validity of these assumptions are often used to evaluate the performance of fitted models.

### 3.10.1 Error Metrics

A number of different methods are available to summarise the overall error performance of a fitted model. The mean squared error (MSE) for the fitted model is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2. \tag{3.50}$$

Squared error based metrics penalise larger errors more heavily than smaller errors. The MSE for a model has units equal to the square of the original quantity units, and so another metric, which is expressed in the original model output units, is the root mean squared error (RMSE), given by

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2}. \tag{3.51}$$

The RMSE is more easily and more intuitively interpreted than the MSE as it is quoted in the same units as the original output variable.

The mean absolute percentage error (MAPE) is the mean error given as a percentage of the original value. This metric is calculated as

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\|\epsilon_i\|}{\|y_i\|}. \tag{3.52}$$

The coefficient of determination, $R^2$, is often employed as a measure of how well estimated values follow the variations in the real data, or the "goodness of fit" of the model. The $R^2$ statistic has values between 0 and 1 and is defined as

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}, \text{ where,} \tag{3.53}$$

$$SSE = \sum_{i=1}^{n}\epsilon_i^2 \text{ and } SST = \sum_{i=1}^{n}(y_i - \bar{\mathbf{y}})^2, \tag{3.54}$$

with $\bar{\mathbf{y}}$ denoting the mean of the observed output values and $SSE$ and $SST$ representing the *sum of squared errors* (see Equation (3.30)) and the *total sum of squares* respectively. An $R^2$ value of 1 indicates that the model captures the output variable variance perfectly. Mathematically, the $R^2$ statistic is equivalent to the square of the linear correlation between the real and fitted values, $y$ and $\hat{y}$.

The mean and standard deviation values of both the training data values and the estimation errors can be useful metrics. It is generally expected to see an estimation error mean extremely close to zero. The estimation error standard deviation is also a significant value: if no better than the training data standard deviation, then the model has performed no better than a simple mean estimator. A ratio of the estimation error standard deviation to the training data standard deviation significantly below 1.0 indicates good regression performance, with a level below 0.1 often said (heuristically) to indicate good regression. This regression ratio (or, more accurately, one minus this ratio) is sometimes referred to as the explained variance of the model.

### 3.10.2 Graphical analysis

A graphical analysis of the residual signals can be useful to diagnose problems with models, to compare the performance of different model types, and to check the validity of the assumptions made about the residual values. For example, by examining plots of the residual values at each observation, as shown in Figure 3.11, it can be verified that the residuals have approximately constant variance and that the mean of the residuals is zero. Plots of the residual against the fitted values and against the regressor variables should also show no visually detectable correlations or patterns.

Deviations from the normal distribution can be detected using a *normal probability plot*, which is a plot of the residuals ranked in numerical order against the cumulative probability $P_i = (i - 0.5)/n, i = 1, 2, \cdots, n$. This plot is done on normal probability axes, on which the vertical axis scale is skewed such that the spacing of divisions becomes larger further from the center point 0.5 towards the upper and lower points, 1 and 0 respectively [47]. If the residuals come from a normal distribution, the resulting points will lie approximately along a straight line. Deviations from this straight line indicate values that do not arise from a normal distribution. Figure 3.12 shows examples of normal distribution plots that arise from normal and non-normal error distributions. Histograms can also be used to visualise the distribution of the model residuals. An example of a histogram with a superimposed fitted normal distribution is shown in Figure 3.13.

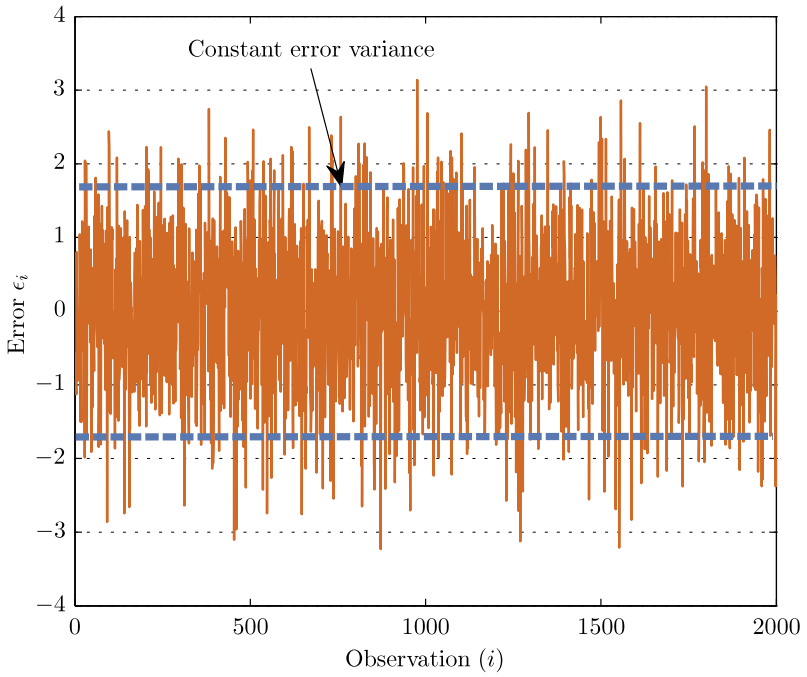



FIGURE 3.11: Example residuals from a linear modelling experiment. A plot of the residuals is used to verify that the residuals have a constant variance, and a mean value of approximately zero. Deviations from this behaviour are indicators of problems with the model fitted.

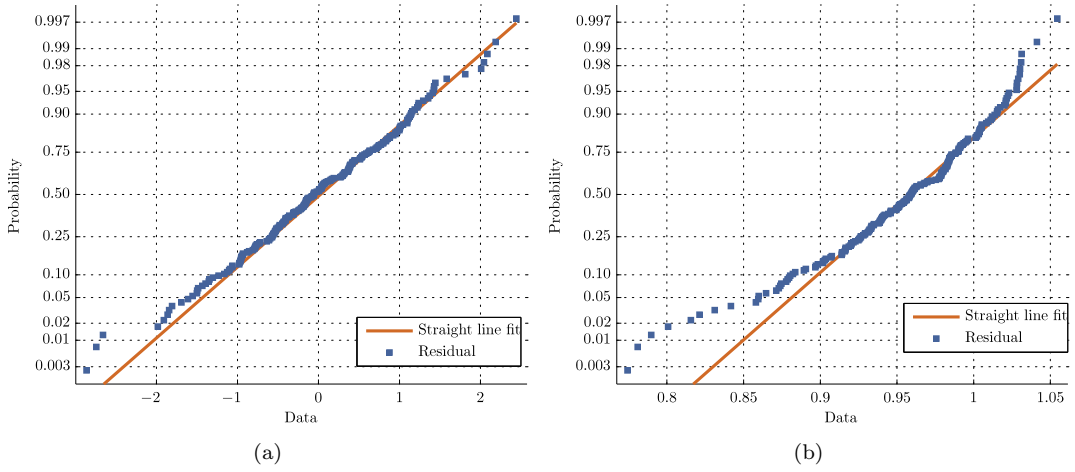



FIGURE 3.12: Normal probability plots for (a) normally distributed data and (b) non-normally distributed data. Non-normal data does not form a straight line on the normal probability plot. Note how the vertical axis scale is adjusted for these plots to suit the normal distribution.

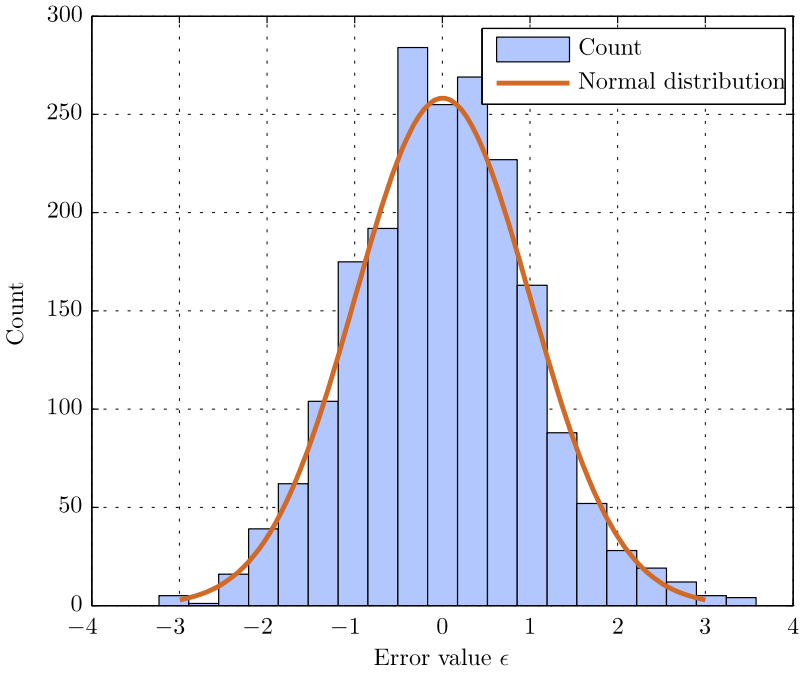



FIGURE 3.13: Histogram of normally distributed residuals, with a superimposed fitted normal distribution.

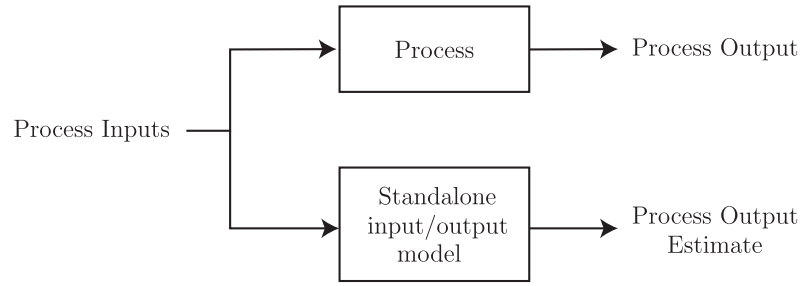
# Chapter 4

# Virtual metrology and modelling of plasma etch - a review

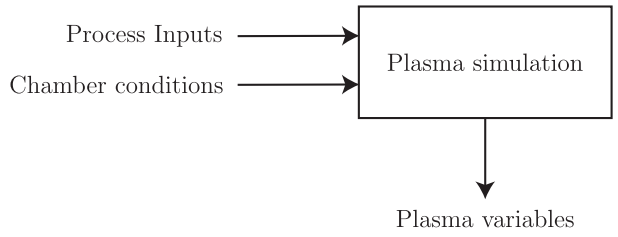
This chapter provides an overview of the research in virtual metrology (VM) and modelling of plasma etch processes in semiconductor manufacture, using an extensive sample of work from the public domain. The research is broadly divided into three sections in this review: standalone input/output modelling, plasma simulations, and virtual metrology modelling. The key differences between these approaches is depicted in Figure 4.1, and each is now discussed in turn:

1. **Standalone input/output models**: This section examines research modelling input/output relationships for plasma etch processes. An input/output relationship is the relationship between the process input variables that are normally specified as set points to plasma chambers (such as input gas flows or chamber pressure) and the process output variables of interest (etch rate, electron density, etc.). Input-output models are typically used to examine the process response to varying inputs during process optimisation, and to estimate process outputs for hypothetical input variable combinations not applied to the actual etch process. Typically, standalone input/output models are operated in an open-loop manner, using only the process input variable set points to estimate the process outputs.

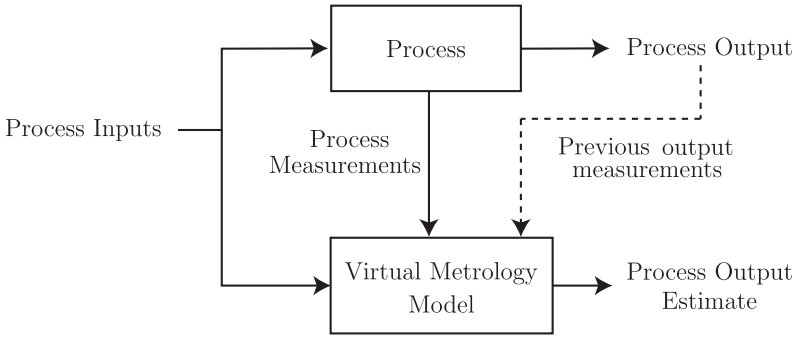
2. **Plasma simulation**: Plasma simulation research focusses on the development of accurate representations of the fundamental physical and chemical processes that occur in plasmas. Such simulations are used to generate estimates of the evolution of plasma variables in a plasma chamber for given conditions (chamber dimensions, gas, power, pressure etc.). Simulation results are typically spatially resolved and simulations are rarely computable in real-time due to the complexity




(a) Standalone input/output model




(b) Plasma simulations




(c) Virtual metrology model

FIGURE 4.1: Division of VM and modelling research. The literature review splits the published research into three sections (a) Standalone input/output models, (b) plasma simulation, and (c) virtual metrology models.

of the calculations involved. Plasma simulations are typically considered for simple plasma discharges and rarely completed for etch processes due to the complexities of the interactions at the wafer surface and the more complicated chemical make up of the plasma.

3. **Virtual metrology models**: The third section concentrates on research using models for plasma etch processes that estimate process output variables using data recorded in-situ *during* process runs, for example optical emission spectroscopy (OES) data. Virtual metrology (VM) models can be used to estimate inaccessible process variables during or after the process, reducing the requirement for regular real metrology. Estimates from VM models, in contrast to the standalone

input/output models that only use the process input set points as model inputs, take the actual state of the process into account during estimation because they rely on measurements generated by the etch process. VM models cannot be used to generate estimates of output variables without actually running the process. Previous measurements of the process output or state can also be used to aid the estimation effort.

Particular focus is given to VM models, as such research is most aligned to the contributions of this thesis. Examples of input-output research and plasma simulation research is included for background context information, and to ensure a rounded review of the plasma modelling literature is provided. To finish the review, a brief discussion on fab-wide VM schemes is presented.

## 4.1 Standalone input/output models

The development of input-output models for semiconductor etch can be subdivided into two main approaches, analytical modelling and empirical modelling. Research using each of these methodologies is addressed in turn.

### 4.1.1 Analytical models

Analytical models are based on first-principals knowledge of the physical, chemical, and/or electrical behaviour of plasma discharges. In this particular context, analytical models are defined as models that relate the process inputs to the process outputs by means of single or multiple analytical equations. The level of detail in the model equations can be customised for each specific application. Simplifications result in lower computational demands but are usually accompanied by reductions in the resolution and accuracy of the model estimates. Analytically derived equations can usually be solved relatively easily such that analytical models can be operated with modest computing power. A number of example applications of analytical modelling are now described.

An excellent review performed by Badgwell *et al.* [116] in 1995 addresses analytical modelling for a number of semiconductor manufacturing process, and contains a number of further examples of analytical models for the etch process.

### First-principals analytical models

An example of a purely analytical, kinetic model of the plasma etching process is provided in a 1982 work by Kushner [117] describing silicon etching in $C_nF_m/H_2$ and $C_nF_m/O_2$ plasmas. Kushner's model is one of the first kinetic models of etching processes, and considers electron-impact events, ion chemistry, neutral gas-phase reactions, diffusion, space-charge effects, adsorption, desorption and the etching process reactions using physical relationships. A limited set of chemical species and electron/ion reactions are included, and rate constants for the reactions are estimated or obtained from published literature. An analytical equation for the plasma etch rate is determined and good agreement with experimental data is achieved. Similar work by Lieberman [118] shows a self-consistent solution for the dynamics of a capacitive RF plasma sheath, taking into account ion collisions. Lieberman develops analytical expressions for ion current, sheath capacitance per unit area, ratio of the DC voltage to the peak sheath oscillating voltage, $2^{nd}$ and $3^{rd}$ voltage harmonic amplitudes, and the conductance per unit area for stochastic heating.

Analytical models for the evolution of etch profiles on wafer surfaces have been developed by some researchers. An example of such a model is seen in work by Berg *et al.* [119] where *curve evolution* is used to model surface development for simulated isotropic etch processes. Extensions to the model to cater for anisotropic etch are also suggested. Abdollahi-Alibeik *et al.* [120] simulate a chlorine trench etch process, taking special consideration of ion reflection from the sidewalls of the etched trenches, and the resulting microtrenching effects on the bottom. A microtrench is an indentation on the bottom of etched trenches, usually close to the trench sidewalls. Simulations were carried out using SPEEDIE (a Stanford etching and deposition profile simulator), using an analytical calculation for particle fluxes, etching and deposition and a Monte Carlo calculation to calculate the effect of the plasma sheath on the etching ions.

Abraham-Shrauner [121] uses first principle models to model the formation of etch profiles at the sides of open areas on the wafer surface (*half-trench* profiles) for a $CHF_3$ / $CF_4$ / Ar plasma in a magnetically-enhanced reactive ion etch (MERIE) reactor. Contrary to the work by Abdollahi-Alibeik [120], Abraham-Shrauner concludes that reflected ions from the sidewalls do not contribute significantly to microtrenching. Analytical equations are derived for the slope of the trench sidewall and the etch rates inside the trench.

Liu and Abraham-Shrauner [122] present a theoretical plasma etching model for contact holes etched in $SiO_2$ with $CF_4/CHF_3/Ar$ in an MERIE reactor. Analytical

expressions are derived for ion fluxes and etch rates in etched contact holes, and the evolution of the etch profile is then simulated using MATLAB ® software.

Models exist that use fluid-solid interaction principles to model the etch rate in plasma processing. For example, Bray and Rhinehart [123] use a mechanistic model, along with the Hougen and Watson method for modelling gas-solid reactions [124], to derive an expression for the etch rate of novolac-based polymers (often used as photoresist) in an $O_2$/Ar inductively coupled plasma (ICP). In [123], the model form is determined *a priori* and the parameters are estimated from experimental data. The model is based on the assumption that a limited number of etch sites (areas suitable for chemical reactions) exist on the wafer surface. The absorbtion, desorption, and reaction rates of etchant species are taken into account to arrive at an analytical expression for etch rate. Similar work is carried out by Gottscho *et al.* [125] using the incident fluxes of etchants, along with sticking coefficients, sputter yields and evaporation (desorption) rates to analyse etch rate. Scanning electron microscope (SEM) images and Langmuir probes are used to determine some important parameters for the model, allowing estimation of the profiles of etched structures in aluminium and oxide etch processes.

Abrokwah *et al.* [126] develop a physics-based model, the parameters of which are estimated from experimental data, that predicts etch rate in response to pattern density effects on the wafer surface. Because pattern density influences the consumption of etchant species, variations in pattern density across wafer surfaces lead to localised areas of differing etch rates (see Section 2.3.4). Abrokwah *et al.* [126] incorporate feature-dependent and pattern-density effects into a combined analytical model resulting in a 4.4% RMSE for etch rate estimates.

#### Equivalent circuit models

In the analysis of electrical signals such as voltage, current, phase, and any harmonics of these recorded from a plasma chamber, many researchers model plasmas using an electrical circuit that represents the electrical load presented by the plasma, otherwise known as an *equivalent circuit model*. The degree of detail included in the circuit model is dependent on the application and on the accuracy required of the model.

Early work by Van Roosmalen [127] details a method of estimating fundamental plasma variables, such as ion bombardment energy, electron density, and ion flux density, using the plasma impedance value that is calculated from the settings of the matching network components. A simple equivalent circuit model of the plasma is used, where a resistor represents the bulk plasma and capacitors represent the sheaths. However, a

series of simplifications are made to arrive at the results. A similar equivalent circuit model is used by Bushman *et al.* [128] where, again, the plasma sheaths are approximated as capacitances and the bulk plasma is represented using a resistor. The model is used to estimate power, sheath thickness, and sheath capacitances.

An example of an equivalent circuit model with some further detail is found in work by Kanoh *et al.* [37] where the equivalent circuit for a parallel-plate, capacitively-coupled, plasma etching system is developed to determine the plasma impedance. Again, the plasma is approximated by a conductor, and the sheaths are approximated by capacitors. The equivalent circuit elements for each of the etch chamber parts are shown in Figure 4.2. Using a series of simplifications outlined in [37], an expression for plasma impedance is obtained and the model estimates of impedance are found to agree with experiment.

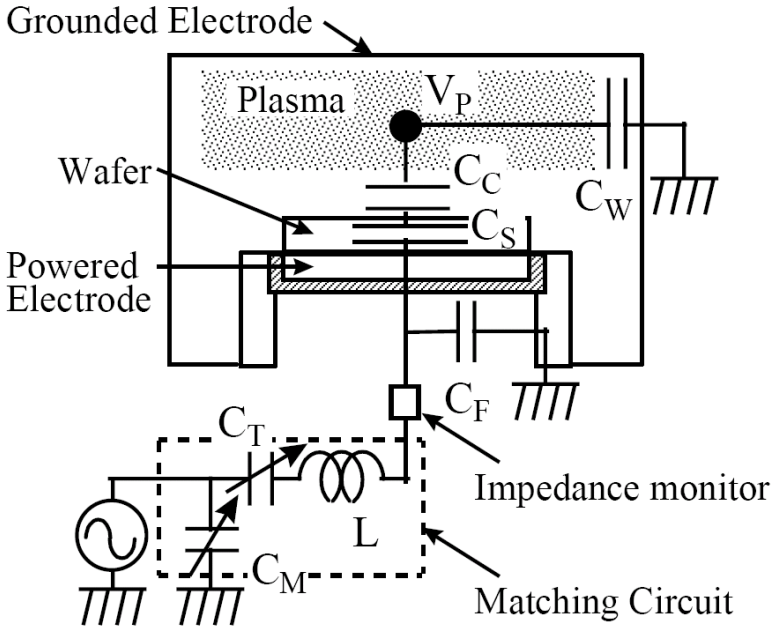



FIGURE 4.2: Schematic of plasma reactor, with equivalent circuit parameters overlaid [37]. $V_p$ is the plasma potential, $C_F$, $C_S$, $C_C$ and $C_W$ are the floating capacitances of the powered electrode, wafer, powered-electrode sheath and wall sheath respectively. $C_M$, $C_T$ and $L$ are the capacitance values of capacitors and the inductance in the matching network.

Colpo *et al.* [129] determine an equivalent circuit for an ICP source, improving upon previously-reported more basic models that rely on a transformer model to represent the power coupling between the plasma and powered RF antenna. The researchers simulate the plasma physically using a conductive dummy load made of an aluminium cylinder having the same volume as the plasma and use a Bode analysis to identify the equivalent circuit. Two series LRC circuits are found to be associated with the impedance profiles

found. Accurate estimates (15% error on reactance values over a wide operating range) of plasma resistance, inductance, and conductivity are illustrated with the series LRC circuit model, and the potential for non-invasive monitoring of other plasma variables is highlighted.

It should be noted that equivalent circuits are not only used for characterisation of the plasma and chamber themselves but are also used to represent stray capacitances and impedances that can be troublesome in plasma processing equipment using high frequency RF power. The parasitic impedances of plasma chambers increase with frequency [128], so at the high frequencies in use during plasma processing (not to mention the harmonics of these), relatively small stray capacitances and inductances become influential to the process. Sobolewski [130] uses equivalent circuit models to correct RF measurements to their correct values in his research [131, 132]. Bushman *et al.* [128] also reports on the characterisation of losses due to parasitic impedances, working on the losses sustained from transmission lines between sensors and the electrodes.

### 4.1.2 Empirical models

The word empirical denotes information gained by means of observation, experience, or experiment. Empirical data are data that are produced by an experiment or observation, and empirical models refer to models based upon the analysis of these data, rather than on knowledge of the underlying physical or chemical properties of the process. Because of the inherent complexity of the plasma etch process, empirical techniques are often used for model building, rather than attempting to describe all of the chemical, electrical, and magnetic interactions between the constituent gases, chamber components, and wafer surface.

Empirical input-output models can be created using data collected during production. However, for semiconductor manufacturing, variations in the process inputs are typically minimal during production activities, and so such models typically explore only a small range of the input space.

Hence, it is more typical that data for empirical input/output models are gathered from designed experiments, or factorial experiments [133]. During factorial experiments, the process input variables are varied over a number of discrete levels to determine their effects on the system output. Factorial experiments are designed using techniques cumulatively known as *design of experiment* (DOE) techniques, whereby the principal operating space of the process is explored and the effects of each system input variable

on the output variables of interest are captured. Process output data are recorded for set values of input variables and these data are then used, for example,

- to determine which input variables have most impact on output variables,
- to determine the optimal values of the input variables for a desired process output,
- to determine the optimal values of the input variables so that the effects of external disturbances are minimised, or
- to determine an empirical model of the system so that future output estimates can be made for different combinations of the input variables.

In the vast majority of situations, two discrete levels are chosen for each input variable, and it is assumed that the relationships between the input and output variables are linear. Such experiments are denoted as $2^p$ factorial designs, where $p$ is the number of input variables, and they require $2^p$ experimental runs to test all combinations of input variable set points levels. As the number of input variables and input variable set points increases, the number of experiments required rises exponentially, and it can become logistically infeasible to perform large factorial experiments in cases where experiments are expensive in terms of time or money. In cases where it is impossible to explore all input combinations, *fractional-factorial* experiments are used, where some of the input variable combinations are omitted, but, as a result, complete information about some of the high-order effects between the input and output variables is not captured [133]. However, typically, high-order effects are smaller than the main effects [28], and hence, fractional-factorial experiments provide an affordable and practical alternative to full-factorial experiments where experimentation is expensive or time consuming. A half-fractional experiment with $p$ factors requires $2^{p-1}$ experiments, and is denoted a $2^{p-1}$ fractional-factorial experiment.

In cases where higher-order models are required to fit a process response, more complex experimental designs, such as a central composite design, can be used [133]. A central composite experiment is useful for building a second-order (quadratic) model for the process output variable without the requirement for a complete three-level factorial experiment. Generally, central composite designs consist of a $2^p$ factorial experiment, with $2p$ axial or star runs, and a number of center runs as depicted in Figure 4.3.

For creation of plasma etch rate models, Klimecky *et al.* [134] attempt to maximise the number of factorial experimental points explored during one experimental run using in-situ etch rate measurements. The aim of this work is to reduce the cost of running factorial experiments by reducing the number of wafers required to explore the operating

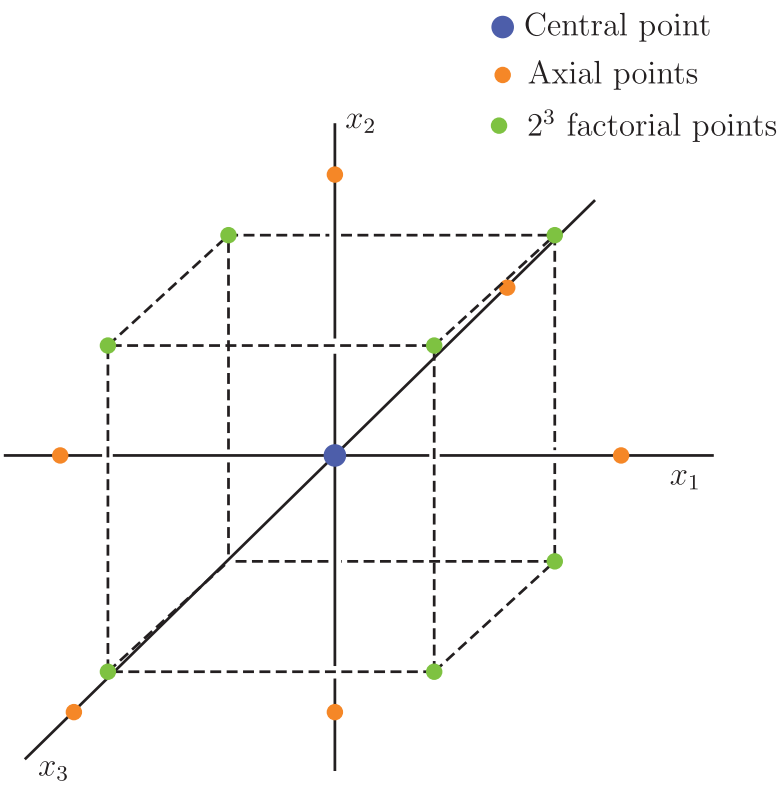



FIGURE 4.3: Central composite design for $p = 3$ [133]. Central and axial points are labelled.

space of the process. The in-situ measurements are performed using laser reflectometry along with an extended Kalman filter technique developed by Vincent *et al.* [135]. Typically, in-situ measurements are not used, and one experimental wafer is used for each experimental point explored.

The two most common approaches for creating input/output models based upon factorial experiment data are linear regression and non-linear artificial neural networks.

### Linear empirical modelling

Research by May *et al.* [136] in 1991 is an example of the type of factorial experiments required to model a plasma etch process. In [136], a $2^{6-1}$ fractional-factorial experiment is performed using RF power, pressure, electrode spacing, $CCl_4$ flow, He flow, and $O_2$ flow as input variables and examining output variables etch rate, selectivity, and uniformity. Nonlinearities are found in nearly all responses, and linear regression models are produced for each of the outputs using the data from a central-composite circumscribed Box-Wilson designed experiment. The models are then used to successfully determine an optimised process recipe.

Linear regression analysis can be used with data from designed experiments to produce a *response surface* of the process being examined. *Response surface methodology* (RSM) is a general technique used in the empirical study of process input/output relationships [28]. A response surface is usually presented in the form of a contour plot showing how output variables vary in response to input variable changes, as exemplified in Figure 4.4. Tan *et al.* [137] employ designed experiments, varying RF power, chamber pressure, and $SF_6$ flow rate to characterise an $SF_6$/Ar etch process and then stepwise regression to select variables related to etch rate, uniformity and anisotropy. McLaughlin *et al.* [138, 139] vary RF power, pressure, gas composition, and flow rates in a central-composite experimental design to collect data for models estimating etch rate, selectivity, uniformity, and anisotropy for $SiO_2$ etching in $CF_4$-based plasmas. The effects of the input variables on process variables such as species emissions and DC bias are also examined. Mozumder and Barna [140] use a $2^3$ factorial experiment to characterise the uniformity response of a silicon nitride etch process using chamber pressure, RF power, and $CHF_3$ flow. A 31-point experimental design (a "D-optimal" design) is also used to develop a process model for etch rate, line width loss, and uniformity in response to five process inputs comprising the RF power, chamber pressure, and three gas flow rates. After experimentation, the model is successfully used to develop a run-to-run control system whereby recipes are specified in terms of etch performance variables. Work by Hong *et al.* [141] and Himmel and May [82] include a number of further references to work employing RSM techniques with plasma etch processes.

### Non-linear empirical models

The non-linear nature of plasma processes means that linear modelling techniques can fail to capture process variation accurately. Due to their non-linear modelling capabilities, artificial neural networks (ANNs) have been adopted by many researchers for etch process modelling, and a number of authors have shown that ANN models outperform linear statistical techniques when creating empirical models from DOE data [82, 142–144]. Himmel and May [82] directly compare ANNs to quadratic-based RSM techniques when modelling etch rate, selectivity, and uniformity for polysilicon etch in $CCl_4/He/O_2$ plasma. The ANN process models exhibit significantly superior performance, with improvements in etch rate estimation accuracy of 38% ($\sim 5\%$ absolute error reported) and selectivity estimation accuracy of over 62%.

A significant amount of research is performed where ANN models are developed with data from designed experiments, followed by the use of these models to investigate the effects of input variables on the process outputs over the experimental range. Feed-forward multi-layer perceptron networks trained using the backpropagation algorithm

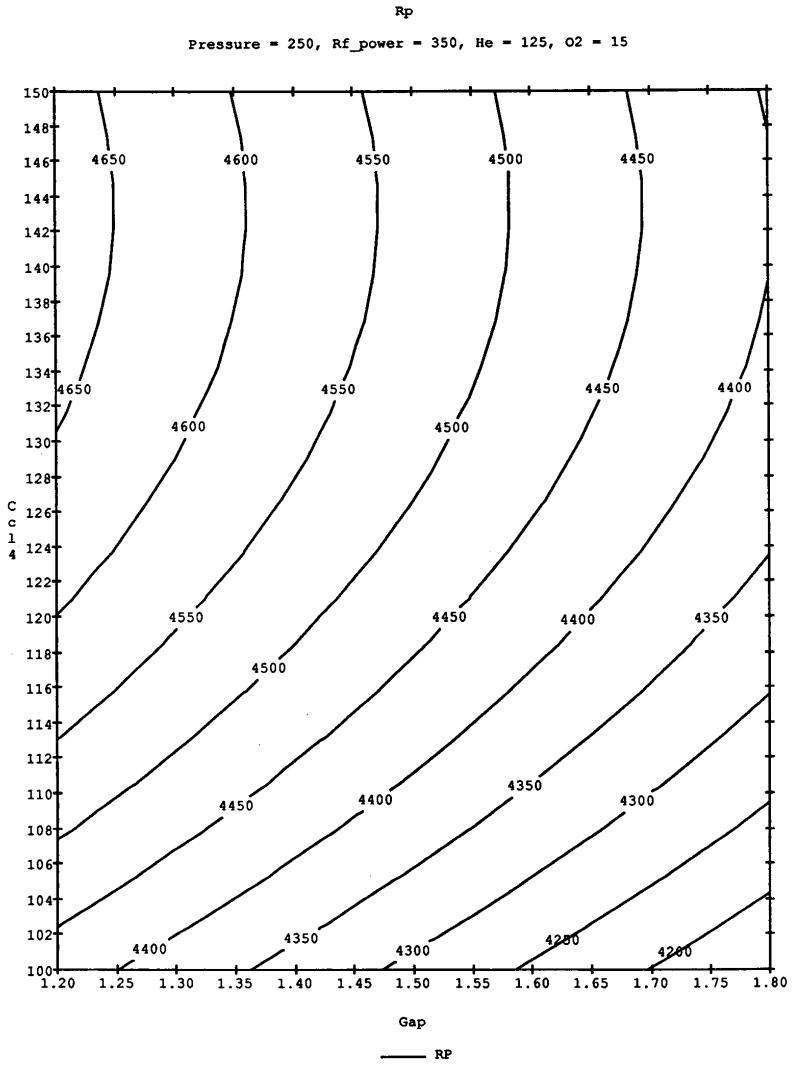



FIGURE 4.4: Contour plot of polysilicon etch rate versus $CCl_4$ flow and electrode spacing [136] produced using linear regression and data from a designed experiment.

were reported as the most generally applicable and most popular networks employed in semiconductor manufacturing in 1994 [75] and are still popular in the literature. For example, Rietman *et al.* [144] use an ANN trained with the backpropagation algorithm on production data from a plasma gate etch process to perform a sensitivity analysis of the effects of various input variables on etch rate, identifying influential variables and influential process steps for etch rate. Kim *et al.* [86, 145–148] from Sejong University have conducted a number of studies where ANNs are trained on DOE data using the backpropagation algorithm and are employed for a number of purposes:

- Modelling of etch rate, selectivity, anisotropy and critical dimension bias in a MERIE process using $Cl_2$, $BCl_3$, $N_2$ flow rates, pressure, RF power, and magnetic field strength as input variables [145] in a $2^6$ factorial design. Increases in accuracy

of up to 36% for anisotropy estimates are achieved through optimisation of the ANN activation function parameters.

- Modelling of oxide film etch rate, uniformity, and etch profile angle in a MERIE chamber with $CHF_3/CF_4$ plasma using RF power, pressure and gas flow rates as input variables [86, 146, 147] (errors reported for etch rate of between 3 – 6 %).
- Modelling and study of the effects of plasma density on the etch rate of silicon carbide in a $C_2F_8$ ICP etch system using RF bias power, pressure, $O_2$ fraction, and gap spacing as input variables in a $2^5$ factorial experiment [148].

More recently, Xia *et al.* [90] compared ANNs trained using the BP algorithm with ANNs trained using the BFGS algorithm for modelling plasma etch rate of silicon carbide in an ECR chamber. The microwave power, DC bias, process pressure and $O_2$ fraction are used as input variables in a 25-point Box-Wilson central composite designed experiment to gather data for modelling. The ANNs trained using BFGS required fewer hidden neurons and less iterations to obtain the same training results as those trained using BP. Further investigations using the ANN model to gain insight into the silicon carbide etch process and to develop optimised etching recipes are presented in [149].

ANN structures apart from multi-layer perceptron ANNs are examined by different researchers. Kim and Kim [150] employ a generalised regression neural network (GRNN) to model profile roughness in a $CHF_3/CF_4$ silicon etch process. A GRNN is an ANN with four layers: an input layer, a pattern layer, a summation layer, and an output layer. GRNNs are similar to GPR models in that the network output for a specific test input is created using a weighted sum of the training data outputs, depending on the proximity in the input space of the test input to the training inputs. Each summation layer neuron defines a *spread*, usually a Gaussian function, which determines how the training outputs are combined. There is one pattern neuron per training point, and hence GRNN networks suffer from the curse of dimensionality. Kim and Kim [150] collect data using a $2^3$ factorial experiment and use a genetic algorithm (GA) to optimise the spread values in the summation layer to enhance the predictive capability of the model, achieving a 54.6% improvement in RMSE for estimation of profile roughness.

Kim *et al.* [151] use a GRNN model to estimate the discrepancy in the sidewall bottom etch rate with respect to the center etch rate in etched trenches. Again, optimisation of the GRNN spread values using a GA decreased the RMSE of the models by 32.6%. Data were collected using a $2^{4-1}$ fractional-factorial experiment, varying RF power, pressure, and two gas flow rates as input variables. Kim and Lee [142] compare the GA-GRNN modelling technique to statistical regression models containing square

and input variable interaction terms for estimation of silicon oxynitride etch rate, varying RF power, bias power, chamber pressure, and $C_2F_6$ flow rate in a $2^4$ full-factorial experiment. The GA-GRNN model yields better predictions of etch rate, demonstrating a 52% reduction in RMSE.

Kim and Park [143] show that radial-basis function networks (RBFNs) produce superior prediction results over statistical regression models for prediction of etch rate and surface roughness of silicon carbide films in an $NF_3$ ICP etch system. Data are gathered using a $2^4$ factorial experiment, varying electrode RF power, wafer bias power, chamber pressure, and $NF_3$ gas concentration. Etch rate estimation accuracy improvements of 40% are reported for the RBFN models over the statistical models. RBFNs are similar to GRNNs but contain only three layers, the input layer, the pattern layer, and the output layer. Each unit in the pattern layer represents a receptive field, which is an area in the input space that activates the local pattern units. The pattern centers and widths and the output weights are optimised during training to obtain the best predictions.

Kim *et al.* [152] use a polynomial neural network to estimate etch rate and wafer DC bias using source power, bias power, pressure, $O_2$ fraction, and electrode spacing as input variables in a $2^5$ factorial experiment. Compared to ANNs trained using BP and a statistical regression model, the polynomial neural networks demonstrate improvements in etch rate estimation accuracy of 34.3% and 55.7% respectively, requiring less computation during training but proving difficult to optimise correctly.

## 4.2 Plasma simulation

Simulation of plasma processes using physics-based knowledge and validated mathematical relationships can yield accurate and complete information on discharge processes, useful for process optimisation, troubleshooting, and chamber design. The research considered in this section provides detailed spatial and time resolved information on plasma processes. Such simulation typically considers as much detail of the fundamental physical and chemical interactions between the components of the plasma system as possible. The complexity of the etch process means that, correspondingly, accurate simulations are extremely complex, and a balance between computational complexity and simulation accuracy must be found depending on the application at hand. Only a basic overview of plasma simulation techniques is provided here as such research is quite different from the core subject matter of this thesis. Representative research from the literature is referenced.

Broadly, plasma simulation can be divided into three sections that are now examined in turn: particle-in-cell (PIC), fluid, and hybrid models. A brief review of the main properties of each simulation type is presented. The choice of model used by researchers or industrial practitioners: PIC, fluid, or hybrid, depends on the application in question and the level of detail required in the results. The correct simulation type must be chosen for the best accuracy [153], and a thorough understanding of the limitations of each model type, along with the assumptions and simplifications made within, is required for this choice. In general, physics-based simulations of plasma-based manufacturing processes are not suitable for real-time operation due to limitations in computational power. However, the development of plasma simulations is an active area of research, and future advances in both simulation practices and computing power may someday enable real-time computation of such models.

### 4.2.1 Particle-in-cell models

Particle-in-cell simulations provide a self-consistent, fully kinetic representation of general plasmas. PIC models take advantage of the collective behaviour of charged particles in plasma to model the kinetics of various species by simulating the behaviour of a reduced number of computer particles (or *super-particles*) [153]. The kinetics of each species in the plasma are simulated by solving fundamental physics equations (Maxwell's and Newton-Lorentz equations). Typically, super-particles representing charged particles such as electrons and ions are simulated, while neutral species are assumed to be uniformly distributed throughout the plasma. The first PIC models, containing $10^3$-$10^4$ super-particles, appeared in the 1950's and included solutions of Coulomb's law with calculations of particle trajectories (requiring $N^2$ calculations per $N$ particles). Modern PIC simulations contain between $10^6$ and $10^8$ super-particles and typically include Monte Carlo collision (MCC) schemes, accurate Coulomb collision modelling, secondary electron emission effects, details of the external circuit impedances, noise control, and radiation transport effects. A thorough review of the art of PIC simulations, and a study of recent advances is provided in the work of Verboncoeur [154].

The actual implementation of a PIC simulation requires the following steps [153]:

1. Initially distribute the super particles in the simulation domain and velocity space.
2. Solve Maxwell's equations to determine electric and magnetic fields at grid points in the simulation domain.
3. At each time step, update the position and velocity of the particles by solving the Newton-Lorentz equations.

4. Incorporate collisions using the MCC scheme.

5. Incorporate the interactions between the surfaces and particles - reflection, absorption and emission.

6. Return to step number 2.

PIC models provide an accurate representation of species behaviour in a plasma, and are often used as benchmarks for testing the results of more approximate models. However, PIC-MCC simulations usually require a large amount of computing power, even more so for plasmas with complex chemistry. The number of super-particles must be kept to a sufficient level to avoid statistical under representation, numerical heating, and noise [153]. Improved models for reactive gas mixtures, multiple excited states, and ionisation pathways are required for increased accuracy. However, such improvements will further increase the computational requirements of the models and it is estimated that a multi-petaflop computer would be required for full simulation of a plasma fusion reactor [154]. Such demanding computational requirements disqualify the use of PIC-type models in real-time applications.

A one-dimensional PIC-MCC model is used by Krimke and Urbassek [155] to simulate an Argon discharge in an ECR chamber and is compared and shown to agree with analytical theory and experimental data from previous work. Osano and Ono [156] use an atomic-scale cellular model to simulate etch profile evolution during chlorine-based etch processes, using a Monte Carlo calculation to simulate the transport of ions and neutrals and placing great emphasis on plasma-surface interactions during etching. Lee *et al.* [157] describe one dimensional PIC-MCC models of capacitively coupled oxygen-argon plasmas, and find that the simulation data agree well with experimental data. The authors also compared the model outputs to the results produced by fluid models (discussed in the Section 4.2.2).

### 4.2.2 Fluid models

Fluid models describe plasmas based on the density, mean velocity, and mean energy of constituent species [153]. These values are obtained by solving the continuity, flux, and energy equations for each species in the plasma. The electrons, ions, and neutrals are essentially treated as interpenetrating fluids [158]. The velocity moment of the Boltzmann equation provides the fluid equations, and Maxwell's equations are coupled with the fluid equations to obtain a self-consistent representation of the electric and magnetic fields. The fluid equations are solved numerically after converting the set of coupled

differential equations that make up the fluid model into a set of finite difference equations. Typically, a particular velocity distribution for each species in the plasma must be assumed so that a closed set of fluid equations can be obtained; often, Maxwellian energy distributions are assumed, or it is assumed that charged particles are in local equilibrium with the electric field. However, both assumptions are approximate and are not accurate for all circumstances [153]. The inherent approximations in the derivation of fluid models limit their use to high pressure plasmas, hence restricting their use in plasma processing applications. The main advantage of fluid models over PIC simulations is a lower computational demand. Additionally, complicated chemistries can be included, catering for numerous reactions.

Kim *et al.* [153] demonstrate that PIC modelling outperforms fluid modelling for capacitively coupled plasmas (CCPs), due to non-local electron kinetics that cannot be captured by fluid models assuming specific velocity distributions, playing a key role in the discharge. Lee *et al.* [157] compare fluid and PIC-MCC simulation techniques for both capacitively and inductively coupled plasmas (ICPs). It is found that a one-dimensional PIC simulation agrees better with experimental data than two-dimensional fluid simulations for a CCP, again due to the importance of non-local kinetics in these plasmas.

Wilcoxin and Manousiouthakis [158] use fluid models to identify the steady state gains between manipulated variables and the plasma properties that affect plasma etching. A simplified "argon-like" DC plasma is simulated, and the effects of pressure and applied voltage on ion flux energy are studied. Goglides *et al.* [159] examine fluid modes for a $CF_4$ discharge, and in particular investigate the role of $CF_x$ radicals and negative ions in etching processes.

Meeks and Won Shon [160] use a continuous-flow, well-stirred tank reactor (CSTR) approximation to model an ICP plasma reactor. CSTR approximations are commonly used for modelling of chemical reaction mechanisms, determining the spatially and time-averaged species compositions in the reactor through solution of species, mass, and electron-energy balance equations. Surface kinetics, deposition, and etch rates are also modelled in [160]. The model has small computational demands and its results are found to agree satisfactorily with experimental values.

A more recent example is found in work by Sfikas *et al.* [161], where a fluid model is created and compared to the PIC model described by Krimke and Urbassek [155]. The electron density, temperature, ion flux, and electrostatic field distributions are examined. Good agreement was found with the accurate PIC-MCC simulations for all but ion flux

density. This difference was attributed to the electrostatic field calculations close to the chamber walls.

Sobolewski [131] uses a numerical model for the sheaths in a plasma discharge. The one-dimensional model uses fluid equations for ion momentum conservation and is applied to an Argon discharge in an ICP gaseous electronics conference reference cell [162]. The model is valid for all frequencies, assumes a Maxwell-Boltzmann electron velocity distribution and neglects secondary electron emission.

### 4.2.3 Hybrid PIC/fluid models

Hybrid models aim to combine the individual strengths of fluid and PIC models, ideally incorporating the fast computation speed of fluid simulations with the accurate particle kinetics of PIC simulations.

Variations exist in the hybrid schemes that are used. For example, Sommerer and Kushner [163] model ions as a fluid while electrons are treated using a PIC scheme. Conversely, Porteous and Graves [164] simulate ion energy distributions on wafer surfaces using a PIC scheme, and consider electrons as a Maxwellion fluid. Vasenkov and Kushner [165] use a particle module for electron-dependent properties and a fluid module to determine kinetics of heavy particles.

Yang and Kushner [166, 167] use a two-dimensional fluid hydrodynamics simulation that incorporates a Monte Carlo simulation for secondary electron emission to investigate plasma behaviour in single frequency and dual-frequency MERIE reactors. Hoekstra and Kushner [168] use a hybrid model to model three-dimensional etch features in high-density plasma chambers with a Monte Carlo scheme for modelling of feature profiles and a hybrid plasma equipment model to determine ion and neutral angular energy distributions. The effect of the model parameters on microtrenching and sidewall reflection are investigated.

The review by Kim *et al.* [153] examines the application of hybrid models to ICP chambers, and a number of further examples of both fluid and hybrid models are referenced in the review by Badgwell *et al.* [116].

## 4.3 Virtual metrology models

VM models are models that use in-situ data collected during process operation to infer information about process variables of interest that are inaccessible or difficult to measure at the time of processing.

During the standalone input/output process modelling described in Section 4.1.1, typically a small number of model input variables (approximately 4/5) are used, and these input variables are predetermined before modelling commences. However, during VM model development, large numbers of candidate variables can be available as VM model input variables, depending on the quantity and type of in-situ sensors that are installed on the process tool. Vast numbers of samples of each variable can also be collected. As such, much of the VM research undertaken shares a number of commonalities in terms of data preprocessing steps.

Typically, the first step in preprocessing for wafer or lot-level VM models is the extraction of summary statistics from time series traces of variables recorded during each individual process step, each wafer, or each lot of wafers. The summary statistics may include statistics such as mean, variance, kurtosis, range, etc. The summarisation step is taken to reduce the volume of data prior to VM modelling, hence easing computation and storage demands.

Feature extraction and feature selection are further preprocessing steps typically undertaken prior to VM modelling.

#### Feature extraction

Feature extraction or data reduction techniques are methods that transform available variables such that the information contained within them is summarised by a reduced set of new variables. Because process databases often consist of large numbers of intercorrelated variables, the effective dimension of the space in which the variables move can be small compared to the original number of variables. After data reduction, analysis is performed in the reduced subspace of the new derived variables.

In general, latent variable methods such as PCA and PLS are popular in VM literature for data reduction. PCA is widely employed to reduce the dimensionality of OES data because many of the wavelength signals are correlated in time, and hence can be effectively represented by relatively few principal components. While PCA and PLS are linear techniques, non-linear techniques, such as auto-encoder neural networks

(AENNs) or kernel PCA, can also be employed. Kourti [59] provides an overview of different latent variable methods, along with their recursive counterparts, emphasising that process knowledge is a must in any such application, since process operators may need to select appropriate weights and transformations for available variables.

### Feature selection

Feature selection or variable selection techniques are methods that aim to find a subset of the available variables that are most relevant to the process output and hence should be included as VM model input variables. Performing feature selection prior to modelling reduces problems with the curse of dimensionality and may speed up the model training procedure.

The most popular feature selection technique is stepwise regression, which is a model performance based technique that adds the best variable or removes the worst variable from a LSR model on an iterative basis (see Section 3.3). Genetic algorithms [169] (GAs) are employed by some researchers for feature selection, where the VM model input selection is evolved over successive generations using the model error performance as a fitness function. Feature selection can also be achieved by ranking the candidate variables by some metric (for example, linear correlation coefficient), and discarding those that do not achieve an adequate score.

Combinations of feature extraction and feature selection can be employed. For example, PCA can be used to extract a number of principal component score variables from a data set, and feature selection techniques such as stepwise regression can then be used to select a subset of the principal components most related to the process output variable.

### Division of VM literature

The research on VM modelling is divided into four categories, depending on the aim of the particular VM implementation:

1. Endpoint detection. Endpoint detection concerns the estimation of the time of process step termination as an etch process proceeds.

2. Fault detection and classification. Fault detection and classification concerns the analysis of process data to aid the timely detection and diagnosis of faulty process operation.

3. Plasma variable estimation. Plasma variables include fundamental properties of the process plasma such as electron temperature, species concentrations, ion densities, etc.

4. Etch process variable estimation. Etch process variables are variables that relate to the etching performance of the chamber, such as etch rate, uniformity, etch profile, etc.

### 4.3.1 Endpoint detection

Endpoint time in a plasma etch process is the time at which a layer on the surface of a wafer is completely removed by the etching plasma. Typically, the etching layer is not completely removed at all locations on the wafer surface at the same time, and it is required to over-etch to some degree to ensure that all of the required material is removed at all locations. Excessive over-etching may remove the film underneath the target layer, possibly causing device failure and yield reduction [170]. Therefore, in many processes it is critical to detect the endpoint time correctly and cease etching at the appropriate moment. Endpoint detection is generally achieved by monitoring process data during etch processes and analysing the data in real time to find *change-points* or specific patterns that are known to represent endpoint events. Change-points are points in time where the system transitions between different behaviours. Endpoint detection falls under the umbrella definition of VM modelling in this thesis because it uses in-situ measurements from the etch chamber to estimate the optimal endpoint, which is not measured directly.

#### Optical-based endpoint detection

OES (See Section 2.5.1) is a popular technique for diagnosis and monitoring of plasma processes. The relatively low cost and the portability of OES devices ensure that practitioners favour their use in many VM applications. The data recorded by optical emission spectrometers is rich in information about the etching plasma state and OES data are by far the most extensively used measurements for plasma etch endpoint detection. By tracking the optical emission of chemical species important to the etching process over time, changes in the species concentrations that occur at endpoint can be detected. OES data exhibit relatively fast response times to changes in plasma chemistry, and settings such as integration times can be changed to optimise this response time along with the signal-to-noise ratio [171].

Processes with relatively large open areas (> 10%) for etching typically produce clear endpoint signals in plasma optical emissions. Endpoint for such processes can sometimes be detected by monitoring single wavelengths from the optical emission spectrum. For example, Dolins *et al.* [172] diagnose problems and detect endpoint for a plasma etch process by monitoring variations in a single wavelength and comparing the variations to known "normal" traces. Litvak [173] demonstrates a similar technique using monochromators with $SiO_2$ and polysilicon etch processes (using RIE, HDP, and chemical downstream chambers) and use a number of signal conditioning techniques to increase the detection accuracy, successfully detecting endpoint for wafers with open areas as low as 0.25 %. A ratio of different channels is used to cancel background noise effects in the OES data. However, the use of single-wavelength endpoint detection techniques is not typical for processes with such a low open area for etching.

With increased miniaturisation of semiconductor devices, the critical dimensions of devices are shrinking, the open areas on wafers are diminishing, and more stringent control of plasma etch endpoint is required. Endpoint detection using single wavelength techniques for modern plasma etch process presents several difficulties [174]:

1. Plasma etchant gases are often complex molecular compounds. Molecular optical emission is generally more complex than atomic emission, presenting a continuous spectrum due to the many different excitation and de-excitation states possible. A continuous spectrum makes the identification of single wavelengths containing endpoint information difficult.

2. Modern processes using high density plasmas often yield different spectra than older processes due to different excitation processes in the plasma, and hence experience and knowledge from older processes cannot be easily transferred to newer product lines.

3. Etching of the mask materials coating the wafer corrupts the optical emission from the process and adds strong background noise to OES data.

4. As the open area percentage of wafers gets smaller, the signal-to-noise ratio of the endpoint signals produced during the transition from one layer to another decreases significantly, preventing reliable endpoint detection.

As a result of these challenges, many authors use multivariate techniques examining the full optical spectrum recorded from the plasma to detect endpoint. As previously discussed at the start of Section 4.3, PCA is commonly used to reduce the dimensionality of OES data. For example, White *et al.* [62] use PCA to compress the information

contained in OES data from a high-density plasma, oxide etch system and then employ a modified $T^2$ statistic to detect the change in behaviour that corresponds to endpoint. An exponentially weighted moving average (EWMA) estimator is used to update the PCA model means for normalisation of new wafer data and a $Q$-statistic is monitored to identify process changes that require a re-computation of the PCA model. However, the authors struggled to implement the endpoint system for wafers with open areas less than 10%. Yue *et al.* [174] use PCA as a variable selection technique to select important wavelengths from an array of three spectrometers operating over the UV, visible, and IR spectral ranges. Generally, the second and third principal components were found to contain the endpoint information, with the first capturing a linear trend in the OES data. Important wavelengths are selected using windowed-PCA analysis of the OES data with a thresholding technique on the principal component loadings. SEM micrographs are used to verify that endpoint is detected correctly.

Han *et al.* [175] successfully demonstrate endpoint detection using the principal components of OES data for a $SiO_2$ etch process in $CF_4/CHF_3/O_2/Ar$ plasma, using wafers with as low as 0.4 % open area. The endpoint signal could not be detected using single wavelength techniques. To increase the speed of the algorithm for execution in real time, the OES data is not normalised before the calculation of the principal components. The sensitivity of the technique is further increased in [170] using a ratio of different principal components. Further work by Han *et al.* in [176] uses PCA combined with support vector machines (SVMs) to detect endpoint in the same process, demonstrating considerable improvement over the basic PCA-based endpoint system.

Rangan *et al.* [177] suggest a method whereby the OES principal components recorded during an aluminium etch process (using $Cl_2/BCl_3$ plasma) are modelled as outputs of a two-state linear system. Jump linear filtering is used to estimate the system states using real-time OES data, and significant shifts in the system states are found to be indicative of etch endpoint.

Ragnoli *et al.* [178] introduce non-negative matrix factorisation (NMF) as an alternative technique for endpoint detection in OES data. NMF operates similarly to PCA in that it determines a matrix factorisation for an input data matrix. The NMF calculation can be constrained to produce sparse loading vectors, which is useful to quickly determine which wavelengths are most important in the etch process. The NMF components are shown to capture the same endpoint feature as found using PCA.

### RF-based endpoint detection

After OES data, electrical data are the second most widely used for endpoint detection. Early research using electrical measurements for endpoint detection was completed by Ukai and Hanazawa [179] in 1979, where the plasma impedance is shown to change appreciably at endpoint in an etch process. Fortunato [180] in 1987 focusses on the change in reflected power during etch to detect endpoint, but notes that the technique can only work for processes where a discontinuity in the plasma impedance occurs at the end of the etching process.

Later, Kanoh *et al.* [37] investigate the relationship between the change in plasma impedance at endpoint, the applied RF power, and the open area ratio of a $SiO_2$ wafer, etching with $CHF_3$. It is found that the change in impedance increases with increased open area and increased RF power. The use of digital filters where possible is proposed to filter out noise and increase the SNR of the RF signals when small open area wafers are being etched. An equivalent circuit model (see Section 4.1.1) of the plasma is used to determine the variables most affecting impedance. Rietman *et al.* [181] demonstrate the use a Fourier series created using the magnitudes of both OES and RF signals to form characteristic "blob" shapes that can be used by operators to identify endpoint times in an intuitive way. Koh *et al.* [182] perform PCA on RF harmonic signals from a plasma process as a photoresist layer is etched, demonstrating how the principal component scores change substantially when endpoint is reached and the underlying layer is exposed.

Patel *et al.* [183] use the RF voltage and current on each electrode, the phase angle between these two signals, and the induced DC bias on the electrodes to determine the endpoint for $SiO_2$, $Si_3$, and Al etch processes in diode, triode, and magnetic multipole enhanced triode reactor configurations. Triode reactors are RIE-type reactors where both electrodes are powered and the chamber walls are grounded. A magnetic bucket is added to one powered electrode in the magnetic-multipole enhanced triode configuration. Because the endpoint detection is based on a transition from one steady-state plasma condition to another, the first derivative of the signal recorded is used to find the change point.

Bose *et al.* [184] investigate the use of plasma RF variables for endpoint detection but conclude that optical measurements are usually more sensitive for endpoint detection in the $SF_6$ polysilicon etch process investigated. Some enhancements in sensitivity are achieved using ratios of RF variables. Maynard *et al.* [185] train an ANN classifier to identify endpoint situations from RF signals including reflected powers, matchbox capacitor positions, and DC bias, from a process etching TiN. The ANN is operated

as a binary classifier to determine whether the process has reached endpoint or not at each sample. The classifier is trained using data that is manually classified by a human operator using ellipsometry. The authors find the method to work only when the electrical signals change significantly at endpoint, and the ANN is found to be as accurate as the operator used to gather the training data.

Law *et al.* [186] use a remote-coupled dual-directional coupler to collect information on the harmonics of the fundamental applied RF frequency for endpoint detection. It is found that the harmonic containing endpoint information can be predicted through an analysis of the frequency response of the chamber prior to processing. Successful endpoint detection for two photoresist stripping processes and a cleaning process are demonstrated. The relative change in RF signals for cleaning processes is generally found to be larger than for etch processes due to the increased effective etch open area comprising the chamber walls. In a similar study, Bonner and Clark [187] demonstrate the use of RF-based sensors to optimise a plasma-enhanced CVD chamber cleaning procedure, arguing that RF-based sensors offer superior performance to OES-based endpoint detection techniques.

Dewan *et al.* [36] use a commercial plasma impedance monitor, the Scientific Systems Smart PIM unit, to monitor the fundamental and first four harmonic components of the RF power signals from a $SiO_2$ CCP etch process using $SF_6$ plasma. The phase angle of the first harmonic is found to be the most responsive signal to endpoint. The final value of the phase angle is modelled as a polynomial function of the etch system inputs. By monitoring the actual phase angle signal in real time, endpoint is successfully detected when the signal reaches the modelled value.

#### Mass spectrometric-based endpoint detection

A less commonly applied technique for endpoint detection, most likely due to the bulky equipment required, is mass spectrometry (MS). Characteristic endpoint signals appear in MS signals when the chemical constituents of the etching plasma change upon breakthrough to a new layer. Thomas III *et al.* [171] perform an analysis of the endpoint response times for a single-wavelength OES and a MS technique. The OES signal is found to respond in 0.2 seconds, whereas the MS signal responded in 0.9 seconds. Because the MS equipment is installed downstream from the plasma process, the response time is limited by a the transit time of the etch products from the wafer surface to the detector. Li *et al.* [188] compare a downstream and a direct line-of-sight mass spectrometer and show that the downstream MS can be used for endpoint detection, yielding

similar variations in signal intensities at endpoint as single-wavelength OES for a relatively large open area of 26 %. Line-of-sight MS is tested but deemed too sophisticated for practical implementation in an industrial environment.

For further references concerning endpoint detection outside of the scope of this thesis, the reader is guided to work by Ma [189] and Goodlin [190].

### 4.3.2 Fault detection and classification

Fault detection and classification (FDC) is the use of on-line data from a manufacturing process to warn operators of errors in production (fault detection) and identify the most likely fault type (fault classification). FDC may include detection of sensor failure, equipment failure, presence of unusual disturbances, and any other situation that does not correspond to routine operation [59]. FDC is essentially a use for VM, whereby process variables are analysed to extract meaningful variables using VM techniques, and then classification or further modelling techniques are used to detect and identify faults. FDC research is discussed in this section because the application of FDC techniques to a process typically involves the use of similar data sources, processing and modelling techniques to VM applications. This subsection presents a representative sample of published work in the area of FDC for plasma-etch processes.

A common approach for monitoring semiconductor manufacturing processes is to apply statistical processing control (SPC) rules to important variables. The central idea in SPC is that controlled processes operate in the presence of only chance causes of variation. Significant deviations in monitored variables that cannot be assigned to chance signify that the process is operating out of control. A cause-and-effect relationship between the monitored variables and the process quality is assumed. Control charts are used to monitor each variable, and it is assumed that the monitored variables are identically, independently and normally distributed (IIND). Because semiconductor manufacturing processes typically generate large numbers of inter-correlated variables for process monitoring, faulty conditions can result in alarms on several SPC charts if univariate SPC is applied to each one individually. Also, as the number of monitored variables increases, the chances of false alarms also increases. Multivariate SPC is used to reduce the number of control charts and the number of false alarms in such situations, and also take interactions between variables into account if required. Multivariate SPC involves the calculation of multivariate statistics such as Hotelling's $T^2$ statistic using all or a subset of the monitored variables and the applications of univariate SPC rules to the multivariate statistic. In this manner, the process can be monitored using a single control chart. Data preprocessing and variable selection is typically carried out to

ensure that the variables are IIND before carrying out multivariate SPC in this manner. Work by May and Spanos [28] provides a broad overview of the application of SPC in semiconductor production.

### Optical-based FDC

PCA is a popular technique for data reduction prior to performing FDC on multivariate manufacturing processes. Yue *et al.* [60] propose to use multiway PCA with the $T^2$ statistic and $Q$-statistic on OES data to detect faulty conditions on a contact etch process with $1-2\%$ open area. The authors use a "sphere" threshold technique on principal component loadings to select important variables from a single spectrometer. Fault classification is performed by examining principal component score plots and comparing the directions of faulty wafers in the principal component space to known faulty conditions. It is suggested that a library of faulty principal component directions can be stored to aid future fault identification, similar to techniques employed in chemometrics research [191]. Han *et al.* [170] also propose a fault detection algorithm based on the principal components of OES data for a $BCl_3/Cl_2$ etch process. The researchers compare the principal component scores at each sample time for processing wafers to the principal components scores of previously processed wafers. Abnormal changes in the principal components are used to detect faulty conditions.

Hong and May [192] propose the use of multi-layer perceptron (MLP) ANNs with OES and MS data as model input variables to estimate the *process input variables* for the purposes of fault detection in an etch process. PCA is used as a variable selection technique for the OES data and variable selection is carried out on the MS data using process knowledge and manual selection of atomic masses. The ANNs are trained to estimate the process inputs, RF power, pressure, and gas flow, and faults are detected by identifying deviations between the estimated process inputs and the actual process inputs. Satisfactory performance is recorded for detection of faults that are artificially induced by introducing 10% deviations in input variables. Similar work is described by Shadmer *et al.* [193] where ANNs are shown to outperform linear techniques in estimating the process inputs using OES and MS measurements from a $CHF_3/O_2$ plasma with no wafer present. Again, faults are detected and root causes are isolated by analysing discrepancies between the estimated and actual process inputs.

Sarmiento *et al.* [194] employ one-class SVMs to detect faults in RIE processes using PCA-reduced OES data. The one-class SVM system successfully identified all artificially excited faults in RF power using 35 selected wavelengths (using PCA) for an etch process. It is argued that the use of ANNs for fault detection requires large amounts of faulty

data to successfully train a network, whereas the one-class SVM technique requires only data collected during normal equipment operation.

### RF-based FDC

Guo and Spanos *et al.* [35, 195] demonstrate the use of RF-based measurements to monitor a plasma etch chamber, collecting five process variables - the position of the RF tune vane and RF load coil in the electrical match unit, the RF phase error, the plasma impedance, and the peak-to-peak voltage across the electrodes. The $T^2$ statistic is calculated from these variables and monitored using SPC control charts, successfully detecting artificially induced faulty conditions. To ensure the variables are IIND prior to SPC, auto-regressive models are first fit to the process data, and the $T^2$ statistic is calculated using the model residual signals.

An adaptive $T^2$ statistic, based on four RF-based measurements along with temperature and pressure measurements, is employed by Spitzlsperger *et al.* [61] to monitor a drifting plasma etch process. The adaptive technique is employed to reduce the number of false alarms caused by slow drifts in the covariance structure of the process variables as a result of chamber conditioning and process drift. The adaptation is implemented by updating the mean and covariance values for the monitored variables using an exponentially weighted moving average (EWMA) technique (based on a similar method used by Chamness *et al.* [196]). Domain knowledge is incorporated to distinguish between normally drifting variables and drifts that indicate faults.

Wise *et al.* [34] compare PCA, multiway PCA, trilinear decomposition, and parallel factor analysis for fault detection on an Al etch process with $BCl_3/Cl_2$ plasma using RF, OES, and *machine state variables*. Machine state variables include data that are measured typically directly from the plasma chamber, for example, chamber pressure, temperature, gas flow rates, power set points etc. *Local* models, built on subsets of data, and *global* models, incorporating more data and hence more process variation, are trained. Local models were found to capture more faults than global models in the tests performed. The best combination of variables for FDC is found to be machine state information along with RF signals. For the data set investigated, parallel factor analysis is found to perform marginally better than PCA and trilinear decomposition.

Chen *et al.* [197] use RF sensors and a comprehensive fault library to diagnose faulty conditions, using pattern matching software to compare signals from each wafer to the library of faulty signals. However, the fault library must be constructed using a DOE where faults are introduced manually, and this library may be prohibitively

large. Law and Macgearailt [198] use ratios of reflected and incident powers from RF and microwave sources in an ECR plasma etch chamber to characterise normal etching operations and aid fault detection. K-means clustering is used to define data clusters representing normal operation, and power source faults are detected when sensor data are observed that do not fit into the predefined clusters.

### Machine state-based FDC

The monitoring of machine state variables is perhaps the most cost effective method of process monitoring, considering that no additional sensors beyond the original processing tool sensors are required. Successful FDC schemes using machine state variables have been implemented by a number of authors.

For example, Imai [199] uses a PLS model with various machine state variables as inputs to estimate changes in plasma electron density that are indicative of stray contaminant particles in the etch chamber. Stray particles destroy integrated circuitry when they come in contact with wafer surfaces. The particles acquire a negative charge be collecting free electrons from the plasma, resulting in a detectable change in the overall plasma electron density in the chamber. In [199], such changes in electron density are reflected in the recorded machine state variables, and a PLS model is used to estimate the number of particles present. The authors report $R^2$ values of 0.75 for the model estimates. Potential fault occurrences are deduced from the particle predictions.

Chang [200] uses SPC techniques and machine state variables to detect fault situations. Similar to the work in [35], the IIND residuals from a process model are monitored for deviations from normal operations. However, instead of more typical linear time-series modelling, Chang [200] uses radial basis function (RBF) neural networks to develop time-series models of plasma etch process output variations. The RBF networks operate on a sliding-window basis to accommodate process drift. The models are trained to estimate the current time sample (in seconds) of the running process using the chamber state variables as inputs.

Gallagher and Wise [63] demonstrate how static PCA models are *not* robust to process drift, maintenance, cleaning, and equipment installation events, yielding false alarms after such events occur. To maintain the validity of PCA models over many wafer etch cycles, and hence avoid false alarms from FDC systems as monitored variables drift outside of preset limits, several researchers employ adaptive or recursive PCA techniques. EWMA techniques are employed in [63] to update the mean and covariances of PCA models to cater for process drifts and shifts and these adaptive models are shown to

remain robust for over three months of production data. However, only 8 of 21 induced faults are identified. Chamness [196] uses a similar EWMA update scheme for the mean values of a PCA model and an exact recursive standard deviation update to keep PCA models robust over 3000 wafers from a plasma etch process. Chamness uses 21 different chamber state variables including pressure, matching network component positions, temperatures at the electrodes and chamber walls, and coolant gas flows. Errors are detected using a threshold on the $Q$-statistic values, and variable contribution plots for the $Q$-statistic are used to classify faults once detected.

Many semiconductor manufacturing processes violate the PCA assumptions of linearity, Guassianity, and uni-modality, posing difficulties for PCA-based FDC techniques and resulting in misleading $T^2$ and $Q$ statistic values. He and Wang [58] propose a fault detection technique based on the k-nearest neighbours (kNN) method of clustering and demonstrate successful fault detection for the Al stack etch process investigated by Wise *et al.* [34]. The stack etch process is both non-linear and multi-modal, and 19 machine state variables are used for FDC. The technique is based upon the premise that normal production data from wafers will be numerically close together and form clusters that faulty data will not fit into. The kNN technique is shown to outperform $T^2$ and $Q$-statistic based fault detection schemes. The computational complexity and storage requirements of the kNN fault detection technique are improved in [201] where PCA is used first as a data reduction technique and faults are then detected using kNN clustering in the reduced principal component space.

A number of authors also use independent component analysis (ICA) for data preprocessing prior to fault detection. ICA is a multivariate statistical tool that can extract statistically independent components from observed data [202]. ICA has the advantage over PCA when used for FDC since ICA does not assume a Gaussian distribution of data. However, the determination of how many independent components to extract from the process data is non-trivial, the extracted independent components are randomly ordered, and the algorithm results are dependent on initial conditions, often requiring several executions to determine optimal solutions. Lee *et al.* [203] propose a technique to address these issues and demonstrate the effectiveness of the technique on industrial data sets, including the semiconductor etch machine state variables examined by Wise *et al.* [34]. The proposed technique is shown to be more effective than PCA for fault detection for the data set used. Similar to the research in [34], global and local models are investigated, and the best fault detection accuracy is achieved using the local models.

Ge and Song [204] propose an adaptive PCA-based method that is claimed to outperform the PCA-kNN scheme of He *et al.* [201], and uses simpler calculations than the ICA method of Lee *et al.* [203]. A monitoring statistic, based on a one class support

vector data description (similar to that previously used in [194]), is developed and shown to outperform the kNN-based fault detection technique for the same etch process and machine state variables as in [34].

### 4.3.3 Estimation of plasma variables

A number of researchers focus on the estimation of plasma variables from real-time measurements. Such research is not always concerned with plasma etch processes alone, but rather on plasma discharges in general.

#### Optical-based plasma variable estimation

OES data can be used to gain a measure of species densities in plasma through the use of *actinometry* techniques [29]. Actinometry, as mentioned in Section 2.5.1, involves the addition of a known quantity of an inert gas to a plasma and the use of the relative emission intensities of the added gas to determine the concentration of other species of interest. Hanish *et al.* [205] demonstrate the use of actinometry techniques to estimate atomic chlorine concentration in a plasma for control purposes.

Malyshev and Donnelly [206, 207] develop a similar technique to actinometry, trace-rare-gas OES (TRG-OES), in which the emission lines from small concentrations of five inert gases, He, Ne, Ar, Kr, and Xe, are used to determine plasma electron temperatures and in some cases plasma electron density. The results from TRG-OES are found to agree with Langmuir probe measurements for low pressure plasmas, with accuracy reducing at higher pressures ($> 5$ mTorr) due to variations in the electron energy density function. Hence, TRG-OES can be used as a non-intrusive alternative to Langmuir probes for low-pressure plasmas. The application of the technique for silicon-based semiconductor processing is considered in [208], and a vertically movable OES device is employed in [209] using TRG-OES to obtain spatially resolved measurements of electron temperatures, chlorine density, and approximate electron-energy distribution functions in a plasma etch processing tool.

#### RF-based plasma variable estimation

Electrical measurements from RF systems are also used as inputs to VM models, relating waveforms for voltage, current, and phase to plasma variables. The advantage of these techniques is that, similar to OES, typically no invasive hardware is required, allowing

VM without contamination risks. Research by Sobolewski exemplifies this area. An investigation into three methods to determine ion current from RF current and voltage measurements is presented by Sobolweski in [210]:

1. **Power/voltage method**: This method entails the division of measured RF power by the measured RF voltage to determine ion current. However, the assumptions of non-sinusoidal sheath voltages, zero time variances in ion current, along with other simplifications lead to relatively inaccurate results.

2. **Analytical method**: This method uses an analytical equation for the current-voltage relation of the plasma sheath (described in [211]) to determine ion current. Again, assumptions in the model, especially the exclusion of time variation in the ion current, lead to errors in this method, which are more pronounced at higher frequencies.

3. **Numerical sheath method**: This method uses a numerical model for the plasma sheath as described in [131]. The sheath is modelled in one dimension using fluid equations for the ion dynamics (see Section 4.2.2). The model includes time-dependent ion current and ion density effects and simulates the ground sheath and the sheath adjacent to the RF biased electrode simultaneously. The model makes approximations to allow fast computation and hence real-time application to plasma chambers.

Sobolewski produced several publications detailing the use of the numerical sheath model described in [131], demonstrating the use of the model to monitor sheath voltages and ion energies in generic plasma [212] and ion current drift due to deposition in etch chambers [132]. A more recent work investigates the real-time, non-invasive monitoring of ion energy and ion current at a wafer surface during an etching process [40]. The monitoring technique involves varying the inputs to a plasma sheath model using a sum of squares fitting algorithm such that the model outputs match the actual RF waveforms recorded from the plasma chamber. In [40], the responses of the ion current and energy in response to changes in source power, gas flow and pressure are analysed.

Yamashita *et al.* [213] investigates the use of RF waveform analysis to monitor ion energy and ion flux density for a RIE chamber. The authors use a capacitive approximation in a model of the plasma sheaths, and also assume that the sheaths are collisionless. The model results are used to minimise etch-induced damage and contamination by finding optimum conditions for operation. Essentially, ion flux is determined using the power/voltage method as described in [210].

### SEERS-based estimation of plasma variables

The Hercules® [214–216] system is a measurement system based on self-excited electron resonance spectroscopy (SEERS) signals. Hercules and SEERS use a passive in-situ sensor on the plasma chamber wall and an equivalent circuit model for the plasma sheath to monitor plasma variables such as power coupling, electron collision rate, and electron density in the plasma in real time. The plasma sheath model uses a hydrodynamic approach for the plasma, the inert mass of electrons is treated as an inductance, and collisions with neutrals, including power dissipation in the expanding sheath, as a resistance. When the capacitive behaviour of the sheath is taken into account, the plasma behaves as a damped oscillation circuit [214].

Figure 4.5 shows the equivalent circuit used by the Hercules system. The left side of the figure shows the external excitation of the system, and the plasma, represented on the right side, is modelled as a damped oscillation circuit. The non-linear behaviour of the sheath creates harmonics of the driving frequency to excite the oscillating circuit [215].

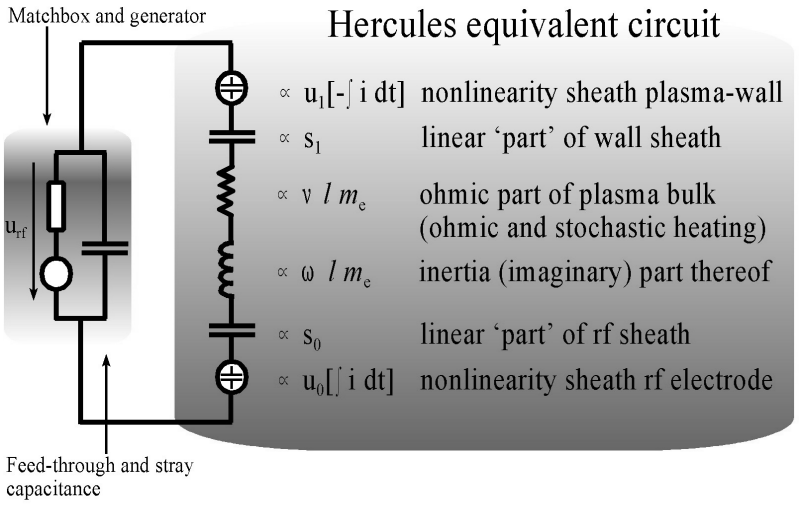



FIGURE 4.5: Hercules equivalent circuit [215].

Steinbach *et al.* [215] show how the plasma variable estimates from the Hercules system are useful for endpoint detection, fault detection, process optimisation, estimation of etch variables, and chamber matching applications. The Hercules monitoring system is capable of operation with contamination on the surface of its sensor, enabling reliable measurements throughout complete maintenance cycles.

Wurm *et al.* [214] performed similar studies with the Hercules system, analysing electron density and power dissipation effects, system matching applications, cleaning cycle optimisation, and long-term process monitoring. Tegeder *et al.* [217] use SEERS and Hercules to create process fingerprints for fault detection and analyse chamber conditioning effects. The relationships between the plasma variable estimates and downstream electrical test measurements are also examined to create an early warning system for faults. In a presentation by Steinmetz *et al.* [216], the Hercules measurements are shown to be useful for investigation of first wafer effects, wafer fault analysis and characterisation of chamber cleaning steps.

Despite the advantages of the Hercules and SEERS system, it has not found widespread application in VM research, possibly due to need for adaptation of the chamber to allow installation of the sensor. The sensor requires physical access to the plasma, which can be difficult to implement in industrial environments.

### 4.3.4 Estimation of etch variables

The estimation of etch performance variables during processing using in-situ measurements enables tighter etch process control and allows the introduction of run-to-run APC, or even real-time APC, to the plasma etch process [218]. VM models have advantages over the standalone input/output models discussed in Section 4.1 when used for estimation since actual process data that reflect changes in the process, rather than constant system input set points, are used to estimate the variables of interest. For wafer-level VM models, statistical summaries of time series traces for each wafer (or part thereof) are typically used as VM model input variables. Research by Kim and Kim [147] compare VM models using OES data to conventional standalone input/output models for estimation of etch rate, profile angle, and etch rate uniformity and finds that the VM models produce superior estimates.

Generally, VM models are created using empirical modelling techniques because first-principals physics-based models are unable to produce estimates in a timely fashion without broad simplifications, and, in many cases, the fundamental relationships between the measured process variables and the variables to be estimated are complex or poorly understood. Although a great deal of work has been published detailing VM research, a significant amount of further research has been potentially completed by private companies, and such work remains confidential and inaccessible to the public.

**Optical-based etch variable estimation**

Camara and Zekentes [219] monitor a single emission line corresponding to atomic fluorine (704 nm) to indirectly monitor the etch rate of a silicon carbide wafer in $SF_6$/Ar plasmas. The etch rate is found to be directly related to the fluorine peak intensity. Similarly, Mozumder and Barna [140] use the slope of a single emission line to monitor etch uniformity and etch rate in a nitride etch for control purposes.

Zhang *et al.* [81] use eight selected OES wavelengths along with chamber state variables to model the expected ellipsometry signal from an etch tool in order to estimate etch rate and etch depth simultaneously. Finite impulse response models that using up to 56 past process measurements to estimate the ellipsometry signal are employed. Model parameter estimation is accomplished using numerous methods, the best being a discrete wavelet transform. The modelling results are shown to be superior to previous work by Rietman *et al.* [220] who used ANN models on the same data set.

Kim and Kim [147] compare the use of different OES-based VM models and a standalone input/output model. ANNs are trained to estimate etch rate, profile angle, and etch non-uniformity using four different combinations of input variables recorded during a $2^{4-1}$ fractional-factorial experiment:

1. Expert recommended wavelengths. A set of wavelengths is selected based on the advice of engineers with extensive process knowledge.
2. Expert recommended wavelengths along with the wafer DC bias.
3. Etch process inputs. The etch process inputs are the input set points for the plasma chamber, and correspond to the type of inputs normally used for standalone input/output models seen in Section 4.1.
4. The main principal component scores extracted from the OES data.

The first two models outperform the traditional standalone input/output model by 10.1% and 39.3% respectively for etch rate estimation accuracy. When combined with the expert recommended wavelengths, the DC bias is found to improve the estimates for etch rate by 32%, but degrade model performance when estimating non-uniformity and profile angle by 41.8% and 23% respectively. The authors conclude that the DC bias is not an influential variable for wafer non-uniformity and profile angle. The models using principal component scores as inputs perform poorly for this data set.

Chen *et al.* [221] examine the use of PCA and PLS to compress OES data for modelling of etch rate, uniformity and aspect-ratio dependent etching. Data for the PCA

and PLS models are obtained using a 26-run, five-factor, central-composite designed experiment with one spectral observation per wafer, and the resulting estimates are compared with estimates from a LSR model using manually selected spectral lines as input variables. The PCA and PLS models are found to be more effective for modelling of etch rate ($R^2 = 0.87$) and uniformity ($R^2 = 0.95$) than the LSR models with manually selected spectral lines. PLS models are generally found to use fewer parameters than the PCA models. Modelling of aspect-ratio dependent etching is generally found to be inaccurate ($R^2 = 0.5 - 0.66$), and this inaccuracy is attributed to the accuracy of the metrology used to originally train the models.

Ragnoli *et al.* [222] compare PCR, PLS regression, forward selection component analysis (FSCA) and forward selection regression (FSR) to model etch rate from OES data for a production data set. FSCA is a statistical technique that attempts to summarise the variance in a set of input data by extracting key variables that best represent the information contained in the input data. FSCA is similar to PCA in that it is especially suitable for input data that contains many correlated variables. However, the extracted components correspond directly to individual variables from the original input data, rather than linear combinations of all variables as is the case with PCA. Such one-to-one correspondence can be useful in cases where analysts are searching for VM input variables that are influential to the process output. The authors conclude that FSCA and FSR are more effective than PCA and PLS for feature selection in the examined data set. $R^2$ values of 0.94 for the etch rate estimates are achieved for an interleaved data set using FSR.

Hong *et al.* [141] use ANNs to model etch rate, uniformity, selectivity, and anisotropy using OES data from a $2^4$ factorial experiment carried out on a benzocyclobutene (BCB) plasma etch process. PCA and auto-encoder neural networks (AENNs) are used to extract features from the OES data. The AENNs extract features that produce estimates 1.14 % better on average than the models based on features extracted using PCA, albeit with a significant increase in computational complexity.

White *et al.* [223] examine the use of multiple optical emission spectrometers to determine spatially resolved estimates of plasma etch variables. Three separate physical locations on one plasma chamber are used to collect OES data during factorial experiments, and PCR, PLS, and ANN models are trained using the principal components of the collected OES data to model metal line widths across the wafer. The principal components extracted from the OES data are found to be correlated to the input variables of the etch chamber. The addition of machine state variables such as endpoint traces and dc bias signals are *not* found to increase the estimation accuracy of the models. The authors conclude, as expected, that linear PLS and PCR techniques are most suited for

experiments operated tightly around a center point, but that non-linear ANN techniques provide better estimation accuracy over larger ranges. A 50 % reduction in estimation error is achieved through the use of multiple-beam OES data over single wavelength methods.

### RF-based etch variable estimation

As well as being used for endpoint detection (Section 4.3.1), RF based sensors are shown by Tsunami [224] to be useful for oxide etch rate estimation. A commercial impedance sensor, the Straatum SmartPIM plasma impedance monitoring device, is used to estimate etch rate using the fundamental and some harmonics of current, voltage, and phase from the power supplies of a dual frequency etch system. Summary statistics are extracted from the voltage, current, real power, complex power, phase, impedance, resistance, and reactance signals, and stepwise regression is used to model the etch rate using these statistics, achieving an $R^2$ value of 0.96 for a production data set. The VM model is constant, and, as a result, does not remain accurate for large data sets where there is considerable process drift.

Garvin and Grizzle [225] demonstrate the use of a broadband RF sensor that examines the reflections from an antenna in an ICP plasma excited by a sweep of frequencies between 1–2.25 GHz. A purely empirical model is compared to a parametric approach for estimation of etch rate. The parametric approach involves the determination of series RLC equivalent circuits that match the response of the broadband sensor and empirically relating the circuit variables to the plasma etch rate. Stepwise regression is used as a modelling tool and the empirical model is found to estimate etch rate with slightly better accuracy ($R^2 = 0.997$) than the parametric model ($R^2 = 0.962$).

### Machine state-based etch variable estimation

Some of the first publications concerning the VM of etch variables were produced in 1994 by Lee and Spanos [83, 226] in which machine state variables are used to estimate etch rates, selectivity, and uniformity of etched wafers. Empirical models are created using LSR, PCR, PLS and ANNs using a training data set collected from a $32-$run designed experiment, and tested using a data set recorded four weeks later. Uniformity estimates are inaccurate because none of the input variables were spatially resolved. Polysilicon, oxide, and photoresist etch rates are estimated with accuracies of approximately $5-10\%$ by the PCR, PLS and ANN models, the results of which are shown to be statistically equivalent to each other and significantly better than the MLR results.

Card *et al.* [227] use ANNs to predict etch rate and selectivity using similar measurements to Lee and Spanos, but from production data. Prior to modelling, Card *et al.* [227] perform variable selection using GAs. GAs are used to optimise the selection of a subset of variables to be used as VM model input variables. The VM model is then used in a run-to-run process controller. Because a fixed VM model is used, the authors recommend complete retraining on regular intervals to maintain model currency.

Zeng *et al.* [228, 229] estimate etch bias using PCR, PLS, and ANN models with chamber state variables. Different variable selection techniques are examined for VM models of *etch bias*, that is the difference between the critical dimensions of lithography patterns and etched trenches. Stepwise selection, random modelling, and GAs are investigated as model performance-based variable selection techniques, and a linear correlation test is used as an information theory-based technique. Stepwise selection is found to be as effective as the more complex algorithms for variable selection, and ANNs produce the most accurate estimates of etch bias with $R^2$ values of 0.74.

Kang *et al.* [230, 231] discuss the use of stepwise selection, decision trees, GAs, PCA and kernel-PCA for dimensionality reduction and variable selection. In [230], the authors examine four empirical modelling techniques, MLR, ANNs, kNN-regression, and support vector regression (SVR) for VM implementations on two plasma etch processes using chamber state variables. Stepwise selection is reported as the best variable selection technique out of those investigated. The best regression algorithms for the etch process are found to be LSR and SVR. However, the small test data set employed did not span enough time to capture process drifts or preventative maintenance (PM) events.

Lin *et al.* [232] develop an ANN-based variable selection algorithm to create VM models for estimating a critical dimension in an etch process using chamber state variables as inputs. During execution of the algorithm, multiple ANN models are developed, and the partial F-statistics are used to determine when to add or subtract variables from the ANN models. An input selection using linear stepwise selection is used to seed the algorithm which produces ANN models with superior estimation capability to those created using expert-recommended or stepwise-selected variables alone. Through the variable selection algorithm, an original set of 66 variables is reduced to approximately 10 variables and the VM models estimated the critical dimension within approximately 1% mean absolute percentage error (MAPE).

Cheng *et al.* [233, 234] present a method to evaluate the reliability of VM predictions by calculating a "reliance index" (RI) that quantifies the reliability of each VM estimation. The calculation of the RI is based upon the degree of overlap between the probability distributions of estimates arising from two process models: a linear MLR

"reference" model, and a non-linear ANN "conjecture" model. A RI threshold value is defined, below which estimates are deemed unreliable. To assist the evaluation of each estimate, similarity indices with corresponding thresholds are used. A global similarity index (GSI) uses the Mahalanhobis distance to gauge whether new data from the process is similar to that used during model training, and individual similarity indices (ISIs) are calculated for each VM model input variable to identify potential causes for reliability issues. A red/orange/green lighting system is implemented to allow quick interpretation of the different indices by system operators. The system is used to monitor a data set of 24 wafers in [234] from a production etch process successfully, but the details of the input and output variables are kept confidential.

In other research, Cheng *et al.* [235] describe the development of a dual-phase VM scheme in which preliminary estimates of process outputs are made using the process data, and these estimates are updated at a later stage when real metrology values from neighbouring wafers become available. The VM estimates are considered more reliable after this update. The accompanying RI and GSI of each VM estimate are also calculated to gauge the reliability of the estimates and determine if model training or further measurements are required. An industrial CVD process is used to demonstrate the effectiveness of the scheme. Su *et al.* [236] investigated the training and estimation speed of different model types for the dual-phase system in [235] using industrial CVD and etch process data and concluded that double-layered ANNs are too computationally demanding for real-time operation and MLR models do not provide enough accuracy. Single layer ANNs are recommended for the VM scheme. Again, the input and output variable details are kept confidential.

## 4.4 Fab-wide virtual metrology

When accurate VM models are developed for each processing step in the semiconductor manufacturing cycle, advanced process control (APC) can be implemented at all processing steps using feed-forward and feed-back control loops. A number of researchers have concentrated on the development of fab-wide VM frameworks suitable for semiconductor manufacturing. However, most of the published studies are based on the simulation of production environments, since actual fab-wide implementation of VM systems remains an aspiration. A sample of such research is presented here.

Generic approaches to the development of VM models for manufacturing processes have been proposed. For example, Ferreira *et al.* [237] suggest a generic nine-step method for creating VM models for individual semiconductor manufacturing processes.

Overall, the method is broken into three procedures; data pre-processing (including variable selection), VM model development, and VM model implementation. An industrial CVD process is used as an example application, and a VM model using ANNs is successfully implemented.

Su *et al.* [238] stress the importance of feed-forward and feed-back control (between different manufacturing steps) in semiconductor manufacturing. A control architecture with multiple layers in a cascade structure is suggested. A dual-VM scheme, consisting of a tool-level estimator for estimation of wafer characteristics and a product-level estimator for estimation of the electrical performance of finished products, is presented. The electrical performance of a wafer is referred to as the wafer acceptance test (WAT). A schematic of the proposed fab-wide VM architecture is shown in Figure 4.6.

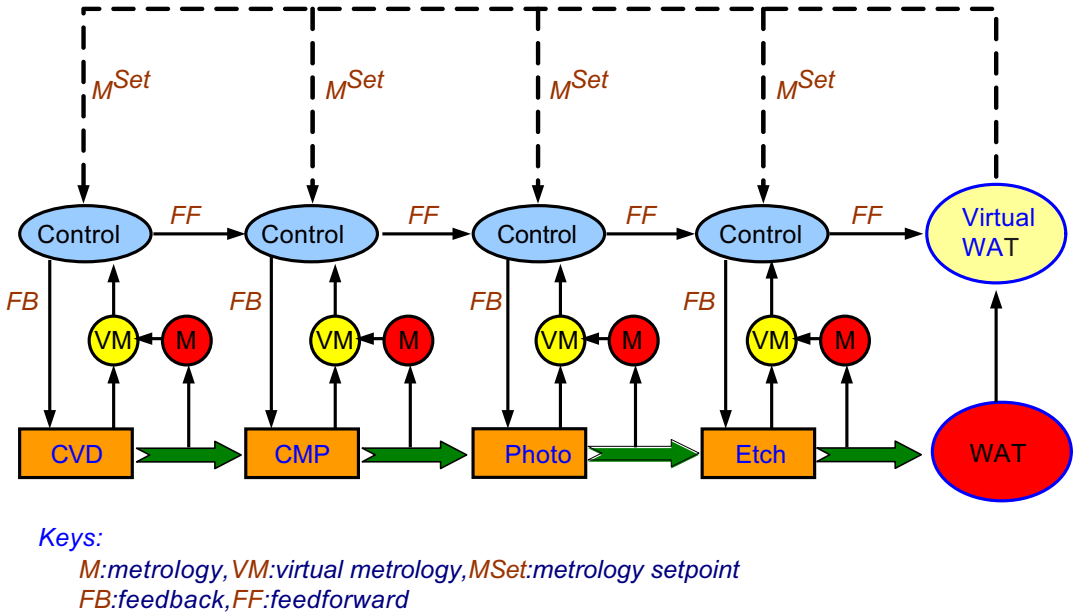



FIGURE 4.6: Detailed feed-forward/feed-back control structure incorporated with VM and VM-WAT [238].

Pasadyn and Edgar [239] examine a Kalman filter-based state estimation scheme for the manufacturing process that views the manufacturing area with all the tools, products, and processes contained within as a single interrelated system. The estimator is designed to operate even in the case of missing or delayed measurements and is shown in [239] to work with a run-to-run controller on a simulated process.

Khan *et al.* [16, 17] present a wafer-to-wafer control scheme using VM on a factory level as illustrated in Figure 4.7. Similar to Ferreira *et al.* [237], Khan *et al.* [16] suggest a generic set of steps to implement VM, recommending the use of designed experiments for data collection and PLS models for estimation. A simulation study of two tandem

processes illustrates the effectiveness of the scheme using a double exponentially weighted moving average (dEWMA) wafer-to-wafer controller and showing improved performance over traditional lot-to-lot controllers that use only actual metrology data. Issues such as poor data quality, metrology delays, maintenance events, and multi-step processes are identified as challenges facing the successful implementation of fab-wide VM and control in the near future. Details on the recursive PLS algorithm and dEWMA controller that are used, along with simulation results on a simulated multi-input multi-output (MIMO) process, are presented by Khan *et al.* in [11].

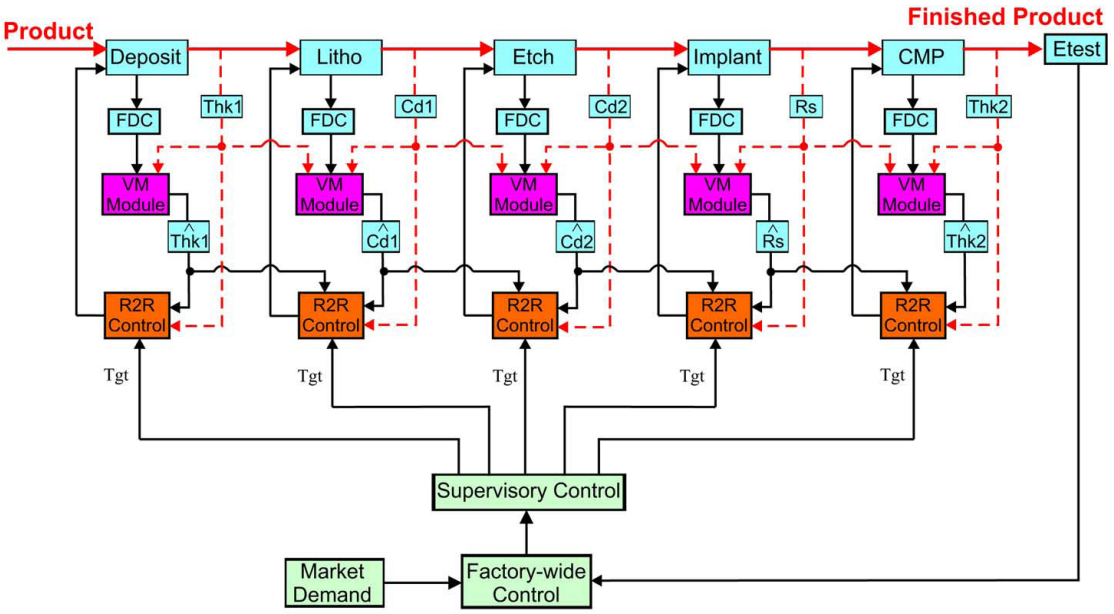



FIGURE 4.7: Fab-wide control using VM at individual processing tools [17].

## 4.5 Discussion

Direct comparison of results between different VM and modelling research is made difficult by the wide range of processes, data sources, techniques, and applications used by different researchers. However, some broad conclusions can be taken from the research examined in this chapter.

Standalone input/output models are less effective for process monitoring purposes than VM models using real-time process data. Standalone empirical input/output models do not take the plasma chamber state into account during production, and this state can have a significant influence on the process output. Analytical models require too much simplification, with resultant losses in accuracy, for real-time applicable to complex plasma etch processes. Accurate simulations of the etch chemistry cannot be achieved in a timely fashion.

Data quality is important for successful VM implementation and it is important to use sensors capable of measuring process variations that are relevant to the system output. Factorial experiments can be used to collect the data for model building if available, but such data is not always representative of the variations observed in production processes. The reality of many production environments is that measurement frequency is relatively low, and the etch process suffers from extensive process drift and regular unpredictable process shifts due to maintenance events.

The estimation accuracy of VM models can be reduced in situations where there are patterns in the process data that are not represented in the training data used during model creation. In general, extrapolation of a model operating space can produce unpredictable estimates, especially for non-linear models. As such, it is useful to obtain a measure of the level of confidence in each estimate, especially in cases where VM estimates are used for process control.

No single modelling or variable selection techniques stand out as particularly beneficial for VM modelling of etch processes. Fab-wide implementation of APC using VM schemes is in its infancy, but the potential performance gains of the realisation of such systems have been identified. Only through further development of VM models for each semiconductor manufacturing process can a VM-empowered APC system be implemented.

By no means is there a widely accepted standardised method for VM modelling of the plasma etch process. Relatively few researchers have provided comprehensive information on methods to develop VM models that can cater for the peculiarities of etch process dynamics. Chapters 6 and 7 now strive to provide a comprehensive examination of VM model development for a production etch process. Among other techniques, Gaussian process regression (GPR) models that produce estimate confidence intervals as part of the estimation procedure are examined as a potential VM models for the etch process. Chapter 8 examines the use of VM techniques for real-time control, and further literature in the area of control is examined at that point.

The literature review detailed in this chapter formed part of a publication published in the IEEE Transactions on Semiconductor manufacturing [8].

# Chapter 5

# Case study: Virtual metrology for the VASIMR engine

## 5.1 Introduction

In this chapter, a case study is examined in which virtual metrology (VM) techniques are used for temperature estimation in the variable specific impulse magnetoplasma rocket (VASIMR) engine. The VASIMR engine is one of a new generation of electrically mediated propulsion systems for spacecraft that produces thrust from a fuel in plasma form using electric energy. The use of electric energy is in contrast to the vast majority of existing rocket engines which rely exclusively on chemical combustion to achieve thrust in the vacuum of space.

An empirical state-space model relating the VASIMR engine temperatures to the plasma engine inputs is developed for temperature estimation. The model uses optical emission spectroscopy (OES) measurements (see Chapter 2) as correction terms to enhance closed-loop estimation accuracy. Results are presented for both open-loop and closed-loop estimation strategies.

## 5.2 The VASIMR engine

### 5.2.1 Basic operation

The VASIMR engine is an electric propulsion technology being developed by the Ad Astra Rocket Company at its facilities in Houston, Texas, USA, and in Liberia, Guanacaste, Costa Rica. The VASIMR engine uses a highly ionised plasma accelerated using magnetic fields to produce thrust. In the VASIMR engine, plasma is generated using a helicon antenna driven by radio frequency (RF) waveforms at the frequency 13.56 MHz. Helicon discharges are known to be efficient methods for plasma production [240].

The VASIMR engine produces thrust in three different stages and is depicted in Figure 5.1. Initially, electromagnetic waves from the RF-powered helicon antenna energise free electrons present as a result of ambient thermal energy in a neutral gas fuel. These electrons then ionise atoms in the gas through energetic collisions to create a plasma. This first stage is the helicon stage of the VASIMR engine.

Charged particles in plasmas follow helical paths along magnetic field lines due to the Lorentz force, which is proportional to the charge on the particles, the velocity of the particles, and the strength of the magnetic field. In the VASIMR engine, carefully designed electromagnets form magnetic field patterns that confine the plasma and move it from the helicon stage to the ion cyclotron resonant heating (ICRH) stage of the rocket.

In the ICRH stage, a second helicon antenna excites the ions in the plasma at their gyrofrequency, further energising the propellant gas [241]. The gyrofrequency, or cyclotron frequency, of an ion is the frequency of rotation of the ion as it spirals in a magnetic field, and is given by

$$\omega_{ci} = \frac{q_i B}{m_i} \text{ rad/sec}, \tag{5.1}$$

where $\omega_{ci}$ is the ion cyclotron frequency, $q_i$ is the ion charge, $B$ is the magnetic field strength, and $m_i$ is the mass of the ion. The ICRH antenna operates on the same principal as the electron cyclotron resonance (ECR) plasma chambers discussed in Section 2.4.3, heating ions instead of electrons by supplying energy at their cyclotron frequency.

In the final stage of the rocket, a group of magnets, comprising the magnetic nozzle, accelerate the plasma away from the craft along expanding magnetic field lines. It is at

this point that the energised plasma physically detaches from the engine and its magnetic field, thus creating thrust. The exhaust velocities from VASIMR are expected to reach as high as 120 km/s [242] in final prototypes. For near-term applications of VASIMR, large solar arrays are expected to generate electric power for the rocket [243].

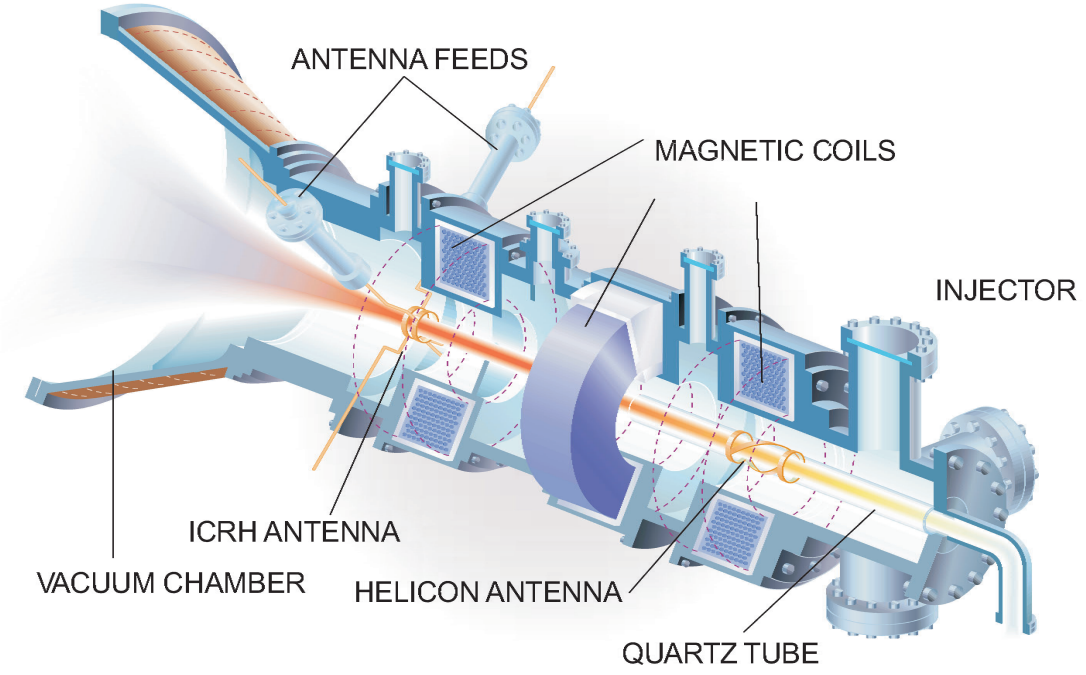



FIGURE 5.1: Schematic of the variable specific impulse magnetoplasma rocket (VASIMR) engine. The VASIMR consists of three main stages, namely, a helicon plasma source, an ion cyclotron resonance heating (ICRH) plasma accelerator, and a magnetic nozzle that accelerates the plasma away from the craft to produce final thrust [241]. For prototypes on Earth, a 5-$m^3$ vacuum chamber is used to simulate the vacuum of outer space. Image courtesy of NASA.

### 5.2.2 Advantages of the VASIMR engine

Compared to alternative electric propulsion technologies, the VASIMR engine has many advantages as a potential space propulsion system. Electrodes are not required to be in contact with the plasma since power is inductively coupled to the plasma in the engine. This electrode-less design allows greater power densities to be sustained for longer periods of time than conventional chemical combustion or ion engines, without fear of electrodes becoming damaged or wearing out. This design is essential for missions requiring months or years of continuous rocket operation. The propellant gas, which is Argon, is inexpensive, chemically inert, and widely available.

The VASIMR engine produces thrust very efficiently compared to chemical rockets. The efficiency of a jet or rocket engine is measured by the momentum change effected per

unit weight of propellant consumed. This characteristic, which is measured in seconds, is termed the *specific impulse* of the engine. Used as a measure of economy for rocket engines, specific impulse is comparable to a miles-per-gallon, or litres-per-100 km, rating for a motor vehicle.

Chemical rockets used in spacecraft typically produce thrusts of 60000–70000 N at a specific impulse of 300 s [244]. High thrusts from chemical rockets provide a large acceleration, but with a rapid consumption of fuel. The VASIMR engine produces relatively low thrusts of 5–10 N, but with a high specific impulse of between 5000 and 15000 s [245]. With the ability to sustain thrust for prolonged periods, the relatively small acceleration achieved using low thrusts ultimately achieves higher spacecraft velocities for a given fuel supply. Figure 5.2 illustrates that over the vast distances involved in interplanetary travel, a VASIMR-equipped spacecraft can reach distant destinations in less time than a spacecraft using chemical rockets for propulsion, with the same quantity of fuel.

The VASIMR engine gets its name from its capability to vary its specific impulse performance in order to produce more or less thrust as required. By varying the amount of energy dedicated to the ICRH and helicon sections, and varying the amount of propellant delivered for plasma generation, VASIMR is capable of either generating low-thrust, high-specific impulse exhaust or relatively high-thrust, low-specific impulse exhaust [242].

Although significantly more than 10 N of thrust is required to produce a substantial fraction of the gravitational force felt on Earth, the constant acceleration from continuous VASIMR operation results in an artificial gravity effect on board spacecraft, reducing the physiological effects that weightless environments have on the human body.

### 5.2.3 Problem statement

Heat is an undesirable by-product of helicon plasma production. The helicon stage of the VASIMR engine comprises a quartz gas containment tube surrounded by a helicon antenna. Because the ionisation mechanisms are not completely efficient, some neutral atoms do not acquire sufficient energy from collisions to expel electrons into the plasma. Although these neutral atoms might achieve an excited state temporarily from such collisions, they eventually return to their base configuration, releasing energy as photons in the visible, IR, and UV spectra as described in Section 2.5.1. The released energy radiates away and is absorbed by the gas tube and other nearby engine elements.

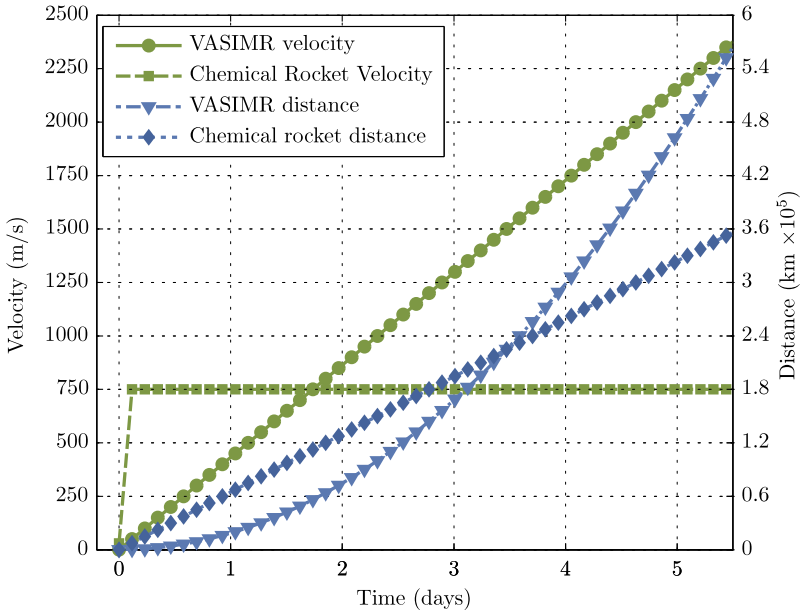



FIGURE 5.2: Example performance of a VASIMR engine-propelled spacecraft compared to a chemical rocket-propelled spacecraft. This figure compares the velocity and distance profiles for a hypothetical 2000 kg mass accelerated using a thrust of 60000 N from chemical rocket with specific impulse of 300 seconds, and a thrust of 10 N from VASIMR with a specific impulse of 5000 seconds. Both spacecraft start with zero initial velocity and 500 kg of fuel. The chemical rocket burns its fuel in 25 seconds to achieve maximum velocity. After 3.5 days, the VASIMR engine-powered craft passes the conventional craft travelling twice as fast, and still having used only 61 kg of its fuel reserves. The VASIMR engine-equipped craft can continue to accelerate for 28 days using the initial 500kg of fuel, reaching a final velocity of 12.3 km/s.

In addition, high velocity neutrals can be created as a result of energetic collisions between particles. These neutral atoms are not affected by the magnetic field lines and continue on their original paths at high velocities, ultimately colliding with other particles or the gas containment tube. Furthermore, since the gas containment tube is not completely transparent to radio-frequency energy, it absorbs a portion of the energy transmitted by the antenna. All of these effects result in significant and rapid heating of the gas tube as shown in Figure 5.3.

Temperature control of the gas tube is critical to the VASIMR engine design since the quartz tube can potentially reach absolute temperatures and achieve temperature gradients beyond its allowable limits. In addition, the VASIMR engine prototype uses superconducting magnets located close to the gas tube to produce the strong magnetic fields required to propel the plasma from the engine to produce thrust. These magnets operate at cryogenic temperatures that must not be affected by the heat produced during the plasma production process.

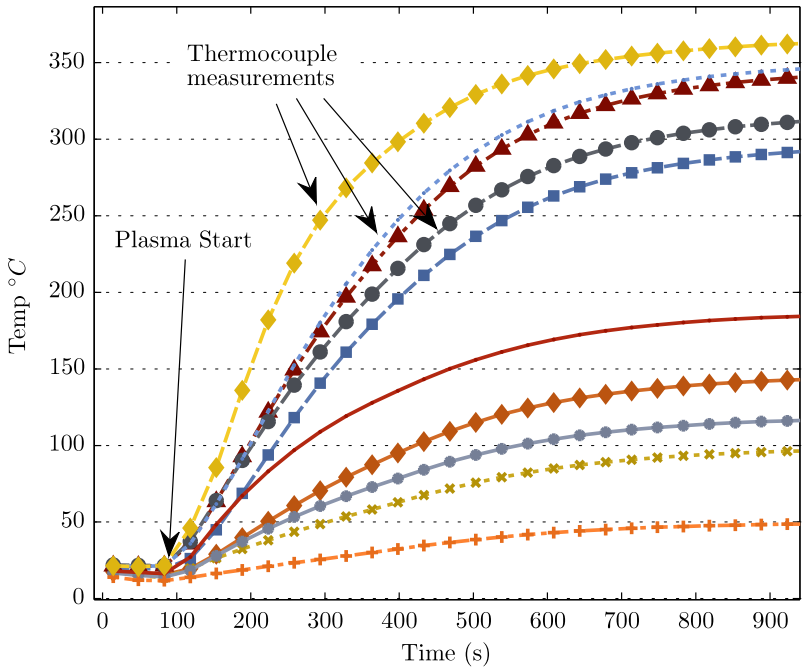



FIGURE 5.3: Gas containment tube surface temperatures measured at 10 different locations during plasma startup. Temperatures rise quickly after a stable plasma is established. Note the large variation in temperatures on the surface of the gas containment tube in steady state. The nonuniform distribution occurs since the heat deposited by the plasma varies with position due to the mechanisms of energy transfer to the plasma by the helicon antenna.

Although laboratory prototypes of the rocket use thermocouples to monitor the temperatures of the gas containment tube, thermocouples cannot be used in the final flight design because they would obstruct cooling designs that are in development. Thermocouple temperature signals are also subject to electromagnetic interference from the helicon antenna and, furthermore, the thermocouples themselves are physically fragile.

The goal of this case study is to develop a virtual metrology system for estimating the temperature distribution on the gas containment tube in the helicon section of the VASIMR engine, in the absence of direct temperature measurements from thermocouples. In particular, a state-space prediction model, that employs OES measurements from the plasma for temperature-estimate correction, is used. Because OES measurements directly correlate to the excitation of non-ionized neutrals in the plasma, and since these neutrals contribute to the heating of the gas containment tube, it is conjectured that OES data can be used to assist in temperature estimation.

## 5.3 The helicon antenna system

As discussed in Section 5.2.1, high-density plasma is produced using a helicon wave source in the VASIMR engine. The helicon section of the engine, depicted in Figure 5.4, consists of a helicon antenna wrapped around a quartz tube through which neutral gas is flowing. Electromagnetic coils, which maintain a magnetic field parallel to the gas flow, surround the quartz tube.

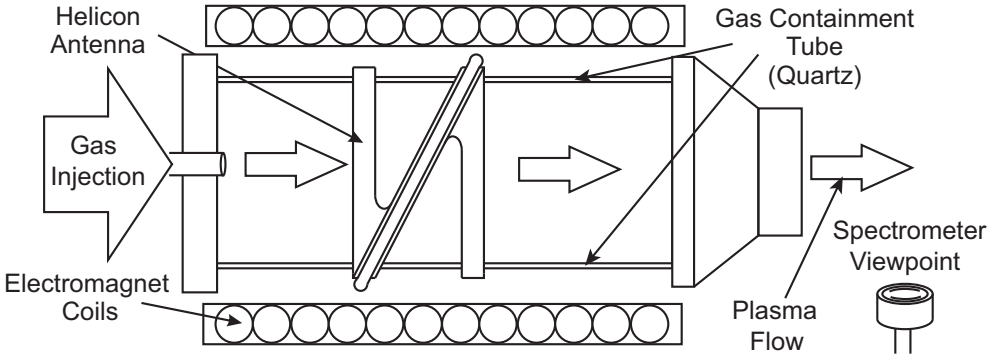



FIGURE 5.4: Helicon section of the VASIMR engine. A helicon discharge uses a right-hand circularly polarised wave to efficiently ionise a neutral gas to plasma state. The helicon is surrounded by electromagnetic coils that maintain a magnetic field along the axis of the antenna to assist in high-density plasma production.

Helicon discharges are a form of inductively coupled plasmas (ICPs) where a magnetic field is used to assist in the production of high-density plasma. The magnetic field has three main functions in such plasmas.

1. The magnetic field increases how far an electromagnetic wave penetrates into the plasma, also known as the *skin depth*. With the magnetic field in place in a helicon discharge, the electromagnetic waves can penetrate into the entire plasma.

2. The magnetic field helps to confine the electrons in the plasma for an extended time allowing more collisions to occur.

3. The magnetic field gives the operator the ability to vary plasma properties such as the plasma density uniformity [22].

In the VASIMR engine, the magnetic field also confines the plasma to the center of the quartz tube, and guides the plasma flow to the next section of the engine.

A helicon wave is defined as a right-handed polarised wave that propagates in a radially confined magnetised plasma for frequencies $\omega_{ci} \ll \omega \ll \omega_{ce}$, where $\omega_{ci}$ is the

ion cyclotron frequency, $\omega$ is the frequency of the helicon wave, and $\omega_{ce}$ is the electron cyclotron frequency [246]. A detailed review of the discovery and advances in helicon research is provided in [246] and [247].

When helicon input conditions such as pressure, power, and magnetic field strength are varied over a broad range, helicon discharges operate in several distinct modes [248] namely capacitive, inductive, and helicon-wave modes [249]. Jumps between modes, which are accompanied by dramatic changes in plasma density (by factors of 2 or 3), can arise during smooth variations in the system input variables. The experiments described in this chapter have power settings of 0.8–1.4 kW, where the system operates in an inductive mode. Within each operational mode, the use of a linear estimator model is justified, while multiple linear models could be employed to cover a range of modes.

For the purposes of this study, the helicon plasma source is operated as an isolated plasma production system. There is no ICRH antenna or magnetic funnel on the prototype engine used for experimentation. The flow of Argon gas into the quartz tube, the dc current in the electromagnets, and the RF power delivered to the helicon antenna can all be varied independently. Changes in these input variables result in changes in the plasma generated, with consequent variations in the optical emission from the plasma and heat distribution on the surface of the quartz tube.

## 5.4 Modelling the VASIMR engine

### 5.4.1 Optical data preprocessing

OES data are collected from the plasma downstream from the helicon section, as depicted in Figure 5.4. Several steps are undertaken to extract the features of interest and to restructure the collected data into a form that is useful for temperature estimation.

An Ocean Optics S2000 spectrometer, sampled once per second, is used to collect the OES data from the plasma exhaust of the helicon section of the VASIMR engine. At each sampling instant, the intensity of the plasma optical emission is recorded at 2047 wavelengths between 177 nm and 880 nm with an integration time of 200 ms (see Section 2.5.1). A sample spectrum from the VASIMR plasma is shown in Figure 5.5. Although only Argon is present in the VASIMR plasma, many spectral lines are observed due to different ionisation levels of the argon atoms in the helicon section of the engine. An analysis of the time evolution of the spectral intensity lines reveals that many lines are highly correlated in time, with correlation coefficients greater than 0.75. Due to

the high levels of correlation between the time series of the intensity measurements at each wavelength, principal component analysis (PCA) is used to identify the main uncorrelated, or independent, components that contribute to the variance in these time series.

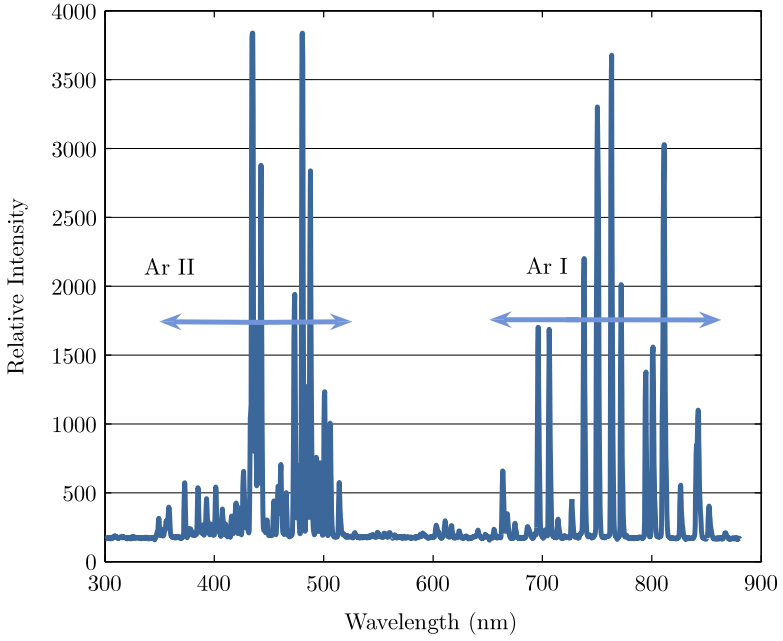



FIGURE 5.5: Spectrum for Argon plasma in the VASIMR engine. Two distinct groups of emission lines can be seen in this figure. Ar I denotes the first ionisation level of an Argon atom. Ar II, the second ionisation level, releases higher energy photons, and hence corresponds to lower emission wavelengths (from Equation (2.7)).

Typically, normalisation to unit variance is performed when the original data have multiple amplitude scales to treat all variables with equal importance during PCA. However, unit variance normalisation is not used during this study since all OES wavelengths are recorded on the same intensity scale. It is not desirable to give the same importance to low-intensity spectral lines with a low signal to noise ratio (SNR) as other, high-intensity spectral lines with higher SNRs.

For the OES data recorded during the VASIMR engine experiments, it is found that just three principal components are capable of representing 97% of the original data variance. Reducing the OES data set from 2047 correlated variables to only 3 orthogonal principal components that can represent the majority of the data variance significantly reduces computational requirements during temperature estimation and shows that the underlying process driving OES variation can be adequately described by three independent time series.

### 5.4.2 Model form

The VASIMR helicon section has three manipulated inputs, namely, the gas flow rate, the electromagnet current, and the RF power delivered to the antenna. A state-space model of the form

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k), \tag{5.2}$$

$$\mathbf{y}(k) = \mathbf{C}\mathbf{x}(k), \tag{5.3}$$

is used to model the system. The system inputs $\mathbf{u} \in \mathbb{R}^{3\times 1}$ correspond to the inputs to the helicon (gas flow rate, electromagnet current, and RF power), the outputs $\mathbf{y} \in \mathbb{R}^{3\times 1}$ are the three principal components arising from the PCA analysis of the OES data which as discussed in Section 5.4.1, and finally, the state vector $\mathbf{x} \in \mathbb{R}^{18\times 1}$ represents the temperatures measured at 18 locations on the gas containment tube. Hence, $\mathbf{A} \in \mathbb{R}^{18\times 18}$, $\mathbf{B} \in \mathbb{R}^{18\times 3}$, and $\mathbf{C} \in \mathbb{R}^{3\times 18}$.

The temperatures are measured using thermocouples bonded to the outside surface of the gas containment tube. Figure 5.6 shows the layout of the thermocouple array. Fifteen thermocouples are arranged along three longitudinal lines of five thermocouples, at angular locations $\theta = \pi/3$, $\pi$, and $5\pi/3$ radians, while three thermocouples are positioned at intermediate angles (0, $2\pi/3$, and $4\pi/3$ radians) between the longitudinal lines. Figure 5.7 shows temperatures recorded from the thermocouples arranged in the longitudinal lines at one particular point in time. The hottest part of the gas tube is near the center in the x-direction, in the region surrounded by the helicon antenna which corresponds to the region of plasma production. Thermocouples are sampled at 1 Hz using a National Instruments analog-to-digital convertor (ADC) interfaced with a LabView$^{\text{TM}}$ software control system for the VASIMR engine.

### 5.4.3 Model identification

For model identification, data records for $\mathbf{u}$, $\mathbf{y}$, and $\mathbf{x}$ are available for various system excitations. The model parameters are determined by first expanding Equation (5.2)

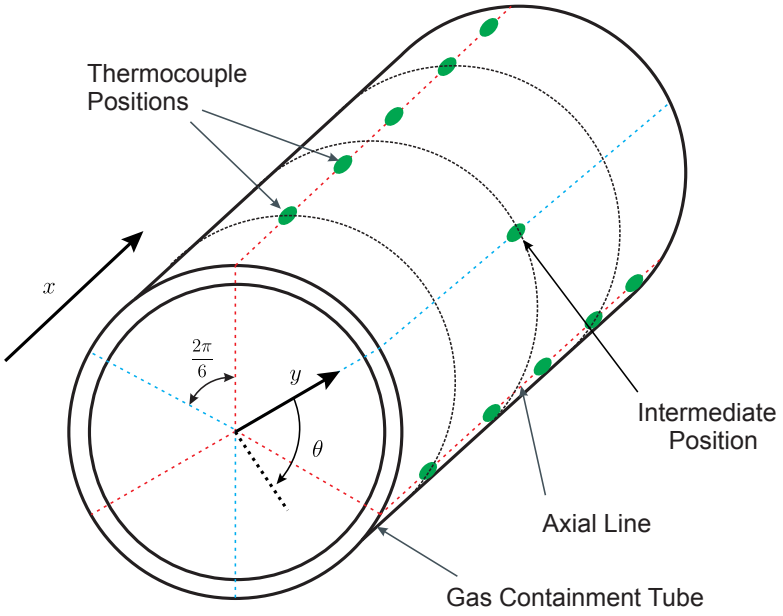



FIGURE 5.6: Thermocouple positions on the gas containment tube. An array of 18 thermocouples is used to record temperature information from the outside surface of the gas tube. The thermocouples are arranged in three longitudinal lines of 5 thermocouples, with 3 extra thermocouples placed in intermediate positions between these lines.

for sample $k$ as

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ \vdots \\ x_n(k+1) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ \vdots \\ x_n(k) \end{bmatrix} \cdots$$
$$\cdots + \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{bmatrix} \begin{bmatrix} u_1(k) \\ u_2(k) \\ \vdots \\ u_m(k) \end{bmatrix}, \tag{5.4}$$

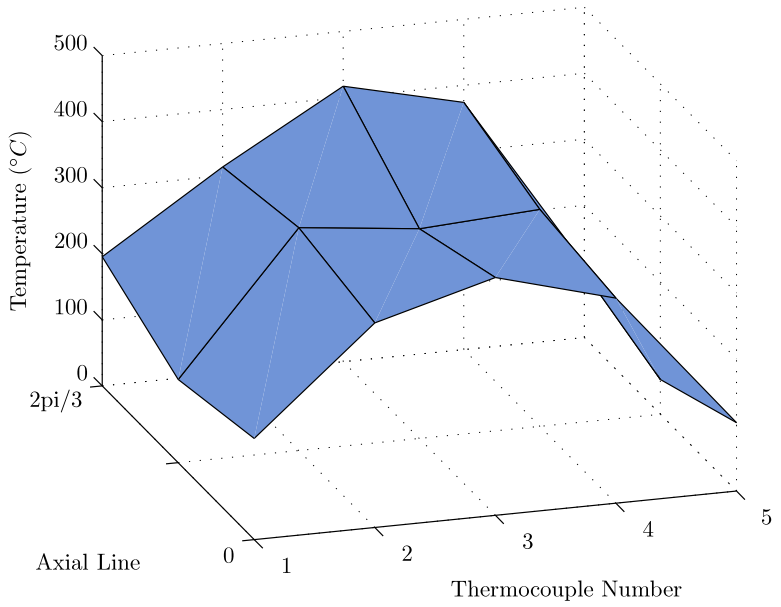



FIGURE 5.7: Sample temperature values recorded from thermocouple array. The region near the center of the tube, near thermocouple positions 2, 3, and 4, corresponds to the region inside the helicon antenna, where plasma is produced and, consequently, where the highest temperatures are recorded.

where $n$ is the number of system states and $m$ is the number of system inputs. The first row of Equation (5.4) can be written, for $k+1, k+2, \cdots, k+N$, as

$$\begin{bmatrix} x_1(k+1) \\ x_1(k+2) \\ \vdots \\ x_1(k+N) \end{bmatrix} = \begin{bmatrix} x_1(k) & \cdots & x_n(k) & & \\ x_1(k+1) & \cdots & x_n(k+1) & \cdots & \\ \vdots & \vdots & \vdots & \cdots & \\ x_1(k+N-1) & \cdots & x_n(k+N-1) & & \end{bmatrix}$$

$$\begin{bmatrix} & u_1(k) & \cdots & u_m(k) & \\ \cdots & u_1(k+1) & \cdots & u_m(k+1) & \\ & \vdots & \vdots & \vdots & \\ \cdots & u_1(k+N-1) & \cdots & u_m(k+N-1) & \end{bmatrix} \begin{bmatrix} a_{11} \\ \vdots \\ a_{1n} \\ b_{11} \\ \vdots \\ b_{1m} \end{bmatrix}. \tag{5.5}$$

Equation (5.5) is of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}, \tag{5.6}$$

which has the least squares solution (as shown in Section 3.1)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{5.7}$$

In this application, $\hat{\boldsymbol{\beta}}$ represents the elements of the first row of the $\mathbf{A}$ and $\mathbf{B}$ matrices of the state-space model in Equation (5.2). To find the $i^{th}$ row of $\mathbf{A}$ and $\mathbf{B}$, the least squares problem is

$$\begin{bmatrix} x_i(k+1) \\ x_i(k+2) \\ \vdots \\ x_i(k+N) \end{bmatrix} = \begin{bmatrix} x_1(k) & \cdots & x_n(k) & & \\ x_1(k+1) & \cdots & x_n(k+1) & \cdots & \\ \vdots & \vdots & \vdots & \cdots & \\ x_1(k+N-1) & \cdots & x_n(k+N-1) & & \end{bmatrix}$$
$$\begin{bmatrix} & u_1(k) & \cdots & u_m(k) & \\ \cdots & u_1(k+1) & \cdots & u_m(k+1) & \\ & \vdots & \vdots & \vdots & \\ \cdots & u_1(k+N-1) & \cdots & u_m(k+N-1) & \end{bmatrix} \begin{bmatrix} a_{i1} \\ \vdots \\ a_{in} \\ b_{i1} \\ \vdots \\ b_{im} \end{bmatrix}. \tag{5.8}$$

To obtain all rows of the estimates $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ of $\mathbf{A}$ and $\mathbf{B}$, a total of $n$ least-squares problems are solved. A similar formulation is used to find $\hat{\mathbf{C}}$, the estimation of $\mathbf{C}$. Expanding Equation (5.3) for sample $k$ as

$$\begin{bmatrix} y_1(k) \\ y_2(k) \\ \vdots \\ y_p(k) \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pn} \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ \vdots \\ x_n(k) \end{bmatrix}. \tag{5.9}$$

The $i^{th}$ row of Equation 5.9 can be written, for $k+1, k+2, \cdots, k+N$, as

$$\begin{bmatrix} y_i(k+1) \\ y_i(k+2) \\ \vdots \\ y_i(k+N) \end{bmatrix} = \begin{bmatrix} x_1(k) & x_1(k) & \cdots & x_n(k) \\ x_1(k+1) & x_1(k+1) & \cdots & x_n(k+1) \\ \vdots & \vdots & \vdots & \vdots \\ x_1(k+N) & x_1(k+N) & \cdots & x_n(k+N) \end{bmatrix} \begin{bmatrix} c_{i1} \\ c_{i2} \\ \vdots \\ c_{in} \end{bmatrix}, \tag{5.10}$$

which is again of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ and can be solved via linear least squares. To solve for the complete $\hat{\mathbf{C}}$ matrix, a total of $p$ such least square problems must be solved, where $p$ is the number of outputs in the state-space model.

### 5.4.4 State estimation

Measurements of only $\mathbf{u}(k)$ and $\mathbf{y}(k)$ are available during normal VASIMR operation, when thermocouple measurements $\mathbf{x}(k)$ are not available. The input/output measurements are used to estimate the state vector $\mathbf{x}(k)$ of gas tube temperatures. In both operational and experimental modes of the VASIMR engine, the values of the input vector $\mathbf{u}$, which are the gas flow rate, the electromagnet current, and the antenna RF power, are set in real time.

With the model described by Equations (5.2) and (5.3), we can predict the state vector $\mathbf{x}(k)$ for a known input sequence $\mathbf{u}(k)$, assuming knowledge of the initial system state $\mathbf{x}(0)$. However, due to inaccuracies in both the model structure and parameters, unmeasured disturbances to the process, and significant uncertainty in the initial system states, state predictions from an open-loop estimation model in this form are rarely of practical value [250].

To decrease sensitivity to inaccurate or unknown initial conditions, a Luenberger observer is used to asymptotically estimate the state. The Luenberger observer [251] incorporates a correction term to the state estimate that is based on the error between the modelled system output $\hat{\mathbf{C}}\hat{\mathbf{x}}(k)$ and the measured output $\mathbf{y}(k)$. The correction term is incorporated in the state equation such that

$$\hat{\mathbf{x}}(k+1) = \hat{\mathbf{A}}\hat{\mathbf{x}}(k) + \hat{\mathbf{B}}\mathbf{u}(k) + \mathbf{L}(\mathbf{y}(k) - \hat{\mathbf{C}}\hat{\mathbf{x}}(k)), \tag{5.11}$$

where $\hat{\mathbf{x}}(k)$ is the estimated state, and $\mathbf{L}(\mathbf{y}(k) - \hat{\mathbf{C}}\hat{\mathbf{x}}(k))$ is the correction term. $\mathbf{L} \in \mathbb{R}^{n\times p}$ is a gain matrix, adjusted to achieve satisfactory estimation error dynamics. The estimator structure is shown in Figure 5.8, where $\hat{\mathbf{x}}^p$ is used to denote the state estimate before correction, that is,

$$\hat{\mathbf{x}}^p(k+1) = \hat{\mathbf{A}}\hat{\mathbf{x}}(k) + \hat{\mathbf{B}}\mathbf{u}(k). \tag{5.12}$$

With the estimation error defined as $\mathbf{e}(k) \triangleq \mathbf{x}(k) - \hat{\mathbf{x}}(k)$, the error dynamics are found by subtracting the estimate of Equation (5.11) from the state shown in Equation (5.2) [250] to give

$$\mathbf{e}(k+1) = (\hat{\mathbf{A}} - \mathbf{L}\hat{\mathbf{C}})\mathbf{e}(k), \tag{5.13}$$

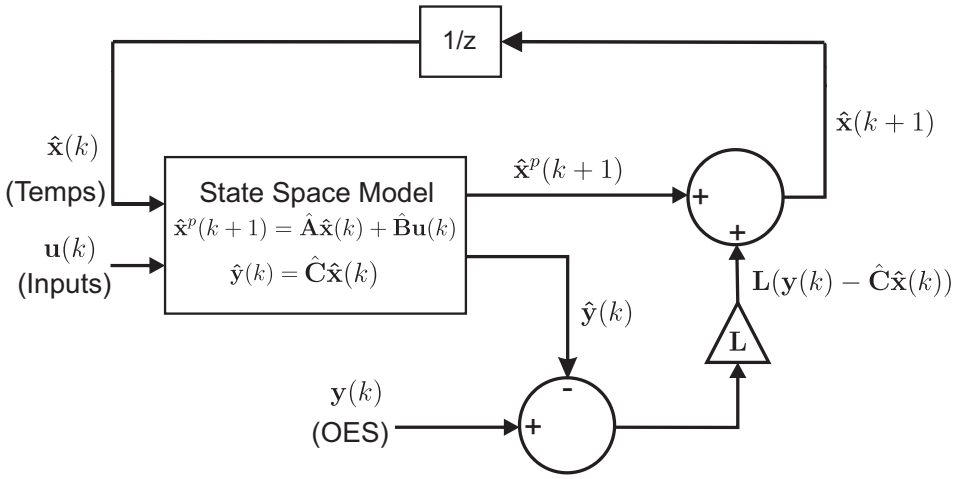



FIGURE 5.8: State-space model with estimation feedback. Errors between the estimated outputs and the measured outputs are used to update the estimated state vectors. In this diagram, $\hat{x}^p$ denotes the state estimate before correction.

assuming that the model $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}$ is sufficiently close to the real $\mathbf{A}, \mathbf{B}, \mathbf{C}$. The characteristic equation of the error is given by

$$\det[z\mathbf{I} - (\hat{\mathbf{A}} - \mathbf{L}\hat{\mathbf{C}})] = 0. \tag{5.14}$$

The desired location of the estimator error poles, which determine the error decay speeds and noise rejection properties, can be specified as

$$z_i = \zeta_1, \zeta_2, \zeta_3, ....\zeta_n. \tag{5.15}$$

The specified desired poles yield a desired characteristic equation

$$\alpha_e(z) = (z - \zeta_1)(z - \zeta_2)(z - \zeta_3)....(z - \zeta_n) \tag{5.16}$$

and by comparing the coefficients in the expansion of Equation (5.16) with the coefficients yielded from Equation (5.14), the desired value of $\mathbf{L}$ can be determined.

An alternative method of determining $\mathbf{L}$ for a desired error response is the use of Ackermann's formula [250] in estimator form, given by

| Value | Low | Mid | High |
|---|---|---|---|
| Antenna Power (W) | 800 | 1100 | 1400 |
| Magnets (A) | 800 | 1000 | 1200 |
| Gas Flow Rate (sccm*) | 100 | N/A | 300 |

TABLE 5.1: Table of VASIMR experimental input levels. Experimental levels for antenna power, electromagnet current, and gas flow are shown. The variations in experimental inputs are deliberately kept small in order to avoid helicon mode changes. All combinations of the levels shown are explored, requiring 18 experiments in total. No mid value is used for the Argon gas flow rate due to operational constraints. (*sccm = standard cubic centimeter per minute)

$$\mathbf{L} = \alpha_e(\hat{\mathbf{A}})\mathbf{O}^{-1}\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}, \text{ where } \mathbf{O} = \begin{bmatrix} \hat{\mathbf{C}} \\ \hat{\mathbf{C}}\hat{\mathbf{A}} \\ \vdots \\ \hat{\mathbf{C}}\hat{\mathbf{A}}^{n-1} \end{bmatrix}. \tag{5.17}$$

Here, $\alpha_e(\hat{\mathbf{A}})$ is the matrix formed by substituting $\hat{\mathbf{A}}$ into Equation (5.16), the characteristic polynomial of the desired closed-loop estimator, and $\mathbf{O}$ is the observability matrix. Even with (relatively small) modelling errors such that the model parameters $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, and $\hat{\mathbf{C}}$ are not exactly equal to $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, $\mathbf{L}$ can be determined to achieve stable error dynamics with acceptably small errors [250].

### 5.4.5 Experimentation

A factorial experiment consisting of 18 experimental input set points is carried out to gather data for model identification. The input vectors are chosen such that the helicon remains in the same operational mode. Perturbations in antenna power, gas flow, and magnet current are introduced as described in Table 5.1. Each input vector corresponds to a different combination of input-variable values, and the resultant gas tube temperatures, monitored using the 18 thermocouples, are allowed to reach steady state where possible, as shown in Figure 5.9. Four of the experiments are repeated to ensure that consistent temperature and OES readings are recorded for repeated input conditions, leading to 22 experiments in total.

The data from the factorial experiment is to be used to create the state-space model as described in Section 5.4.3, which is configured in closed-loop estimator form as shown in Figure 5.8. However, due to differences in the dynamic responses of the OES and the thermocouple data, some signal processing is applied to the OES data prior to modelling, as described in Section 5.4.6.

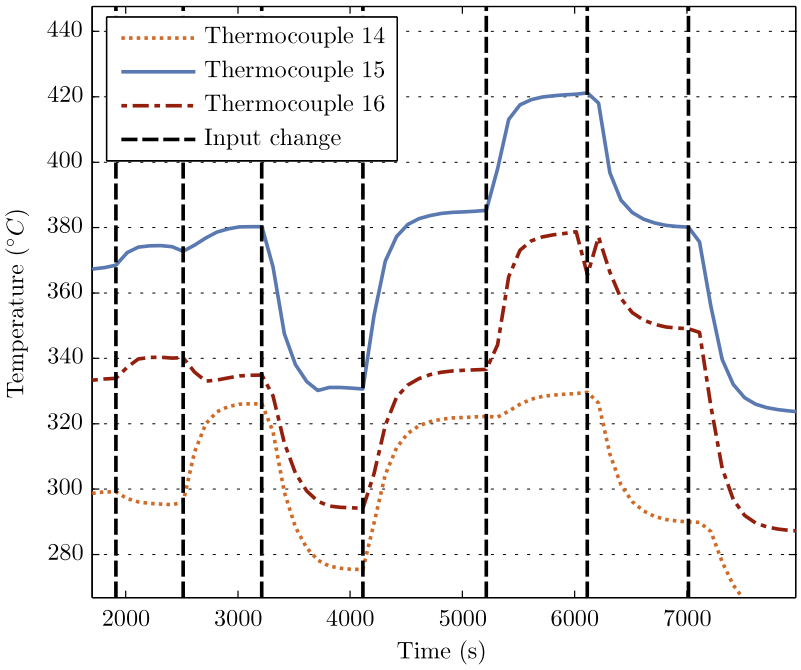



FIGURE 5.9: Example thermocouple measurements during DOE. Abrupt changes in temperatures and temperature distribution are observed when system inputs are changed. Temperatures are allowed to reach steady state before the inputs are changed again.

### 5.4.6 Adjustment of OES time constant

The transients in the OES principal components are relatively fast compared to those of the thermocouples, because the plasma, and correspondingly its light emission, reacts almost instantly to changes in the system inputs, while the gas tube temperatures reach steady state at a slower pace due to thermal inertia of the gas containment tube. The time response for one thermocouple reading is shown in Figure 5.10. The time constant $\tau$ of a system represents the time it takes the step response of the system to reach 63.2% of its final value, and is shown to be approximately 254 seconds for the signal shown in Figure 5.10. The time constant for the OES data, in contrast, is approximately 2.5 seconds.

The discrepancy between the OES and temperature data time constants leads to difficulties in determining a satisfactory output matrix $\mathbf{C}$ for the model. The output state-space equation (Equation 5.3) describes a static, time-invariant linear relationship between the gas tube temperatures in $\mathbf{x}$ and the OES principal components in $\mathbf{y}$. However, while this linear relationship may exist for the steady-state points of operation, it cannot exist for all sample points if the time constants for the OES and temperature data are different.

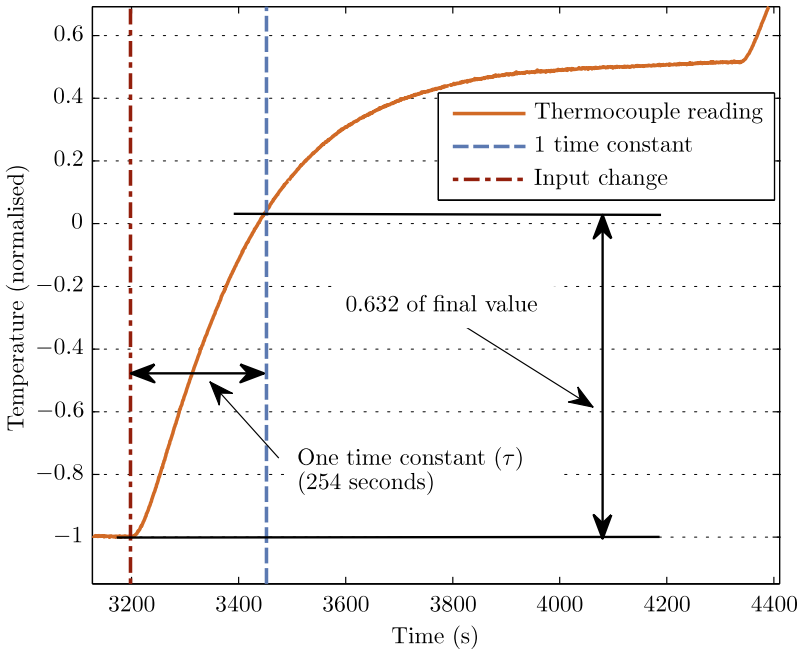



FIGURE 5.10: Time response for thermocouple signal. The time constant $\tau$ for the thermocouple signal shown is approximately 254 seconds.

In order to assist the identification of a satisfactory linear relationship for all samples, the dynamic response of the spectral data is slowed using an exponentially weighted moving average (EWMA) filter. The filtered signal from the EWMA filter is denoted $S(k)$, and

$$S(k+1) = \alpha_{ewma} S(k) + (1 - \alpha_{ewma}) y(k), \tag{5.18}$$

where $y(k)$ is the original OES principal component signal. The principal component signals are slowed to have similar time constants to those of the temperature readings, allowing consistent estimation of a constant output matrix $\mathbf{C}$. Figure 5.11 depicts the effect of several different $\alpha_{ewma}$ values on an OES principal component signal. A filter coefficient of $\alpha_{ewma} = 0.995$ is chosen for the best match between the principal component time constant and the time constant of the thermocouple signal. The time constant for each thermocouple on the gas containment tube is found to dependant on its exact location relative to the antenna. However, a value of $\alpha_{ewma} = 0.995$ suitably filters the principal component signals to suit the average thermocouple time constant values. The EWMA filter also removes noise and erroneous fluctuations from the OES principal component signals, as depicted in Figure 5.12.

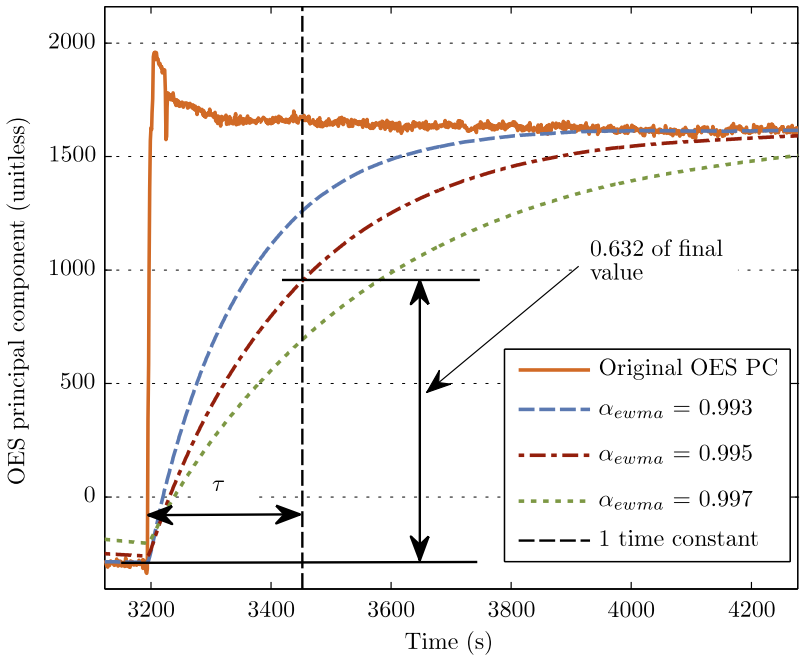



FIGURE 5.11: Effect of different $\alpha_{ewma} = 0.995$ on an OES principal component. A value of $\alpha_{ewma} = 0.995$ is ultimately chosen to adjust the principal component time constants to be close to that of the temperature signals.

## 5.5 Performance and results

### 5.5.1 Output equation validation

To validate the the output equation (Equation (5.3)) of the state-space model, Figure 5.13 compares two predicted principal components $\hat{\mathbf{y}}$ produced by the Equation (5.3) when driven with the recorded temperature data $\mathbf{x}(k)$, to the real principal components of the OES data recorded, $\mathbf{y}$. The relationship between the states and the components is adequately represented by the linear relationship $\mathbf{y}(k) = \hat{\mathbf{C}}\mathbf{x}(k)$, given the quality of the model/data match in the figure. The agreement between the estimated outputs and the real outputs confirms the existence of an approximate static and linear relationship between the gas tube temperatures and the EWMA-filtered OES principal components, simultaneously validating the choice of $\alpha_{ewma}$.

### 5.5.2 Multi-step prediction performance

As a further test of the state-space model performance, the model is configured in an open-loop manner. In this configuration, no feedback term is included to correct the state estimates, corresponding to $L = 0$ in Equation (5.11). With $L = 0$, the state

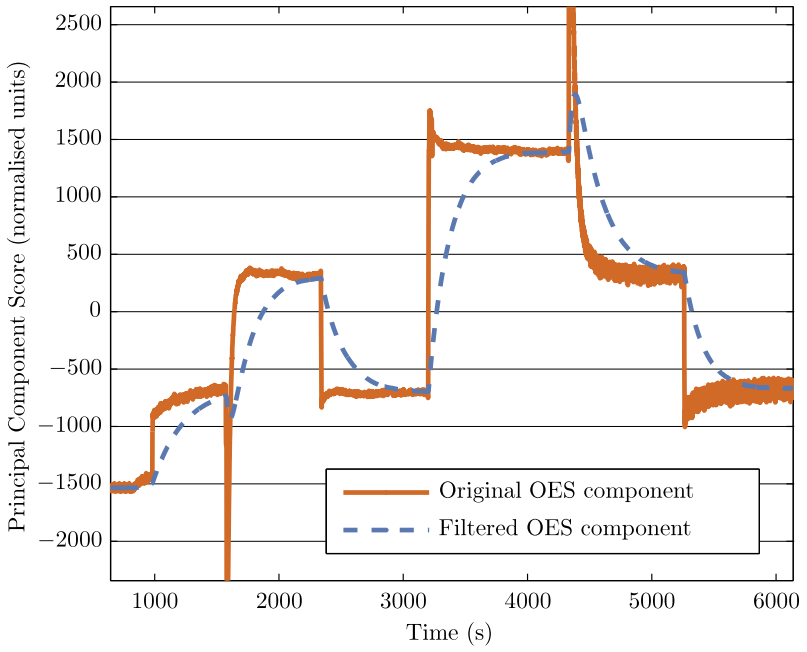



FIGURE 5.12: Application of exponentially weighted moving average filter to optical emission spectrometer (OES) data. The filter serves to slow the transients of the OES data, allowing a linear relationship between OES and temperature data to be found.

estimate $\hat{\mathbf{x}}(k+1)$ is independent of the measured output $\mathbf{y}(k)$ and depends only on the VASIMR engine inputs $\mathbf{u}(k)$ and the previous state estimate $\hat{\mathbf{x}}(k)$, such that

$$\hat{\mathbf{x}}(k+1) = \hat{\mathbf{A}}\hat{\mathbf{x}}(k) + \hat{\mathbf{B}}\mathbf{u}(k) \tag{5.19}$$

Two conditions are tested. Firstly, when configured in this open-loop manner and given accurate initial conditions such that $\hat{\mathbf{x}}(0) = \mathbf{x}(0)$, the temperature estimates are found to remain reasonably accurate with changes in inputs for $k > 0$. The state-space model can estimate future temperatures with a root mean squared error (RMSE) of 2.1%.

However, in a real VASIMR engine implementation, the precise initial temperatures of the system are not known, since no absolute measurement of temperature is available. The second test simulates a situation with unknown initial conditions. The model is tested with a random initialisation of $\hat{\mathbf{x}}(0)$ to investigate the evolution of the state estimates over time. As expected, larger errors are observed for unknown initial conditions, with temperature estimates remaining inaccurate for the duration of the test, and an approximately constant offset error is observed.

The two conditions are demonstrated in Figure 5.14, which shows the evolution of two estimated system states as examples, representing two thermocouple temperature

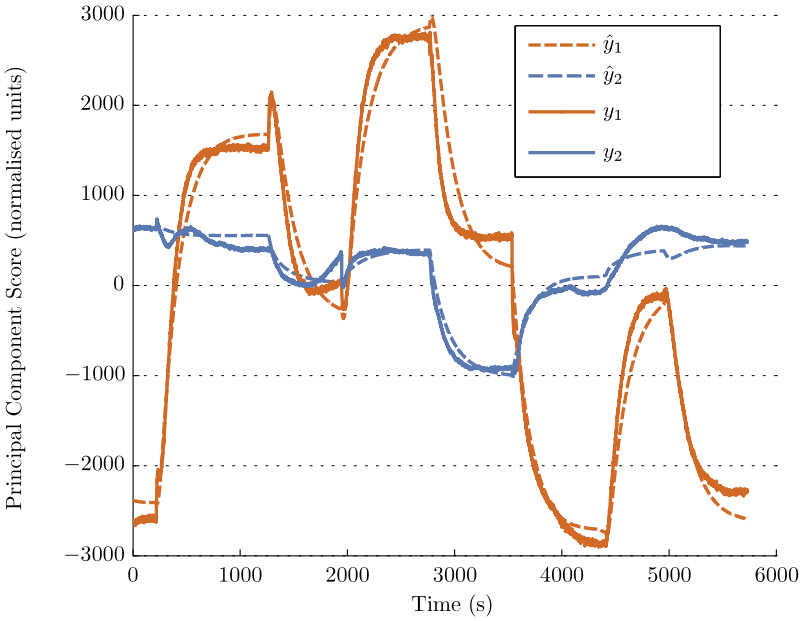



FIGURE 5.13: Comparison of the model outputs $\hat{\mathbf{y}}$ in response to real state vectors $\mathbf{x}$ with the recorded system outputs $\mathbf{y}$. The system outputs are the principal components of the optical emission spectroscopy data. The model output equation $\mathbf{y}(k) = \hat{\mathbf{C}}\mathbf{x}(k)$ emulates the optical emission principal component scores when driven by the real system states.

measurements on the gas containment tube. Figure 5.15 shows the open-loop behaviour of the system outputs (the OES measurements) subjected to the same tests, depicting two of the three OES principal components on the same axis. As is the case with the estimated state evolution, the output estimates are inaccurate for unknown initial conditions for $k > 0$.

### 5.5.3 Closed-loop estimation

To accurately correct the state estimates when the initial states of the system are unknown, feed-back of the measured system outputs is introduced. Equation (5.11) is used to update the state estimate $\hat{\mathbf{x}}(k+1)$ in response to the model output estimation error. The estimator transient response is to be set to be slightly faster than the transients found in the model state matrix $\hat{\mathbf{A}}$ using Ackermann's formula. As in Section 5.5.2, a random initial condition is used to simulate the scenario where the temperature is unknown at estimator startup. Figure 5.16 shows the evolution of a selection of estimated states with the estimator configured in closed-loop form. For comparison, the inaccurate open-loop state estimates are also shown on Figure 5.16. Figure 5.17 shows the corresponding evolution of the model outputs for the system in both open-loop and closed-loop configurations.

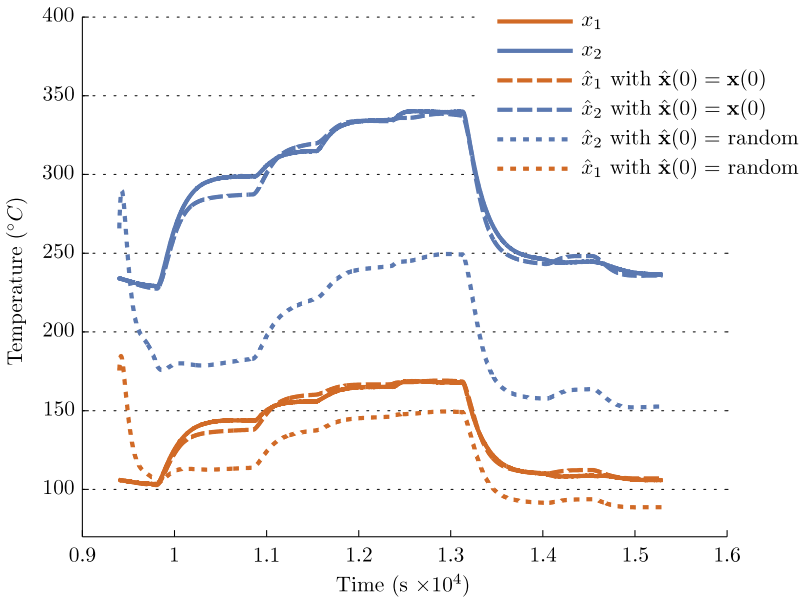



FIGURE 5.14: Evolution of two estimated state vectors for state-space model with differing initial conditions. Greater errors are observed for cases with inaccurate initial conditions. These inaccuracies are expected since no error feedback exists in this configuration, leaving estimates uncorrected.

For the closed-loop results, the estimated outputs and the estimated states converge toward the true values over time as a result of the error feed-back implementation, even with unknown initial conditions. The provision of feed-back correction removes the open-loop requirement of exact initial conditions for estimator accuracy and gives a RMSE of $\sim 2\%$ after the estimator converges.

## 5.6 Summary and conclusions

The VASIMR propulsion system is an plasma propulsion system for space craft that uses magnetic fields to accelerate plasma to produce thrust. Undesired heat produced in the helicon section of VASIMR must be monitored and removed safely to avoid damage to system components, especially when higher power operating regimes are explored. This chapter demonstrates a strategy for virtual metrology of distributed temperatures in VASIMR, based on OES measurements and a state-space model where the states represent the distributed temperature profile. OES provides a non-invasive measurement technique that is used as an output correction term for the VM scheme.

In this application, it is shown that the 2047 OES channels recorded from the VASIMR plasma can be accurately represented by only 3 principal components for temperature estimation. Use of the principal components as corrector terms in the

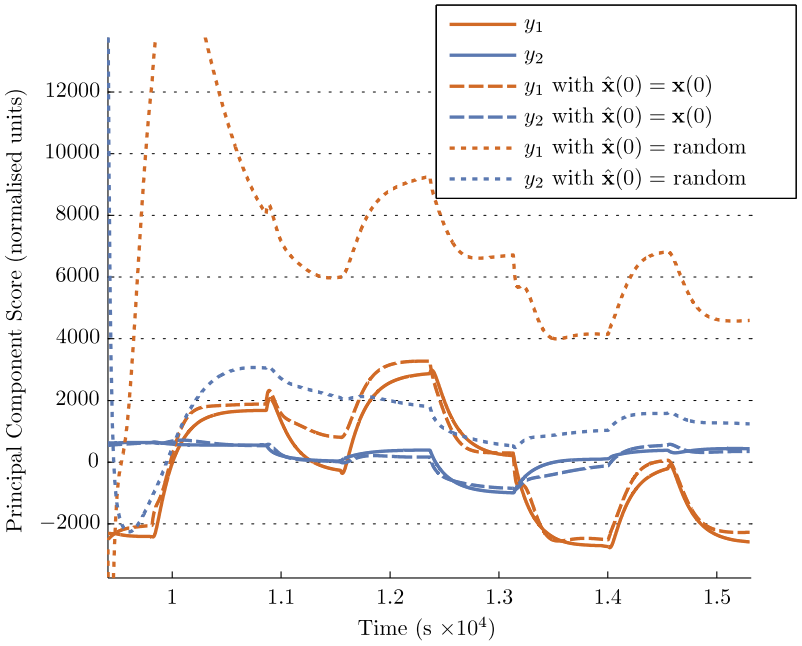



FIGURE 5.15: Evolution of estimated outputs for the state-space model with different initial conditions. With inaccurate initial conditions, the output of the model does not follow the real output.

state-space model dramatically improves the capability of the model to recover from unknown initial conditions and multiple system input changes. Although the experiments completed in this chapter are performed off-line using pre-recorded data, the VM calculations are straightforward computationally and could be performed with relatively few adjustments for real time temperature estimation.

The results of this chapter were published in the Irish Signals and Systems conference (ISSC) [252] and the IEEE Control Systems Magazine [253].

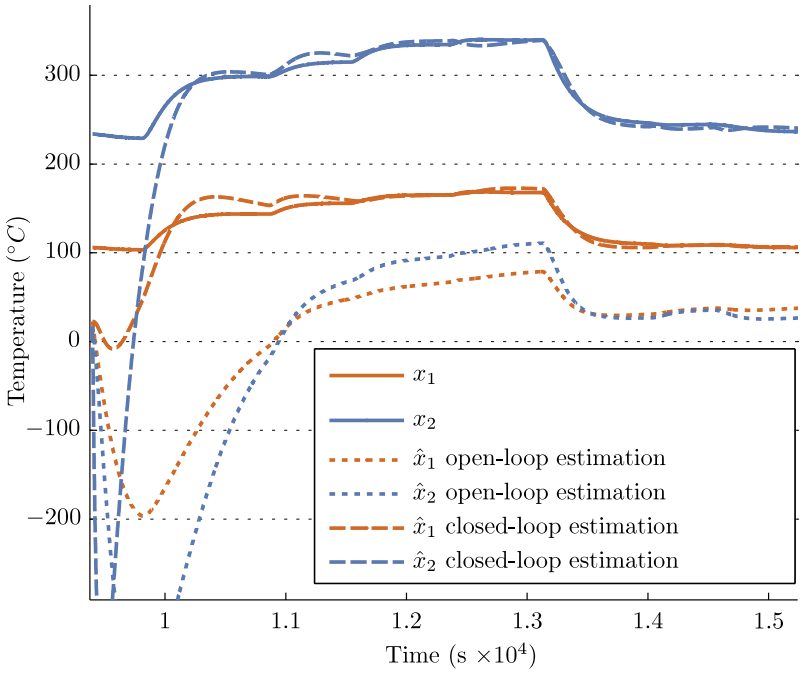



FIGURE 5.16: Comparison of state prediction performance with and without error feedback. With error feedback in place, state estimates converge to the true state values. This case illustrates a realistic condition, where initial conditions are unknown.

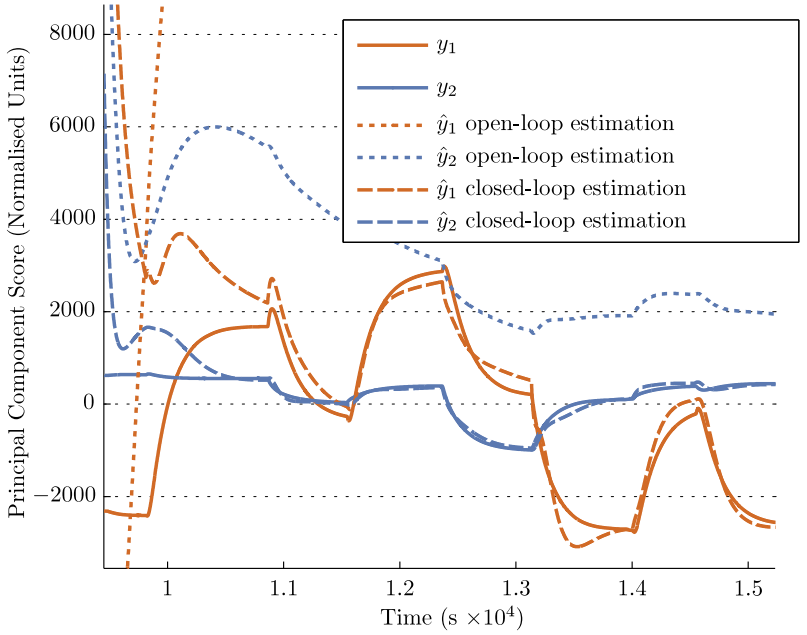



FIGURE 5.17: Comparison of output prediction performance with and without error feedback. Output predictions converge to the true outputs when output feedback is in operation. In contrast, the open-loop estimation scheme produces inaccurate predictions (estimates for $\hat{y}_1$ are off the scale of this figure).

# Chapter 6

# Global modelling of a plasma etch process

In this chapter, the performance of global virtual metrology (VM) models are investigated using an industrial etch process. In this context, *global* models refer to models that are trained on a set of data and are then, without changes to the model, used to estimate etch rate for all further wafers. The etch process studied is a multi-step *shallow trench isolation* (STI) etch process in which trenches are formed on exposed areas of a 200-mm diameter multilayered silicon wafer. Such trenches are ultimately filled with insulating material to prevent electrical current leakage between components that are formed in later processes. STI increases transistor-packing density, allowing for more functionality and speed per unit area on wafer surfaces. In this chapter, details of the etch process are first provided; following this, the global VM estimation performance is examined.

## 6.1 Etch process description

This section provides an overview of the STI etch process being examined, describes where this particular etch process fits into the semiconductor manufacturing cycle, details all of the measurements taken from the etch chamber, and addresses possible sources of disturbances to the etch performance. For confidentiality reasons, details on the exact chemistries and dimensions used in the process are not provided.

### 6.1.1 Pre-etch wafer preparation

Silicon wafers undergo hundreds of process steps over approximately 30–40 days before completion in a typical production cycle and the particular STI etch process examined here is completed within the first ten processes.

Prior to the plasma etch process, layers of material are deposited on the surface of each silicon wafer. First, a relatively thin insulating layer of silicon dioxide (the *pad-oxide* layer) that insulates later layers from the wafer substrate material is thermally grown on the bare silicon. On top of the insulating layer, a relatively thick layer (thicker than the insulating layer by a factor of approximately 40) of silicon nitride is deposited using a plasma-enhanced chemical vapour deposition process. The silicon nitride layer acts as a hard mask during etching. The deposition and thermal growth processes are well controlled, with typical standard deviations of 2% and 0.4% of the overall thickness of the layers for the pad-oxide and nitride layers, respectively.

A layer of photoresist material is deposited on top of the nitride layer by spin coating. Precise patterns are developed in the photoresist using ultraviolet photolithography. Undeveloped photoresist is removed using a developer solution, exposing the underlying nitride material where etching is required. An overview of photolithography technology can be found in [27].

### 6.1.2 Plasma etch

The STI etch process comprises five different steps. Each step entails adjustments in the chamber gas flow rates, the power applied to the powered electrode, and the pressure in the etch chamber. A cross-sectional view of the wafer stack with the aims of the five etch process steps labelled is presented in Figure 6.1 and each step is explained below:

1. A two-gas plasma is used to remove approximately 75% of the exposed nitride layer in a timed etch.

2. A less aggressive etching chemistry is used to etch the remaining nitride layer. Step 2 ceases in response to an endpoint signal from a monochromator installed on the etch chamber. Hence, the time required for this etch step varies depending on the etch performance.

3. Step 3 is an overetch step of a fixed duration. The overetch step ensures that no silicon nitrite remains in the etched trenches so that the insulating pad-oxide layer beneath is fully revealed. Depending on the amount of remaining silicon nitride in

the trenches after Step 2, some of the pad-oxide layer is also removed during this step.

4. The main trench etch is completed during Step 4, where a trench is etched in the underlying silicon. This is a timed etch, the duration of which is decided off-line by operators who attempt to compensate for drifting process parameters by adjusting the etch time manually. A relatively aggressive etch chemistry is used.

5. The final step in the process is a timed etch to round the bottom of the etched trenches and ensure the correct trench profile is achieved in all trenches on the wafer surface. The etching gases are changed for this final step to achieve the desired trench profile.

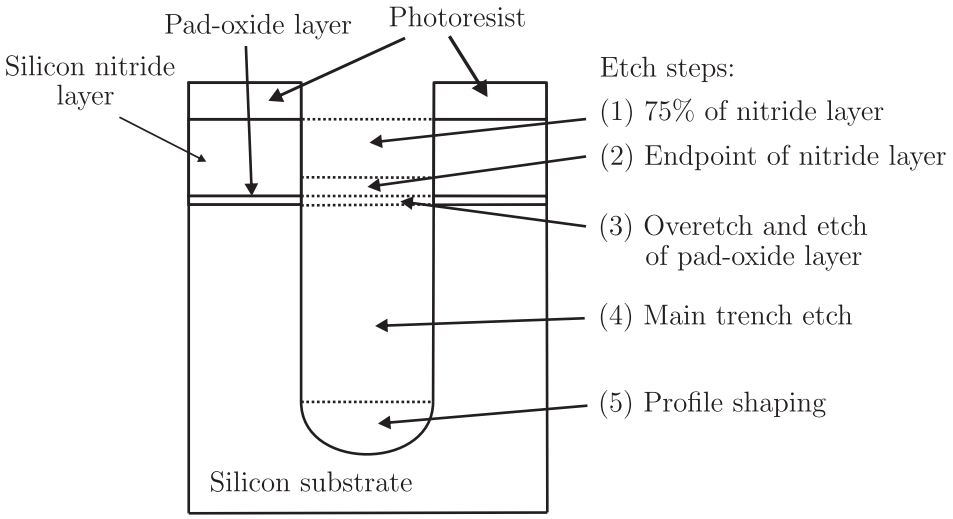



FIGURE 6.1: Cross-sectional view of wafer stack (not to scale). Trenches are etched onto the wafer surface through a number of layered materials using a multi-step etch process. This diagram shows the layers grown on the silicon wafer surface, along with the material removed by each of the process steps.

### 6.1.3 Post etch processing

Each wafer undergoes several further processes after plasma etching before a measurement of the trench depth is carried out. The photoresist material is first removed in the ashing and sulphuric cleaning processes. Several iterations of oxide growth and chemical mechanical planarisation (CMP) are then completed to fill each trench with silicon dioxide. Multiple iterations are used to ensure that trenches are filled uniformly and that no air pockets are inadvertently formed. A phenomenon known as *dishing* can occur during CMP, during which material is removed from trench openings accidentally, leaving a non planar surface at the top of the trench. Finally, the nitride and pad-oxide

layers are stripped (using hydrofluoric and phosphoric acid cleaning processes), leaving the isolation trenches on the wafer surface with areas of silicon between them where transistors are built in later manufacturing processes.

The depth of the trench is measured via reflectometry after the trenches have been filled completely with silicon dioxide. Although dishing can affect the trench depth measurement for some etch processes, the effects are relatively small ($< 1\%$) when compared to the overall trench depth for the process being studied. The trench is filled before depth measurement so that the reflectometrer can be used to identify interference between reflections from the top and bottom of the silicon dioxide.

### 6.1.4 Process drift

The greatest difficulty encountered during modelling and control of plasma etch processes is that etch chamber conditions drift over time, altering process behaviour and yielding unpredictable changes in etch depth. The drift in process behaviour occurs as the chamber becomes conditioned by repeated etch operations. The conditioning process occurs because contaminants from etch products and sputtered material from the wafer surface are deposited on the walls of the chamber during wafer processing. The chamber temperature drifts over each lot (batch of wafers) processed, varying across each lot as more wafers are etched. On longer time scales, physical components on the etch machine behave differently as they age and wear. Periods of idle time where processing is not completed can also significantly contort the chamber performance.

### 6.1.5 Preventative maintenance

In an attempt to maintain a consistent etch performance, the drifting behaviour of the process is counteracted through the use of *preventative maintenance* (PM) operations. PM operations are maintenance operations carried out after a specific period of time, or after a specific number of wafers are processed, even if the chamber is still operating within specifications. PM operations are completed after approximately every 1000 wafers processed on each chamber for the process studied in this thesis.

There are two classifications of PM events that are carried out. Normal PM operations involve a physical clean of the inner chamber walls, the replacement of ceramic chamber components and the chamber electrodes, and a full check on the operation of the chamber. More comprehensive maintenance is carried out on a quarterly basis when the vacuum seals on the chamber and the wafer loading mechanisms are replaced.

During PM events, the etch tools are unavailable for wafer processing for up to fourteen hours, representing a significant overhead in a semiconductor fab.

The aim of the PM operations is to restore the etch chamber to a nominal initial state. Figure 6.2 shows the mean endpoint monochromator output for each wafer over a period of four PM cycles. Deposition of materials inside the plasma chamber clouds the viewing window for the monochromator as wafers are processed, resulting in lower amplitude outputs. During PM operations, the viewing window is cleaned, restoring the clear view of the plasma, creating a sawtooth-like signal shown in Figure 6.2. The conditioning effect is also reflected in electrical measurements from the chamber, as demonstrated in Figure 6.3, where the mean impedance during Step 4 of each wafer is displayed. The changes brought about by PM events present difficulties for construction of consistent etch rate VM models.

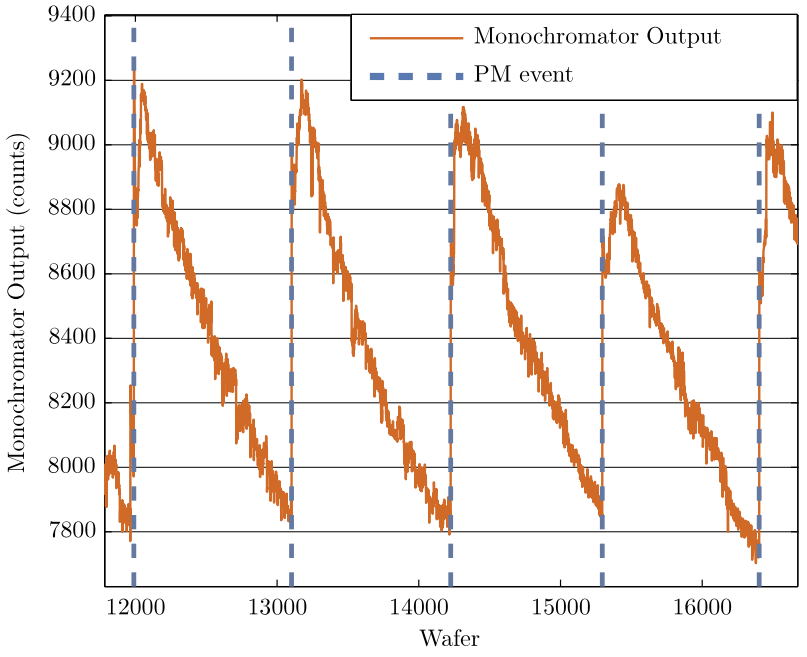



FIGURE 6.2: Endpoint monochromator output over four preventative maintenance (PM) cycles. Sputtered material and etch products deposited on the chamber viewing window attenuate the monochromator signal over each PM cycle. After each PM event, the window returns to normal operation and the monochromator signal increases in intensity.

The product mix also hinders the creation of consistent models of the etch chamber. Wafers processed in the chamber come from different product lines, which are mixed in a relatively random fashion and processed in order of arrival. The same etching chemistry is used to etch each product, but as a result of different open area ratios on the wafer surfaces and hence different etch responses, inconsistent etch depths are observed from product to product.

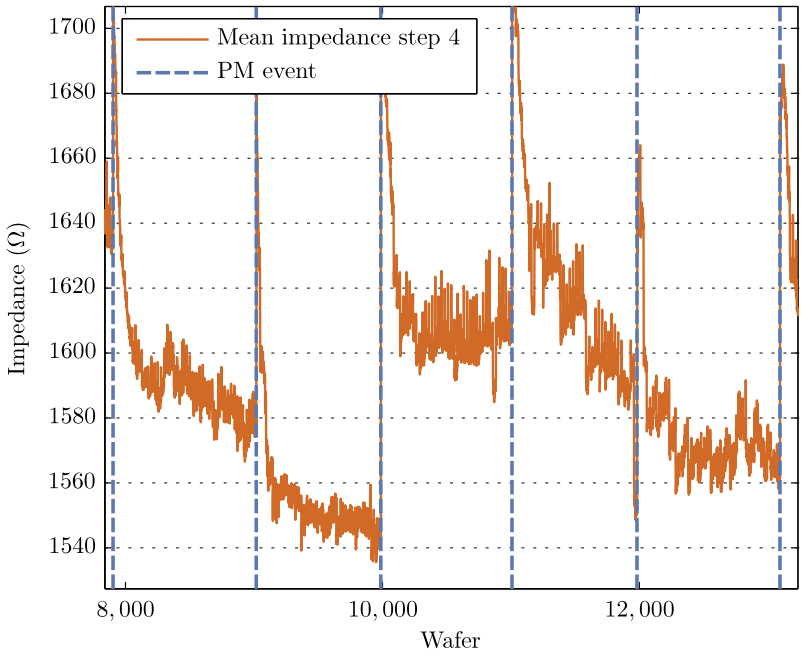



FIGURE 6.3: Mean impedance for Step 4 of each wafer. The impedance values undergo a drifting pattern due to chamber conditioning as wafers are processed along with discontinuities after preventative maintenance (PM) events.

Three mechanisms are in place to reduce the impact that chamber drift and PM cycles have on the etch process behaviour:

1. Etch rate tests are carried out after each PM event to check that the chamber is operating within specification. Wafers with blanket layers of different materials are used for this purpose, where measurements of layer thickness before and after processing, along with the etch time, are used to calculate etch rate. If the chamber operates out of an allowable range, further PM operations are carried out until a satisfactory etch performance is achieved.

2. After each PM event, conditioning wafers are etched in the etch chamber before production material. The conditioning wafers are bare silicon wafers with no photoresist. By etching these wafers the chamber is conditioned before production material is etched. Even with this safeguard in place, in very sensitive processes, up to 400 non-critical product wafers are processed in each chamber after PM events before products requiring stricter control of etch depth are etched.

3. Chemical cleaning procedures are carried out before each lot of wafers is processed. Chemical cleans involve running corrosive plasmas in the etch chamber when empty to partly remove wall deposits from previous etching operations.

In some etch processes, external disturbances to the etch depth achieved can arise from upstream and downstream (pre-measurement) processes. For the process under study in this thesis, the layers of material formed on the product wafers are created using techniques that are reasonably well controlled and assumed to have negligible deviation from expected values. Measurements of these variations are unavailable. The error in etch depth measurements is also assumed to be negligible (see Section 6.2).

Although the procedures outlined above help to keep the chamber etch rate within specifications, completely predictable and consistent etch depths are still not achieved. As discussed in Section 6.1.6, a level of manual and automatic control is used to ensure the minimum of product waste.

### 6.1.6 Current control methods

The plasma etch process under study is controlled using statistical process control (SPC), where a control chart is analysed daily by a trained engineer who makes decisions regarding the operation of the process. SPC is used with process measurements in fabrication plants to warn operators of unusual deviations and trends in the measurements in order to avoid impending product waste and process failure.

One of the principal determinant variables for the performance of the etch process product material is known as the n-well resistance, or $\rho_{n-well}$. The $\rho_{n-well}$ for each wafer is measured electrically at end of line, but there is a linear relationship between $\rho_{n-well}$ and trench depth, given by

$$\rho_{n-well} = C_1 d - C_2, \tag{6.1}$$

where $d$ is the trench depth, and $C_1$ and $C_2$ are constant values that depend on the process. Estimates for $\rho_{n-well}$ are calculated using measurements of the etch depth taken downstream from the plasma etch process (discussed further in Section 6.2) and plotted on control charts, where SPC rules are applied. Deviations from normal behaviour trigger alerts to process engineers who diagnose any problems and take action to keep the process in control.

Figure 6.4 shows the $\rho_{n-well}$ values for a number of measured wafers and the SPC limits used for process monitoring. The SPC limits shown in diagram are not symmetrically distributed around the target $\rho_{n-well}$ value because there is greater tolerance for errors resulting in over etched trenches than for errors resulting in under etched trenches. Trenches that are deeper than specifications may still effectively electrically

isolate transistors, whereas shallow trenches can result in crosstalk between transistors in finished products.

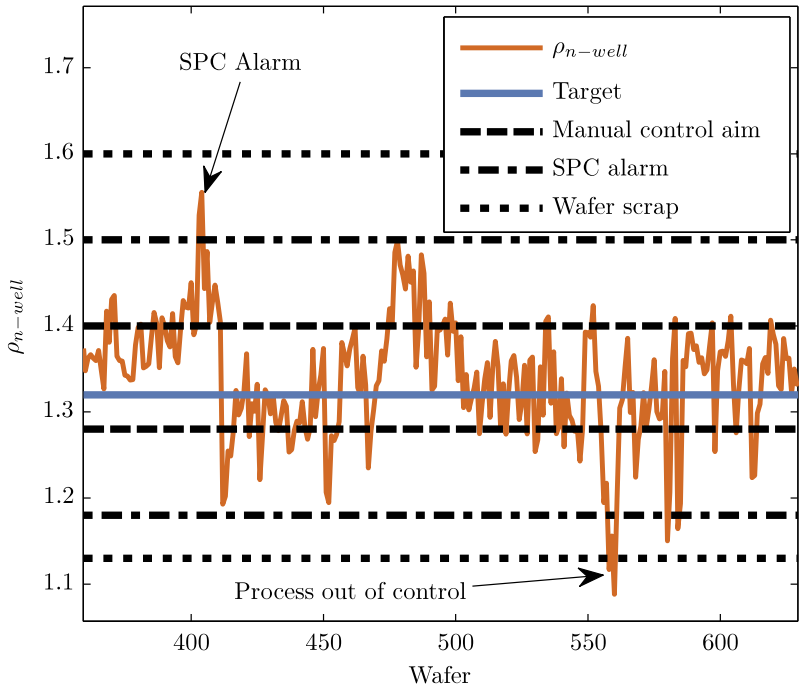



FIGURE 6.4: Estimates of $\rho_{n-well}$ from etch depth for a set of wafers with statistical process control (SPC) limits marked. Manual control is implemented in an effort to keep $\rho_{n-well}$ values within the inner limits. Automatic SPC alarms trigger at the second set of limits, where engineers examine the machine. The outer limits require all wafers to be tested for trench etch depth, and usually result in wafer scrap.

Operators monitor the etch depth achieved by the process and adjust the main trench etch time, that is the Step 4 etch time, to account for drift in the chamber. The main etch time is specified in seconds, and ranges from $73 - 81$ seconds. The Step 4 etch time is also adjusted manually to account for changes in product material, catering for different open areas on the wafer surfaces. This manual, experience-based control of the system is sufficient to maintain the trench depths of the product material within specifications most of the time. However, with a metrology delay in the process of up to two days, significant deviations in etch performance can be costly in terms of wafer scrap before corrective action is implemented.

## 6.2 Measurements recorded

Sensors measure a number of variables during the etch process to monitor the status of the plasma etch chamber and to monitor whether the etch process is behaving within specifications. The data analysed in this chapter are obtained from three sources:

etch process (EP) data recorded from the plasma tool, plasma impedance information recorded from a plasma impedance monitor (PIM) sensor installed on the electric interface to the etch chamber, and optical measurements of etch depth taken some time after the etch process is completed.

### 6.2.1 Etch process data

Sensors pre-installed on the etch processing tool record tool related variables such as the chamber pressure, chamber temperatures, matchbox component positions, and the gas flow rates during the etch process.

Data are sampled at a frequency of approximately 1 Hz and are recorded on a central database in the fabrication plant. These data are used both in an on-line fashion by the chamber to warn of irregularities during operation and in an off-line fashion by operators to discover possible causes of error for product wafers that do not perform to specifications at end-of-line testing.

Summary statistics for each process step are calculated from the time series data to reduce the amount of data for processing and to summarise the chamber behaviour for each process step. The summary statistics are sufficient to identify wafer-level trends and patterns in the etch process. The extracted statistics include mean values, standard deviations, maximum values, minimum values, and signal ranges. These statistics are extracted for each process step after the plasma has been given time to settle, based on predefined time ranges as depicted in Figure 6.5. The time ranges are typically triggered on the applied power signal.

Table 6.1 shows the variables recorded from the etch tool for step 4, the main etch step. A complete listing of the collected variables for all process steps is provided in Appendix A.

### 6.2.2 Plasma impedance monitor (PIM)

As described in Section 2.5.8, a plasma impedance monitor (PIM) is a sensor that is installed between the matching network and the plasma electrodes that measures the harmonics of current and voltage that are generated on the electrical supply lines to the chamber by the non-linear impedance of the etching plasma.

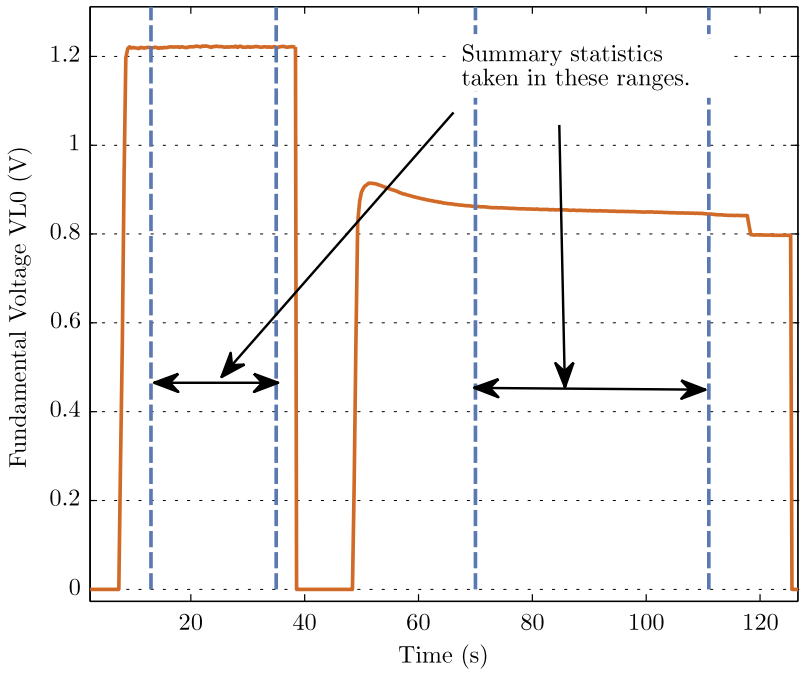



FIGURE 6.5: Example extraction of summary statistics from time series data. This figure shows an example of the sections of PIM time series data used for the extraction of mean and standard deviation statistics. The diagram depicts two steps of the etch process.

In the trench etch process described in this chapter, PIM measurements are available at the fundamental frequency 13.56 MHz, and at 52 harmonics of this frequency for the main etch step of the plasma etch process.

Two separate Fourier analyses are performed by a digital signal processor in the PIM unit as it collects the RF sensor measurements. The first analysis calculates the lower frequency PIM variables, comprising values of voltage ($V$), current ($I$), and phase ($\phi$) for the fundamental frequency and the first four harmonics, collectively named the *lower PIM* variables. The second analysis is based on a different algorithm and produces $V$, $I$, and $\phi$ at the fundamental frequency value, along with 52 harmonics, forming the *upper PIM* variables. Both analyses operate on the same principals, and calculate amplitudes for each harmonic of voltage, current, and the phase angle between them.

Only the values for current and voltage recorded at the fundamental frequency are calibrated such that the recording scale corresponds correctly to units of amps and volts respectively. The first five harmonics in the upper PIM variables are attenuated as a result of a filtering effect in the processing units. To form a complete set of harmonics from the PIM sensor, the lower PIM variables are used as measures of the first five harmonics, and the remaining harmonics, from 5–52, are taken from the upper PIM

| Variable Name | Description |
|---|---|
| RANGE-ENDPOINT_A | Range of endpoint signal recorded from chamber monochromator A |
| MEAN-ENDPOINT_A | Mean value of endpoint signal recorded from chamber monochromator A |
| MEAN-GAS_FLOW2 | Mean value of MFC2 flow rate |
| STD-DEV-GAS_FLOW2 | Standard deviation of MFC2 flow rate |
| MEAN-Gas_FLOW7 | Mean value of MFC7 flow rate |
| STD-DEV-GAS_FLOW7 | Standard deviation of MFC7 flow rate |
| MEAN-GAS_FLOW6 | Mean value of MFC6 flow rate |
| STD-DEV-GAS_FLOW6 | Standard deviation of MFC6 flow rate |
| MEAN-GAP | Mean gap between upper and lower electrodes |
| MEAN-UP_ELECT_TEMP | Mean temperature of upper electrode |
| MEAN-LOW_ELECT_TEMP | Mean temperature of lower electrode |
| MEAN-TV_ANGLE | Mean throttle valve angle |
| STD-DEV-TV_ANGLE | Standard deviation of throttle valve angle |
| MEAN-CHAMBER_PRESS | Mean value of chamber pressure |
| STD-DEV-CHAMBER_PRESS | Standard deviation of chamber pressure |
| MEAN-RF_LINE_IMP | Mean value of RF line impedance |
| MEAN-RF_LOAD_COIL_POS | Mean position of matchbox load coil |
| MEAN-RF_LOAD_MATCH_PH | Phase between pre and post match RF waveforms |
| MEAN-RF_MATCH_1_TUNE | Mean position of matchbox tune coil |
| MEAN-RF_REFLECTED | Mean value of RF reflected from chamber |
| STD-DEV-RF_REFLECTED | Standard deviation of RF power reflected from chamber |
| MEAN-RF_MATCH_1_DC_BIAS | Mean DC Bias recorded on wafer chuck |
| MEAN-RF_FORWARD_GEN | Mean RF power generated at generator |
| STD-DEV-RF_FORWARD_GEN | Standard deviation of RF power generated |

Table 6.1: Recorded parameters for step 4 of trench etch process.

variables. The exact operation of the Fourier transform algorithms used is proprietary information belonging to Lam Research (Ireland) Ltd.

### 6.2.3 Etch depth measurement

Measurements of the etched trench depth are made downstream from the etch process in the production cycle to monitor the etch process performance. There is a metrology

delay of up to two days before the etch depth measurements are carried out because the wafer undergoes further processing steps before metrology is completed, as described in Section 6.1.3. Ultimately, the final product performance depends on measurements of $\rho_{n-well}$ at the end of the production line. However, $\rho_{n_well}$ data is not available for the data sets investigated in this chapter.

As discussed in section 6.1.3, trenches are filled with silicon dioxide after the etch process. Measurement of trench depth is carried out optically using an *Optiprobe* measurement tool. The OptiProbe operates using beam profile reflectometry (BPR) to determine the thickness of silicon dioxide deposited in the etched trenches on the wafer surface. BPR measures the intensity of light reflected from the surface of the wafer as a function of the incoming angle of reflection. The reflected intensity changes in relation to the angle of reflection, optical properties of the measured material and thickness of the material being measured. The material thickness is determined by fitting the parameters of previously defined models to the recorded intensities.

The trench depth is measured at nineteen different points across each wafer and the mean depth of these points is taken as the average etch depth for the wafer.

The measurement of trench depth for every wafer is impossible in a high volume manufacturing environment due to the time overhead required. Hence, etch depth measurements are carried out as intermittently as possible while still providing enough information to monitor process performance. Wafers are processed in lots of 25 wafers in the fabrication plant and while EP and PIM data are recorded for every wafer, trench depth metrology is only performed for two wafers from approximately every third lot, resulting in an average metrology rate of 4 wafers in 75, or approximately 5%. In measured lots, the etch depths for wafers 13 and 25 are measured. Wafers 13 and 25 are chosen as wafers that provide etch depth measurements that are indicative of the average etch depth for the complete lot. Figure 6.6 shows a recorded EP variable and indicates the wafers that undergo etch depth metrology.

The OptiProbe measurement system is an accurate non-destructive method of depth measurement with a typical error of 0.83 nm. With a total trench depth of over 500 nm, this error is negligible as a percentage.

For experimental work, process development, or troubleshooting, further information on the etch process performance can be obtained by physically slicing etched wafers and taking scanning electron microscope (SEM) photographs of the side profile of the etched trenches. Collecting profile information with SEM photographs requires the destruction of product wafers and as a result is seldom carried out.

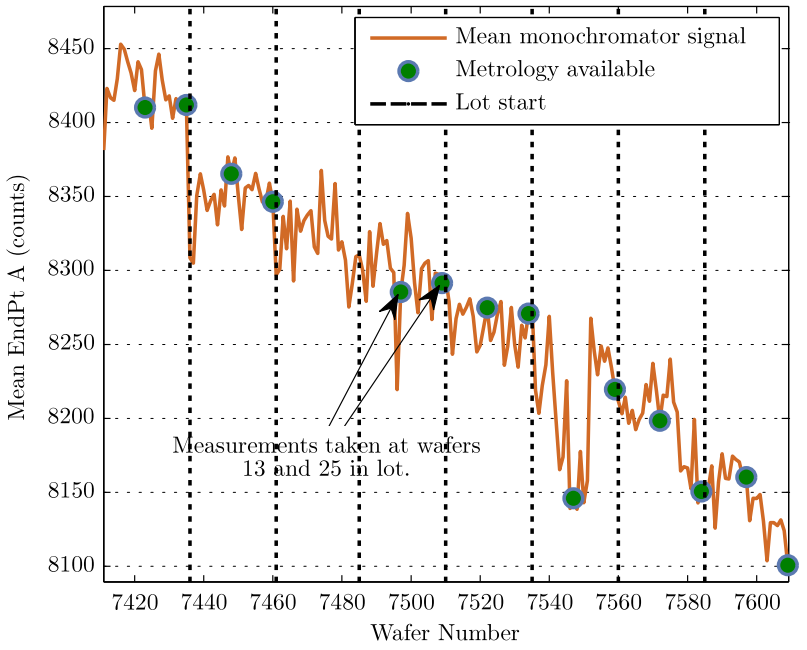



FIGURE 6.6: Frequency of etch depth metrology. This figure depicts an etch process (EP) variable, the mean value of one of the endpoint monochromators installed on the chamber, for every wafer processed. The wafers for which etch depth is measured are marked. This particular section of process data has a relatively high density of etch depth measurements.

## 6.3 Data set description

VM models are constructed for the industrial plasma etch process described in Section 6.1. This section describes the data sets collected from the process for VM and the preprocessing undertaken to prepare the data for modelling.

### 6.3.1 Data collection

Data were collected from two etch chambers over a period of approximately six months. As a result of logging errors and unexpected changes in the manufacturing environment, the data set is not complete, and does not include all of the wafers processed throughout the data collection period. At different times, sensors failed to measure variables accurately, network servers failed to record data, and data became corrupted during processing. Table 6.2 summarises the collected data from each etch chamber.

For Tool 1, EP data for 6238 wafers were recorded without the corresponding PIM data and there are 3330 such wafers in the data set from Tool 2. The missing data are unrecoverable as they were lost at the time of processing due to sensor failures in the

| | **Tool 1** | **Tool 2** |
|---|---|---|
| Start date | 07/06/2008 | 11/6/2008 |
| End date | 20/11/2008 | 26/8/2008 |
| Number of wafers with EP data | 18524 | 10718 |
| Number of wafers with PIM data | 12286 | 7388 |

TABLE 6.2: Summary of etch data collected. Data are collected from two separate plasma etch chambers over a period of approximately six months.

fabrication plant. There are a number of chronological gaps in the data that can be seen in Figure 6.7 where the data are depicted with respect to the time of collection.

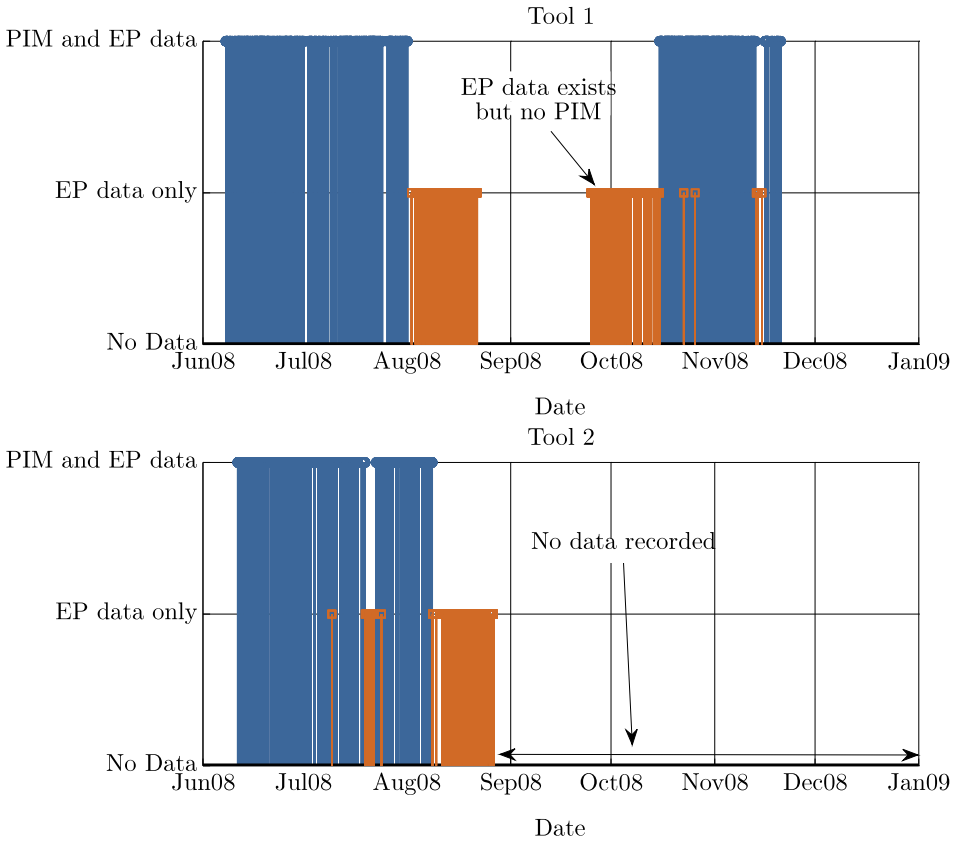



FIGURE 6.7: Data available for virtual metrology. Errors in the fabrication plant data collection system resulted in a large number of wafers for which there was only EP data recorded. Blank spaces represent dates with no data available, even though there may have been wafers processed during these times.

### 6.3.2 Raw data treatment

The sensor data from the plasma etch process are recorded routinely in the fabrication plant and stored for a limited time on network servers. Data are stored in a binary format consisting of time-series traces for all of the variables measured. As described

in Section 6.2.1, summary statistics for variables of interest are extracted from the time series data.

EP data are recorded along with the associated *context data*. Context data are made up of a time-stamp, lot number, and wafer number for each wafer, along with details of the processing tool used. Due to the particular computer infrastructure in the fabrication environment, the PIM data are stored in a different location without any context data apart from a time-stamp for each wafer. For the purposes of VM in this thesis, the EP and PIM data files are imported into the MATLAB® environment and the PIM data are matched to the corresponding EP data on the basis of the time-stamp to create a unified data set for further analysis and processing.

Downstream measurements of trench depth and the corresponding $\rho_{n-well}$ estimates are collected from a third location in the fabrication plant. These data points contain context data for each lot measured, and so can be assigned to the correct wafer data in the data set.

Because the etch time for the main etch step is varied manually by operators in the plant as detailed in Section 6.1.6, the VM models in this thesis focus on the estimation of overall average *etch rate*, which is calculated by dividing the measured trench depth by the time taken for the main etch step.

### 6.3.3 Variable removal

EP data comprising a total of 131 different variables are collected over the five steps in the etch process. Visual inspection of these variables, and an application of engineering process experience, can immediately discount a number of the variables as irrelevant for the purposes of VM of trench etch rate.

Firstly, the trenches are etched in step 4 of the process, and so it is a valid and logical step to discard variables from Steps 1 and 2, where no etching of the silicon trenches is carried out. Secondly, a number of variables can be removed that contain standard deviation statistics and variables that remain constant or have very little variance over the data set, for examples see Figure 6.8. In this manner, using visual analysis and process experience, a further 41 variables are removed from steps 3, 4, and 5. A complete listing of the removed variables is provided in Appendix B.

After the removal of the more obviously superfluous EP variables as described, 184 different variables remain in the data set, comprising 28 EP variables and 156 PIM variables (52 harmonics each of $V$,$I$, and $\phi$).

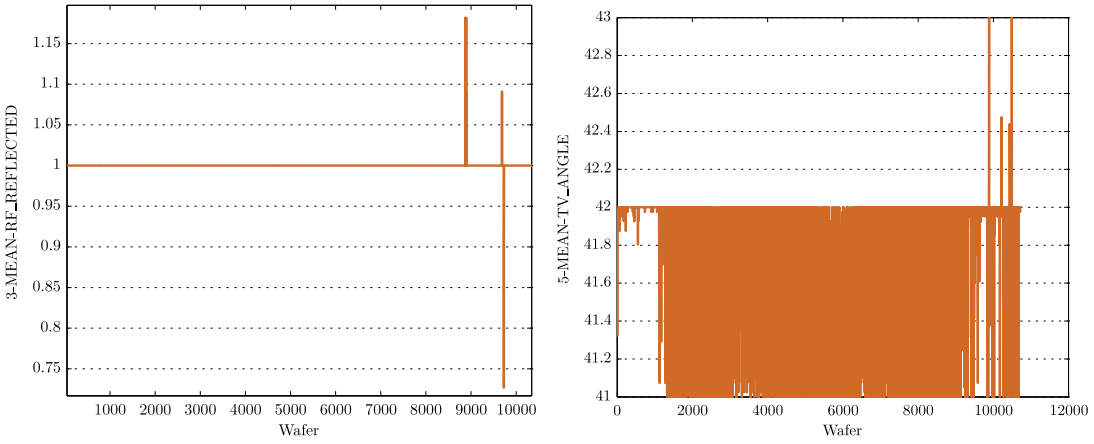



FIGURE 6.8: Examples of variables removed from etch process (EP) data. Variables with little or no information content, or those that relate to parameters not connected with the main etch performance are removed.

### 6.3.4 Outlier Removal

Erroneous variable values are recorded in the data as a result of mechanical or electronic glitches that occur during the operation of the etch tools and the recording processes. Samples that contain errors are normally numerically distant from the majority of the data and are considered to be outliers. The identification and removal of such data points is advantageous as they can result in inaccurate models if used during VM model training.

PCA along with the $T^2$ and $Q$ statistics (see Section 3.5) are used to detect outliers in the trench etch data sets. PCA is carried out on the EP data and the PIM data separately. Samples are deemed as outliers when they exhibit unusually high $T^2$ or $Q$ statistics. The removal thresholds for each statistic are determined manually through inspection of the statistics for each data set.

The EP data is mean centered and the standard deviation of each variable is normalised during the PCA analysis, since the data contains a number of different variables measured on different scales. A principal component model is constructed using the normalised data. Figures 6.9 and 6.10 show the $T^2$ and $Q$-statistics calculated using the EP data for each wafer from Tool 1 and Tool 2 respectively, along with the chosen cutoff threshold for outliers. The $Q$-statistic is calculated using a principal component model that explains 95.4% of the data set variance using 15 of the 28 available components. All variables are included in the calculation of the $T^2$ statistic. Wafers which exhibit substantially higher (by a factor of 5) $T^2$ and $Q$-statistic values than the majority of the other wafers are identified as outliers.

For example, in the EP data, the tool variable 5-MEAN-RF_LOAD_MATCH_PH shown in Figure 6.11 was recorded incorrectly for wafer 4015 from Tool 1. The value recorded in error is clearly not within the normal distribution of the variable, and this deviation results in large $T^2$ and Q values for the wafer in question, as depicted in Figure 6.9. Similar deviations in other EP variables are responsible for the high $Q$ or $T^2$ values of other wafers shown in Figure 6.9.

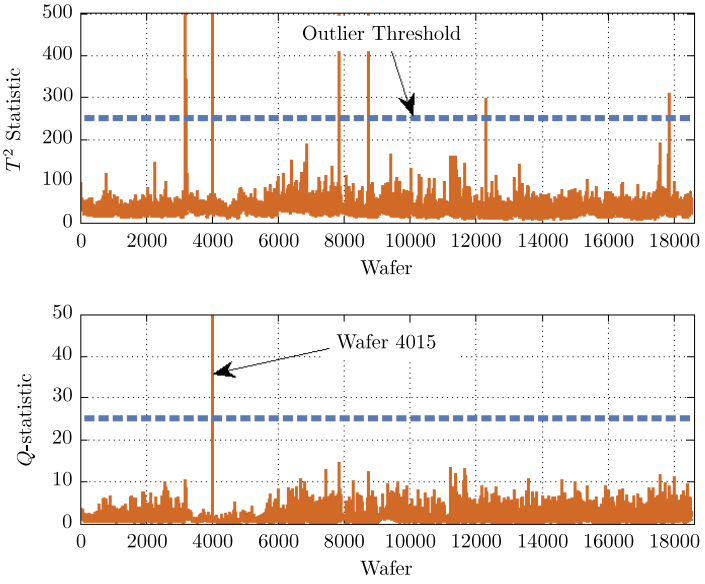



FIGURE 6.9: $Q$ and $T^2$-statistics of each wafer recorded from Tool 1 for EP data. Wafers with unusually high values for $T^2$ or $Q$ values exhibit behaviour that is different to rest of the data set.

The outlier detection process is repeated for the PIM variables. Each PIM variable is mean centered and normalised to have unit variance. A number of wafers appear as outliers as seen in Figure 6.12 and Figure 6.13, with signals such as those shown in Figure 6.14 responsible for the outliers.

The outlying wafers with statistics above the thresholds shown on in Figures 6.9, 6.10, 6.12, and 6.13 are removed from the data set.

### 6.3.5 Final data set contents

After the removal of the variables deemed unnecessary for VM and the removal of outlying wafer data, a data set remains that is used for the development of VM models for plasma etch rate.

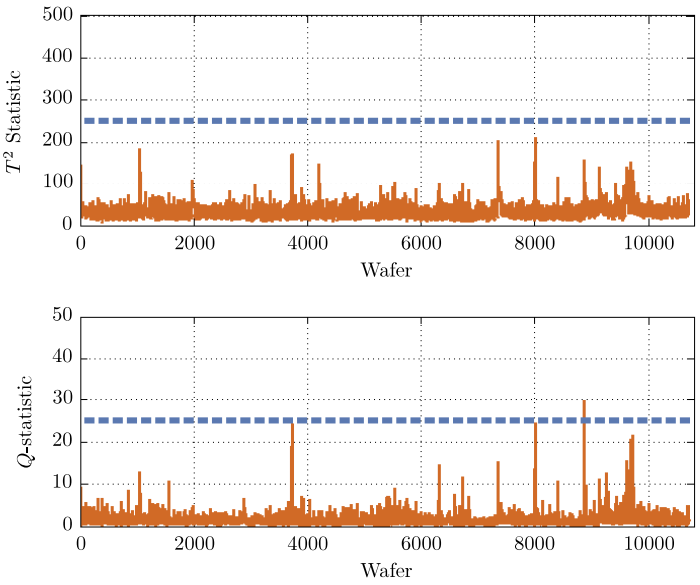



FIGURE 6.10: $Q$ and $T^2$-statistics of each wafer recorded from Tool 2 for EP data. The data from Tool 2 contains fewer extreme outliers than the data from Tool 1 shown in Figure 6.9

Because a substantial number of wafers from each tool are missing the PIM data, the data is divided into those wafers with EP data only, and the subset of those that have both PIM and EP data. Table 6.3 summarises the wafers for which EP data are available, and Table 6.4 summarises those wafers for which both EP and PIM data are available.

| | **Tool 1** | **Tool 2** |
|---|---|---|
| Number of wafers with EP data | 18400 | 10715 |
| Number of etch depth measurements | 789 | 358 |
| Measurement frequency | 4.29 % | 3.34 % |
| Number of PM cycles | 18 | 11 |

TABLE 6.3: Wafer information available with EP data for every wafer.

| | **Tool 1** | **Tool 2** |
|---|---|---|
| Number of wafers with EP and PIM data | 12133 | 7365 |
| Number of etch depth measurements | 529 | 245 |
| Measurement frequency | 4.36 % | 3.33 % |
| Number of PM cycles | 12 | 8 |

TABLE 6.4: Wafer information available with EP and PIM data for every wafer.

The data set containing both EP and PIM data is used during global modelling of the plasma etch rate so that comparisons between the etch rate estimates made using

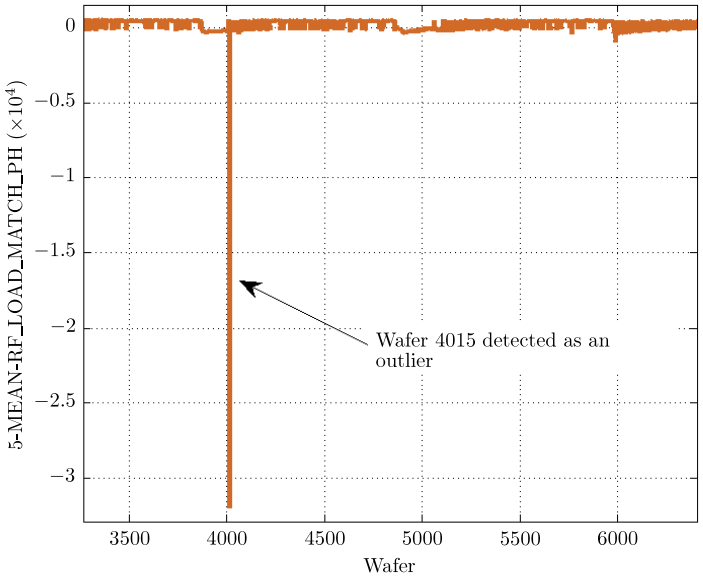



FIGURE 6.11: Etch process parameter "5-MEAN-RF_LOAD_MATCH_PH" for all wafers from Tool 1. The erroneous value for "5-MEAN-RF_LOAD_MATCH_PH" recorded for wafer 4015 causes corresponding high $T^2$ and $Q$-statistic values for this wafer in Figures 6.9 and 6.10.

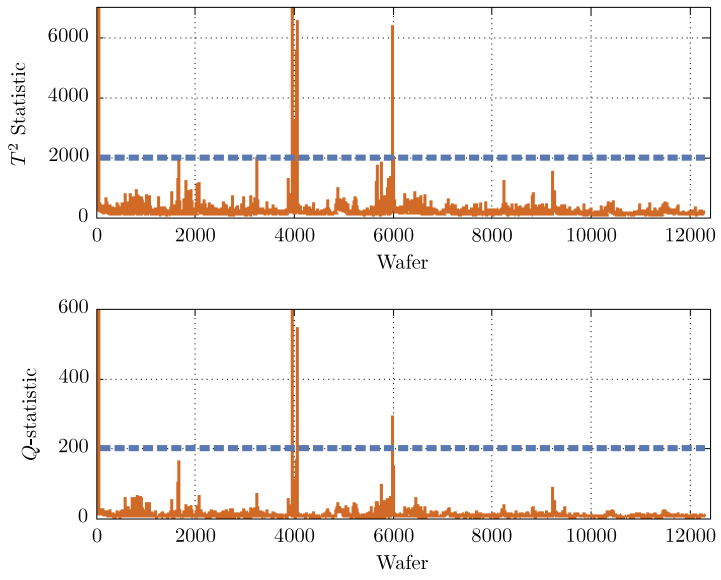



FIGURE 6.12: $Q$ and $T^2$-statistics for PIM data of each wafer recorded from Tool 1

different sensor information can be made.

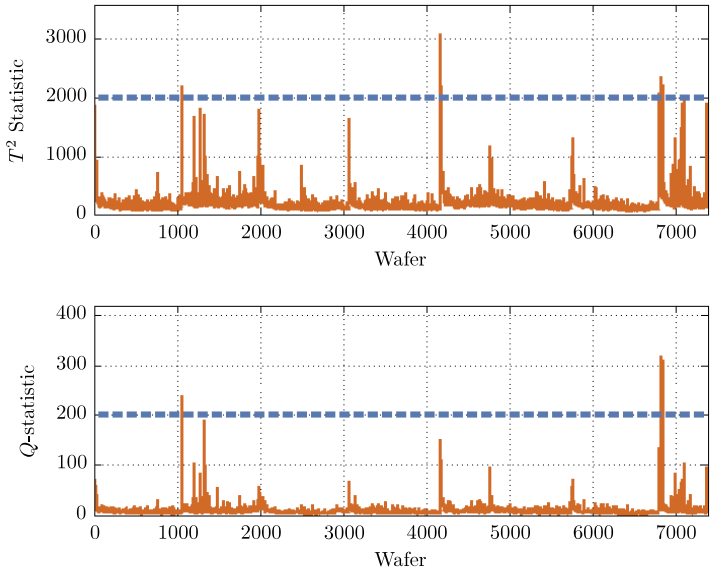



FIGURE 6.13: $Q$ and $T^2$-statistics for PIM data of each wafer recorded from Tool 2

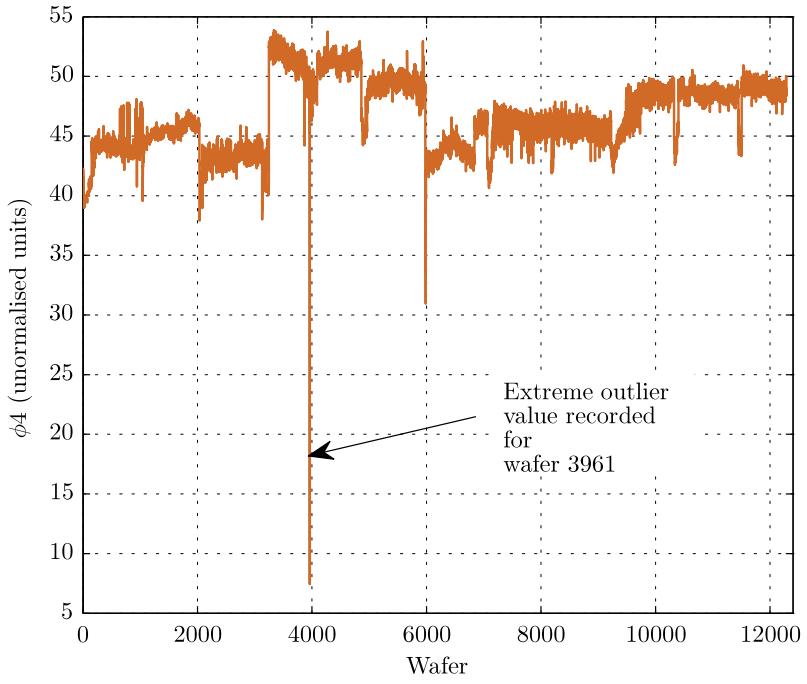



FIGURE 6.14: The fourth harmonic of current from the plasma impedance monitor (PIM) variables $\phi 4$ for Tool 1. The $\phi 4$ variable contains a potentially erroneous value for wafer 3961 leading to extraordinarily high $T^2$ and $Q$ statistics for this wafer.

## 6.4 Virtual metrology approach

Static models are employed to estimate the etch rate from the EP and PIM measurements. The VM models focus on the estimation of etch rate, rather than $\rho_{n-well}$ because

physical measurements of $\rho_{n-well}$ are unavailable. The use of time-series models is precluded because the etch rate measurements are not performed on a uniform basis. The VM models are of the form

$$\hat{r}(k) = f(u_1(k), u_2(k), \cdots, u_m(k)), \tag{6.2}$$

where $\hat{r}(k)$ is the estimated etch rate value for wafer $k$ and $u_1(k), u_2(k), \cdots, u_m(k)$ are the measurements taken from the chamber sensors during etching of wafer $k$. It is assumed that the relationship between the sensor measurements and the plasma etch rate is time-invariant for the global models.

To examine the performance of VM models for different circumstances, fifteen different variable selection/data reduction schemes and two data division schemes are examined.

### 6.4.1 Training, validation, and test data subsets

As described in Chapter 3, model estimation performance can be substantially more accurate when tested using the data upon which the model is trained, whereas this level of accuracy may not be representative of the model performance on previously *unseen* data. Hence, different data are used for model building and for model testing. The data used during the building of models are the *training* data and the unseen data used for model evaluation are the *test data*. A third subset of *validation* data is sometimes extracted from the training data before the training procedure and used to aid early-stopping or cross validation procedures where necessary, as described in Section 3.4.3.

Ideally, the training data for global models are chosen so that they capture information from the complete operating space of the system. Data from specifically designed experiments are often used to train global models [11, 221]. In this work, however, only past production data are used for model training. Although designed experiments are preferable at times to capture more variation, the high value nature of semiconductor processing means that such experiments can be prohibitively expensive in terms of wafer scrap and tool down-time.

The proportion of data points used for the training, validation, and test data sets is a user preference. For the purposes of this thesis, 50% of the wafers are set aside for model training, 20% for model validation, and the remaining 30% are retained as unseen data for model testing during model creation procedures.

### 6.4.2 Input combinations

Fifteen different combinations of recorded variables are created from the data available for use as input variables to virtual metrology models. The different input selections are identified for the remainder of this thesis by the text given in **bold** in the list below.

1. **EP**: EP Data only. This variable selection contains only the variables recorded directly from the etch tool. All PIM-derived variables are ignored. This data set is the most cost effective since no additional sensors beyond the original etch hardware are required. There are 28 EP variables included in this selection from steps 3-5 of the etch process.

2. **EP-Step**: EP Data with stepwise selection. The stepwise selection algorithm is applied to the EP data set before modelling. The reduced set of variables selected by the stepwise algorithm are used as the inputs to the VM models.

3. **PIM$_0$**: PIM data for the fundamental harmonic only. This set of variables contains the PIM data for the fundamental values of voltage, current, and phase from the PIM sensor for Step 4 along with the calculated values for impedance, power, reactance and resistance (Equations (2.34) - (2.37)). The fundamental values of current, voltage, phase, and power listed in Appendix A for steps 3 and 5 of the etch process are also included. In total, 19 variables are included in this combination of variables.

4. **PIM$_0$-Step**: PIM$_0$ variables reduced using stepwise selection. The stepwise selection algorithm is used to identify a subset of the PIM$_0$ variables at the fundamental frequency.

5. **PIM$_5$**: PIM data for the first five harmonics. This corresponds to the mean values of the first five harmonics each of current, voltage and phase that are collected by the PIM sensor for each wafer.

6. **PIM$_5$-PCA**: This data set consists of principal components extracted from the PIM$_5$ variables. Enough principal components are included such that 90% of the variance in the data is explained.

7. **PIM$_5$-Step**: PIM$_5$ variables reduced using stepwise selection. The stepwise selection algorithm is used to identify a subset of the PIM$_5$ data to be used as input variables for the VM models.

8. **PIM**: Complete PIM data. This selection of input variables contains all 52 harmonics current, voltage and phase that are collected by the PIM sensor.

9. **PIM-PCA**: The input combination consists of principal components extracted from the complete set of PIM variables. Sufficient principal components to explain 90% of the data variance are included.

10. **PIM-Step**: Complete set of PIM variables reduced using stepwise selection. Stepwise selection is employed to select a subset of input variables from the complete collection of PIM variables.

11. **XRZP$_5$**: Calculations for reactance (X), resistance (R), impedance (Z), and power (P) for the first five harmonics for the main etch step (Step 4). Equations (2.34) - (2.37) are used to calculate the values.

12. **XRZP$_5$-PCA**: Principal components extracted using PCA analysis of the XRZP$_5$ variables. Sufficient components to explain 90% of the data variance are included as model inputs.

13. **XRZP$_5$-Step**: Stepwise selection of variables from the XRZP$_5$ variables.

14. **EP-PIM$_0$**: Combination of all EP variables with PIM$_0$ variables.

15. **EP-PIM$_0$-Step**: Stepwise selection of the EP-PIM$_0$ variables.

### 6.4.3 Chronological and interleaved data

The training, validation and test data sets are constructed in two separate ways to examine the capabilities of the VM schemes.

1. Firstly, to examine the time extrapolation capabilities of developed models, the data set is kept in chronological order such that the training and validation data are formed using wafers processed before the wafers in the test data, as depicted in Figure 6.15. The first 70% of the wafers form the training and validation data. The validation data is interleaved within the training set, and makes up 20% of the full data set. The test data comprises the final 30% of the wafers.

2. Secondly, to investigate whether model accuracy is improved when information from the same operational region as that of the test wafers is included in the model training data sets, an *interleaved* data set is used, where the training and test wafers are interleaved throughout the complete data set. The data are divided for training, test, and validation processes in the same proportions as for the chronological data set. A depiction of the interleaved data division scheme is presented in Figure 6.16.

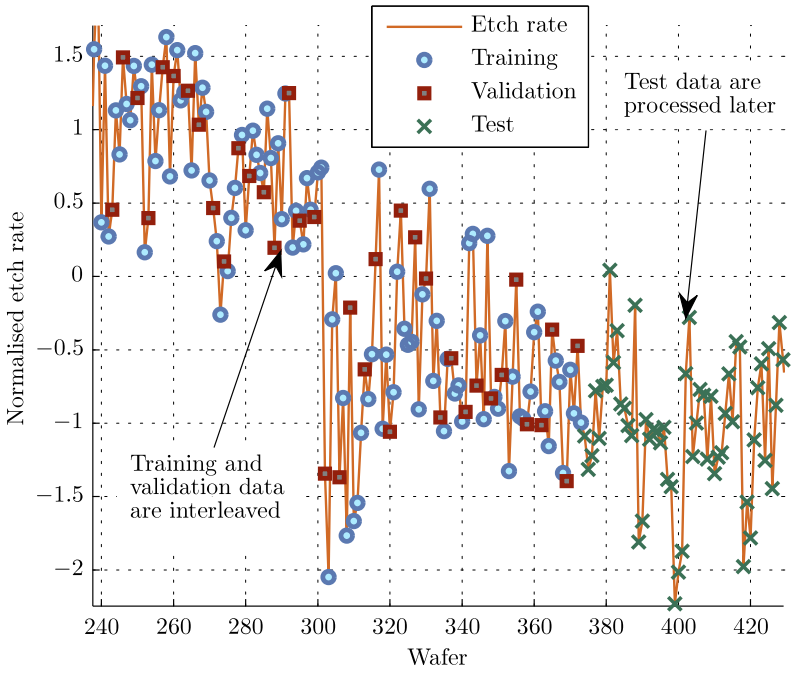



FIGURE 6.15: Chronological data division scheme. Data is divided into training, validation, and test data sets such that the test data is made up of wafers that have been processed after the training and validation wafers.

Discrepancies in etch rate estimation accuracy between models created using the two data sets provides information on how model performance is affected by changes in the system behaviour. The chronological data set investigates whether past data can be used to summarise the operating space of the etch tool in advance of system changes such as PM events. Global models created using the chronological data set extrapolate etch rate estimates beyond the test data range.

The interleaved data set is employed to investigate whether model estimation accuracy is improved when information from the same operational region as that of the test wafers is included in the model training data sets. The interleaved scheme is used purely for investigative purposes since such a scheme is unrealistic in an online system but the scheme gives some measure of the merit of a more comprehensive data logging/metrology philosophy.

## 6.5 Algorithm details

The seven modelling techniques described in Chapter 3 are used to model the global variations in etch rate in the data sets. This section details the implementation details

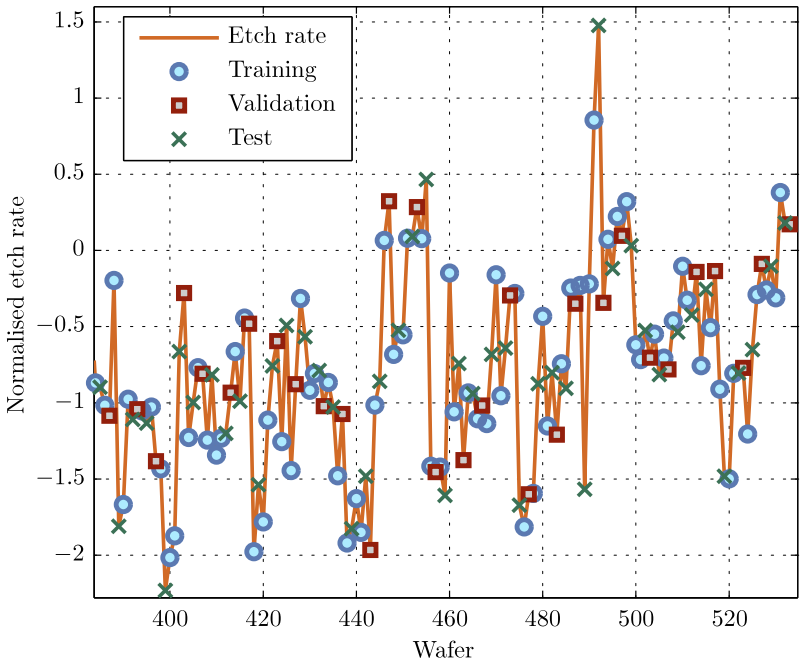



FIGURE 6.16: Interleaved data division scheme. In this data division scheme, the training, test and validation wafers are interleaved throughout the full data set. This data division scheme can reveal whether wafers that are chronologically close to test wafers provide more useful information for modelling than those further apart.

of the modelling techniques for which decisions on model structure and/or training technique are required.

### 6.5.1 Stepwise regression parameters

The stepwise regression models are developed using limiting $p$-values of 0.05 and 0.1 for the $F_{toEnter}$ and $F_{toRemove}$ tests respectively, as recommended in the text by Draper and Smith [47]. Lower $p$-values correspond to higher $F$-test values, and as mentioned in Section 3.3, choosing $F_{toEnter} > F_{toRemove}$ makes it more difficult to add variables to the model than to remove them.

### 6.5.2 ANN structure

Figure 6.17 shows the best validation data estimation mean squared error (MSE) for different 2-layer ANN structures when modelling etch rate using data representative of the etch rate data set. The validation data MSE is the MSE achieved on the validation data set during model training. In such tests with the plasma etch data set, ANNs with a single hidden layer are found to yield similar estimation accuracy to those with two hidden layers. Hence, the ANNs used in this thesis all employ a single hidden layer

of neurons because the single-layered ANNs are less complex and require less training time than ANNs with two hidden layers, while still achieving comparable estimation accuracy.

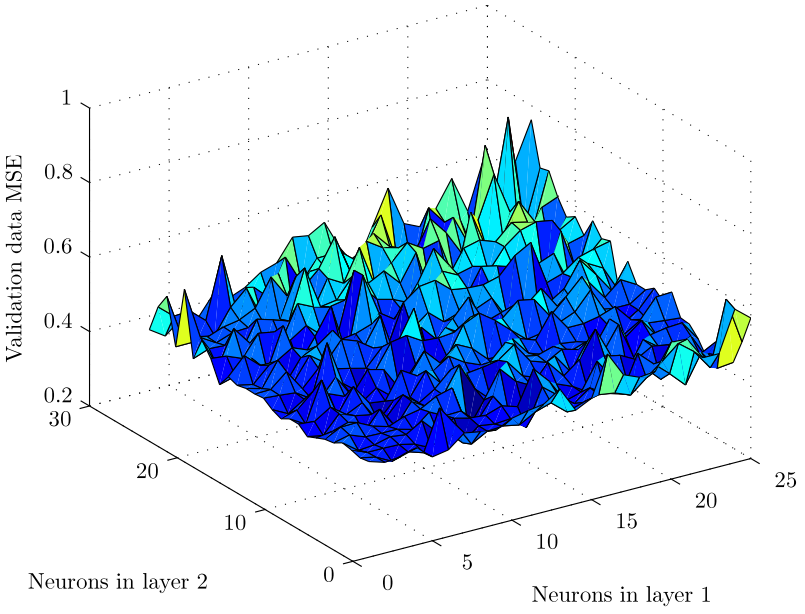



FIGURE 6.17: Validation data mean squared error (MSE) for different ANN structures on sample etch rate data. A general increase in validation data MSE is seen in this figure for ANNs with increasing numbers of neurons.

The ANNs are trained using the Levenberg-Marquardt (LM) optimisation algorithm. Early stopping is employed using separate training and validation data sets to find the optimal training stop-point. The number of hidden neurons in the single hidden layer of each ANN is varied from 5 to 15 in an attempt to find the best network size. Ten random initialisations of the weight matrices are investigated for each network size. The best network is selected by finding the network with the lowest validation data MSE (for etch rate estimation).

### 6.5.3 GPR covariance function

As described in Section 3.8.4, a number of different covariance functions can be specified during the development of GPR models. Ideally, the choice of covariance function is made using *a priori* knowledge of the data to be modelled. Unfortunately, the plasma etch data set is not clearly suited to any particular choice of covariance function because the relationships between the input variables and the plasma etch rate are largely unknown.

An investigation into the estimation accuracy of GPR models trained using different covariance functions is carried out in [254] and it is found that GPR model performance is relatively insensitive to covariance function choice for the plasma etch data set. Exceptionally poor prediction performance is only observed when using squared exponential covariance functions with input selections that include a large number of input variables.

The GPR-based models examined in this thesis employ a covariance function with a constant term, a rational quadratic component, and a random noise component. The rational quadratic (RQ) covariance function is used to allow more flexibility than a squared exponential component. A different length scale is used for each input variable as described in Section 3.8.4. Hence, the covariance function has the form:

$$k(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = \frac{1}{s} + \nu \left(1 + \frac{(\vec{\mathbf{x}}_i - \vec{\mathbf{x}}_j)\mathbf{Q}^{-1}(\vec{\mathbf{x}}_i - \vec{\mathbf{x}}_j)^T}{2\varrho}\right)^{-\varrho} + \sigma_n^2 \delta(i,j), \tag{6.3}$$

where $s$ is the hyperparameter for the constant term, $\mathbf{Q} = \text{diag}(q_{ii}) \in \mathbb{R}^{p \times p}$ is a diagonal matrix of ARD parameters (one for each of the $p$ input variables) for the RQ covariance function, $\nu$ is the hyperparameter determining the degree of variability for the RQ function in the output space, $\varrho$ is a hyperparameter determining the shapes of RQ function, $\sigma_n^2$ is the noise covariance, and $\delta(i,j)$ is the Kronecker delta function such that $\delta(i,j) = 1$ if $i == j, 0$ otherwise. Hence, a total of $p + 4$ hyperparameters must be optimised for input data of dimension $p$.

## 6.6 Virtual metrology results

The plasma etch rate is a challenging variable to model because strong relationships do not exist between the VM input variables and etch rate. The mean value of etch rate is approximately 66 nm/s, with a standard deviation of 1.8 nm/s. Because the mean etch rate is relatively large compared to the etch rate standard deviation, a model that simply produces a constant value at the mean would be reported with a MAPE of only 2.3 %. Although the etch rate estimates from each model are evaluated principally using the the mean absolute percentage error (MAPE), to provide a more detailed insight into the modelling results reported in this section, the coefficient of determination, or $R^2$ value, for the estimates from each model is also reported. Although the MAPE reported for a model that estimates a constant mean value for each wafer is as low as 2.3 %, the $R^2$ value will be 0, thus revealing the inadequacies of the estimation.

The VM results using the chronologically ordered data sets from Tools 1 and 2 are provided in Tables 6.5 and 6.6 respectively, and the results using the interleaved data set

| Tool 1 Chronological | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Model Type | | | | | | | | | | | | | |
| | | LSR | | Stepwise | | LARS | | PCR | | PLS | | NN | | GPR | |
| | Data Type | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE |
| 1 | EP | 0.31 | 1.24 | 0.32 | 1.19 | 0.25 | 1.23 | 0.24 | 1.27 | 0.25 | 1.23 | 0.14 | 1.51 | 0.17 | 1.29 |
| 2 | EP-Step | 0.32 | 1.19 | 0.32 | 1.19 | 0.28 | 1.17 | 0.29 | 1.56 | 0.28 | 1.17 | 0.11 | 2.07 | 0.24 | 1.21 |
| 3 | $PIM_0$ | 0.06 | 1.35 | 0.09 | 1.32 | 0.14 | 1.30 | 0.11 | 1.65 | 0.12 | 1.34 | 0.03 | 1.74 | 0.02 | 1.47 |
| 4 | $PIM_0$-Step | 0.09 | 1.32 | 0.09 | 1.32 | 0.08 | 1.32 | 0.16 | 1.73 | 0.08 | 1.32 | 0.20 | 1.37 | 0.02 | 1.40 |
| 5 | $PIM_5$ | 0.18 | 1.69 | 0.12 | 1.98 | 0.15 | 1.28 | 0.09 | 2.30 | 0.16 | 1.24 | 0.17 | 1.92 | 0.07 | 1.76 |
| 6 | $PIM_5$-PCA | 0.07 | 2.32 | 0.08 | 2.29 | 0.06 | 1.63 | 0.03 | 1.70 | 0.06 | 1.63 | 0.09 | 1.56 | 0.04 | 1.87 |
| 7 | $PIM_5$-Step | 0.12 | 1.98 | 0.12 | 1.98 | 0.12 | 1.56 | 0.07 | 2.22 | 0.15 | 1.47 | 0.14 | 2.10 | 0.09 | 2.03 |
| 8 | PIM | 0.02 | 2.61 | 0.13 | 1.55 | 0.25 | 1.19 | 0.19 | 1.42 | 0.21 | 1.22 | 0.13 | 1.37 | 0.08 | 1.40 |
| 9 | PIM-PCA | 0.23 | 1.44 | 0.24 | 1.42 | 0.27 | 1.20 | 0.10 | 1.45 | 0.21 | 1.24 | 0.09 | 1.68 | 0.07 | 1.47 |
| 10 | PIM-Step | 0.13 | 1.55 | 0.13 | 1.55 | 0.14 | 1.35 | 0.15 | 1.58 | 0.14 | 1.35 | 0.01 | 2.09 | 0.17 | 1.43 |
| 11 | XRZP | 0.16 | 1.47 | 0.10 | 1.31 | 0.19 | 1.22 | 0.15 | 2.07 | 0.21 | 1.21 | 0.10 | 1.47 | 0.05 | 1.46 |
| 12 | XRZP-PCA | 0.15 | 2.11 | 0.15 | 2.17 | 0.15 | 1.57 | 0.15 | 1.74 | 0.15 | 1.54 | 0.08 | 1.61 | 0.13 | 1.41 |
| 13 | XRZP-Step | 0.10 | 1.31 | 0.10 | 1.31 | 0.10 | 1.29 | 0.16 | 1.51 | 0.09 | 1.30 | 0.12 | 1.31 | 0.20 | 1.31 |
| 14 | EP-$PIM_0$ | 0.29 | 1.13 | 0.08 | 1.31 | 0.21 | 1.19 | 0.11 | 1.59 | 0.20 | 1.22 | 0.09 | 1.50 | 0.11 | 1.41 |
| 15 | EP-$PIM_0$-Step | 0.08 | 1.31 | 0.10 | 1.33 | 0.06 | 1.29 | 0.08 | 1.68 | 0.06 | 1.29 | 0.05 | 1.44 | 0.17 | 1.23 |

TABLE 6.5: Global Modelling results for Tool 1 data in chronological order.

are detailed in Tables 6.7 and 6.8. The columns of each table are coloured to highlight the best performing models based on MAPE and to allow the identification of trends in the VM results. The colour ranges from green to red to highlight the best and worst performances, respectively.

The results in Tables 6.5 – 6.8 suggest that there is no modelling technique or input variable selection, among those examined, that produces substantially better results than the others across both of the tools or data set ordering options. However, some trends and patterns in modelling performance are revealed in the results. The results are examined in further detail in Sections 6.6.1 and 6.6.2 before more general conclusions are drawn from the modelling results in Section 6.7.

### 6.6.1 Chronologically ordered data

The EP variables proved to be the most effective input variables for estimation of etch rate for Tool 1 across the majority of the modelling techniques investigated. The best results are obtained through the use of the stepwise selection algorithm to select key EP variables followed by the use of the PLS or LARS modelling algorithms that achieve a test data accuracy of 1.17% MAPE with an $R^2$ value of 0.28. The etch rate estimates from the PLS algorithm are shown in Figure 6.18. Although the estimates shown in

| Tool 2 Chronological | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Model Type | | | | | | | | | | | | | |
| | | LSR | | Stepwise | | LARS | | PCR | | PLS | | NN | | GPR | |
| | Data Type | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE |
| 1 | EP | 0.03 | 2.18 | 0.08 | 2.24 | 0.06 | 1.95 | 0.00 | 1.60 | 0.00 | 1.67 | 0.03 | 2.36 | 0.00 | 1.87 |
| 2 | EP-Step | 0.08 | 2.24 | 0.08 | 2.24 | 0.07 | 2.27 | 0.00 | 1.84 | 0.07 | 2.27 | 0.04 | 2.12 | 0.16 | 2.17 |
| 3 | $PIM_0$ | 0.00 | 2.16 | 0.14 | 1.57 | 0.11 | 1.39 | 0.20 | 1.27 | 0.10 | 1.40 | 0.21 | 1.42 | 0.26 | 1.73 |
| 4 | $PIM_0$-Step | 0.14 | 1.57 | 0.14 | 1.57 | 0.14 | 1.40 | 0.14 | 1.57 | 0.14 | 1.40 | 0.10 | 1.60 | 0.16 | 1.56 |
| 5 | $PIM_5$ | 0.18 | 1.36 | 0.11 | 1.53 | 0.23 | 1.20 | 0.15 | 1.57 | 0.24 | 1.19 | 0.03 | 1.68 | 0.23 | 1.63 |
| 6 | $PIM_5$-PCA | 0.07 | 1.47 | 0.04 | 1.72 | 0.08 | 1.37 | 0.28 | 1.70 | 0.09 | 1.34 | 0.01 | 2.24 | 0.10 | 1.39 |
| 7 | $PIM_5$-Step | 0.11 | 1.53 | 0.11 | 1.53 | 0.16 | 1.35 | 0.13 | 1.51 | 0.11 | 1.43 | 0.12 | 1.51 | 0.30 | 1.12 |
| 8 | PIM | 0.25 | 7.54 | 0.26 | 2.14 | 0.14 | 1.30 | 0.01 | 1.59 | 0.14 | 1.32 | 0.26 | 4.18 | 0.30 | 1.27 |
| 9 | PIM-PCA | 0.07 | 1.47 | 0.12 | 1.41 | 0.10 | 1.31 | 0.28 | 1.13 | 0.18 | 1.25 | 0.08 | 2.21 | 0.16 | 1.37 |
| 10 | PIM-Step | 0.26 | 2.14 | 0.26 | 2.14 | 0.15 | 1.28 | 0.17 | 1.57 | 0.08 | 1.53 | 0.22 | 1.99 | 0.32 | 1.33 |
| 11 | XRZP | 0.23 | 1.32 | 0.28 | 1.20 | 0.38 | 1.04 | 0.20 | 1.21 | 0.30 | 1.09 | 0.21 | 1.49 | 0.21 | 1.39 |
| 12 | XRZP-PCA | 0.17 | 1.23 | 0.27 | 1.16 | 0.28 | 1.16 | 0.38 | 1.10 | 0.25 | 1.18 | 0.09 | 1.28 | 0.07 | 1.36 |
| 13 | XRZP-Step | 0.28 | 1.20 | 0.28 | 1.20 | 0.37 | 1.13 | 0.43 | 1.28 | 0.34 | 1.15 | 0.29 | 1.35 | 0.26 | 1.32 |
| 14 | EP-$PIM_0$ | 0.07 | 2.38 | 0.05 | 2.16 | 0.03 | 1.87 | 0.06 | 1.50 | 0.07 | 1.46 | 0.02 | 1.90 | 0.00 | 1.79 |
| 15 | EP-$PIM_0$-Step | 0.05 | 2.16 | 0.05 | 2.16 | 0.04 | 2.20 | 0.00 | 1.92 | 0.03 | 2.15 | 0.02 | 2.47 | 0.02 | 2.08 |

TABLE 6.6: Global Modelling results for Tool 2 data in chronological order.

| Tool 1 Interleaved | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Model Type | | | | | | | | | | | | | |
| | | LSR | | Stepwise | | LARS | | PCR | | PLS | | NN | | GPR | |
| | Data Type | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE |
| 1 | EP | 0.68 | 1.19 | 0.68 | 1.20 | 0.66 | 1.24 | 0.66 | 1.19 | 0.66 | 1.25 | 0.49 | 1.54 | 0.65 | 1.27 |
| 2 | EP-Step | 0.68 | 1.20 | 0.68 | 1.20 | 0.68 | 1.20 | 0.57 | 1.40 | 0.68 | 1.20 | 0.68 | 1.21 | 0.67 | 1.20 |
| 3 | $PIM_0$ | 0.63 | 1.31 | 0.63 | 1.32 | 0.63 | 1.32 | 0.56 | 1.41 | 0.63 | 1.32 | 0.63 | 1.33 | 0.64 | 1.25 |
| 4 | $PIM_0$-Step | 0.63 | 1.32 | 0.63 | 1.32 | 0.63 | 1.32 | 0.58 | 1.38 | 0.63 | 1.32 | 0.58 | 1.41 | 0.64 | 1.28 |
| 5 | $PIM_5$ | 0.69 | 1.19 | 0.69 | 1.19 | 0.69 | 1.19 | 0.56 | 1.38 | 0.69 | 1.19 | 0.67 | 1.20 | 0.68 | 1.15 |
| 6 | $PIM_5$-PCA | 0.56 | 1.38 | 0.54 | 1.40 | 0.56 | 1.38 | 0.56 | 1.39 | 0.56 | 1.38 | 0.62 | 1.27 | 0.69 | 1.18 |
| 7 | $PIM_5$-Step | 0.69 | 1.19 | 0.69 | 1.19 | 0.68 | 1.20 | 0.53 | 1.42 | 0.69 | 1.19 | 0.70 | 1.20 | 0.69 | 1.17 |
| 8 | PIM | 0.60 | 1.40 | 0.70 | 1.17 | 0.69 | 1.18 | 0.73 | 1.12 | 0.71 | 1.15 | 0.63 | 1.33 | 0.69 | 1.16 |
| 9 | PIM-PCA | 0.73 | 1.12 | 0.72 | 1.15 | 0.70 | 1.16 | 0.68 | 1.21 | 0.71 | 1.15 | 0.64 | 1.29 | 0.72 | 1.11 |
| 10 | PIM-Step | 0.70 | 1.17 | 0.70 | 1.17 | 0.70 | 1.19 | 0.64 | 1.28 | 0.70 | 1.19 | 0.69 | 1.20 | 0.70 | 1.15 |
| 11 | XRZP | 0.69 | 1.17 | 0.67 | 1.22 | 0.69 | 1.17 | 0.57 | 1.39 | 0.68 | 1.17 | 0.71 | 1.13 | 0.70 | 1.17 |
| 12 | XRZP-PCA | 0.57 | 1.39 | 0.57 | 1.39 | 0.57 | 1.37 | 0.57 | 1.36 | 0.57 | 1.37 | 0.69 | 1.21 | 0.70 | 1.18 |
| 13 | XRZP-Step | 0.67 | 1.22 | 0.67 | 1.22 | 0.67 | 1.22 | 0.55 | 1.41 | 0.67 | 1.22 | 0.69 | 1.17 | 0.65 | 1.25 |
| 14 | EP-$PIM_0$ | 0.67 | 1.21 | 0.69 | 1.19 | 0.66 | 1.23 | 0.63 | 1.28 | 0.65 | 1.27 | 0.63 | 1.29 | 0.63 | 1.29 |
| 15 | EP-$PIM_0$-Step | 0.69 | 1.19 | 0.69 | 1.19 | 0.69 | 1.20 | 0.63 | 1.27 | 0.69 | 1.20 | 0.64 | 1.24 | 0.70 | 1.14 |

TABLE 6.7: Global Modelling results for Tool 1 data in interleaved order.

Figure 6.18 accurately capture the large-scale, low frequency etch rate variations in the data, the smaller-scale, high-frequency fluctuations are not modelled accurately. Since

| Tool 2 Interleaved | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Model Type | | | | | | | | | | | | | |
| | | LSR | | Stepwise | | LARS | | PCR | | PLS | | NN | | GPR | |
| | Data Type | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE |
| 1 | EP | 0.26 | 1.35 | 0.28 | 1.27 | 0.33 | 1.21 | 0.29 | 1.28 | 0.32 | 1.26 | 0.31 | 1.40 | 0.26 | 1.24 |
| 2 | EP-Step | 0.28 | 1.27 | 0.28 | 1.27 | 0.27 | 1.25 | 0.27 | 1.26 | 0.27 | 1.25 | 0.22 | 1.44 | 0.33 | 1.21 |
| 3 | $PIM_0$ | 0.38 | 1.21 | 0.31 | 1.18 | 0.36 | 1.18 | 0.29 | 1.19 | 0.37 | 1.17 | 0.52 | 1.15 | 0.63 | 0.81 |
| 4 | $PIM_0$-Step | 0.31 | 1.18 | 0.31 | 1.18 | 0.31 | 1.16 | 0.31 | 1.18 | 0.31 | 1.16 | 0.36 | 1.18 | 0.45 | 1.04 |
| 5 | $PIM_5$ | 0.45 | 1.08 | 0.42 | 1.12 | 0.36 | 1.13 | 0.23 | 1.27 | 0.43 | 1.11 | 0.42 | 1.11 | 0.57 | 0.91 |
| 6 | $PIM_5$-PCA | 0.22 | 1.28 | 0.22 | 1.28 | 0.25 | 1.23 | 0.19 | 1.32 | 0.25 | 1.23 | 0.35 | 1.21 | 0.57 | 0.93 |
| 7 | $PIM_5$-Step | 0.42 | 1.12 | 0.42 | 1.12 | 0.43 | 1.10 | 0.21 | 1.30 | 0.43 | 1.10 | 0.14 | 1.39 | 0.49 | 1.05 |
| 8 | PIM | 0.12 | 3.68 | 0.36 | 1.33 | 0.43 | 1.10 | 0.40 | 1.15 | 0.52 | 1.11 | 0.35 | 1.25 | 0.43 | 1.06 |
| 9 | PIM-PCA | 0.38 | 1.18 | 0.38 | 1.17 | 0.37 | 1.16 | 0.36 | 1.17 | 0.37 | 1.16 | 0.43 | 1.12 | 0.48 | 1.02 |
| 10 | PIM-Step | 0.36 | 1.33 | 0.36 | 1.33 | 0.37 | 1.26 | 0.37 | 1.18 | 0.37 | 1.26 | 0.41 | 1.15 | 0.44 | 1.06 |
| 11 | XRZP | 0.48 | 1.11 | 0.38 | 1.18 | 0.49 | 1.10 | 0.42 | 1.12 | 0.49 | 1.10 | 0.53 | 1.09 | 0.63 | 0.83 |
| 12 | XRZP-PCA | 0.42 | 1.12 | 0.41 | 1.11 | 0.42 | 1.07 | 0.42 | 1.12 | 0.44 | 1.10 | 0.20 | 1.31 | 0.55 | 1.01 |
| 13 | XRZP-Step | 0.38 | 1.18 | 0.38 | 1.18 | 0.38 | 1.16 | 0.37 | 1.15 | 0.38 | 1.16 | 0.38 | 1.16 | 0.39 | 1.12 |
| 14 | EP-$PIM_0$ | 0.31 | 1.49 | 0.31 | 1.20 | 0.37 | 1.13 | 0.31 | 1.25 | 0.34 | 1.19 | 0.33 | 1.15 | 0.48 | 0.97 |
| 15 | EP-$PIM_0$-Step | 0.31 | 1.20 | 0.31 | 1.20 | 0.30 | 1.18 | 0.32 | 1.17 | 0.30 | 1.18 | 0.40 | 1.11 | 0.51 | 1.04 |

TABLE 6.8: Global Modelling results for Tool 2 data in interleaved order.

the test data do not exhibit a great deal of low frequency variation and the high frequency variations are not captured, the reported $R^2$ values are relatively low.

For Tool 2, the results in Table 6.6 suggest that the models using information from the PIM sensor produce superior estimates of etch rate for chronologically ordered data than models using the EP variables. In particular, all of the modelling techniques, except for GPR, estimate the etch rate values most accurately when using the $XRZP_5$ variables as model inputs. The best modelling result is shown in Figure 6.19 and is achieved using a LARS model with the $XRZP_5$ variables, achieving a MAPE of 1.04 and $R^2$ of 0.38.

The poor estimation results produced by the models using EP variables for Tool 2 occur partly as a result of an substantial change in etch rate behaviour due to a PM event in the test data. The estimates from a LARS model using the EP data set as VM input variables are shown in Figure 6.20, where the final PM event occurs at wafer 225. Inaccurate estimation of the plasma etch rate is observed after this PM event for all VM models using the EP data from Tool 2, suggesting that, contrary to the PIM variables, the EP variables do not contain sufficient information to model the changes caused by the final PM event in the data.

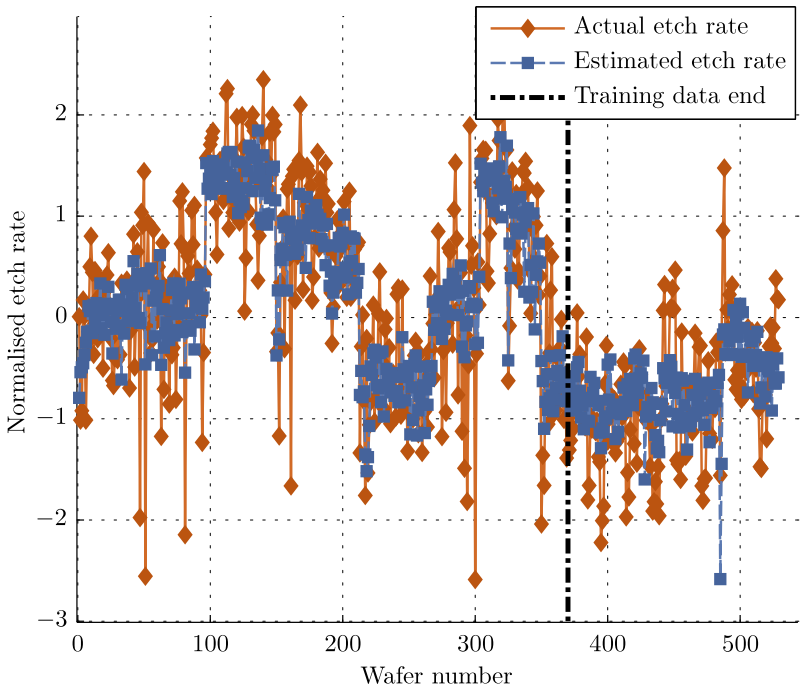



FIGURE 6.18: PLS model etch rate estimation for Tool 1 chronological data. While the general trend of the etch rate variations is captured by the model, high frequency changes are not.

### 6.6.2 Interleaved data

The estimation results from models using the interleaved data are substantially different to the results from those using the chronological data. For Tool 1, Table 6.7 suggests that the predictions from models using the PIM variables are more effective for etch rate estimation. Most of the modelling techniques investigated perform with similar accuracy, with the exceptions of PCR, which produces poor results for all input variable combinations but PIM, and GPR regression, which performs relatively reliably for the majority of input variable combinations. The best overall result for the interleaved data from Tool 1 is achieved using a GPR-based model using PCA-reduced PIM variables (MAPE of 1.11 % and $R^2$ 0.72). The estimates from this model are shown in Figure 6.21. The interleaved estimation results for Tool 2 shown in Table 6.8 also highlight GPR as a superior modelling technique to the others investigated since it produces the most accurate predictions for 13 out of the 15 input variable combinations.

The fact that models perform well using interleaved data sets is to be expected since, unlike the chronologically ordered data sets, in the interleaved sets the training data are numerically close to the test data. This numerical proximity arises as the training and test data consist of wafers that are from the same operational spaces.

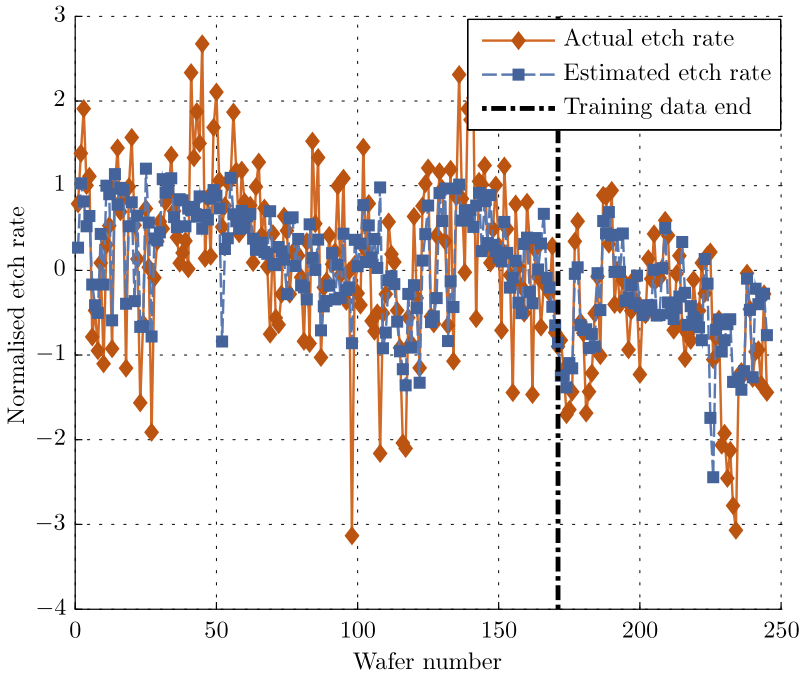



FIGURE 6.19: LARS model etch rate estimation for Tool 2 chronological data. There is less overall variation and fewer wafers measured in the data recorded from Tool 2 than from Tool 1.

The best estimates for the interleaved data from Tool 2 are produced by a GPR model using the $PIM_0$ variables as model inputs; The estimates from this GPR model follow etch rate with a MAPE of 0.81 and $R^2$ value of 0.63, as shown in Figure 6.22

## 6.7 Discussion

### 6.7.1 Input combinations

There does not appear to be one input selection technique that demonstrates significantly more accurate VM performance across all of the tool and modelling techniques investigated for the data considered, but rather, particular input selections appear to suit different situations. For Tool 1, the EP data produces consistently good results for the majority of modelling techniques when the data is in chronological order, but this is bettered by the inputs using PIM data for interleaved data. The calculated values for the reactance, resistance, impedance, and power ($XRZP_5$ data) lead to accurate etch rate estimates for the Tool 2 data in chronological order, but this superiority is not as clear cut during the interleaved data investigations.

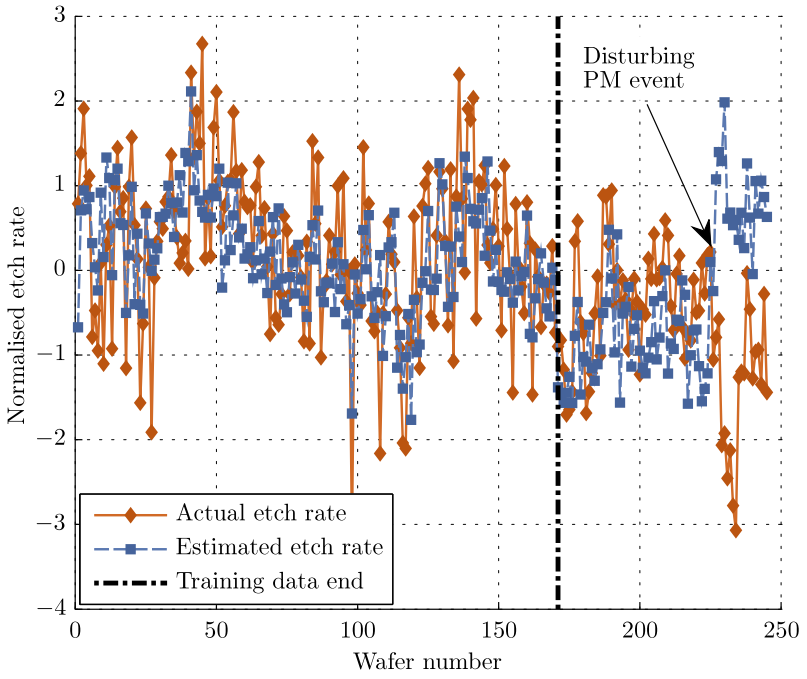



FIGURE 6.20: LARS model using EP variables for Tool 2. The etch rate estimates for the final 25 wafers of the test data set are inaccurate. This inaccuracy is a result of a PM event that changes the chamber characteristics in such a way that is not reflected in the EP variables. Similar results are seen for all models using EP variables to model etch rate variation in the Tool 2 data set.

It could be argued on the basis of the Tool 1 chronological results that the considerable cost of PIM sensor installation is not offset by appreciable increases in etch rate prediction accuracy using the PIM measurements. However, it is only the models that use PIM variables that successfully estimate the changes in etch rate caused by the final PM event in the Tool 2 test data, as shown in Figure 6.20.

Since the data set available is relatively small, it is impossible to draw comprehensive conclusions that will apply to all situations. However, the global modelling results can be used to gain an understanding of the achievable accuracy within the limitations of the measurement frequency in the data.

### 6.7.2 Modelling techniques

Similarly, no modelling technique stands out with substantially better performance across all of the tool and input variable combinations. The LARS and PLS algorithms produce relatively good results over the course of the investigations. ANNs perform on par with the other techniques except for the Tool 1 chronological data. Principal component regression performs poorly on the interleaved data sets, and when used on the EP data in chronological form. Poor PCR performance is expected with the EP

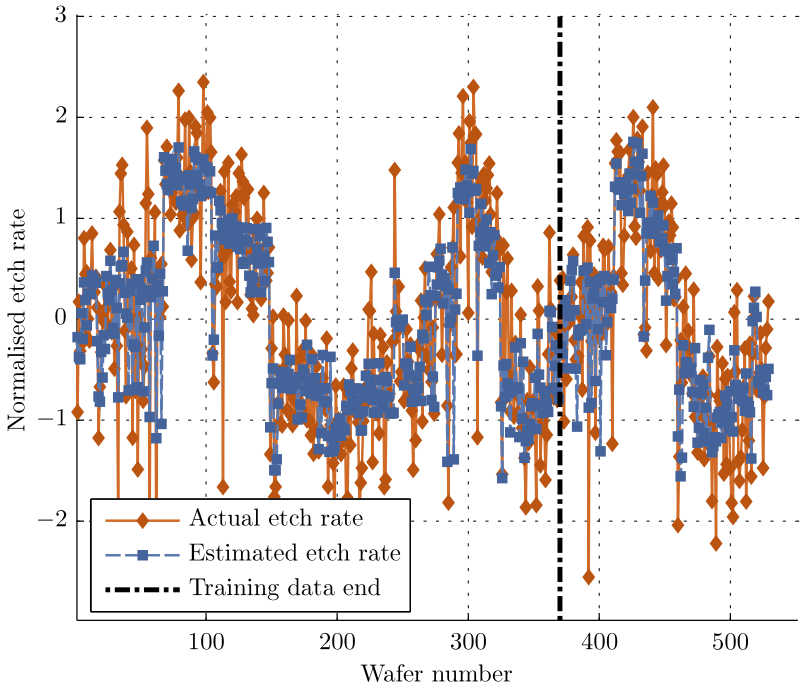



FIGURE 6.21: GPR model etch rate estimates using PIM-PCA inputs on interleaved data from Tool 1. Models using interleaved data are generally more accurate than those using chronologically ordered data since the training data contains information more relevant to the test data. As with the results produced by models using chronological data, the high frequency fluctuations in etch rate are not modelled accurately.

variables as many of the EP variables are uncorrelated and hence not ideally suited for compression using PCA.

As expected, LSR fails when a large number of input regressor variables are presented during training due to rank deficiency in the input data matrix. For example, inaccurate results, due to poorly conditioned input matrices, can be seen for LSR models using the full set of PIM data. Between the two LSR-based techniques with variable selection, LARS models perform better than the stepwise regression models for almost all input selections.

GPR models emerge as the most accurate modelling technique for the interleaved data. For the chronological data, the GPR models do not produce the most accurate etch rate estimates, but have the added advantage over other techniques of easily calculable confidence limits on all estimates. It is important to have an measure of the degree of confidence in the VM estimates, in addition to the estimation itself [234]. GPR models naturally permit confidence intervals to be established based on the amount of training data available in a given operational space. If an estimation is required in an area with little training data, confidence intervals are correspondingly large. 95% confidence

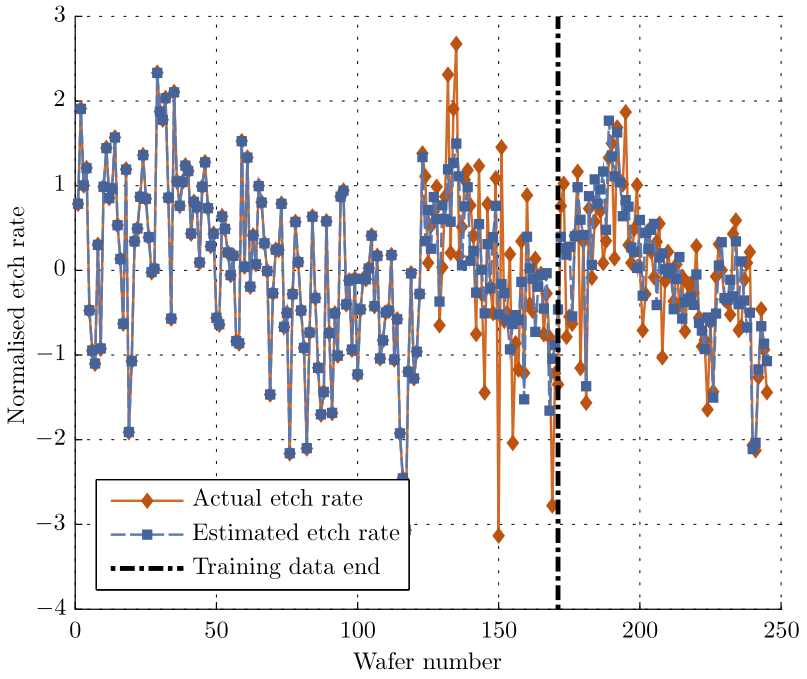



FIGURE 6.22: GPR model using $PIM_0$ variables to model interleaved data from Tool 2.

limits on the estimates calculated using the GPR models generally encapsulate the high frequency fluctuations in etch rate for the chronological data, as shown in Fig. 6.23.

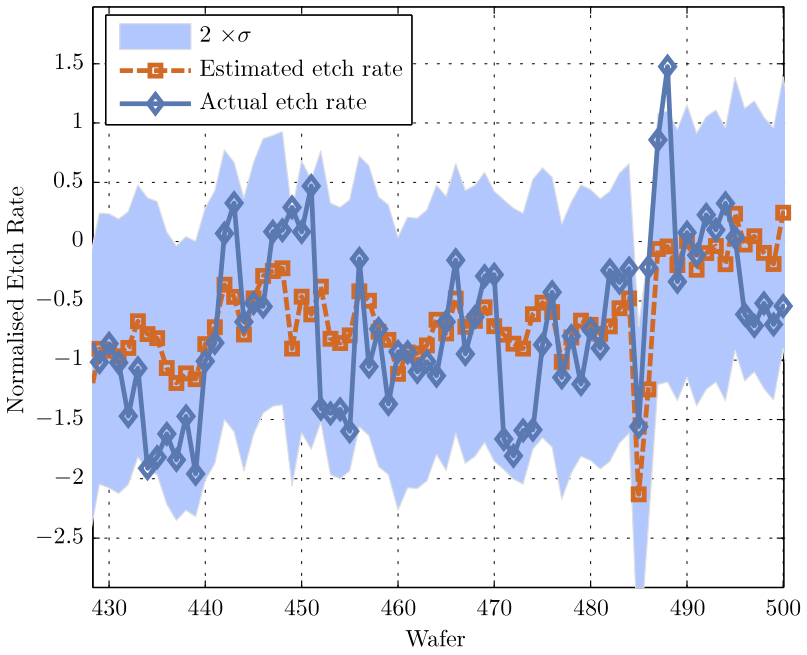



FIGURE 6.23: Test wafer etch rate estimates from a global GPR model using stepwise selected EP variables as model inputs. 95% confidence intervals on the estimates are also depicted.

### 6.7.3 Data ordering

Model MAPE performance is better when interleaved data sets are used to build and test the VM models than when chronological data sets are used. For the interleaved data sets, the effect of PM events on the model performance is reduced because the training data originates in the same operational space (PM cycles) as the test data. The improvement in estimation accuracy lends itself to the assumption that inter-PM variation for each wafer is captured in the neighbouring wafers from the same PM cycle.

In comparison to models trained using interleaved data sets, models that use chronologically ordered data sets are trained on wafer information that contains relatively little information about the operational space of the wafers in the test data set. The chronological models fail to accurately model the etch rate after a significant maintenance event occurs. As seen in Section 6.6.1, the effect of an influential PM event is partly responsible for the poor prediction results produced by models using EP data from Tool 2. The inability to accurately predict etch rate behaviour in future PM cycles is one of the main disadvantages of static global-based models in etch process modelling.

Relatively large improvements in $R^2$ values are seen for models using the interleaved data sets when compared to those using the chronological data sets. The improvement in $R^2$ is somewhat misleading as it is not wholly a result of improved model accuracy, but is instead primarily an artifact of the test data sets used to evaluate model performance. As depicted in Figure 6.24, for Tool 1, the etch rate for the interleaved test data span a much larger range than the etch rate for the chronological test data. Because the VM models can follow large-scale, low-frequency fluctuations in etch rate, but fail at accurately estimating smaller, high-frequency fluctuations, the larger range of etch rate values contained in the interleaved test data increases the reported $R^2$ values. This increase in $R^2$ value is more pronounced in the Tool 1 results than in the Tool 2 results because the first section of etch rate measurements from Tool 1 contains considerably more variation than the later sections. As a result, the $R^2$ values listed in Tables 6.5 and 6.6 are not directly comparable to those in 6.7 and 6.8. The MAPE values, however, are comparable between the data sets.

### 6.7.4 Inter-machine compatibility

The issue of inter-machine compatibility is highlighted by the fact that models using the same input variables perform very differently on different etch chambers. Although both of the tools used in the analysis are physically identical, processing the same mix of wafers, and operating in the same fabrication environment, the relationships between

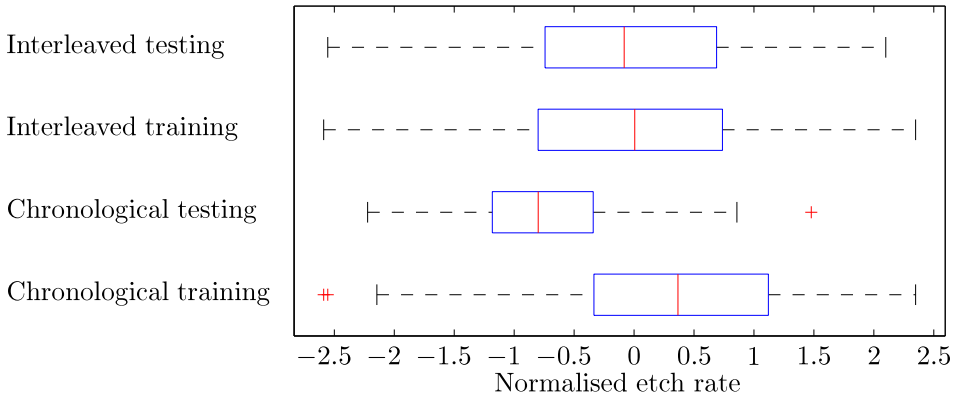



FIGURE 6.24: Etch rate distributions in training and test data sets from Tool 1. Boxes show 25th to 75th quartile ranges, whiskers extend to 2.7 $\sigma$, and other values are marked with crosses. The chronological test data etch rate is restricted to a smaller range of etch rate values than the other data sets.

| **Tool 1** | **Tool 2** |
|---|---|
| 3-MEAN-ENDPT_STEP_TIME | 3-MEAN-RF_LINE_IMP |
| 3-MEAN-RF_LOAD_COIL_POS | 4-MEAN-ENDPOINT_A |
| 4-MEAN-LOW_ELECT_TEMP | 4-MEAN-RF_MATCH_1_DC_BIAS |
| 4-MEAN-RF_MATCH_1_TUNE | 5-MEAN-LOW_ELECT_TEMP |
| 4-MEAN-UP_ELECT_TEMP | 5-MEAN-RF_LOAD_MATCH_PH |
| 5-MEAN-LOW_ELECT_TEMP | |
| 5-MEAN-UP_ELECT_TEMP | |
| 5-MEAN-RF_MATCH_1_TUNE | |
| 5-MEAN-RF_LOAD_COIL_POS | |
| 5-MEAN-RF_LINE_IMP | |

TABLE 6.9: Stepwise selected EP variables for Tool 1 and Tool 2. There is only one variable in common between the two sets of selected variables.

the sensor recordings and the etch performance of each is remarkably different. Table 6.9 shows the variables selected from the EP data from both tools by the stepwise selection algorithm for modelling of etch rate using chronologically ordered data. In Table 6.9, only one variable appears in the models for both machines. This lack of model similarity suggests that there is only a weak relationship between the etch rate and the model input variables. Hence, the structural relationship between the input variables and the etch rate is sensitive to relatively small changes in the chamber characteristics.

The lack of similarity in the variables recorded in both machines prevents models developed on one machine being transferred to another. One would expect differences in the parameter values between models, but not large structural changes. This is one of the most difficult challenges faced in VM modelling of such complex systems. Machine specific models, or models capable of adapting to individual machine performance are required to overcome this challenge.

### 6.7.5 Summary

While the overall trends in etch rate are approximated by the estimates from the best models discussed in Sections 6.6.1 and 6.6.2, no model accurately follows the high frequency fluctuations in etch rate observed in the data. These high frequency etch rate changes are either not reflected in the variables that are used as VM model inputs, or the variations arise elsewhere in the manufacturing process. GPR-based models, however, do generally contain these etch rate variations within 95% confidence intervals.

An interleaved data set is used to examine the effect of including data from the same operational space as the test data in the training data for VM models. Substantial increases in accuracy are observed for models using the interleaved data over models using the chronologically ordered data, demonstrating the importance of using comprehensive training data sets that span the operational space of the test data during the development of global models where possible. The VM models using the interleaved data sets provide some insight on the advantages of a more comprehensive data set and metrology philosophy.

The results suggest that there is no single ideal solution to the problem of globally modelling the complex plasma etch process examined here. The best achievable accuracy using seven different modelling techniques, on two different plasma etch machines, with 15 different input combinations has been investigated and found to be approximately 1.1% MAPE. This level of accuracy is achieved with a relatively low metrology frequency compared to the frequency of process disturbing PM events (approximately 40 etch rate measurements per PM cycle).

Global models based on larger data sets could be constructed if more data was available, but these models may still fail after a PM event in the future that results in behaviour outside of the operational space of the training data. In order to maximise etch rate estimation accuracy across PM events, a number of local etch-rate modelling strategies are considered in the next chapter.

# Chapter 7

# Local modelling of a plasma etch process

The global models explored in Chapter 6 failed to consistently and accurately estimate etch rate variations after preventative maintenance (PM) events are carried out on the etch chamber in the process studied. The validity of the global models is also brought into question as the process drifts due to repeated etch operations that condition the chamber. While global models trained using larger data sets than the data sets investigated in Chapter 6 could be created, such models may still fail to maintain adequate estimation accuracy after maintenance events in the future that change the system behaviour in such a manner that has not been previously captured in the training data.

In general, global models have the advantage that only one model is required for use, hence involve a relatively simple training and set up procedure. However, large global models incorporating data from a broad operating space may become overly general, and not capture specific behaviour for different input conditions. In general, greater amounts of data may lead to a better average approximation, but using less data may amount to better accuracy over smaller regions of the operating space.

Local modelling of complex non-linear processes through the division of the complete operating space into smaller sections is a well known "divide and conquer" technique to avoid the curse of dimensionality and lack of transparency for global models [255]. Different local models are created to model different regions of the process operating space. However, difficulties related to the partitioning of the operating space, structure determination, and local model identification are the main drawbacks of local modelling approaches [96]. For time-varying systems such as the plasma etch system, Su *et al.* [236] noted that the inclusion of data from operational spaces before and after process

| | EP data | EP and PIM data |
|---|---|---|
| Number of Wafers | 18513 | 12133 |
| Measured Wafers | 793 | 529 |
| PM Cycles | 18 | 12 |
| Measurement Frequency | 4.3 % | 4.4 % |

TABLE 7.1: Description of data collected from Tool 1.

drift serves to improve the accuracy of virtual metrology (VM) models. Wise *et al.* [34] compare the use of global and local models for FDC of a plasma etch process, finding the local modelling approach more successful.

This chapter explores three local modelling schemes that strive to improve upon the accuracy of the global models of Chapter 6. The development of local models involves the partitioning of collected data into subsets of data points and the creation of multiple VM models using these subsets.

## 7.1 Local modelling approach

### 7.1.1 Data set for local modelling

As described in Section 6.3, the etch process data available for VM testing is collected from two plasma etch machines. There are substantially more metrology data available in the data collected from Tool 1 than in the data collected from Tool 2. The data from Tool 1 also contains more variation and is sampled more regularly than the data from Tool 2. Because the division of the data into smaller sections for local modelling restricts the number of samples available for model creation, it is decided to use data from Tool 1 only during the local model investigations to maximise the number of samples available. A summary of the data from Tool 1 is presented in Table 7.1.

Four different combinations of the variables available are considered VM input variables for the local modelling scheme to provide a cross-section of information from the different sensors available. Descriptions for each of the variable selections are given in Section 6.4.2. Where possible, the input selections examined during the local model analysis are:

1. EP data,
2. $PIM_5$ data,
3. XZRP data, and

4. EP-PIM$_0$ data.

### 7.1.2 Modelling techniques for local modelling

Five different modelling techniques are chosen for local model investigations on the basis of their performance during global modelling of the etch rate data set in Chapter 6. The modelling techniques investigated are:

- Stepwise selection with least squares regression (LSR), i.e. stepwise regression, is chosen as a modelling technique that incorporates variable selection into to the modelling procedure.
- LARS modelling is included for comparison with the better-known stepwise regression technique, and to verify if improved modelling performance over stepwise regression is observed for local models, as was seen for the majority of global models in Chapter 6.
- Partial least squares (PLS) regression is chosen as a latent-variable-based technique that can handle large numbers of input variables and restricted numbers of training samples without numerical difficulties by extracting explanatory variables from the original input data. PLS performed well during global model investigations.
- Artificial neural networks (ANNs) are included as a non-linear modelling technique. While neural networks performed well during global modelling tests, difficulties during local modelling may arise due to the limited number of training points available for each model. Stepwise selected input variables (based on linear model performance according to the stepwise regression algorithm) are used as inputs to the ANN models.
- Gaussian process regression (GPR) modelling is included as a non-parametric technique with both linear and non-linear capabilities. Gaussian process models typically require fewer training points than ANN models for accurate estimation and so may be more suited to local modelling. Stepwise selected input variables are used as inputs to the GPR models.

These five techniques are selected as a representative subset of the model types examined and include techniques with both linear and non-linear modelling capabilities. Details on each technique can be found in Chapter 3 and where applicable, any further data processing carried out on the input variables before modelling in this chapter is explained.

Principal component regression (PCR) is not examined since the global PCR models performed poorly during global modelling of similar data sets in Chapter 6. Least squares regression (LSR) is not considered as a modelling technique because the training data sets available for the local models may have insufficient samples to obtain reliable models. Small training data sets causes the number of variables to become comparable or greater than the number of training samples and such ill-conditioned input matrices cause computational instabilities when LSR is employed to form a predictive model.

## 7.2 Regional PM cycle models

The first of the local modelling approaches is a method to partition the data set into separate bins depending on their original *temporal* position within each PM cycle. A different VM model is then constructed for each region within a PM cycle, and the models are switched when estimating output variables for unseen wafer data, depending on the PM position of the unseen data. VM models built on these data are termed *regional PM cycle models*. The PM intervals are assumed to be regular, as is the case with most semiconductor manufacturing processes. The regional PM cycle modelling scheme is illustrated in Figure 7.1.

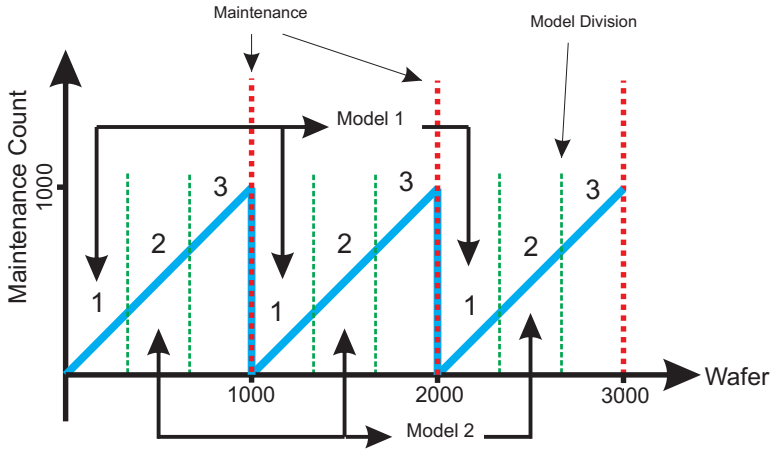



FIGURE 7.1: Regional PM cycle modelling scheme with three models used to cover the PM range. The diagram shows how wafers from similar sections of PM cycles are used to form models and perform virtual metrology on wafers from the same range in future PM cycles.

It is conjectured that the beginning sections, middle sections, and end sections of individual PM cycles may be more similar to the corresponding sections in other PM cycles than to the other sections of the same PM cycle. Lending weight to this hypothesis is the fact that several of the measurements recorded from the etch chamber (that are used as VM model input variables) exhibit repeatable patterns over the course of each

PM cycle, and it is known that chambers undergo repetitive cleaning and conditioning procedures as wafers are processed in each PM cycle. Figure 6.2 shows an EP variable that exhibits repeated patterns throughout each PM cycle, and Figure 7.2 shows a similar pattern recorded in a PIM variable. The purpose of the regional PM cycle modelling scheme is to examine whether similarities between the plasma etch data at different stages of PM cycles can be exploited to boost etch rate estimation accuracy. It is thought that if similarities exist between similar stages in different PM cycles, the regional PM cycle modelling scheme may lead to more accurate estimation of etch rate in unseen data by switching VM models as wafers are processed in new PM cycles.

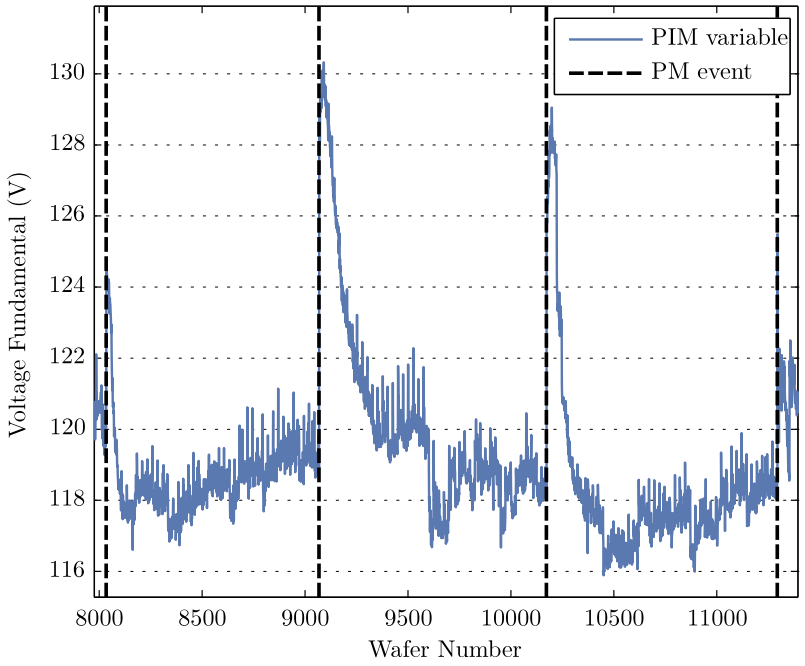



FIGURE 7.2: PIM variable (VL0) that shows a repeatable pattern over multiple PM cycles. Such patterns suggest that similarities exist between the data from similar regions in each PM cycle.

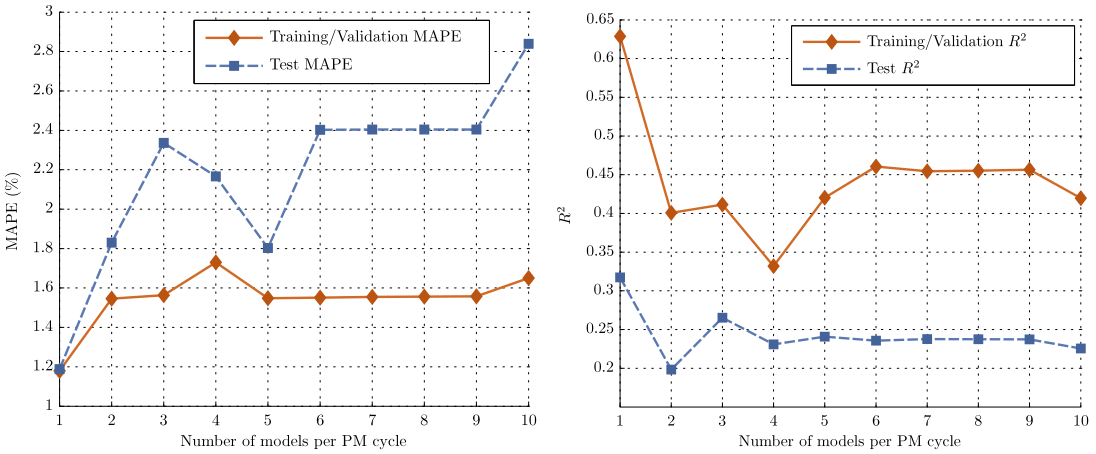
### 7.2.1 Modelling results

The etch rate estimation performance of the regional PM cycle models is tested using wafer data that contains both EP and PIM data to allow performance comparisons between models using data from different sources. The first 70% of the data are used for training and validation of models as described in Section 6.4.1 (50% for training, 20% for validation). The test data set (30%), used to judge model performance on unseen data, occurs chronologically later than the training and validation data, offering data from unseen PM cycles as a test of performance.

The training data are separated into sections according to their position within each PM cycle as depicted in Figure 7.1. The validation data for each section are interleaved

uniformly among the training data. The data in each section are mean centered and normalised to unit variance, using statistics calculated using the training data points. Data are normalised differently for each section depending on the means and standard deviations of the variables in the training data for that particular section.

The number of models created per PM cycle is varied from one to ten to determine an appropriate level of granularity and highlight the effect of the regional modelling technique on estimation accuracy. Figures 7.3, 7.4, 7.5, 7.6, and 7.7 show the regional PM cycle modelling results for etch process (EP) data (see Section 6.4.2) using stepwise regression, LARS, PLS, ANNs, and GPR models respectively. The first point on each figure, where one model is used used to model the full range of the PM cycles, is equivalent to a chronological global modelling scheme where all training data are used to form one VM model. To evaluate model performance, the mean absolute percentage error (MAPE) and $R^2$ values on the de-normalised etch rate estimates are examined.




(a) (b)

FIGURE 7.3: (a) Mean absolute percentage error (MAPE) and (b) $R^2$ results for regional PM cycle modelling technique using EP data with varying numbers of regional stepwise regression models per PM cycle.

The results in Figures 7.3 – 7.7 show that the use of regional PM cycle models does not improve the estimation accuracy on the unseen test data, compared to the global modelling results (1 model per PM cycle). Figures 7.8 and 7.9 demonstrate that regional models using plasma impedance monitor (PIM) data as input variables behave similarly to those using the EP data, with model estimation performance worsening as the number of models used to represent the full PM cycle range is increased for both PLS and GPR-based models.

Training data estimation performance however, shows improved performance with increasing numbers of models in some cases, as demonstrated in Figures 7.5 and 7.6. The increase in training data estimation accuracy is a result of the decrease in the number of

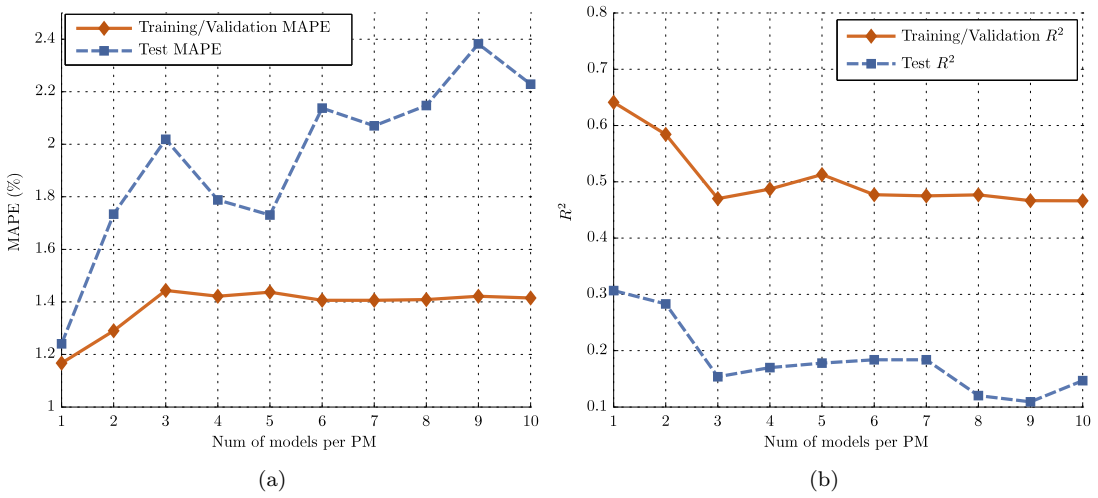



FIGURE 7.4: (a) Mean absolute percentage error (MAPE) and (b) $R^2$ results for regional PM cycle modelling technique using EP data with varying numbers of regional LARS models per PM cycle.

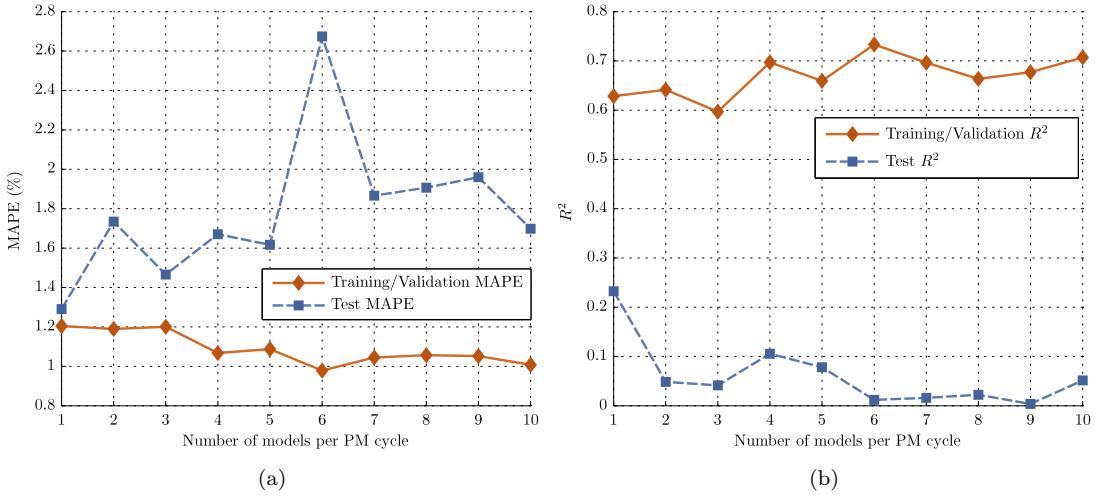



FIGURE 7.5: (a) Mean absolute percentage error (MAPE) and (b) $R^2$ results for regional PM cycle modelling technique using EP data with varying numbers of regional PLS models per PM cycle.

training samples available per model when the data are further divided, as depicted in Figure 7.10. With fewer samples available during model training, models fit the training data effectively but demonstrate poor estimation capability on the test data.

### 7.2.2 Summary

This section has used a regional modelling scheme to divide PM cycles into chronological sections and subsequently build VM models on each section separately in an attempt to improve upon the estimation accuracy accuracy of global VM models. The results in

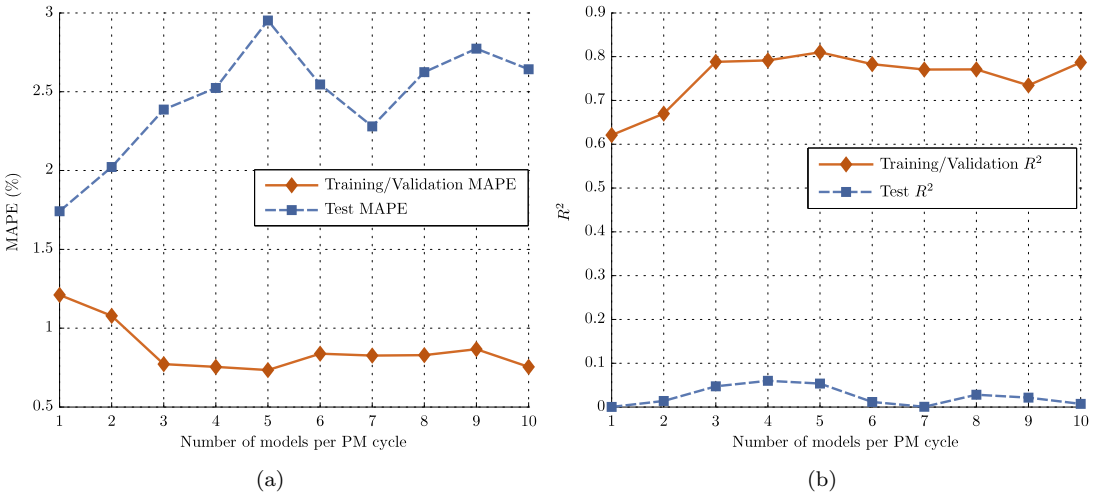



FIGURE 7.6: (a) Mean absolute percentage error (MAPE) and (b) $R^2$ results for regional PM cycle modelling technique using EP data with varying numbers of regional ANN models per PM cycle.

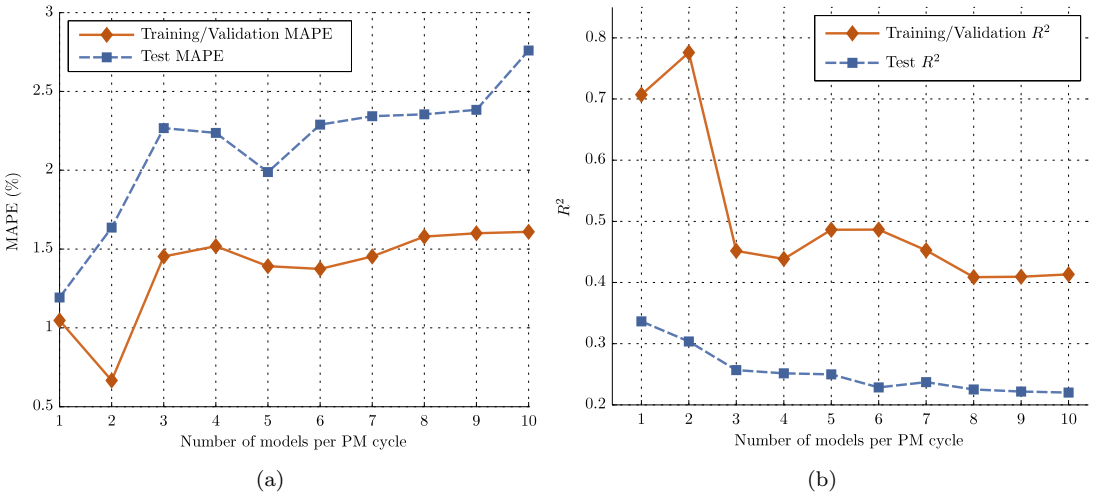
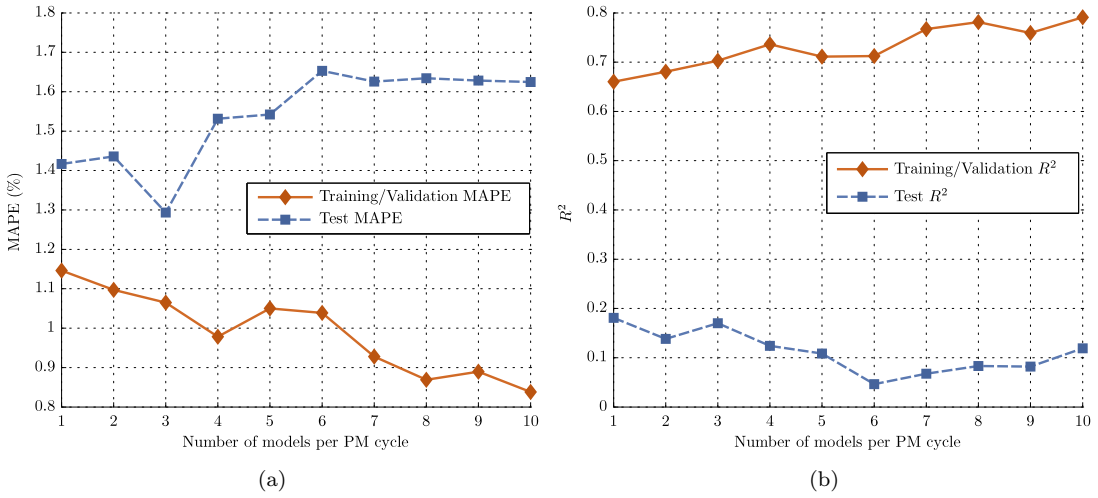


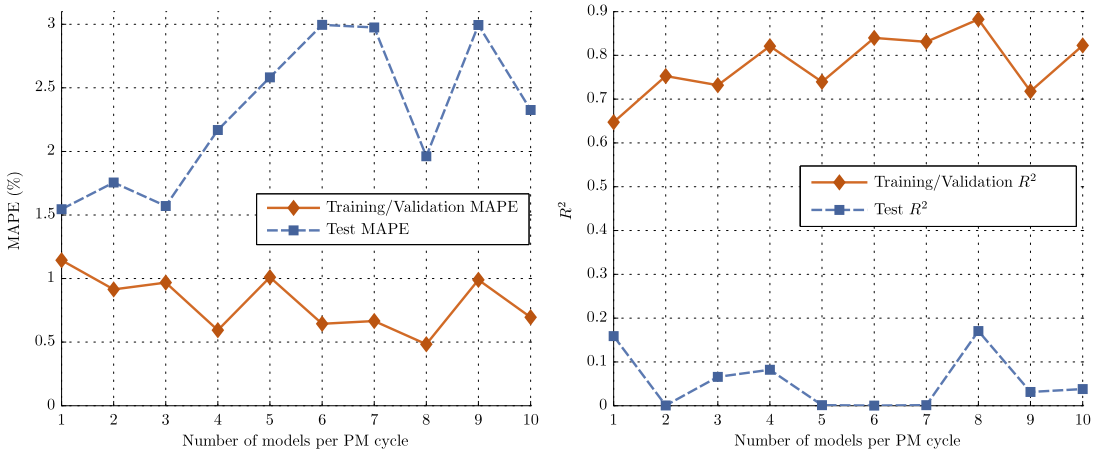

FIGURE 7.7: (a) Mean absolute percentage error (MAPE) and (b) $R^2$ results for regional PM cycle modelling technique using EP data with varying numbers of regional GPR models per PM cycle.

Section 7.2.1 demonstrate that regional PM cycle models *do not improve* the estimation accuracy of the VM models examined. The similarities that exist between corresponding regions of different PM cycles are not sufficient to aid in etch rate estimation, suggesting instead that *inter-PM variations may be more influential than intra-PM variations*. In Section 7.3, a second method of partitioning the process data to exploit similarities between different wafers to aid estimation accuracy is examined.




(a) (b)

FIGURE 7.8: (a) Mean absolute percentage error (MAPE) and (b) $R^2$ results for regional PM cycle modelling technique with varying numbers of regional PLS models per PM cycle. PIM data are used as the source data for these PLS models.




(a) (b)

FIGURE 7.9: (a) Mean absolute percentage error (MAPE) and (b) $R^2$ results for regional PM cycle modelling technique with varying numbers of regional GPR models per PM cycle. PIM data are used as the source data for the GPR models.

## 7.3 PM cycle clustering

The work of Section 7.2 examined the possibility of using similarities between corresponding regions of different PM cycles to improve etch rate estimation accuracy over that of global modelling techniques. However, the inaccurate results of Section 7.2.1 and the inaccurate chronological etch rate estimation after PM events demonstrated in Chapter 6 suggest that significant differences exist between the different PM cycles in the data set. The aim of this section is to investigate these differences and test whether any similarities between different PM cycles can be found and harnessed to improve etch rate estimation accuracy.

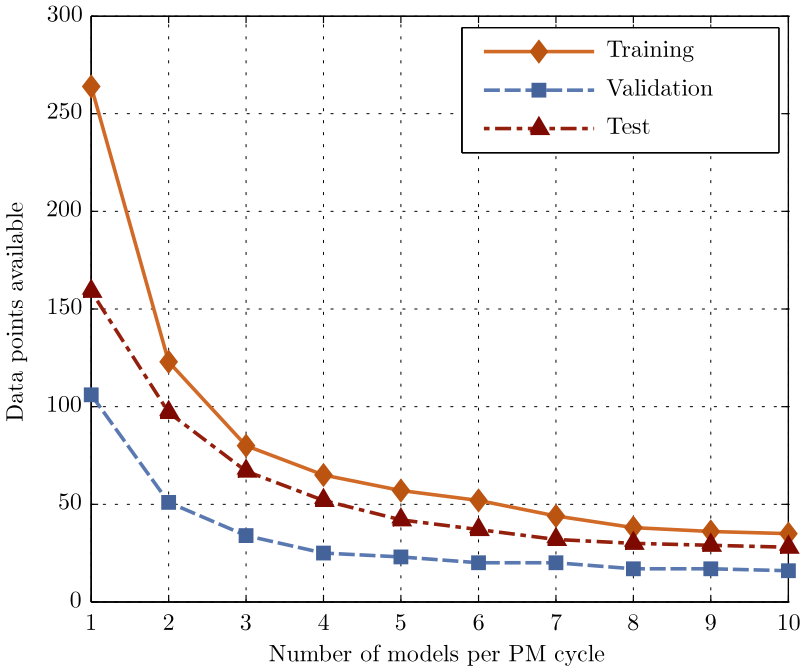



FIGURE 7.10: Variation in number of data points available for model training, validation, and testing as the number of regional models per PM cycle is increased.

### 7.3.1 Data set clusters

The first step in this investigation is the identification of similar data points in the etch data set. To allow comparisons of different sensor types, and to include as much wafer information as possible during data analysis, data containing both EP and PIM measurements for each wafer are used.

The EP variables are investigated first. Principal component analysis (PCA) is carried out on the 28 variables of the EP data after mean centering and standard deviation normalisation. PCA is used here as a data reduction technique, where the largest sources of variance in the EP data are described by the first principal components. The first three principal components, explaining 60.4% of the EP data variance, are shown in Figure 7.11, where different PM cycles are represented by different coloured points. The data set contains data from twelve different PM cycles where EP and PIM data were recorded simultaneously. The analysis is considered unsupervised because the etch depth data are not included.

Three distinct clusters exist in the EP data as shown in Figure 7.11, where, for ease of visualisation, only every fifth wafer is shown. This work is not the first occasion where clusters have been observed in semiconductor etch data. The clusters are similar in form to the clusters seen in work by Zeng and Spanos [229], where clustering algorithms are applied to an etch bias data set from two different machines. Similar patterns are seen

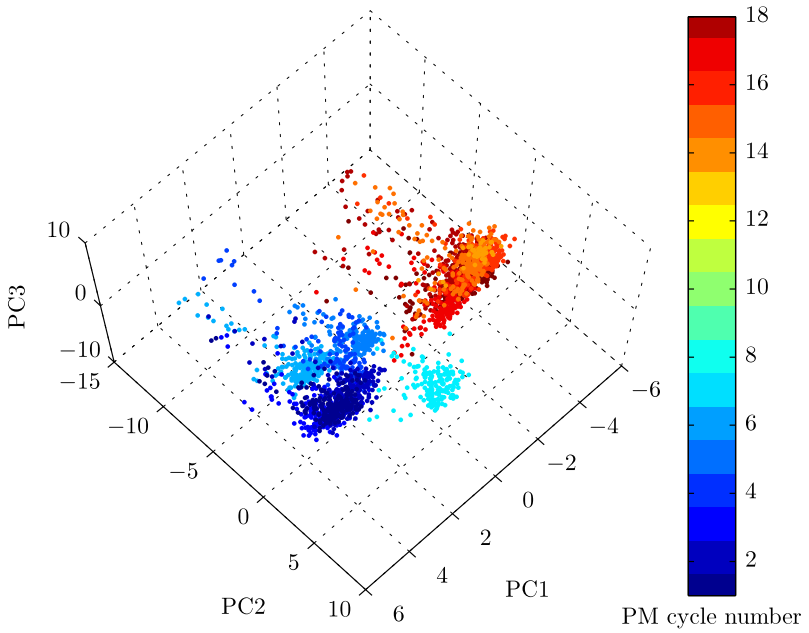



FIGURE 7.11: First three principal components (PCs) of etch process (EP) data. Data from each preventative maintenance (PM) cycle are coloured differently. Three separate clusters are visible in the data set.

in a stack etch process by He and Wang [58, 201] during work on fault detection and classification. Data clusters are also observed through an analysis of the data from Tool 2 used in Chapter 6.

The first three principal components of the PIM data which explain 64.2% of the PIM data variance are shown in Figure 7.12(a) and also reveal a number of distinct clusters, ruling out EP sensor errors as the source of the EP data clusters. The clusters are found through a PCA decomposition and truncation of the PIM voltage, current, and phase signals. Because the phase values recorded are relatively erratic signals that are not well correlated together, they are not very well suited to compression via PCA. To incorporate the phase measurements in the analysis in a more structured manner, PCA is performed on the reactance (X) and resistance (R) values for each of the 52 harmonics, calculated using Equations (2.36) and (2.37). The principal components arising from the X and R data (XR data) are shown in Figure 7.12(b); four distinctly separated clusters are found in the data.

To include as much independently measured data as possible in the cluster formation, the principal components from both the EP and the XR data are used to form Figure 7.13, resulting in more distinctly separated clusters. In Figure 7.13, the first principal component of the EP data is plotted against the first two components of the XR data to reveal four distinct clusters, where each cluster is made up of data from one or more PM cycles.

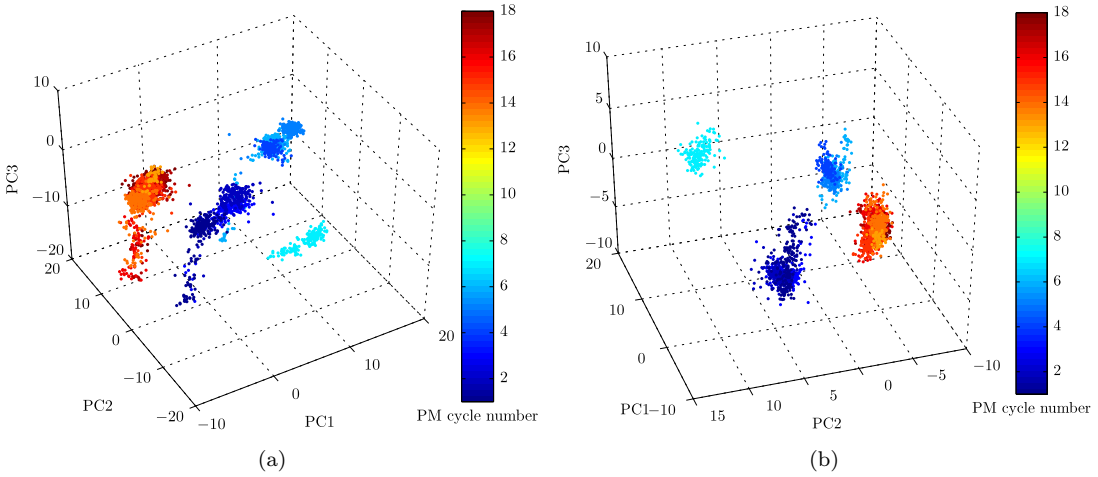



FIGURE 7.12: (a) First three principal components (PCs) of plasma impedance monitor (PIM) data and (b) first three principal components of plasma reactance (X) and resistance (R) data. Each predictive maintenance (PM) cycle is coloured differently. Four distinct clusters are visible in the data set.

The existence of the clusters in the data suggests that the etch process moves between distinct operating modes over the course of the complete data set. The changes in cluster, indicating changes of operating point, appear to be brought about by PM events. This modal operation may explain the lack of fit achieved during global modelling of the full data set, and the poor estimation results for the regional PM cycle models. The main sources of variation in the EP data, which are assigned the largest loadings during PCA, are the matching network variables, the chamber electrode temperatures, and the RF power supplied (a function of the matching network performance). For the XR data, the variance is relatively evenly spread among the harmonic data. The existence of the clusters is thought to arise from the electrical characteristics of components that are replaced during PM events and sensitivity to the precise positioning of other components by engineers.

It is important to note that the number of clusters observed is less than the number of PM cycles in the data set. This lends weight to the hypothesis the recorded data can be divided into clusters of data with similar characteristics, and it is conjectured that as the number of PM cycles of data in the data set tends towards infinity, a finite number of data clusters will be observed. Should the complete principal component space be filled when additional data is available such that individual clusters are difficult to isolate, the data can still be divided into local modes of operation using a clustering algorithm.

The objective of the following sections is to investigate whether the use of multiple specialised cluster models that concentrate on the operational space of each cluster separately is better for estimation of etch rate than the use of models with global scope.

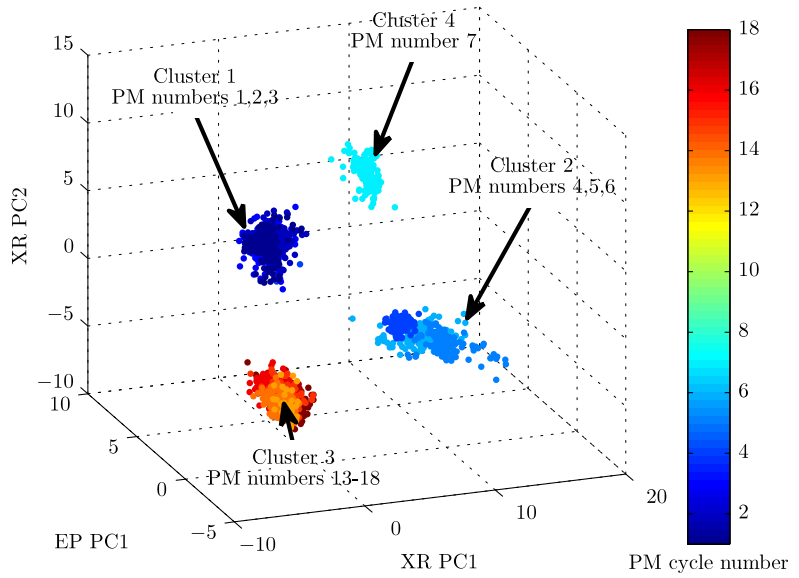



FIGURE 7.13: First principal component (PC) of EP data plotted against first two PCs of plasma reactance and resistance (XR) harmonic data. Each PM cycle is coloured differently.

### 7.3.2 Data set preparation

The wafer data containing both PIM and EP data originate from 12 different PM cycles, numbered 1–7 and 13–18. The missing PM cycles numbered 8-12 are are excluded from this analysis as they contain data with no PIM data recorded (see Table 7.1) which prevents the calculation of reactance and resistance values for the clustering identification process.

The principal components of the wafer data are shown in Figure 7.13. The four clusters are given cluster numbers for ease of reference:

1. Cluster 1: A cluster containing PM cycles 1,2, and 3. The data points in this cluster are relatively close to each other.
2. Cluster 2: A cluster containing PM cycles 4,5, and 6. The data in Cluster 2 are more spread out than the data in other clusters.
3. Cluster 3: A second tightly contained cluster containing PM cycles 13–18.
4. Cluster 4: A cluster that contains only the $7^{th}$ PM cycle.

The clusters observed in the EP and XR data are echoed in the etch rate measurements. Figure 7.14 shows the wafer etch rate with the cluster and PM divisions marked. There are differences in the average etch rate between each cluster and these are further highlighted by the box plot in Figure 7.15. An analysis of variance (ANOVA) of the

etch rates in each cluster rejects the null hypothesis that the mean etch rates in each cluster are equal at the $\alpha = 0.01$ significance level.

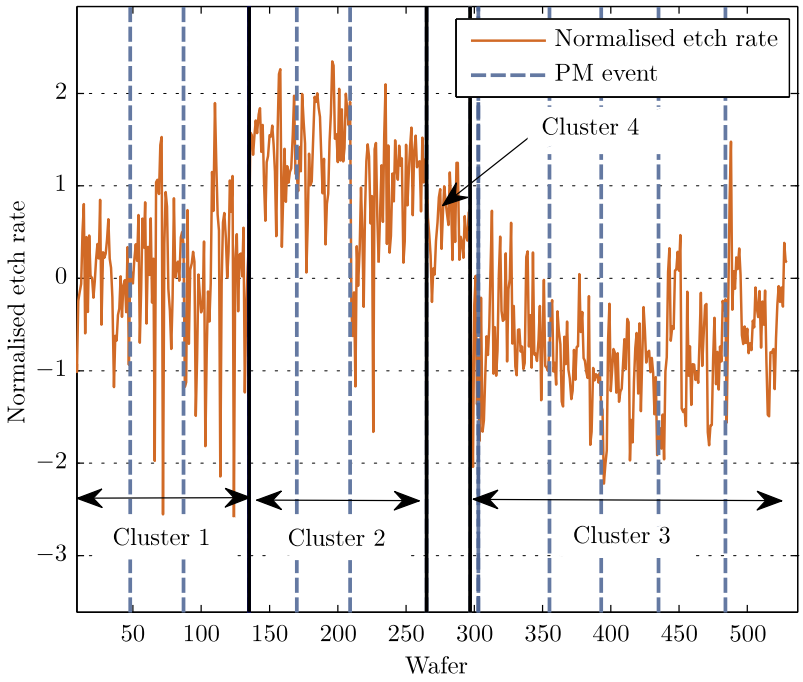



FIGURE 7.14: Normalised etch rate with cluster boundaries and PM events marked. Differences in the mean etch rate in each cluster are evident in the figure.

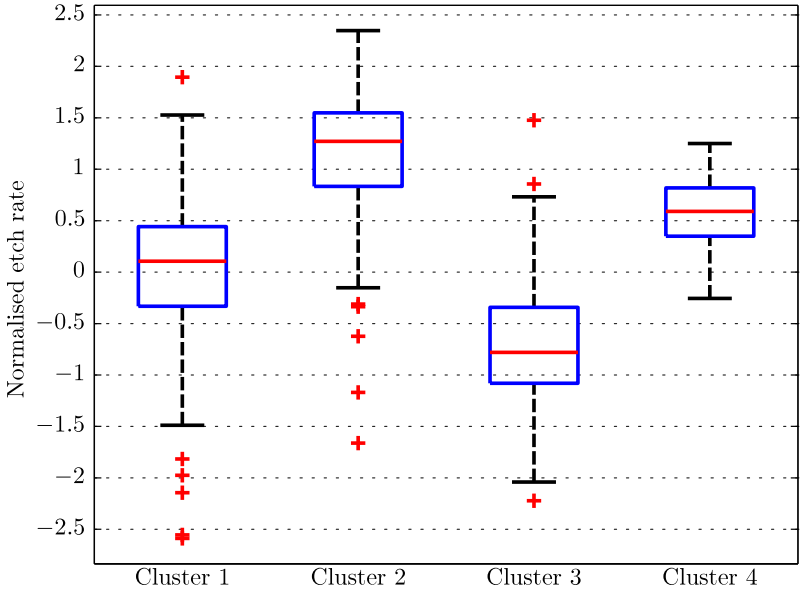



FIGURE 7.15: Box plot showing etch rate distributions in each cluster. The bottom and top of the box mark the the upper and lower quartiles of the etch rate and the band in the middle of each box marks the median etch rate for each cluster. The box whiskers extend to $+/-2.7\sigma$ and outlying points beyond these limits are marked with a '+'.

To test the performance of cluster-based modelling, data from one PM cycle is extracted from each cluster to act as test data during model evaluation. PM cycles centered

in each cluster are chosen such that the test data set comprises data from PM cycles 3,6,17, and 7, representing clusters 1,2,3, and 4 respectively. Since Cluster 4 comprises data from a single PM cycle, that is PM cycle 7, no VM model is created for the cluster. Hence, the data in Cluster 4 provide an opportunity to explore VM strategies during etching of wafers where previously defined cluster models are not available. Figures 7.16 and 7.17 show the wafer data with the training data shown in black and test data shown in red.

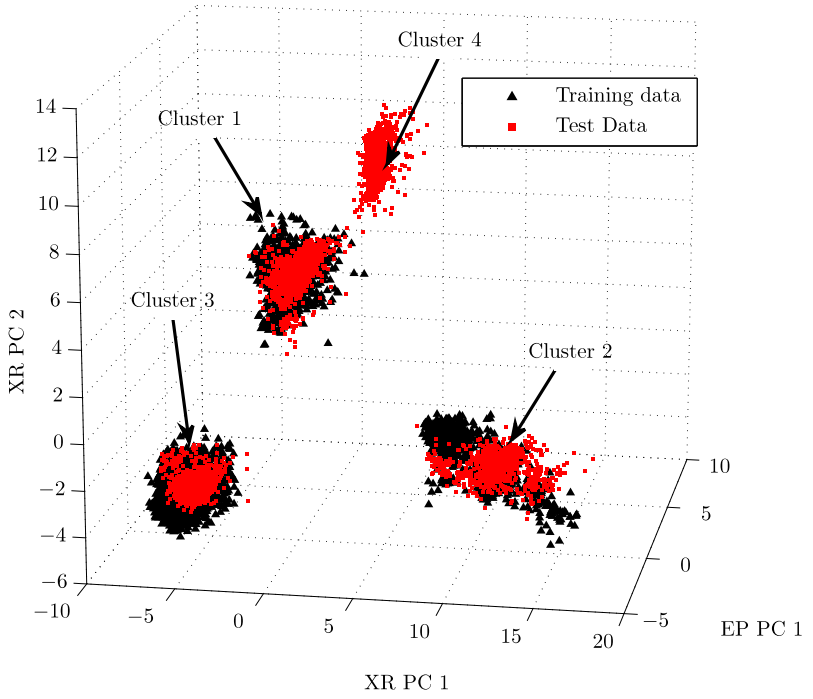



FIGURE 7.16: First principal components (PCs) of EP data plotted against first two PCs of plasma reactance and resistance harmonic data (XR data). Data from one preventative maintenance (PM) cycle in each cluster is chosen as test data and coloured in red.

For each modelling technique in turn, four different models are created; three cluster models are created, one for each of the training data sets from Clusters 1,2, and 3, and one global model is trained using the training data from all three clusters. Each cluster model is then tested using the test data points from the relevant cluster. The global model is tested on the test data from all of the clusters.

### 7.3.3 Results

The mean absolute percentage error (MAPE) on the de-normalised etch rate estimates is used to compare model results during this investigation because each cluster uses different mean and standard deviation values to normalise test data. The $R^2$ value for

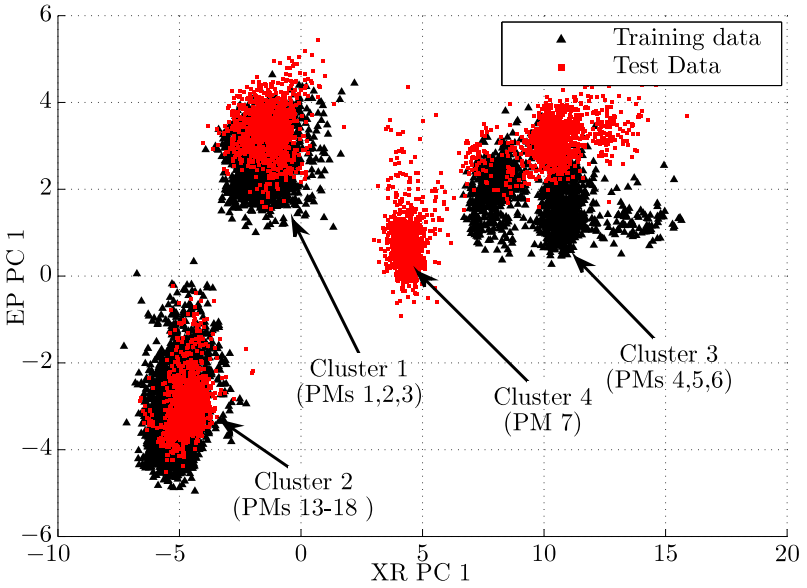



FIGURE 7.17: First principal component (PC) of EP data plotted against first PC of plasma reactance and resistance data (XR data). Again, the four separate clusters are visible in the data set, where data from one preventative maintenance (PM) cycle in each cluster is chosen as test data and coloured in red.

each model is also included to provide extra information on the quality of the estimations, since, as outlined in Section 6.6, the MAPE value alone can be misleading in some cases.

As found with the global modelling results in Chapter 6, no modelling technique or input variable combination stands out with exceptional performance when compared to the others during cluster-based modelling. Tables 7.2, 7.3, 7.4, and 7.5 compare the cluster and global model results for EP, $PIM_5$, XRZP, and EP-$PIM_0$ data respectively. The modelling results shown are based on model performance on test data from all of the different clusters. Cluster model performance is calculated by combining the estimates from each cluster model when tested on test data from that same cluster, i.e. predictions from cluster model 1 on test data from Cluster 1, predictions from cluster model 2 on test data from Cluster 2 etc. The cluster with no training data, Cluster 4, is not included in the results of Tables 7.2 – 7.5. The cluster models marginally outperform the global models in 14/20 cases.

ANN models are found to produce relatively poor results throughout cluster model testing. The poor estimation capabilities of the ANN models are attributed to the relatively small training data sets used for cluster model training; approximately 40 samples are available per PM cycle such that training data for Clusters 1 and 2 have only approximately 80 samples available for model training. Since a larger number of data points are available during global model training than for cluster model training, global ANN models generally outperform cluster ANN models regardless of input variable

| EP Data | | | | | |
|---|---|---|---|---|---|
| | Global Model | | | Cluster Models | |
| Model Typ | MAPE | | R2 | MAPE | R2 |
| PLS | ✗ | 1.79 | 0.53 | ✓ 1.60 | 0.52 |
| LARS | ✗ | 1.77 | 0.54 | ✓ 1.64 | 0.47 |
| Stepwise | ✗ | 2.58 | 0.28 | ✓ 1.65 | 0.48 |
| ANN Step | ✓ | 1.54 | 0.54 | ✗ 1.83 | 0.31 |
| GPR Step | ✗ | 1.75 | 0.36 | ✓ 1.73 | 0.45 |

TABLE 7.2: Comparison of global and cluster model performance on unseen data from each cluster using etch process (EP) data. A correct mark indicates whether the global or local model demonstrates a better MAPE.

| $PIM_5$ Data | | | | | |
|---|---|---|---|---|---|
| | Global Model | | | Cluster Models | |
| Model Typ | MAPE | | R2 | MAPE | R2 |
| PLS | ✗ | 1.65 | 0.50 | ✓ 1.57 | 0.51 |
| LARS | ✓ | 1.59 | 0.53 | ✗ 1.74 | 0.31 |
| Stepwise | ✗ | 6.11 | 0.41 | ✓ 1.62 | 0.44 |
| ANN Step | ✗ | 1.85 | 0.42 | ✓ 1.69 | 0.42 |
| GPR Step | ✓ | 1.60 | 0.52 | ✗ 1.70 | 0.46 |

TABLE 7.3: Comparison of global and cluster model performance on unseen data from each cluster using $PIM_5$ data.

| XZRP data | | | | | |
|---|---|---|---|---|---|
| | Global Model | | | Cluster Models | |
| Model Typ | MAPE | | R2 | MAPE | R2 |
| PLS | ✗ | 1.61 | 0.47 | ✓ 1.58 | 0.46 |
| LARS | ✗ | 1.60 | 0.47 | ✓ 1.59 | 0.48 |
| Stepwise | ✗ | 133.78 | 0.34 | ✓ 1.65 | 0.43 |
| ANN Step | ✓ | 1.68 | 0.50 | ✗ 2.43 | 0.43 |
| GPR Step | ✗ | 1.69 | 0.49 | ✓ 1.66 | 0.46 |

TABLE 7.4: Comparison of global and cluster model performance on unseen data from each cluster XRZP data.

selection. These results are in line with the general perception that ANNs are data-hungry [91], typically requiring a relatively large number of samples to develop useful models. Hence, the scarcity of training data in the cluster model exercise restricts the use of ANN models for clusters where few data points are available. If the ANN model results are excluded from the analysis, the cluster models outperform the global models in 13/16 of the remaining combinations of input variables and modelling techniques.

To demonstrate the degree of localisation, the three cluster models and the global

| EP-PIM$_0$ data | | | | |
|---|---|---|---|---|
| | Global Model | | Cluster Models | |
| Model Typ | MAPE | R2 | MAPE | R2 |
| PLS | ✗ 1.96 | 0.52 | ✓ 1.70 | 0.52 |
| LARS | ✗ 1.85 | 0.52 | ✓ 1.59 | 0.52 |
| Stepwise | ✗ 41.13 | 0.36 | ✓ 1.77 | 0.43 |
| ANN Step | ✓ 1.65 | 0.49 | ✗ 1.86 | 0.45 |
| GPR Step | ✓ 1.57 | 0.53 | ✗ 1.67 | 0.47 |

TABLE 7.5: Comparison of global and cluster model performance on unseen data from each cluster using EP-PIM$_0$ data.

model are tested with the test data from every cluster. The modelling results for clustered PLS and stepwise regression models using EP data are shown in Tables 7.6 and 7.7, respectively. For ease of comparison, the errors on each row of each table are colour coded from red to green indicating the worst to best performing models. As expected, cluster models perform best when tested on data from the same cluster upon which they are trained. The model performances for other input selections demonstrate similar behaviour, as detailed in [256]. Similar to the effect described in Section 6.7.3, the reduced range of etch rate in each cluster of data lead to relatively low $R^2$ values for the cluster models individually.

The model estimates from the PLS and stepwise regression models using EP data are shown in Figures 7.18 and 7.19. While global model estimates follow the trend of the actual etch rate over the complete data set, the cluster model estimates are inaccurate outside of the input space upon which each cluster model is trained. Results similar to those of Tables 7.6 and 7.7 are repeated across all of the input variable and model combinations.

In tests completed for Cluster 4, for which no cluster model was built, none of the models from the other clusters are capable of accurately estimating etch rate. However, in general, the global models yield the most accurate estimates. Therefore, in a production environment, in cases where etch tools operate in a previously unseen operating space, optical measurements of etch depth can be taken with greater frequency than before to allow model identification for the new cluster and to ensure that the process is operating within specifications. If estimates must be made, the global models are most likely to provide a reasonable estimation of the real etch rate.

Although Cluster 1 is a relatively closely packed cluster compared to Cluster 2, modelling results for Cluster 1 are relatively poor. The lack of fit between the estimated etch rate and real etch rate in Cluster 1 is seen in both global and local model predictions,

| PLS Cluster Models using EP data | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Model Source | | | | | | | |
| | | Cluster 1 | | Cluster 2 | | Cluster 3 | | Global | |
| | | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 |
| Data Source | Cluster 1 | 1.85 | 0.14 | 4.03 | 0.00 | 2.83 | 0.01 | 1.89 | 0.11 |
| | Cluster 2 | 4.06 | 0.01 | 1.53 | 0.12 | 3.11 | 0.03 | 1.98 | 0.17 |
| | Cluster 3 | 4.53 | 0.00 | 1.54 | 0.14 | 1.43 | 0.16 | 1.49 | 0.24 |
| | Cluster 4 | 2.31 | 0.06 | 1.82 | 0.12 | 1.97 | 0.03 | 1.84 | 0.01 |

TABLE 7.6: Detailed results from clustered PLS modelling of etch rate data using EP data. Coloured cells are used to allow easy comparison between model results. Cluster models perform best when tested with unseen data from the same cluster from which they were trained, while global models provide reasonable estimates throughout the data set.

| Stepwise Cluster Models using EP data | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Model Source | | | | | | | |
| | | Cluster 1 | | Cluster 2 | | Cluster 3 | | Global | |
| | | MAPE | R2 | MAPE | R2 | MAPE | R2 | MAPE | R2 |
| Data Source | Cluster 1 | 1.84 | 0.18 | 7.67 | 0.00 | 6.54 | 0.01 | 3.23 | 0.03 |
| | Cluster 2 | 9.79 | 0.09 | 1.68 | 0.03 | 6.58 | 0.01 | 2.82 | 0.11 |
| | Cluster 3 | 67.95 | 0.06 | 19.22 | 0.00 | 1.42 | 0.15 | 1.67 | 0.07 |
| | Cluster 4 | 5.02 | 0.09 | 2.52 | 0.01 | 8.93 | 0.02 | 3.25 | 0.05 |

TABLE 7.7: Detailed results from clustered stepwise regression modelling of etch rate data using EP data.

with errors close to 2 % MAPE in many cases. The high MAPE values are due to erratic changes in etch rate in the wafers contained within Cluster 1. The test wafer etch rate variance in Cluster 1 (2.6 $(\text{nm/sec})^2$) is noticeably higher than for Clusters 2 and 3 (1.7 and 1.3 $(\text{nm/sec})^2$, respectively). The high frequency deviations in etch rate in the cluster do not appear to be reflected in the VM model input variables.

Poor estimation quality for Cluster 2 is recorded for all of the input selection and modelling techniques examined. Both Cluster 1 and Cluster 2 have fewer samples for training than Cluster 3 and, additionally, the data points contained within Cluster 2 are not closely packed in the principal component space as seen in Figures 7.16 and 7.17. The larger distances between the wafers of Cluster 2 indicate a dissimilarity between each wafer measured that presents a difficulty for the VM models.

Cluster 3 is more accurately estimated than the etch rate in the other clusters throughout the study. The elevated accuracy seen for Cluster 3 is attributed to the fact that more data points are available for Cluster 3 than any other cluster, and in the principal component space of Figure 7.16, the data points of Cluster 3 are very closely packed.

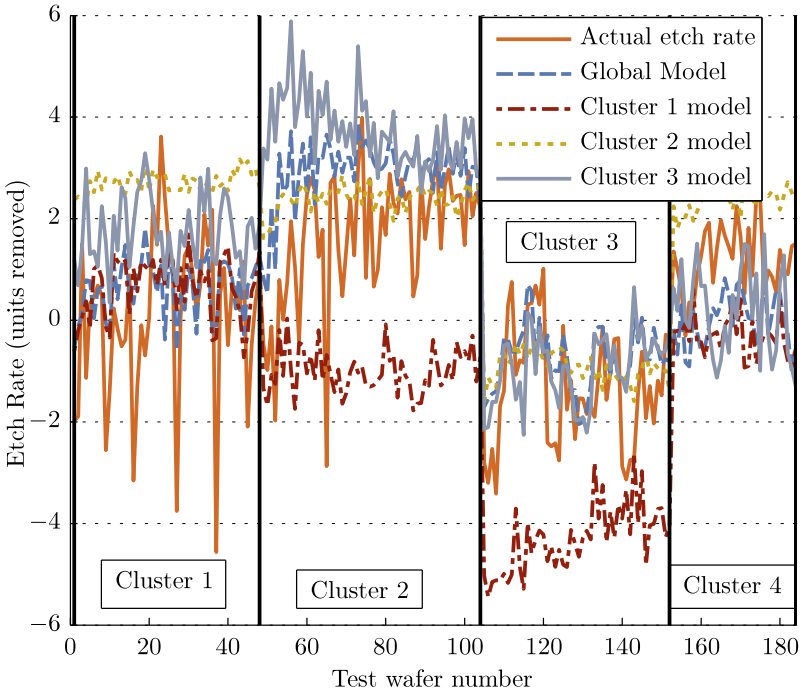



FIGURE 7.18: Estimates from clustered PLS modelling of etch rate data using etch process (EP) data. Cluster models are only useful for etch rate prediction on wafers from the same cluster upon which they were trained.

### 7.3.4 Automatic determination of clusters

The identification of different clusters within the data set can be automated by means of a clustering algorithm. For the purposes of demonstration, the well-documented k-means clustering algorithm is applied to the principal component scores shown in Figure 7.16.

K-means clustering is an unsupervised method of clustering analysis that separates a data set $\mathbf{X} \in \mathbb{R}^{n \times p}$, of $n$ observations and $p$ variables, into $k$ clusters where each observation in $\mathbf{X}$ is assigned to the cluster with the closest mean. Proximity is determined using a pre-defined distance metric in $p$-dimensional space, such as the Euclidean or Mahalanobis distance. The mean of each cluster is called the cluster *centroid*. The k-means algorithm operates to minimise the within-cluster sum of squares (WCSS), given by

$$WCSS = \sum_{i=1}^{k} \sum_{\vec{\mathbf{x}}_j \in S_i} \|\vec{\mathbf{x}}_j - \vec{\boldsymbol{\mu}}_i\|^2 \tag{7.1}$$

where $\mathbf{X}$, made up of row vectors $\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_2, \cdots \vec{\mathbf{x}}_n$, is split into $k$ clusters, $\mathbf{S} = \{S_1, S_2, \cdots S_k\}$ with centroids $\vec{\boldsymbol{\mu}}_1, \vec{\boldsymbol{\mu}}_2, \cdots \vec{\boldsymbol{\mu}}_k$. K-means is an iterative algorithm that adjusts centroid locations over each iteration to minimise the WCSS. The algorithm proceeds as follows:

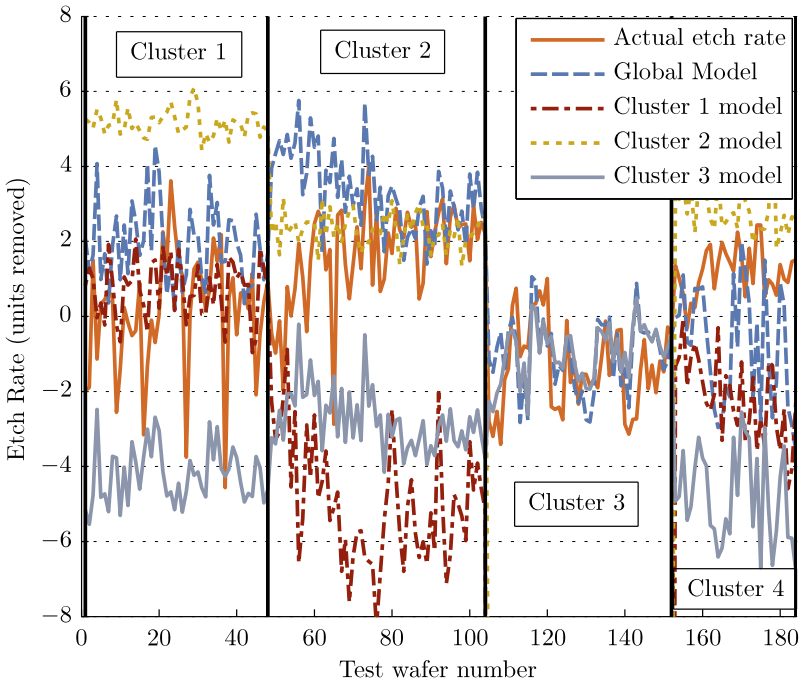



FIGURE 7.19: Estimates from clustered stepwise regression modelling of etch rate data using etch process (EP) data.

1. Choose $k$ observations (rows) at random from $\mathbf{X}$ as centroid locations.
2. Assign each observation in $\mathbf{X}$ to the closest centroid to form preliminary clusters. The analyst can choose the distance metric depending on the application.
3. Update the centroid locations to the mean of the newly formed clusters.
4. Steps 2 and 3 are repeated until the centroid assignments settle in fixed locations and the algorithm has converged.

The final solution found by the k-means algorithm is dependent on the choice of starting centroid locations, and as a result, it is common to repeat the analysis a number of times before settling on a solution. K-means clustering is a heuristic algorithm and there is no guarantee that the final solution is the global optimum. The main drawback of k-means clustering is that the number of clusters must be supplied to the algorithm before it is executed. However, if the clustering process is to be automated, statistical coefficients such as the silhouette coefficient [257] can be used to estimate the optimum number of clusters.

The application of the k-means algorithm to the etch rate training data set effectively identifies the three clusters in the training data described in Section 7.3.1, using the three principal component variables from the EP data and the XR harmonic data. The test data points can be assigned to their correct clusters using the distance from each data

point to the cluster centroids. The test data centroid distances are shown in Figure 7.20. The correct cluster is identifiable from the first wafer of each new PM cycle, allowing the application of the correct cluster model upon the start of each PM cycle. The data points from Cluster 4 belong to none of the identified clusters, as is evident from Figure 7.20, where no cluster centroid is close to the Cluster 4 data points.

In a hypothetical on-line application of the clustering method, the centroid distances can be used to determine when recently processed wafers belong to completely new clusters. Wafer data that does not belong to an existing cluster requires the use of the global model for estimation, and modelling of new cluster information when etch rate metrology becomes available.

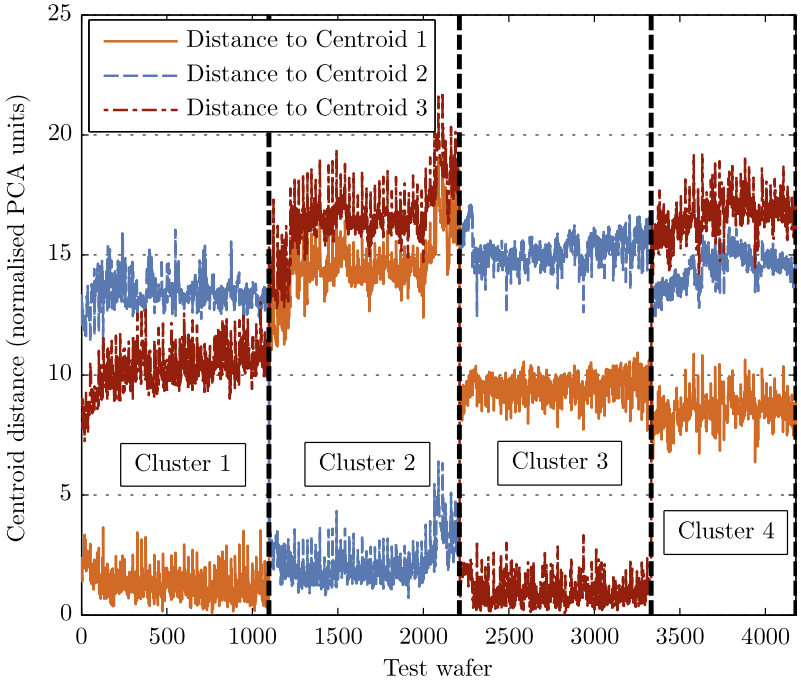



FIGURE 7.20: Euclidean distances from test points to cluster centroids. These distances are calculated in 3-D space using principal components from the EP and X&R data.

Attempts to further improve estimation accuracy using weighted combinations of the cluster and global model estimates, depending on each samples proximity to cluster centroids, are documented in [256]. Based on the premise that data points within a threshold distance of the center of each cluster are more suited to cluster models than wafers on the outskirts of each cluster, a fuzzy weighting scheme is devised as depicted in Figure 7.21. However, no significant increases in accuracy are achieved using this scheme.

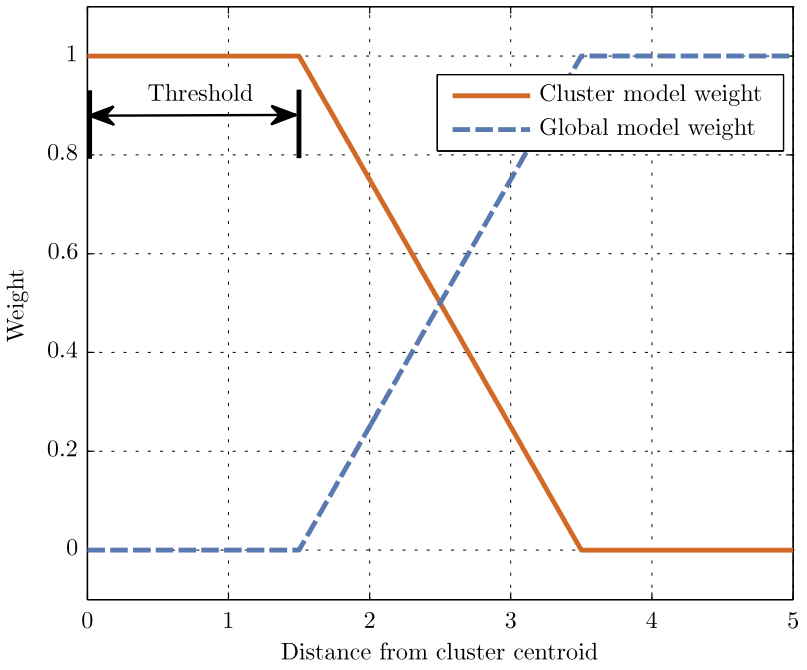



FIGURE 7.21: Fuzzy weighting scheme for cluster modelling. Estimates for wafers further from the cluster centroid than a threshold value are made up of a weighted sum of the estimates from the cluster and global models.

### 7.3.5 Summary

The wafers containing the available PIM and EP data from Tool 1 can be segmented into four separate clusters identifiable through PCA of the recorded process variables from the chamber. Three of the clusters contain data from multiple PM cycles while the fourth cluster only contains data from one PM cycle. Although the data set in this case does not revisit a previously existing cluster when the wafers are ordered chronologically, only 12 PM cycles are examined. With data spanning a greater number of PM cycles, it is hypothesised that later PM cycles will exhibit behaviour similar to previously seen clusters allowing previously defined cluster models to be used for etch rate estimation. As the number of PM cycles observed increases, it is expected that a finite number of clusters will approach a finite constant.

For the majority of input variable and modelling technique combinations, specialised cluster models trained on each cluster are shown to produce more accurate etch rate estimates than globally defined models for unseen test data points. However, overall, the most accurate estimates are achieved using an ANN-based global model with EP data. While reasons for inaccurate etch rate estimates for data from Clusters 1 and 2 are explained in Section 7.3.3, such data is representative of production data, and such difficulties are to be expected in a production environment.

Due to the lack of consistent results and the considerable additional implementation complexities of the clustering technique for relatively small improvements in estimation accuracy, global modelling is preferred to clustered modelling for the data set investigated. However, the feasibility of the clustering technique is demonstrated, along with the potential for improvements in VM accuracy should larger data sets that capture data from more operational spaces be available.

## 7.4 Sliding window models

Continuously updated models that maintain etch rate estimation accuracy in the presence of fluctuating chamber characteristics are explored in this section. The models use stepwise regression, LARS, PLS, ANN and GPR models, as described in Sections 3.3 – 3.8, to estimate etch rate.

### 7.4.1 Introduction to windowed models

*Windowed models* are VM models that are created using an initial set of training data and used to estimate the outputs of further wafers as they are processed until new metrology data becomes available. When new metrology becomes available, the oldest information in the training data set is disregarded and the model is retrained or updated, taking the new measurements into account. The training, estimation, and updating procedures are repeated, creating a moving *window* of training data samples. The number of data samples used as training data for the VM model is specified by the *window length*.

Windowed models can be used to maintain model currency and relevance in time-varying systems. The use of such schemes is not atypical among modelling practitioners concerned with systems that exhibit time-varying properties. For example, Qin *et al.* [258] propose a moving-window PLS scheme with a forgetting factor applied to a catalytic reformer. Malinowski [259] use windowed data sets in conjunction with factor analysis to determine concentrations of components in chemical processes. Dayal and McGregor [260] describe a recursive PLS modelling system that incorporates a forgetting factor to model a continuously-stirred tank reactor and an industrial flotation circuit. Li *et al.* [261] use a recursively updating PCA model to monitor a thermal annealing process.

Windowed models have also been successfully applied to VM applications in semiconductor manufacturing. For example, Khan *et al.* [11] describe a VM and run-to-run control strategy that uses a continuously updating windowed PLS model in a simulated

semiconductor manufacturing process. While the windowed PLS system described in [11] is applied to a simulated process that incorporates an exactly defined drift term, the models described in this chapter are created and tested using actual *production* etch data. Kang *et al.* [231] apply windowed models for VM of photolithography variables, finding that model accuracy improves for longer window lengths (using up to 5 months of data comprising over 1700 wafers).

Techniques exist for incorporating information from new data samples efficiently into PLS [11, 59, 260] and PCA [196, 261] models. Such techniques are often referred to as recursive or adaptive PCA and PLS methods. The training times of the plasma etch VM models under investigation in this chapter are of the order of seconds and, as a result, models can be completely retrained using the new window of training samples at every iteration without the need for recursive algorithms. In a production system, if window lengths become so large so as to become restrictive in terms of training times, recursive modelling techniques can be employed to allow faster model refreshing.

In Section 7.4.2, a weighting factor is introduced to produce a novel weighted PLS algorithm, where the sample weights are assigned according to the maintenance history of the plasma chamber. The weighted PLS algorithm and the determination of the sample weights is described in Section 7.4.2. A description of the models investigated in this chapter is presented in Section 7.4.3 and the windowed modelling results are presented in Section 7.4.4.

### 7.4.2 Weighted PLS algorithm

#### Weighted PLS

It is shown in Section 3.2 that a vector of weighted least squares model coefficients can be calculated as

$$\hat{\boldsymbol{\beta}}_W = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}, \tag{7.2}$$

where $\hat{\boldsymbol{\beta}}_W \in \mathbb{R}^{p\times 1}$ are the weighted least squares regression coefficients, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a matrix of variable observations, with $n$ rows corresponding to samples and $p$ columns corresponding to input variables, $\mathbf{W} \in \mathbb{R}^{n\times n}$ is a diagonal matrix of weights applied to each sample in $\mathbf{X}$, and $\mathbf{y} \in \mathbb{R}^{n\times 1}$ is a vector of system outputs.

A combination of PLS and weighted least squares can be implemented. For systems with a single output variable such as the plasma etch system, PLS can be visualised as a regression procedure where the PLS components $\mathbf{T} \in \mathbb{R}^{n\times l}$ ($l$ represents the number

of components kept in the model) generated from the input matrix $\mathbf{X}$ are used as input variables to a LSR model to estimate the single system output variable contained in $\mathbf{Y} \in \mathbb{R}^{n\times 1}$. In the case of etch rate or etch depth estimation, $\mathbf{Y}$ contains only one variable and is not decomposed into $\mathbf{Y}$-loading and $\mathbf{Y}$-component matrices as described in Section 3.6. $\mathbf{Y}$ becomes a vector $\mathbf{y}$. Hence, the inner-relation for the PLS algorithm (Equation (3.29)) can be described as

$$\mathbf{y} = \mathbf{T}\bar{\boldsymbol{\beta}}, \tag{7.3}$$

where $\bar{\boldsymbol{\beta}}$ is found using the LSR algorithm. A weighted-PLS system can be developed by introducing weights to this LSR problem. Each row of $\mathbf{T}$ can be assigned a weight specified in $\mathbf{W}$, and the model parameters $\bar{\boldsymbol{\beta}}$ can be determined using weighted least squares regression. Hence, $\bar{\boldsymbol{\beta}}$ is given by

$$\bar{\boldsymbol{\beta}} = (\mathbf{T}^T\mathbf{W}\mathbf{T})^{-1}\mathbf{T}^T\mathbf{W}\mathbf{y}, \tag{7.4}$$

where $\mathbf{W}$ is a diagonal matrix of weights relevant to each sample contained in $\mathbf{T}$. The weights can be determined through knowledge of the system being modelled or knowledge of the conditions under which samples were collected.

### Calculation of regression weights

The sample weights used in the weighted-PLS models in this chapter are determined depending on the relevance of each training sample to the data point for which VM is required. This approach is similar to work carried out by Dayal and McGregor [260] where continuously updated PLS models were used to model physical systems. In their work, Dayal and McGregor [260] developed a method to update the PLS model and incorporate a "variable forgetting factor" that discards old data, deemed irrelevant, from the training window. The advantages of using recursive PLS over recursive LSR in the presence of correlated input variables and short data windows is also highlighted in [260]. The idea of weighting individual samples to improve modelling performance is also seen in work by Xu *et al.* [262] who use particle swarm optimisation (PSO) in a black-box manner to determine the best sample weights to optimise model performance.

In the weighted PLS VM models used in this chapter, the sample weights are varied in accordance with the tool maintenance history to satisfy two assumptions. Firstly, due to process drift, it is assumed that more recent samples closer to the *front* of the window are more relevant for estimation of current samples (the *front* of the window is

defined as the most recent sample added to the window). Secondly, it is assumed that samples contained within the same PM cycle as the current sample are more relevant for estimation than samples from previous PM cycles and that data from older PM cycles become less and less relevant as more wafers are processed. Some weight is added to this assumption by the results of Section 6.6.1 where PM events cause inaccuracies in previously defined models and Section 7.3.3 where cluster models estimate etch rate better using models trained using neighbouring, more recent PM cycles (i.e. in the same cluster) than older PM cycles.

A number of steps are taken to determine weights that satisfy these assumptions. Initially, samples are assigned linearly increasing weights across the window, with the sample at the back of the window given a weight of 0 and the sample at the front of the window given a weight of 1. Next, the samples weights are adjusted according to the number of PM cycles spanned by the window, as shown in Figure 7.22. The samples contained in the most recent PM cycle are assigned the largest weights, and older PM cycles are assigned progressively smaller weights. For example, in Figure 7.22, the window spans three PM cycles; 2 is added to the samples in the most recent PM cycle, 1 to the second most recent, and 0 to the oldest PM cycle. The weights are then exponentially transformed to accentuate the effect of the weighting scheme, such that $w_i' = e^{w_i}$, where $w_i$ is the weight of the $i^{th}$ sample in the window and $w_i'$ is that weight after the transformation. The exponential transformation of the weights has the effect of increasing the overall range of the weights, and hence the relative effects of each one, leading to the weighting profile shown in 7.23.

### 7.4.3 Data choice for windowed models

The modelling techniques listed in Section 7.1.2 are investigated as candidate modelling techniques for the windowed modelling scheme, thus including linear techniques (stepwise regression and LARS), a component-based technique (PLS), a non-linear technique (ANNs), and a non-parametric technique (GPR) in the analysis. Two types of models are examined in this section:

Up to this point, *time-invariant* models have been discussed for VM of plasma etch, that is models that are trained once using a set of training data and are used for VM of all future data points. The training algorithm is applied only once at initialisation using the training data set. These models do not update over time and, as described in Chapter 6, are most suitable for systems that do not vary in time or radically change behaviour, and hence, are not useful for long term use in some plasma etch systems [61, 224].

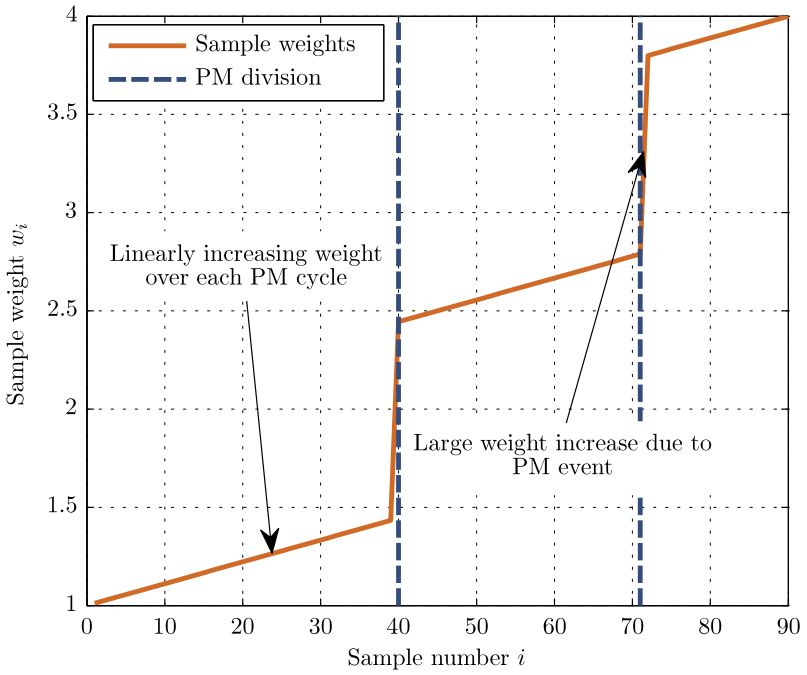



FIGURE 7.22: Weighting profile for window length of 90 samples spanning three PM cycles. The most recent samples, corresponding to the front of the window, are on the right of the figure.

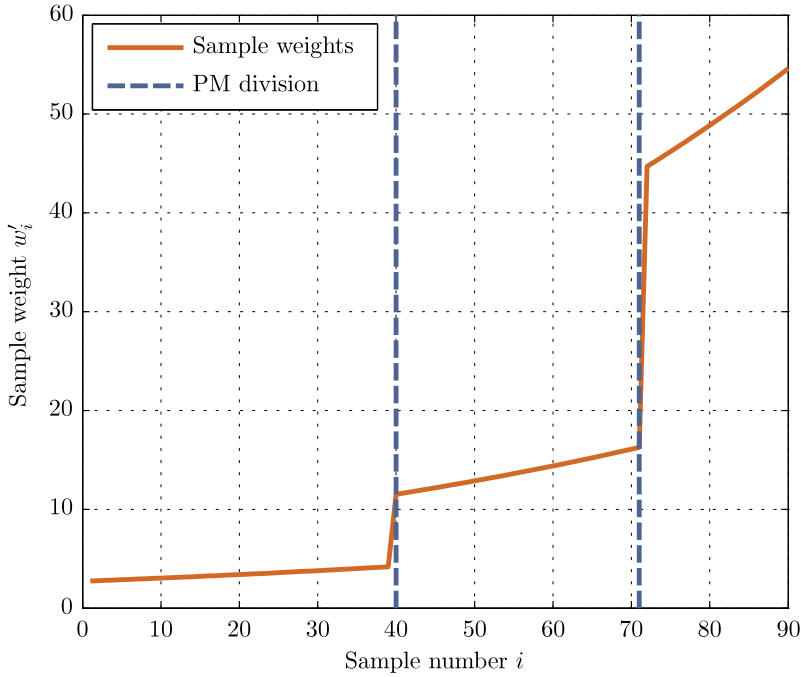



FIGURE 7.23: Weighting profile for window length of 90 samples after exponential transformation.

Windowed models are models that are retrained every time new metrology information becomes available, incorporating the newest information and discarding the oldest information in the training data set.

Because the windowed models depend on the order of the wafers in the data set, contiguous data sets are preferable to those with missing data for testing the algorithms. Hence, the data from Tool 1 containing EP data only is used for windowed model analysis because it is the largest data set comprising over 18000 wafers and the data are almost fully contiguous; there is only one chronological gap of approximately 1 month in the data set. The data set is kept in chronological order throughout the investigation, providing a realistic representation of the type of data collected from etch tools during production. The etch rate variations over the complete data set are shown in Figure 7.24. The modelling results are compared on the basis of MAPE and $R^2$.

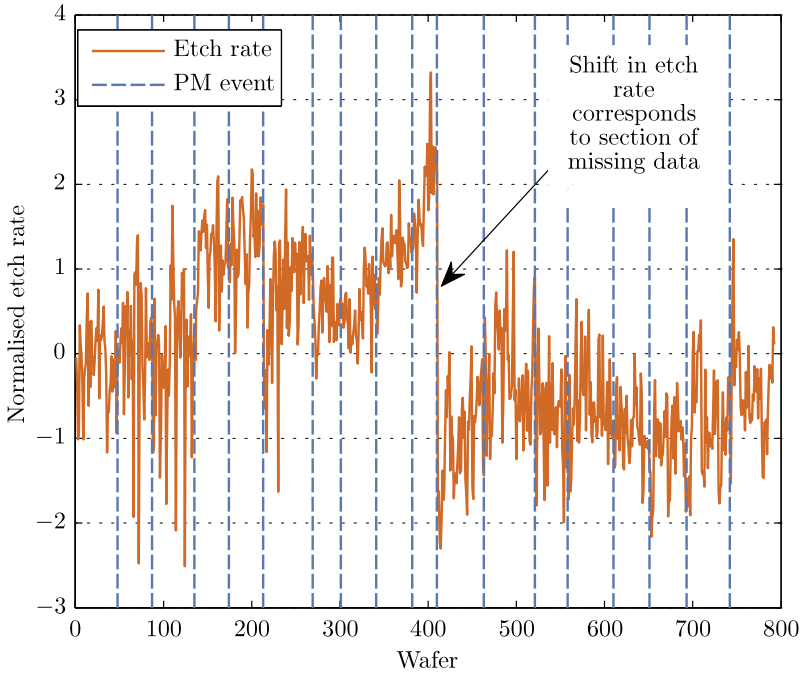



FIGURE 7.24: Optically measured etch rate values for wafers from Tool 1. This data set contains 789 measured wafers and spans 18513 wafers and 18 different PM cycles. One substantial chronological gap in the data exists as labelled, where data collection was disrupted for some time. Only etch process (EP) data is available for all of the wafers.

### 7.4.4 Results

The windowed VM modelling schemes are applied to the etch data set using window lengths ranging from 30 to 300 samples. Wafers after wafer 300 are used as test data points for all model types investigated, with an appropriate portion of the first 300 wafers being used to initialise the first model in each test. Hence, for each window length investigated, 493 different models are trained for each modelling technique, corresponding to one model for each test data point.

Firstly, in Figure 7.25 the MAPEs for windowed PLS models are compared with the MAPEs for time-invariant PLS models for a number of window lengths for comparison purposes. In the case of time-invariant model results presented, the specified window length denotes the number of wafers used to initially create the model, which then remains constant and is used to estimate the etch rate for all of the test data points. MAPEs larger than 3.7% are observed for the time-invariant models for all of the window lengths investigated, demonstrating that the these models fail when estimating unseen etch rate variations. $R^2$ values of below 0.5 are also recorded for the time-invariant models [256]. In a similar fashion, large MAPEs and unreasonable estimates are observed for time-invariant models using the other modelling techniques listed in Section 7.1.2 and hence, such models are not considered for the remainder of this discussion.

The windowed modelling MAPE results for all of the modelling techniques are shown in Figure 7.26. The etch rate estimation accuracy of the PLS and ANN models increases with increasing window lengths. For the stepwise, LARS, and GPR models, window lengths of 60 – 80 samples appear to generate the most accurate estimates.

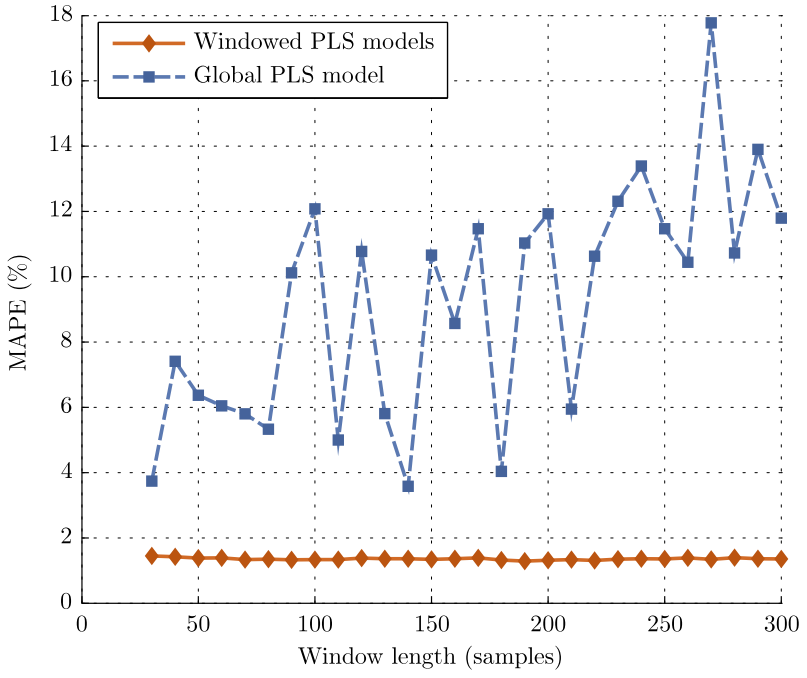



FIGURE 7.25: Mean absolute percentage error (MAPE) for global and windowed PLS models. The global models do not produce accurate results over the test data points.

Figure 7.27 examines the performance of the windowed PLS models with and without the weighting scheme described in Section 7.4.2. Increased accuracy over non-weighted windowed PLS models is achieved using the sample weighting scheme for the majority of window lengths. The increase in accuracy is expected because the maintenance history of the chamber is taken into account during the model creation for each window of training data.

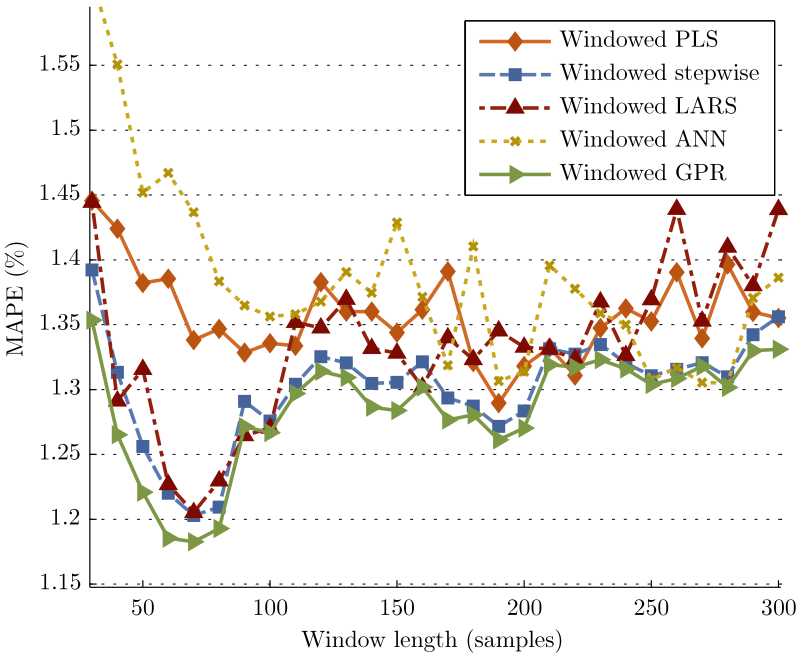



FIGURE 7.26: Mean absolute percentage error (MAPE) for windowed models using different modelling methods. While generally MAPE decreases with increasing window length, an optimal window length of 70 samples appears to generate the best results for three of the modelling techniques.

It is found that further improvements in estimation accuracy can be accomplished with the inclusion of the most recently measured value for etch rate as an input variable in each windowed model. The improvement in estimation accuracy arises because there is an element of autocorrelation in the etch rate measurements. Work comparing windowed PLS models using such *autoregressive* inputs to windowed models without autoregressive inputs is further reported in [263], and the results for the windowed modelling scheme using the autoregressive input are shown in Figure 7.28. The use of two autoregressive inputs (the two most recent etch rate measurements) is found not to appreciably increase the estimation accuracy over the results shown in Figure 7.28.

The GPR and ANN windowed models use stepwise selection (see Section 3.3) to reduce the number of input variables before model training. GPR models perform consistently better than the other modelling techniques, with the best results reported for window lengths of 70 wafers as shown in Figures 7.26 and 7.28. For small window lengths, windowed ANN models perform poorly due to a lack of training data. Increasing the window length improves ANN estimation performance to a level comparable with the other modelling techniques investigated, but the GPR models perform more accurately for all window lengths. The $R^2$ values for the windowed models using the autoregressive inputs are shown in Figure 7.29.

The $R^2$ and MAPE values of the windowed models are substantially better than

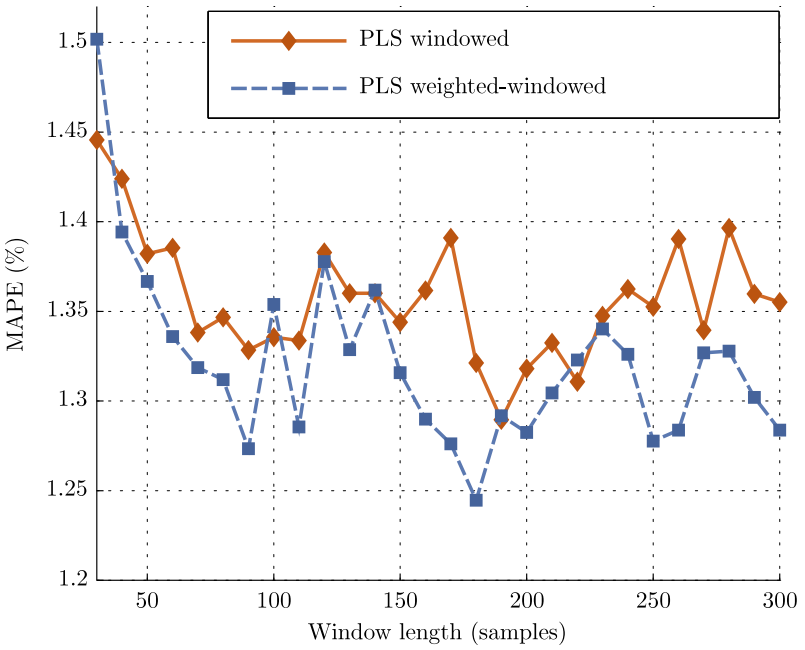



FIGURE 7.27: MAPE for windowed PLS models with and without the maintenance-dependent sample weighting scheme. The etch rate estimation accuracy increases when the weighting scheme is used because the maintenance history of the chamber is accounted for during modelling, weighting each sample according to their relevance to the wafer for which VM is required.

those of the time-invariant models and the global models of Chapter 6. Considering that much smaller amounts of training data are used during windowed model training, and that the models can perform accurately over PM events in the test set, windowed models are preferred.

The best results for the windowed models without the autoregressive term are recorded for a window length of 70 wafers using GPR models with a stepwise selection of the input variables. This model estimates the actual etch rate for the 493 test wafers with a MAPE of 1.18 % and $R^2$ of 0.73. Incorporation of the autoregressive term in the GPR models results in a best MAPE of 1.14 % with an $R^2$ of 0.75, using a window length of 70 wafers. A weighted-windowed PLS model with autoregression and a window length of 200 achieves a MAPE of 1.2 % with $R^2$ of 0.72. Figure 7.30 shows the etch rate estimates for the best PLS-based models and, for comparison, a time-invariant PLS model created using an initial training set of 200 wafers. The time-invariant PLS model fails to produce useful estimations of etch rate due to shifts in the etch performance as wafers are processed and PM events occur.

The high $R^2$ statistics exhibited by the windowed models are not directly comparable to the $R^2$ statistics reported for the chronological global modelling results of Chapter 6. In the windowed model case, the error statistics for each model are calculated on estimates of etch rate that have a greater range than in the global model case. While the

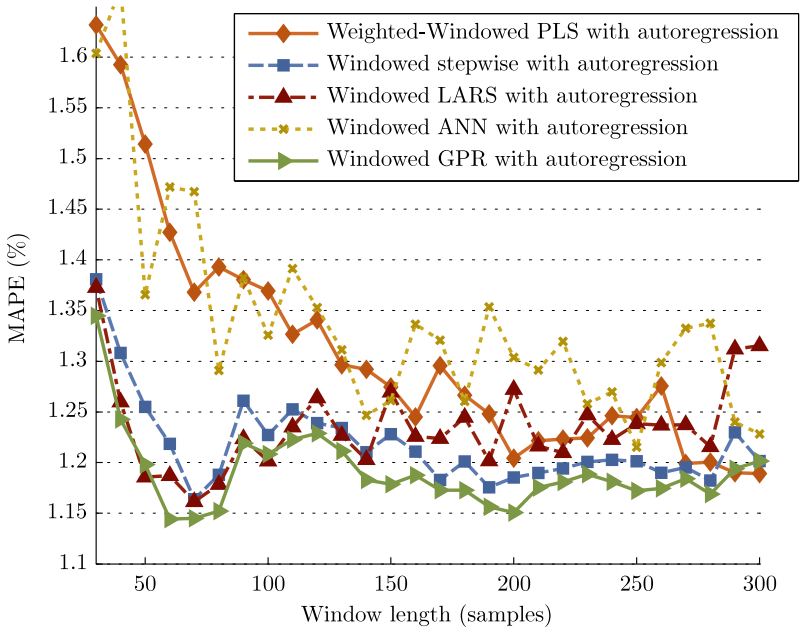



FIGURE 7.28: MAPE for windowed models using an autoregressive input. Inclusion of the most recently measured value of etch rate as an input variable to the windowed models has the effect of increasing the VM accuracy.

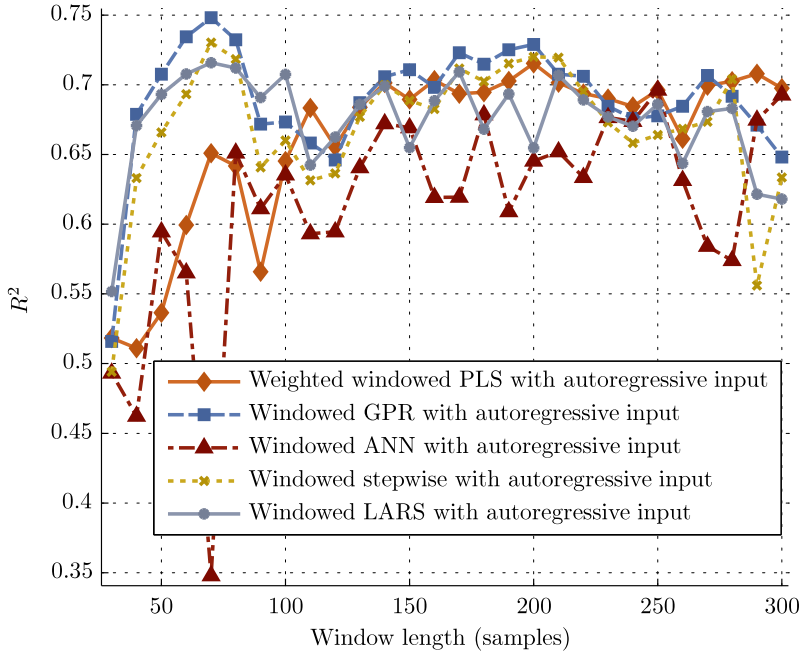



FIGURE 7.29: $R^2$ values for windowed models using autoregressive inputs and window lengths 30-300. As with the MAPE values, model $R^2$ values increase with the introduction of the weighting scheme and autoregressive term. The $R^2$ values for time-invariant models are generally below 0.5 (not shown).

model estimates follow the overall trend of the etch rate variations allowing a relatively high $R^2$ to be achieved, the models still do not accurately model the high frequency fluctuations in the etch rate data, similar to the results described for the global modelling results described in Section 6.7.

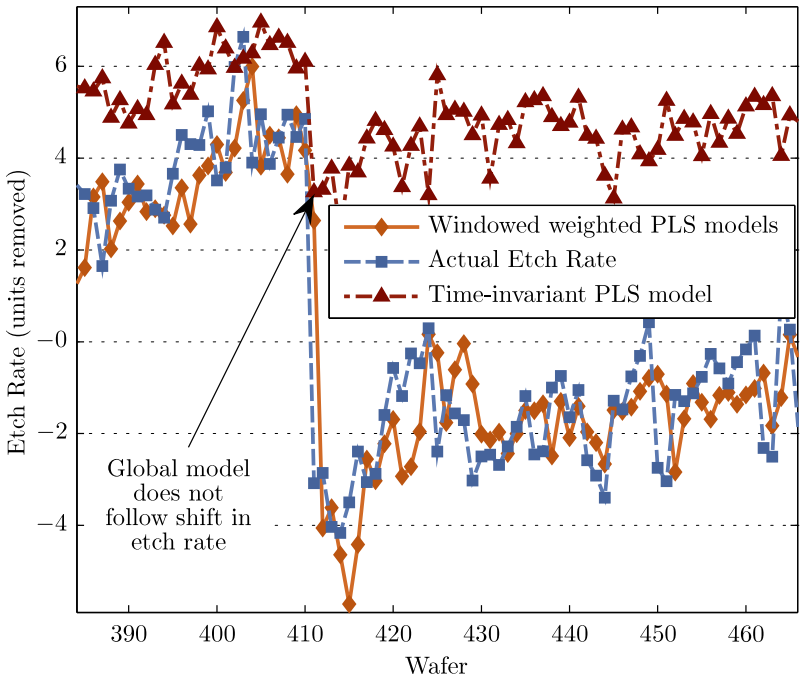



FIGURE 7.30: Section of etch rate predictions from best PLS weighted windowed models. The PLS models use the wafer weighting scheme and a window length of 200 wafers, achieving a MAPE 1.2 % and $R^2$ 0.72. Although the diagram shows how large scale etch rate variations are followed successfully by the estimates, smaller, high-frequency fluctuations in the data are still relatively poorly estimated. The time-invariant PLS model fails to produce accurate VM estimates after a shift in etch rate.

While the estimates from the best windowed GPR and stepwise models are of similar accuracy, one advantage to the use of GPR models is that each etch rate estimate is accompanied by a confidence interval. Figure 7.31 shows the 95 % confidence intervals for each etch rate estimate from the windowed GPR models using a window length of 70 samples. The confidence intervals capture a range around the etch rate estimates which is sufficiently large to encapsulate the measured etch rate value for the majority of wafers. Graphical representations of the residuals produced by the windowed GPR models are shown in Figure 7.32. The model residuals are found to obey a normal distribution.

### 7.4.5 Computational concerns

The computation times for each VM estimate of etch rate must be short enough to be computed in a timely manner for real-time implementation in a semiconductor fabrication plant. While this may not have been a concern for global models because estimates can be computed in less than a second, for windowed models, the time for calculation of the VM estimate may include the time for re-computation of VM model itself. Table 7.8 shows the model training and estimation time for one sample (in seconds) for each modelling technique using a computer with a 2.6 GHz dual-core processor and 2 GB

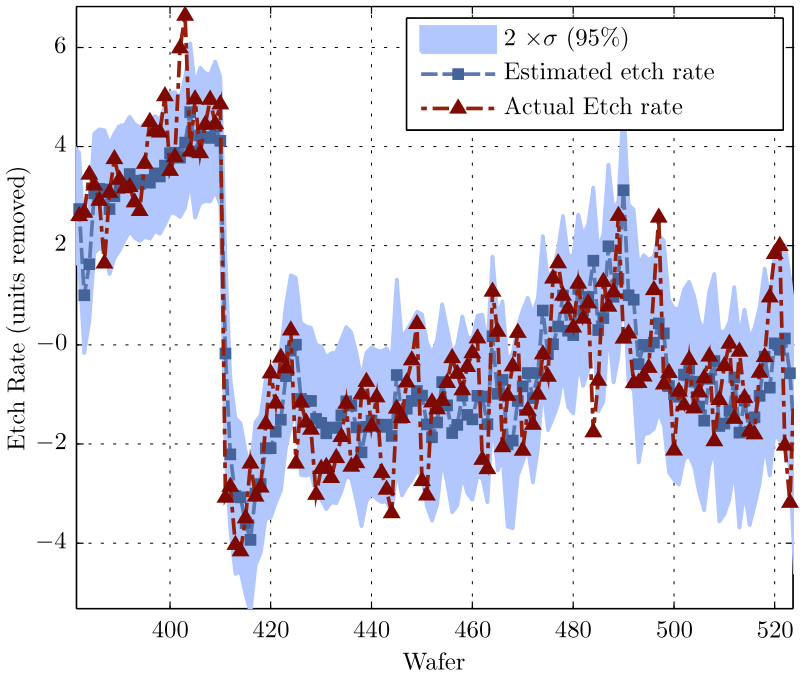



FIGURE 7.31: Windowed GPR model predictions with 95 % ($2\times\sigma$)confidence intervals. One of the benefits of using GPR models is that confidence intervals on predictions can be easily evaluated.

| Window length | Stepwise | LARS | PLS | ANN | GPR |
|---|---|---|---|---|---|
| 70 | 0.0083 | 0.0089 | 0.0185 | 0.9952 | 14.4875 |
| 150 | 0.0144 | 0.0098 | 0.0250 | 1.2321 | 38.9297 |
| 250 | 0.0239 | 0.0110 | 0.0315 | 1.4915 | 118.23 |

TABLE 7.8: Comparison of times required for training and estimation of one window of samples for different model types (in seconds).

RAM. Training times for the stepwise regression, LARS, and PLS regression models are faster than those of ANN and GPR models because the latter two techniques require the use of optimisation techniques and multiple random initialisations during model training procedures.

While early stopping is employed using a validation data set during optimisation of the weights for ANN models, gradient descent is used to optimise the GPR hyperparameters without a validation data set, relying instead on the log-likelihood equation to limit the complexity of the model. The etch processing time for each wafer is approximately 5 minutes, and there is a metrology delay of several hours for etch rate measurements. Hence, according to Table 7.8, any of the three modelling techniques investigated are suitable for real-time implementation of a windowed VM system, in that VM estimates can be made in the seconds between changing wafers, and model training times are negligible compared to the existing metrology delay.

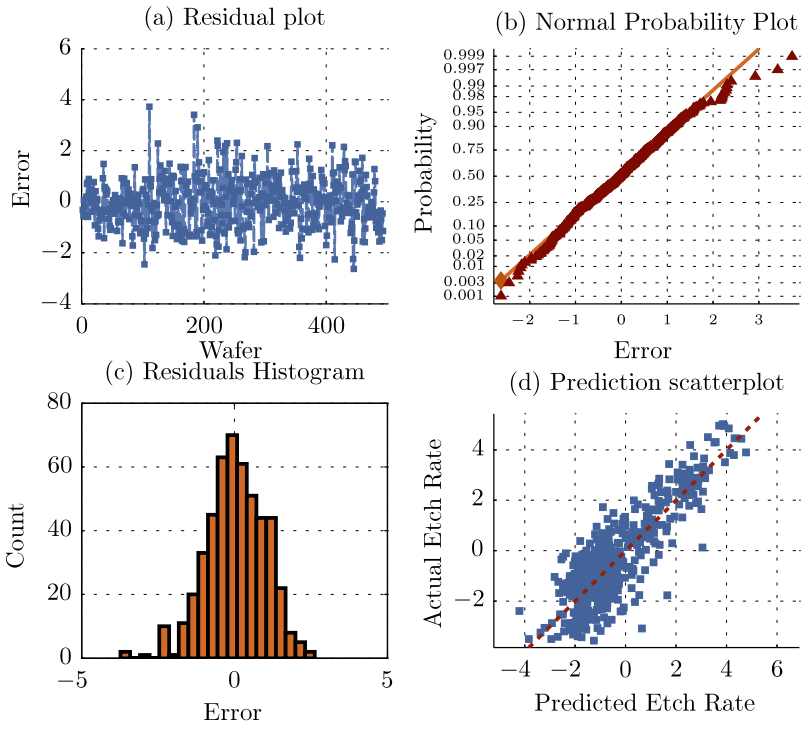



FIGURE 7.32: Graphical analysis of model errors from best GPR model showing (a) error magnitudes, (b) normal probability plot, (c) histogram of error values, and (d) scatter plot of predicted and actual values.

### 7.4.6 Summary

Windowed models take new information into account while discarding past data and, as a result, they are more successful than global models at maintaining model currency and estimation accuracy even with changes in chamber operation that arise due to process drift and preventative maintenance events. Windowed models outperform the time-invariant models for all window lengths examined.

The accuracy of windowed PLS models is improved through the application of a novel sample weighting scheme to the PLS algorithm, where the sample weights are determined in accordance with the maintenance history of the plasma etch chamber. The incorporation of this information leads to an reduction in estimation error for almost all window lengths investigated, as illustrated by Figure 7.27, along with an increase in $R^2$ value.

Further increases in VM accuracy are achieved for all of the modelling techniques through the addition of previous etch rate measurements as input variables to the VM models.

Windowed GPR models were found to perform best out of all of the windowed models, using smaller window lengths to produce more accurate etch rate estimates. Addition

of the autoregressive term to the GPR models further improves performance. The most accurate windowed GPR model estimates the etch rate of 493 unseen test wafers with an $R^2$ of 0.75 and MAPE 1.14 %. The 493 test wafers are dispersed among 11,692 processed wafers, highlighting the robustness of the windowed models to long-term process variations. Although GPR models incur substantial extra computational load to similarly-performing stepwise regression models, typically, the 95% confidence limits accompanying the GPR-model estimated encapsulate the range of etch rate values where the real etch rate lies. Because both GPR and stepwise models can be calculated within the time required for real-time operation, GPR models are preferred due to the added advantage of these confidence intervals.

## 7.5 Local modelling conclusions

This chapter has investigated a number of local modelling techniques to estimate variations in plasma etch rate more accurately than the global models described in Chapter 6. Particular difficulties attach to the utilisation of global VM models across PM boundaries, and a variety of local modelling approaches are explored to tackle this problem.

A regional modelling scheme, described in Section 7.2, is investigated that divides PM cycles into chronologically based regions, and subsequently builds separate local models on each region. The regional modelling scheme fails to improve upon the accuracy of the global model estimates suggesting that there are little or no exploitable similarities between wafer data from similar regions of each PM cycle. The regional modelling scheme is further hindered by a shortage of data with which to train each model as the number of models per PM cycle is increased. However, the results *do not* suggest that more accurate results will be produced using the regional modelling technique on larger data sets.

An analysis of the EP and PIM measurements reveals that the wafer data exhibit modal behaviour, moving between distinct operational spaces. Four clusters of similar data are found to exist in the data set. Each cluster contains wafer data from a whole number of PM cycles. A clustered modelling technique that creates local VM models for each cluster in the data is investigated, where the local VM models are used as appropriate when performing VM for unseen data, depending on which cluster the unseen data belong to. The clusters are identifiable using unsupervised clustering algorithms such as k-means clustering.

Although the clustered models are found to improve modestly upon the accuracy of the global models for the majority of input variable and model combinations investigated

in this chapter, the improvements observed are relatively small and sensitive to the choice of training and test data. On the basis of the results, global models are preferable to cluster models due to more reliable estimates over the full data set and simpler implementation. A larger set of historical data is required to *fully* explore the capabilities of the clustering scheme, but the feasibility of the technique has been demonstrated here.

The third local modelling scheme investigated in this chapter is a windowed modelling scheme. Windowed models outperform unchanging time-invariant models for all window lengths examined. Windowed models, each of which are based on a subset of the full data set, avoid becoming overly general and adapt over time with changes in the chamber characteristics and the arrival of etch rate metrology. Incorporation of chamber maintenance information into the determination of weights for a weighted PLS algorithm improves the accuracy of the etch rate estimates. The error performance of the best windowed models is similar to the performance of the global models trained on *interleaved* data sets examined Chapter 6. Windowed models are implementable in industrial settings.

Although the best windowed models follow the overall trend of etch rate variations, they struggle to accurately model high frequency fluctuations in the data. These high frequency fluctuations do not appear to be reflected in the VM model input variables, and may arise from other related, but unmeasured, manufacturing processes or unmeasured disturbances in the incoming material.

With regard to the modelling techniques used, ANN models are typically outperformed by PLS and GPR models for the local modelling schemes explored as a result of the limited data sets available for model training. LARS models outperform the better-known stepwise regression models for the regional and clustered schemes, but underperform during windowed model tests. GPR models are found to work well with small data sets and produce an accompanying variance value for each etch rate estimate.

In conclusion, the most accurate estimates for the plasma etch process under study can be achieved through the use of a windowed modelling scheme. The windowed modelling scheme can follow variations in etch rate over multiple PM events, and far outperforms the performance of time-invariant and global models in terms of estimation accuracy. Windowed models using GPR produce the most accurate estimates and allow simple calculations of confidence intervals on the etch rate estimates that typically encapsulate the unmodelled variations in etch rate for the data set investigated.

The results of Chapters 6 and 7 have appeared in the Advanced Semiconductor Manufacturing Conference (ASMC) 2009 [264], the IEEE International Conference on

Industrial Technology (ICIT) 2010 [263], the Irish Signals and Systems Conference 2010 [254], and have also been submitted for publication in the IEEE Transactions in Semiconductor Manufacturing.

# Chapter 8

# Virtual metrology and control of plasma electron density

This chapter details the development of a real-time virtual metrology (VM) and model predictive control (MPC) scheme for control of plasma electron density in an industrial plasma etch chamber.

A literature review of related research is presented in Section 8.1. Motivation for the control scheme developed in this chapter is provided in Section 8.2. Following this discussion, a description of the experimental apparatus is provided in Section 8.3, and finally, the implementation details and results from the control algorithms investigated are presented in Sections 8.4 – 8.8.

## 8.1 Control in semiconductor manufacturing

Chapters 6 and 7 of this thesis are primarily concerned with the estimation of wafer state variables, such as etch rate, using process data available during and after each etch process run. Estimation of such variables could be used in either a real-time feed-back loop or in a wafer-to-wafer feed-back control loop to achieve reliable process performance by manipulating etch chamber inputs [231]. As discussed in Chapter 1, a number of different feed-back loops can be implemented using information from plasma processing tools, downstream metrology variables, and estimates from VM models. Three different feed-back loops, operating on different timescales and using different information, are depicted in Figure 8.1. The research in this chapter is focussed on the inner feed-back loop that concerns the real-time regulation of plasma variables, in the case of this

chapter, electron temperature. A brief overview of previously completed research for each of the feed-back control loops is now presented.

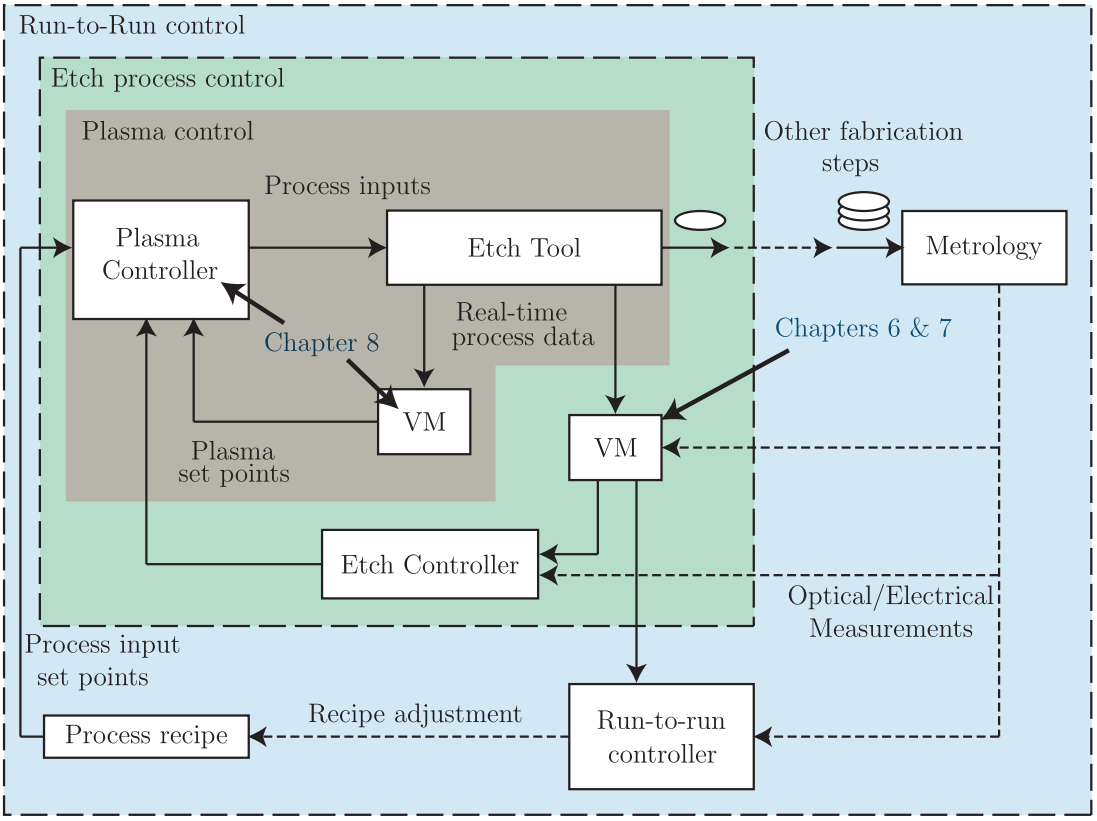



FIGURE 8.1: Etch tool control possibilities with information flow. Three different feed-back loops can be investigated for control.

### 8.1.1 Run-to-run control

The outer loop in Figure 8.1 describes control schemes in which the etch process recipe is adjusted in response to real metrology data collected downstream from the etch process or VM from each wafer processed. Wafer-by-wafer or lot-to-lot control in this manner is known as run-to-run control and is addressed by a number of researchers for different semiconductor manufacturing processes.

For example, Wang and Mahajan [265] demonstrate run-to-run control of a CVD process using an ANN model-based feed-back controller, using an EWMA technique to filter out process noise and detect drifts. Edgar *et al.* [266] focus on the use of state-space models with scheduled observers for run-to-run control of CMP and lithography overlay processes, finding that state-space formulations achieve better set point tracking than EWMA controllers.

Research by Card *et al.* [227, 267] uses ANNs to achieve run-to-run control and to develop a control unit capable of accurately detecting the need for part replacements, cleaning operations, and changes to process settings for etch processes. Rietman and Patel [268] use an adaptive ANN-based controller to regulate the over-etch time for wafers processed and succeed in reducing etch rate standard deviation by up to 40% as compared to a human operator. Similarly, Limanond *et al.* [269] use an optimisation algorithm with goal attainment programming to determine the optimal etch time for wafers so that etch depth variation can be reduced in simulation studies based on production data. Ruegsegger *et al.* [7] demonstrate a run-to-run controller, using feed-forward data from previous manufacturing processes, that reduces etch process variance by adjusting RIE process recipes to compensate for variations in lithography critical dimensions. Khan *et al.* [11] detail a fab-wide VM and run-to-run control scheme with simulated results showing superior control when VM is employed in run-to-run controllers rather than relying on infrequent and delayed real metrology.

Kang *et al.* [230, 231] describe VM schemes that estimate etch process results for each wafer processed using process data. Five input selection algorithms and four regression algorithms were examined as candidate techniques for VM modelling. An EWMA-based run-to-run controller is shown to improve the overall productivity of a simulated photolithography process.

Run-to-run control is not considered further in this thesis. For a complete examination of run-to-run techniques used in semiconductor manufacturing, the reader is directed to the review works by Edgar *et al.* [6], Qin *et al.* [13], Bacelli [270], and Ringwood *et al.* [8].

### 8.1.2 Etch process variable control

The center feed-back loop in Figure 8.1 depicts the control of etch variables such as etch depth, etch rate, uniformity, end-point etc. Control of such variables can be implemented in a real-time or run-to-run fashion depending on the availability of suitable measurements and actuators.

Some of the earliest work in real-time control of plasma etch variables was completed by Butler *et al.* [139] and McLaughlin *et al.* [271], concentrating on the analysis and design of a multi-variable control system for a plasma reactor. Relative gain analysis and the singular value decomposition (SVD) are used to select the manipulated and performance variables, that is to understand the relationship between the manipulated variables (pressure, power, gas flow rates), the process variables (DC self-bias, species

concentrations), and the etch variables (yield, uniformity, anisotropy), and then choose appropriate etch variables that can be controlled in real time and provide a measure of the process quality.

In [135], a Kalman filter is implemented to obtain real-time estimates of etch rate and etch depth from dual-wavelength reflectometry data for real-time proportional-integral (PI) control of etch rate using RF power as the manipulated variable. Accurate end point estimation is also achieved. In [272], real-time control of etch rate is demonstrated using a proportional-integral-derivative (PID) controller and model-based feed-forward action for large changes in etch rate set point. Laser reflectance interferometry (LRI) is employed to measure the etch rate in real time for control purposes and again, RF power is the manipulated variable. Stokes and May [273, 274] describe the use of an indirect adaptive control strategy for control of etch rate, implemented using two neural network models. Data for the system model are gathered using a $2^3$ factorial experiment, varying chamber pressure, RF power, and $BCl_3$. The authors argue that the indirect adaptive control scheme outperforms more conventional linear-quadratic-Gaussian (LQG) methods for controller design. Laser interferometry (LI), residual gas analysis (RGA), and optical emission spectroscopy (OES) data provide process feed-back while RF power, gas flow, and pressure are manipulated to control etch rate and wafer bias voltage on a *simulated* system. In [275], Rosen *et al.* develop real-time feed-back controllers for etch processes based on in-situ spectroscopic ellipsometry measurements of wafer thickness. A combination empirical and first-principals, physics-based, state-space model for the etching dynamics is developed to design the controller. Armaou *et al.* [276] examined the use of three independent PI controllers that use measurements of the etch rate at three locations to manipulate the inlet concentration of etchant gases at spatially distributed locations in a simulated etch process to improve etching uniformity. Parkinson *et al.* [277] demonstrate the use of a multi-input-multi-output (MIMO) control system for controlling critical dimensions (CDs) using scatterometry-based integrated metrology that provides accurate CD measurements for each wafer. Substantial improvements in CD variance are observed for simulation models based on production data.

The biggest drawback for many of the etch process control schemes detailed above is the requirement for bulky and non-portable measurement tools, such as interferometers and ellipsometers, or expensive integrated metrology tools, that are capable of measuring etch rate variables in real time. More modern manufacturing tools are focussing on the integration of metrology solutions for real-time measurement of process variables [12, 13], which will further enable the widespread implementation of real-time and run-to-run etch variable control for new processes.

### 8.1.3 Plasma variable control

The inner control loop in Figure 8.1 concerns the real-time control of fundamental plasma variables such as electron density, electron temperature, ion densities, ion fluxes, etc. These plasma variables are fundamentally related to etch variables such as etch rate and etch uniformity. Hence, regulation of the plasma variables in real time can result in a more stable etch performance.

A number of different approaches to plasma variable control are examined by different authors in the literature. Work by Rashap *et al.* [278, 279] focusses on the real-time control of atomic fluorine density [F] and the wafer bias voltage ($V_{bias}$) in order to regulate etch rate in the presence of a number of disturbances to a RIE process. Chamber pressure and power are the manipulated variables, [F] is determined using actinometry and the real-time etch rate is determined using LRI, but this measurement is not included in the feed-back control loop. Although control of [F] and $V_{bias}$ served to reduce the impact of disturbances compared to open-loop experiments, other uncontrolled factors tended to effect the etch rate apart from [F] and $V_{bias}$. Patterson and Khargonekar [280] use the real-time control of [F] and $V_{bias}$ to demonstrate a reduction of the loading effect (see Section 2.3.4) in the same RIE process, demonstrating an 80% improvement over open-loop methods. Hankinson *et al.* [281] further expand the work by Rashap *et al.* [278, 279] by integrating real-time control of [F] and $V_{bias}$ into a run-to-run control framework using a non-linear controller and Hammerstein model structure. Hanish *et al.* [205] examine the use of actinometry for real-time control of atomic chlorine [Cl], which is more difficult than the measurement of [F] due to the peculiarities of actinometry for chlorine plasmas. An extended Kalman filter is employed to estimate [Cl] in real time for closed-loop PID control.

Park *et al.* [282] focus on the control of the ion energy bombarding the wafer surface during etch processes by manipulating a variable resistor placed in parallel with the blocking capacitor on the RF power line to the wafer electrode. The variable resistor alters the bias voltage $V_{bias}$ of the wafer. The etch rate is measured both off-line using ellipsometry measurements and on-line using a spectral reflectometry system. A controller is designed to regulate $V_{bias}$ and the plasma power to constant set points by manipulating the variable resistor and the RF power generator. The authors demonstrate enhancements in selectivity and improvements in cleaning process effectiveness.

Milosavljevic *et al.* [283] demonstrate real-time, model-based control of ion flux and species densities by manipulating power and oxygen flow rate in a dielectric plasma etch chamber. Ion flux is measured using a specially adapted chamber electrode with a built-in ion flux probe, and species densities are measured using actinometry methods

and a mass spectrometer. The authors demonstrate superior control results using a model-based controller over two single-loop PID controllers.

Lin *et al.* [284] reduce etch rate variations in an ICP etcher by controlling ion current and RF voltage on the target electrode. The manipulated variables are the power from two RF generators that adjust ion density and ion energy, respectively. RF voltage is measured using an impedance meter. The RF peak voltage is chosen as a controlled variable because it is correlated with the sheath voltage and, hence, also the ion energy on the wafer surface. The ion current is controlled because it is related to the ion density. Ion current is estimated, in a form of real-time VM, by dividing the measured RF power by the fundamental RF bias voltage (the "power/voltage" method [210]). The control scheme reduces etch rate variation due to transient wall conditions by a factor of two as compared to open-loop operation.

Later work by Lin *et al.* [285] uses a fuzzy-logic feed-back control scheme to control electron density and ion energy. The first-wafer effect is eliminated for two different etch processes, and electron density is demonstrated as a more effective controlled variable than ion current for etch processes operated at low bias power. A custom-built microwave interferometer measures plasma electron density in real time, but requires adaptations to the wall of the plasma chamber. The fuzzy logic-based controller minimises the number of control actions used, and outperforms a PI controller to which it is compared. Although the first-wafer effect can be eliminated using feed-back control of electron density and ion energy, the authors find that etch disturbances due to pressure changes cannot be compensated for by plasma electron density control alone.

Klimecky *et al.* [286] show that real-time control of electron density can be used to compensate for transient chamber wall conditions, eliminating the first-wafer effect and reducing the etch rate variance, without affecting the resultant etch profile. A PI controller is used and the electron density is monitored in real-time using broadband RF, a microwave resonance based sensor. An ellipsometer is used to measure the etch rate in real time, but is not included in the feed-back control system. The broadband sensor requires modification to the chamber sidewalls, but is reported to not interfere with the etch process.

## 8.2 Motivation

This chapter proposes to use plasma impedance monitor (PIM) measurements for VM of electron density for the purposes of non-invasive electron density control in an industrial plasma chamber. The aim of the work is the development of a closed-loop control

system capable of maintaining a consistent plasma electron density subject to unmeasured disturbances similar to those brought about by PM events. The VM scheme can be implemented relatively easily in a production environment, without the necessity for permanent physical additions or changes to the etch chamber. To the best of the author's knowledge, this work documents the first application of predictive functional control (PFC), a variant of MPC, to real-time control of plasma electron density.

As emphasised by Klimecky *et al.* [286] and Lin *et al.* [285], control of electron density during production operations is desirable for increasing process stability and reducing the effects of drifts and shifts in system characteristics. Steinbach [215] also emphasises that the plasma electron density is a key variable affecting etch process performance. Imai [199] demonstrates the feasibility of using etch process variables for VM of electron density for fault detection purposes.

As shown in Chapters 6 and 7, preventative maintenance (PM) events have a large influence on etch performance and VM model estimation accuracy. PM events involve the routine replacement of components, such as electrodes and ceramic covers, that have been exposed to etchant chemicals from possibly over 1000 wafer etch and cleaning cycles. Although the replacement components are macroscopically identical to those that are removed from the chamber, microscopic differences in the electrical connections made between components when they are replaced change the electrical characteristics of the chamber. Changes in such component connections are more influential as the applied RF frequency increases [128]. At the high frequencies in use during plasma processing, changes in impedance, stray capacitances, and stray inductances cause considerable changes to the electrical behaviour of the chamber. The electrical path between the powered chamber electrode and ground (the ground path) influences plasma variables such as the ion flux to the etching wafer and the DC bias of the wafer in the chamber [282]. Hence, changes in impedance of the path between the powered chamber electrode and ground brought about by PM events can cause the etch performance of the chamber to vary dramatically between maintenance cycles.

For the experiments described in this chapter, a modified match box that allows manual control of impedance is installed on the ground path from the chamber. Hence, variations in the ground impedance can be realised as required, partially simulating the effect of PM events. The ground impedance variations act as unmeasured disturbances to the plasma, changing the plasma variables such as electron density, and affecting the etch performance.

The measurement of electron density is first achieved through the use of a microwave hairpin resonator probe, described in Section 2.5.7. However, the invasive nature of the

microwave probe makes it unsuitable for use in a production environment. To overcome this obstacle, a non-invasive VM system using PIM data is developed to estimate the plasma electron density in real time, as described in Section 8.5. This VM system is integrated into the feed-back control loop using both PI and MPC schemes as detailed in Sections 8.4 – 8.7. Finally, in Section 8.8, an adaptive scheme using recursive least squares is implemented to update the PFC internal model parameters in real time in response to pressure disturbances that change the relationship between RF power and electron density.

## 8.3 Experimental equipment

Figure 8.2 shows all of the connections and tools used to implement the real-time VM and control system. The control calculations are performed using MATLAB® code, where a National Instruments (NI) USB-6009 device is used to take measurements from the PIM sensors and to control the power delivered from the RF generator to the chamber electrodes. Electron density measurements are taken using a hairpin resonator probe in conjunction with a microwave source and an oscilloscope, which are connected to the computer system using a TCP/IP connection. Optical emission spectroscopy (OES) and digital PIM data are stored on a separate database that is not accessible to the control system in real time. A variable matchbox is used to vary the impedance of the ground path from the plasma chamber and hence act as a disturbance to the process. The set up of the plasma chamber, PIM sensors, and hairpin probe are examined in more detail in Sections 8.3.1 – 8.3.3.

### 8.3.1 Plasma etch chamber

A capacitively-coupled, top-powered, parallel-plate plasma etch chamber is used for the analysis presented in this chapter. This production chamber has a full suite of etch gases available for experiments including helium, chlorine, oxygen, $SF_6$, HBr, and $C_2F_6$ that are delivered via mass flow controllers (MFCs). Chamber pressure is automatically controlled to manually specified set points by means of a gate valve between the etch chamber and the vacuum turbo pump. The electrode spacing can be varied from 0.5 – 5 cm.

For simplicity, safety, and cost effectiveness, the experiments carried out in this chapter use helium plasma. The electrode spacing is kept constant at 5 cm, and the chamber temperature is uncontrolled, with experiments carried out when the chamber

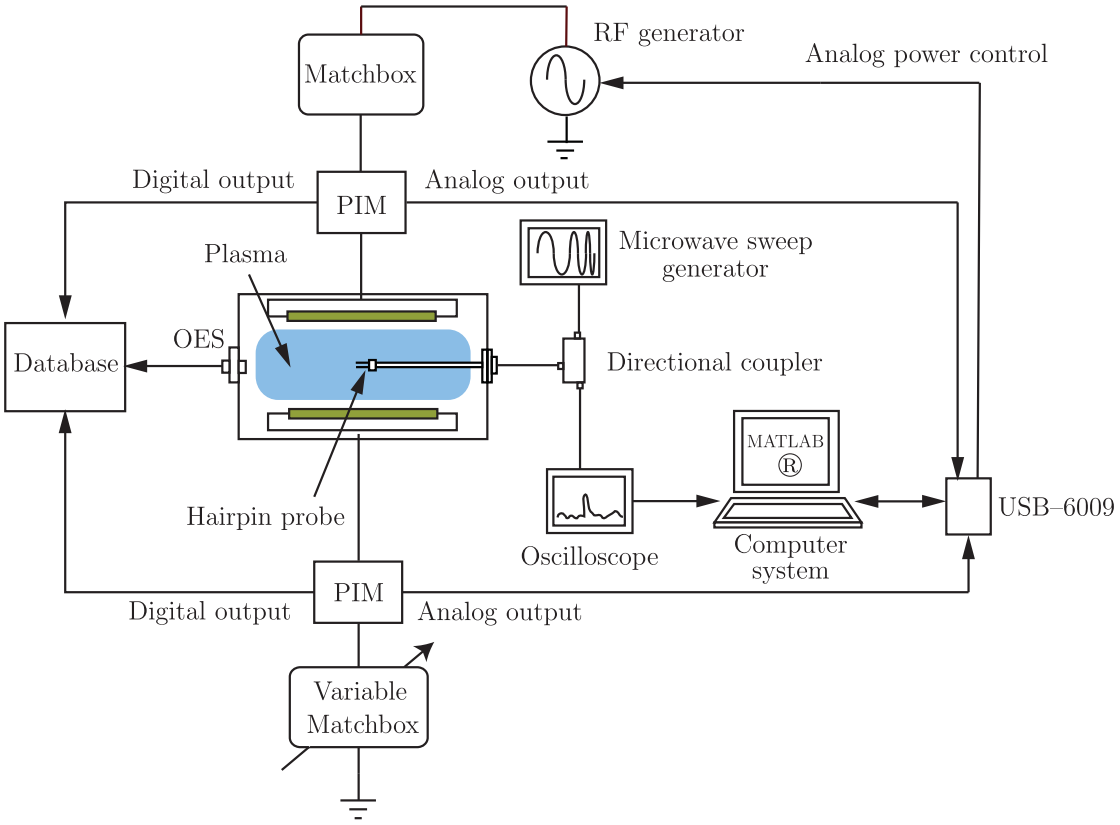



FIGURE 8.2: Virtual metrology and control hardware. This figure depicts the hardware devices and their interconnections used to achieve real time virtual metrology and control of electron density.

is "warm" for consistency, i.e. a plasma is fired for 15 – 20 mins to heat the chamber before experimental data is recorded. The chamber parameters such as pressure, gas flow, and temperature are controlled and monitored by embedded system software on the etch chamber that controls the MFCs, chamber gate valves, and all further necessary subsystems.

### Chamber power control

An RF generator situated beneath the semiconductor factory floor delivers up to 625 W of 13.56 MHz RF power to the topmost chamber electrode. During normal production operations, the amount of delivered power is controlled via a 0 – 10 V reference signal generated by the embedded system on the chamber. The reference signal system operates on a linear scale such that a reference signal of 0 V requests 0 W, and 10 V requests 625 W etc.

To allow automatic control of the chamber power, the reference signal is intercepted in the fabrication environment. A NI USB-6009 data acquisition device is used to

generate an analog reference signal which is then connected to the RF generator. The USB-6009 device is capable of generating a voltage signal in the range 0 – 5 V and hence, a DC amplification circuit with a gain of 2 is used to amplify the reference signal to the desired 0 – 10 V range. Noise attenuation filters are used in the amplifier to ensure that no 13.56 MHz interference is passed to the generator. The USB-6009 is controlled using software scripts implemented in the MATLAB environment.

#### Chamber grounding

The lower (target) electrode in the etch chamber is grounded through a modified match unit that allows the impedance of the path between the lower electrode in the chamber and ground to be varied. This lower match unit is modified such that the position of the load capacitor can be varied remotely, effecting a total impedance of between 0 – 60 Ω. For the remainder of this chapter, the impedance of the path to ground will be referred to as the "ground impedance". Variations in the ground impedance act as disturbances to the plasma in the chamber. As discussed in Section 8.2, these variations are used to simulate possible changes that can occur as a result of a PM operation

A plasma mode change is observed as the ground impedance is raised above approximately 25 Ω, such that the plasma appears to suddenly move away from the grounded electrode and towards the chamber walls. Once the mode changes the ground impedance has negligible further effect on the plasma electron density where the measurement takes place. As a result, the control experiments are restricted to the 0 – 25 Ω range of lower impedance.

### 8.3.2 PIM sensors

As explained in Section 2.5.8, a plasma impedance monitor (PIM) is an electronic sensor that is installed between the matching network and the plasma electrodes. The PIM sensor provides information about the waveforms generated as a result of the non-linear impedance presented by the plasma on the power supply circuitry. Information on the current, voltage, and phase at the fundamental frequency of 13.56 MHz and up to 52 harmonics of this frequency is recorded by the PIM sensors used. For the purposes of the experiments in this chapter, two PIM sensors are used, one installed on each electrode. The upper PIM is installed on the powered or upper electrode of the chamber, and provides information on the applied RF power. The lower PIM sensor records information about the path to ground from the chamber, and is used to measure the the

| PIM sensor | Channel | Variable |
|---|---|---|
| Upper PIM | 1 | VL0 |
| Upper PIM | 2 | IL0 |
| Upper PIM | 3 | PL0 |
| Upper PIM | 4 | - inoperational - |
| Lower PIM | 1 | VL0 |
| Lower PIM | 2 | IL0 |
| Lower PIM | 3 | PL0 |
| Lower PIM | 4 | ZL0 |

TABLE 8.1: Configuration of analog PIM channels. VL0, IL0, PL0, and ZL0 represent the fundamental voltage, current, phase and impedance values respectively.

manual changes in impedance realised by a modified match unit, as described in Section 8.3.1.

Digital data from the PIM sensors are recorded in a database by proprietary software owned by Lam Research. However, the database information is only available after each etch has been completed, preventing real-time access to the digital information for real-time control or VM purposes.

Four programmable analog output channels can provide four analog signals representing PIM variables in real time from each PIM sensor. To allow real-time VM, the analog outputs are sampled using the eight analog input channels of the NI USB-6009 which is interfaced to the control computer. The USB-6009 is capable of sampling each channel at 6 kHz. During control experiments, mean values of the analog signals recorded during each control sampling period are taken to suppress noise.

The analog outputs are assigned to the variables shown in Table 8.1. The analog PIM signals cover a 0 – 5 V range. The analog signals are converted and scaled to the correct units (amps for current, angle for phase etc.) using MATLAB software in the control computer during each sample. The match between the digital and analog signal for the fundamental current (IL0) is shown in Figure 8.3, after the analog signal is converted and scaled to the correct units.

There are two issues with the analog signals that must be taken into account during their use for real-time control. Firstly, the analog outputs experience a delay of approximately 0.5 seconds due to the internal operation of the PIM sensors. This time delay is addressed later in Section 8.4. Secondly, as a result of the small voltage range (0-5V) used to represent a large range of phase angles (0 – 180°), the phase signals are susceptible to noise on the analog outputs. This noise results in noisy power estimates because the power calculations using PIM measurements are especially sensitive to the phase values (see Section 2.5.8) [38]. Erratic values for power are an issue because the

power values are used in Section 8.5 during the development of VM models. An example of the phase signal noise is shown in Figure 8.4.

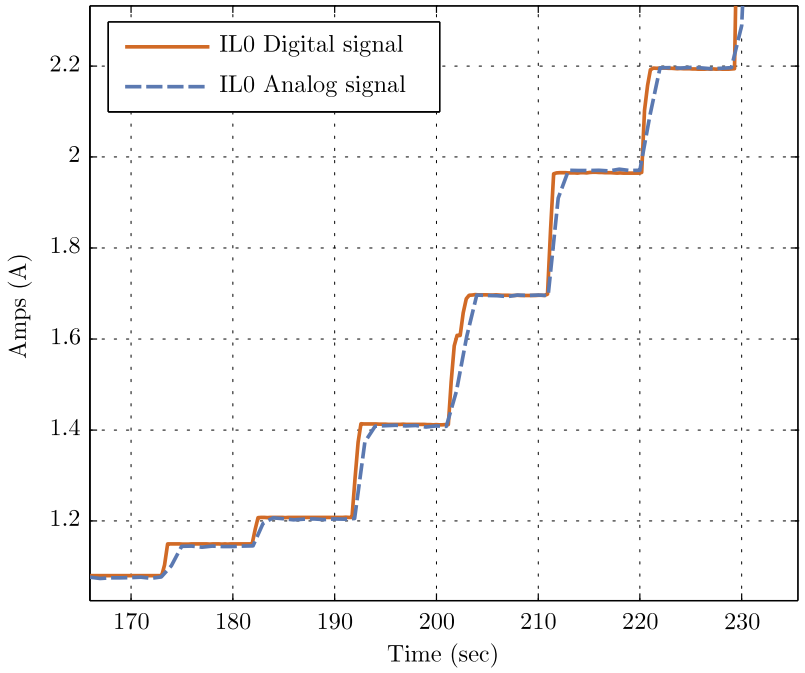



FIGURE 8.3: Fundamental current reading from analog and digital PIM outputs. The digital outputs are not available in real time and thus, they can only be analysed off-line. The analog outputs are delayed by approximately 0.5 s.

### 8.3.3 Hairpin Resonator probe

A hairpin probe is used to provide a real-time measurement of electron density for use in the development of VM models and testing of control algorithms in this Chapter. The operation and construction of the reflection-type hairpin probe is described in Section 2.5.7. For the experiments described in this chapter, the probe reaches the plasma through a modified chamber window fitting on the production etch chamber. Platinum wire is used for the hairpin resonator because of its resistance to etching gases and its high melting point of 1768 °C.

#### Probe setup

The driving current signal for the probe is generated using a microwave source which varies the signal frequency from 2 to 3.8 GHz. The reflected current trace is captured using a Tektronix TDS3034B oscilloscope that is accessed using a TCP/IP direct network connection to the control computer and analysed using MATLAB software. The change in resonance frequency caused by the plasma is determined by comparing the downloaded

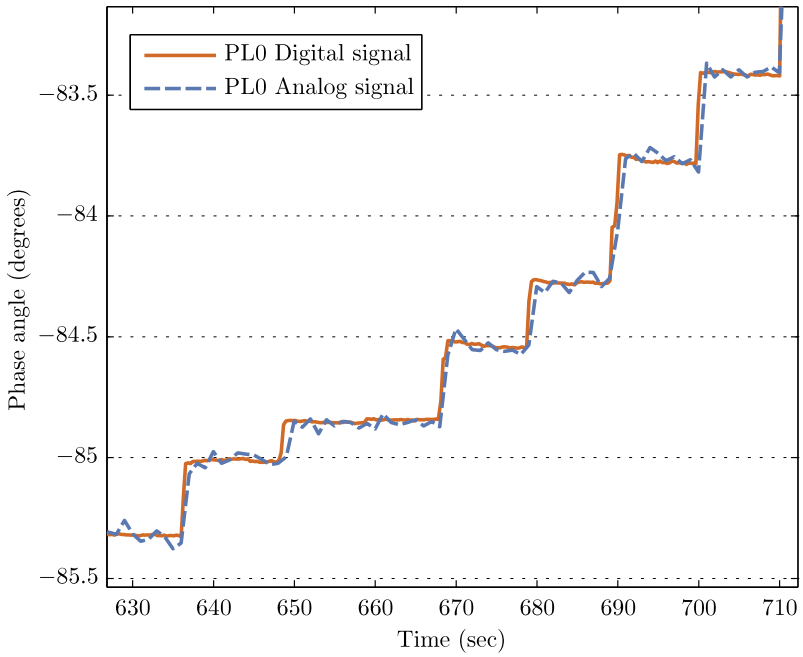



FIGURE 8.4: Fundamental phase reading from analog and digital PIM outputs. The phase value is susceptible to noise because of a scaling effect. Power calculations using the phase value are sensitive to this noise on the analog phase value.

trace to a reference signal recorded with the plasma powered off as shown in Figure 8.5. The change in resonance frequency can the be related to the plasma electron density using Equation (2.33). Similar to the work of Piejak *et al.* [32], the reference signal is updated when required to cater for small but observable changes in the vacuum response of the probe due to thermal distortion. Each trace requires 300 – 400 ms to download from the oscilloscope to the control computer, limiting the sampling rate to approximately 2.5 Hz maximum.

### Error sources

While the hairpin probe provides reliable measurements of electron density the majority of the time, under certain circumstances, care must be taken with the measurement results. It is observed throughout the experiments in this chapter that the height of the resonant peak of the probe diminishes at plasma powers greater than 500 W. Less pronounced peaks can cause inaccurate measurements because the difference between the reflected waveform and the reference waveform is reduced. Such reductions complicate the detection of the peak position because the maximum difference between the waveforms can be reduced to values comparable to the background noise of the system.

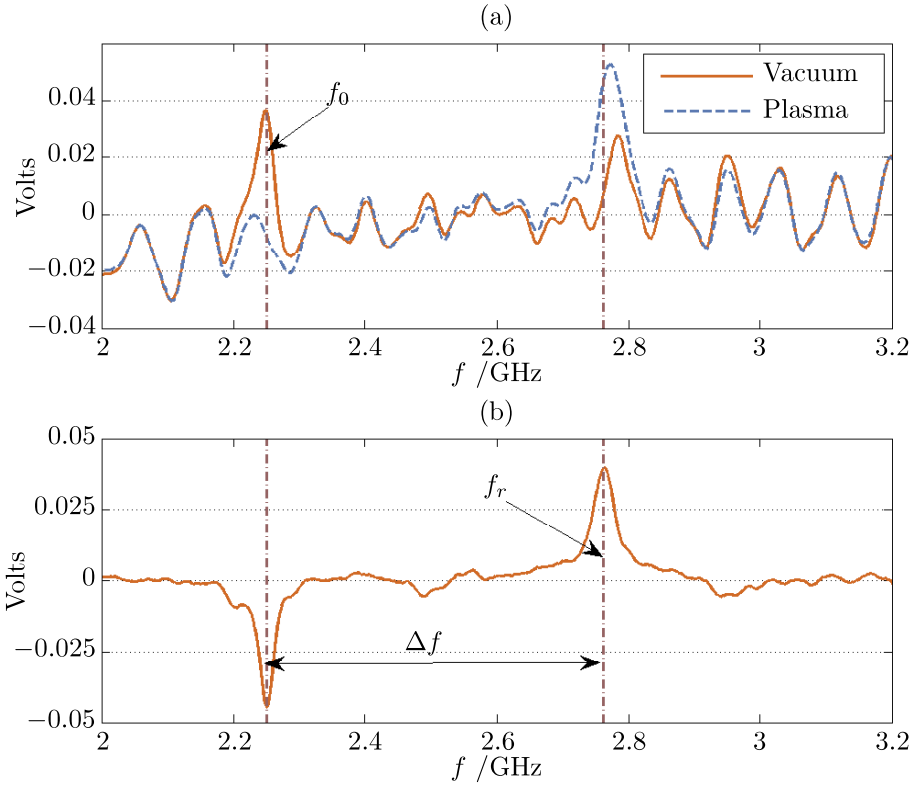



FIGURE 8.5: (a) Reflected signals from hairpin probe in vacuum and during plasma discharge. The resonant frequencies of the probe without and with the plasma are labelled $f_0$ and $f_r$ respectively. (b) Difference between the vacuum and plasma reflect signals. The resonant peak undergoes an upward shift upon plasma ignition due to the change in the dielectric medium between the probe tips.

An example of this phenomenon is shown in Figure 8.6. Where such errors occur during data collection, the erroneous samples are manually corrected via substitution where possible, or otherwise removed after the data is collected. However, during online control of electron density, erroneous measurements of electron density cause oscillations in the manipulated variables, with consequent oscillations in electron density. Hence, operations in high power regimes are avoided where possible.

## 8.4 PI Control of electron density

Using the apparatus described in Section 8.3, the plasma electron density can be controlled using a proportional-integral (PI) controller, while measuring the electron density in real time using the hairpin probe.

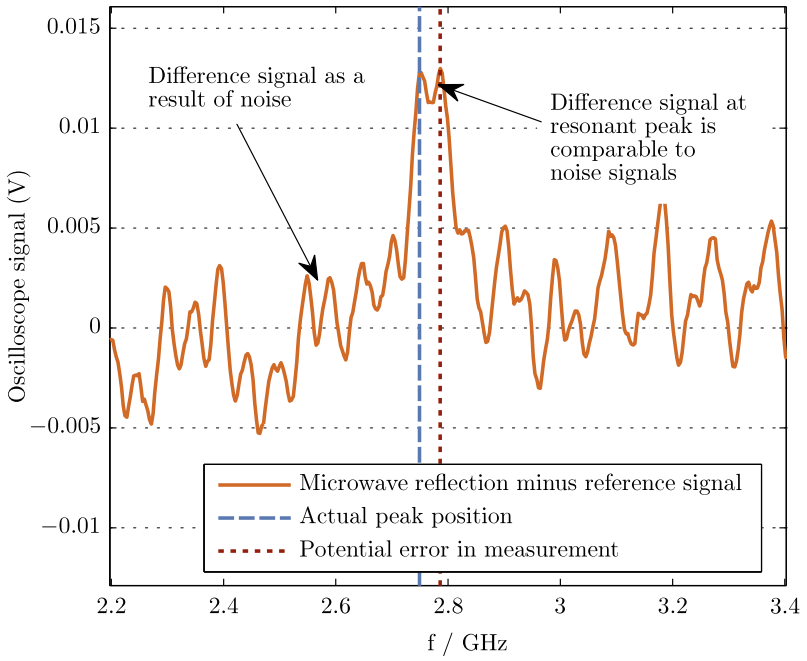



FIGURE 8.6: Erroneous electron density readings from hairpin probe. At high powers, the resonant peak can become diminished such that the difference between the reflected and reference signals is of comparable size to other differences caused by noise. As a result, the peak position can be detected incorrectly, corrupting the electron density measurements.

### 8.4.1 PID control

Proportional-integral-derivative control (PID) control offers one of the simplest and yet most efficient solution to many real-world control problems. Since the introduction of PID control in 1910 and the Ziegler-Nichols tuning methods in 1942 [287], the popularity of PID control has grown immensely. Although a wide variety of other control schemes now exist, controllers based on PID algorithms are still in use in more than 90% of industrial applications [288] and a large range of PID tuning techniques [289], software packages and hardware modules are available to implement PID control [290].

A standard PID controller, also known as a three-term controller, has a transfer function given by

$$C(s) = K_{prop} + K_{int}\frac{1}{s} + K_{der}s \tag{8.1}$$

where $K_{prop}$ is the proportional gain, $K_{int}$ the integral gain, and $K_{der}$ the derivative gain. The controller is implemented in feed-back control systems as shown in Figure 8.7. The proportional term provides a control action proportional to the error signal, the integral term reduces steady-state errors through low-frequency compensation by an integrator, and the derivative term improves the transient response through high-frequency compensation by a differentiator. While, in reality, the controller parameters

| | Rise Time | Overshoot | Settling time | Steady-state error | Stability | Noise immunity |
|---|---|---|---|---|---|---|
| Increasing $K_{prop}$ | Decrease | Increase | Small Decrease | Decrease | Degrade | None |
| Increasing $K_{int}$ | Small Decrease | Increase | Increase | Large Decrease | Degrade | Increase |
| Increasing $K_{der}$ | Small Decrease | Decrease | Decrease | Minor change | Improve | Decrease |

TABLE 8.2: Effects of independent P, I, and D tuning [290].

are mutually dependent in tuning, a basic guide of the individual effects of the three terms is shown in Table 8.2.

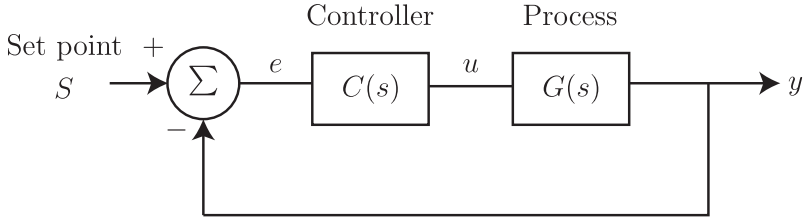



FIGURE 8.7: Feed-back control loop with PID control. The process input $u$ is calculated by the controller $C(s)$ based on the error $e$ between the set point and the process output.

The discrete implementation of the PID controller equation is such that the manipulated variable is given by

$$u(k) = K_{prop}e(k) + \sum_{i=1}^{k} e(i)T_s + K_{der}\left(\frac{e(k) - e(k-1)}{T_s}\right) \tag{8.2}$$

where $T_s$ is the sample period. This discrete implementation employs a discrete integrator using the Euler method of integration and a first-order derivative using a backward finite difference.

Because the derivative term $K_{der}$ can increase sensitivity to measurement noise, which will be a factor when using the VM schemes, it is not used in the electron density control application described here. Hence, the controller used is termed a PI controller.

### 8.4.2 PI control results

The PI controller is first implemented using the hairpin probe measurements for real-time feed-back. As such, the sampling frequency is limited to a maximum of 2 Hz (see Section 8.3.3). Figure 8.8 shows the experimental results of an electron density set point tracking

exercise with the PI controller. In future iterations of the control scheme, the electron density set point can be decided depending on the etch performance required, and may be required to follow changes in set point as different process steps are implemented. However, for the purposes of this work, random electron density set points are used to demonstrate the controller performance. Approximate initial values for the controller parameters are obtained using a MATLAB process simulation and, following this, the controller is tuned by hand with $K_{prop} = 1.7$, $K_{int} = 6.8$ resulting in a relatively slow time constant of 1.35 seconds. The ground impedance is kept constant throughout this experiment. Figure 8.8 also shows the controlled variable, the chamber power.

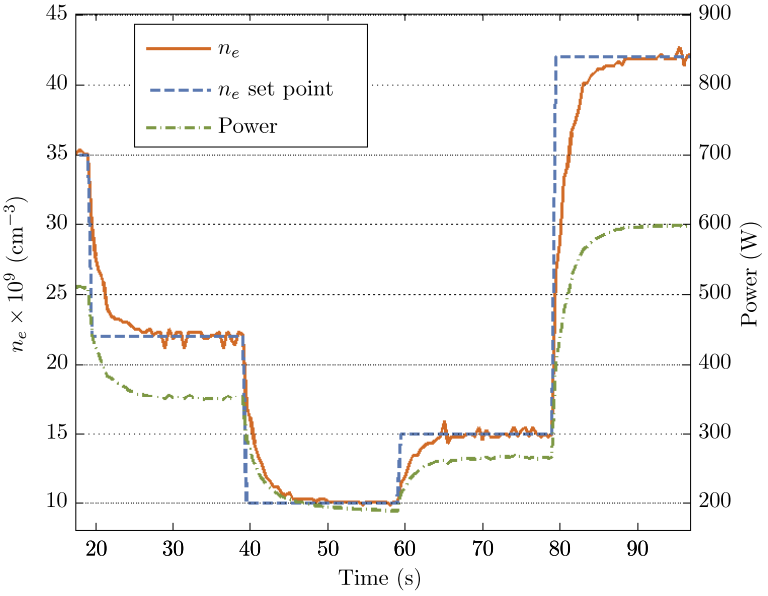



FIGURE 8.8: Electron density control with slow PI tuning. The PI controller is tuned such that $K_{prop} = 1.69$, $K_{int} = 6.78$. The system closed-loop time constant is approximately 1.35 seconds.

The PI controller is capable of countering disturbances in the lower impedance value of the chamber. Changes in the ground impedance disturb the electron density of the plasma. The PI controller returns the electron density back to the required set point value within 5 seconds, as shown in Figure 8.9, where both step and gradual changes in ground impedance are introduced.

The dynamic response of the closed loop system can be speeded up by increasing the proportional and integral gains of the PI controller. Figure 8.10 shows the controlled response for two sets of controller parameters; Figure 8.10(a) shows the response with $K_{prop} = 3$, $K_{int} = 17$ and in Figure 8.10(b), $K_{prop} = 5$, $K_{int} = 30$. The first response in Figure 8.10(a) is faster than the response shown in Figure 8.8 but suffers from slight overshoot and oscillations at some set point changes. However, disturbances introduced

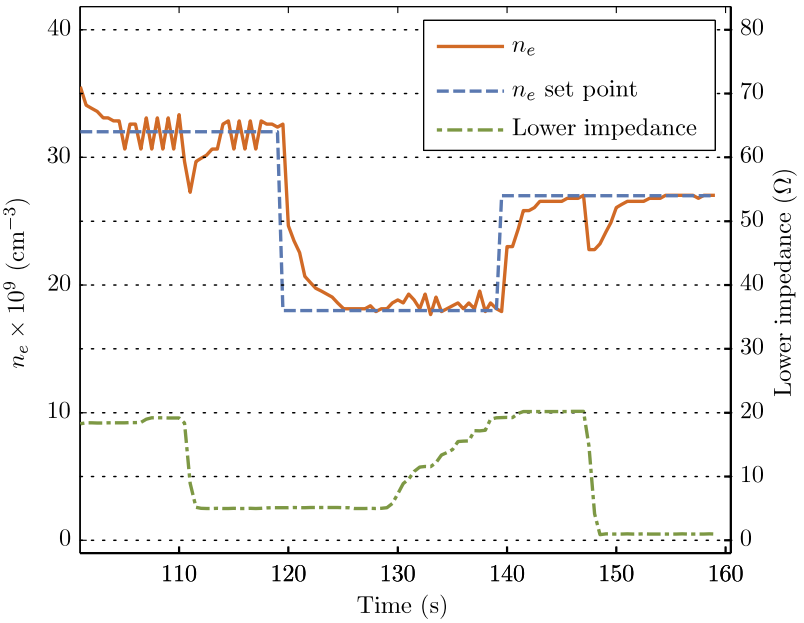
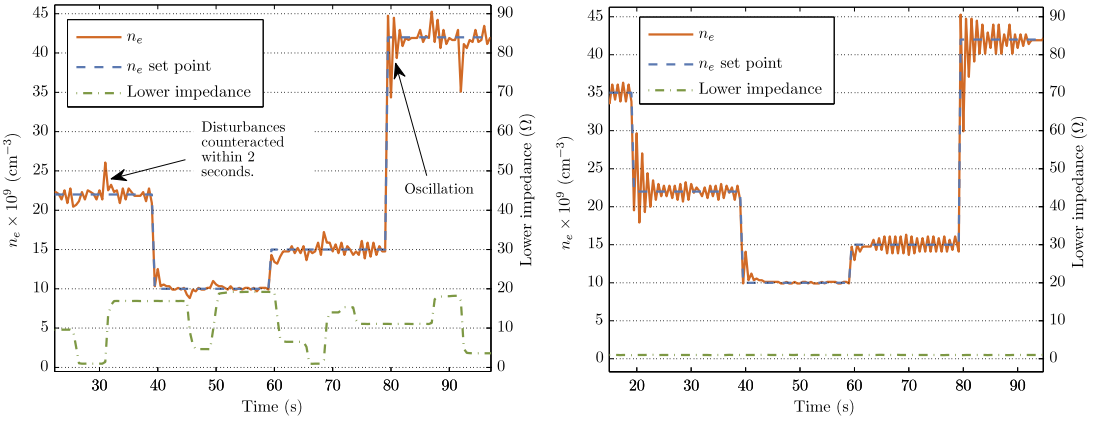



FIGURE 8.9: Electron density control with system disturbances. The PI controller returns the plasma electron density to the electron density set point less than 5 seconds after step changes in the ground impedance value.

using the lower impedance value are compensated for within 2 seconds by the controller. The PI controller parameters used to create the response in Figure 8.10(b) are such that slowly-decaying oscillations appear in the electron density after set point changes and disturbances.




(a) (b)

FIGURE 8.10: Fast control using a PI controller with probe readings. (a) The PI controller parameters are set to $K_{prop} = 3$, $K_{int} = 17$ to produce a fast response with small oscillations at some set point changes. (b) The PI controller parameters are set to $K_{prop} = 5$, $K_{int} = 30$ such that the system is almost unstable with slowly-decaying oscillations in electron density, even without the presence of random disturbances as in 8.10(a).

## 8.5 Virtual metrology of electron density

The hairpin probe is an invasive method of determining the plasma electron density that presents a number of disadvantages when used for real-time control of the electron density in the etch chamber.

- Production wafers cannot be etched while the probe is inserted in the chamber as the plasma is perturbed around the probe body, disturbing the etch process in the space surrounding the probe, reducing the spatial uniformity of the plasma and hence the etch variables.
- The sampling frequency of the probe is limited to 2 Hz due to the time required to download and process the reflected current waveform from the oscilloscope during each sample. This limitation on the sampling frequency limits the speed of the control algorithms that can be implemented. Ideally, approximately 6 – 8 samples should be obtained within the closed-loop rise time. With a sampling frequency of 2 Hz, rise-times of less than 3 – 4 seconds may not be achieved reliably.

Considering these issues, a VM system is proposed to obtain estimates of the plasma electron density. Virtual measurements of the electron density can be obtained in real time using the PIM variables. Although digital information from the PIM sensors is not accessible in real time, the analog outputs can be sampled for VM purposes in real time as the etch process proceeds.

A control system based on the virtual measurements of electron density, rather than the microwave probe, would allow etching of product wafers to proceed unhindered and would also be capable of sampling information from the chamber at a rate faster than 2 Hz. The remainder of Section 8.5 describes the development of a a VM model that can estimate the plasma electron density using the analog measurements from the PIM sensor as VM model input variables.

### 8.5.1 Experimental design

Initially, the VM system is used to estimate the electron density in a helium plasma with helium gas flowing at 200 sccm in the etch chamber at a constant pressure of 250 mTorr. The VM system is designed to operate over a range of ground impedances and applied powers. The electrical dynamics of the etch chamber system are virtually instantaneous, such that a step change in power is reflected instantly in a step change in plasma electron density. As such, static VM models are employed.

The VM models are empirical models created using an experimental data set collected from the chamber. To collect data suitable for VM model creation, electron density measurements are taken using the hairpin probe over a range of powers and ground impedance values. The electrode power is varied from 200 – 600 W and the ground impedance is varied from 0 – 26 Ω. The experimental inputs are shown in Figure 8.11(a) and the corresponding electron density values are shown in Figure 8.11(b). Control experiments are restricted to the 0 – 26 Ω range of ground impedance so that the plasma remains in the same mode over the complete experimental range.

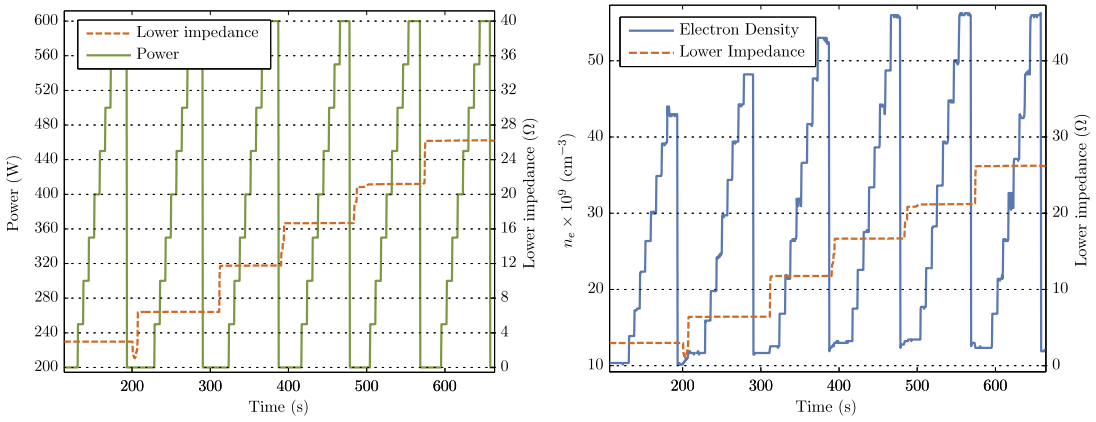



FIGURE 8.11: Designed experiment used for development of electron density VM model. (a) shows the system inputs for the experiment, and (b) shows the corresponding electron density recorded using the hairpin probe.

To avoid biasing the model in favour of lower power operation, from which there are more samples in the collected data, three samples from each unique combination of power and lower impedance are extracted from the data as shown in Figure 8.12 and used for VM modelling. A separate experimental data set containing data from random combinations of power and lower impedance is used for model testing and validation.

### 8.5.2 Modelling results

Least squares regression (LSR), artificial neural networks (ANNs), and Gaussian process regression (GPR) models are investigated as candidate VM modelling techniques (see Chapter 3 for information on these techniques). The ANNs used have a single hidden layer that is varied in size from one to fifteen neurons and initialised five times randomly during model training to determine the optimal network size and best solution. The GPR models use a squared exponential covariance function. The modelling performance of all three techniques is summarised in Table 8.3.

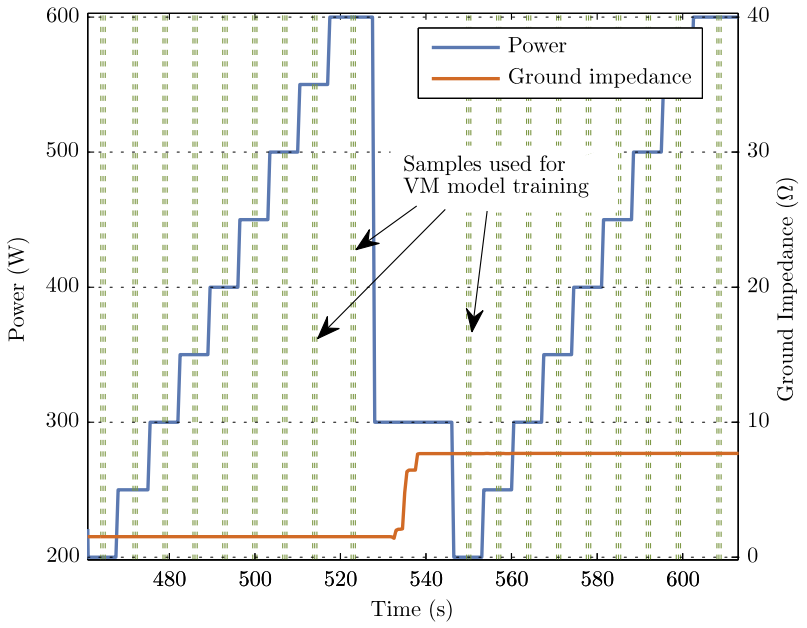



FIGURE 8.12: Samples used for VM modelling. Three samples from each unique combination of ground impedance and power are used to train the electron density VM models.

| | Training Data MSE | Test Data MSE | Max Test Error |
|---|---|---|---|
| LSR | 0.149 | 0.344 | 2.446 |
| ANN | 0.070 | 0.195 | 1.407 |
| GPR | 0.045 | 0.296 | 2.423 |

TABLE 8.3: Electron density VM estimation results. $R^2$ values for all models are greater than 0.99.

The ANN model is the best VM model according to Table 8.3, estimating the chamber electron density for the unseen test data most accurately. Figure 8.13 depicts the match between the model estimates and the recorded electron density for the model test data. Offsets between the estimated and real values of electron density are observed for some data points. These offsets are rarely greater than $1 \times 10^9$ cm$^{-3}$ ($\sim 2-3\%$ absolute error), which is deemed an acceptable level of error for the experimental control work.

### 8.5.3 Virtual metrology delay

The virtual measurement of electron density experiences a delay of approximately 0.5 seconds as a result of the response times of the PIM analog signals. Figure 8.14(a) shows the delay between the VM measurement and the hairpin probe measurement, and Figure 8.14(b) shows the delay between a change in power and the corresponding analog PIM signal response at a sample rate of 8 Hz.

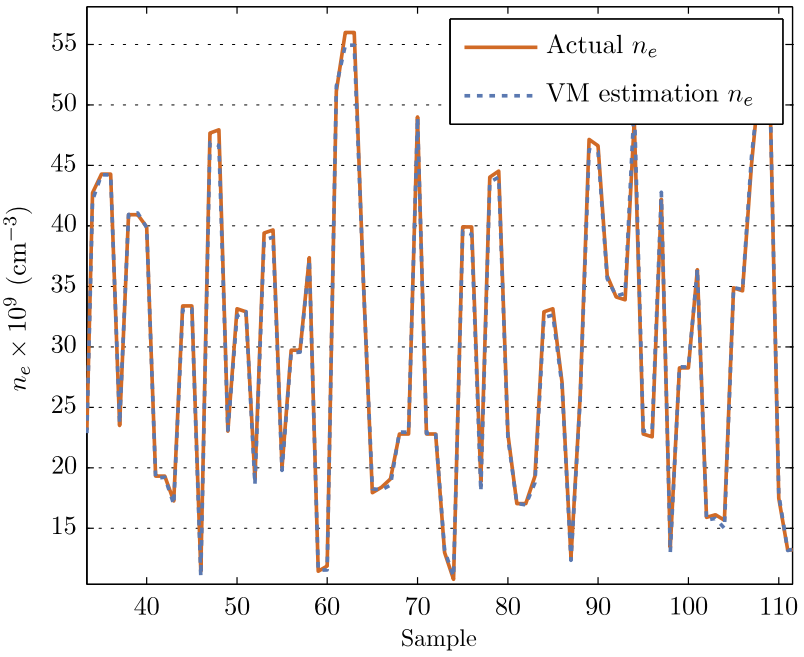



FIGURE 8.13: Electron density estimation using the ANN-based VM model on unseen test data.

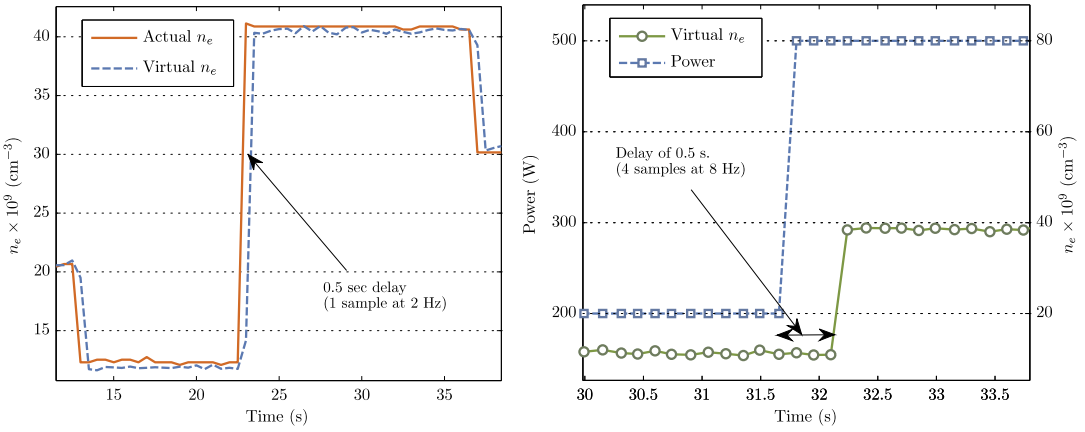



FIGURE 8.14: Delay in VM estimates of electron density. (a) shows actual electron density measurements recorded at the same time as the VM measurements, while (b) shows the delay between a step change in power and a corresponding change in the virtual measurement of electron density at a sampling rate of 8 Hz.

### 8.5.4 PI control using virtual metrology

Control of electron density can be achieved independently of the hairpin probe measurements using the newly developed VM model from Section 8.5. The virtual measurement of electron density is used in the feed-back control loop in place of the probe measurement and used to calculate the electron density error signal for the PI controller. At a sampling frequency of 2 Hz, measurements from the hairpin probe can be taken in

parallel to monitor the actual electron density and validate the VM system. Control results for $K_{prop} = 1.69$, $K_{int} = 6.78$ are shown in Figure 8.15. Disturbances to the system are successfully followed by the VM model and compensated for by the PI control algorithm.

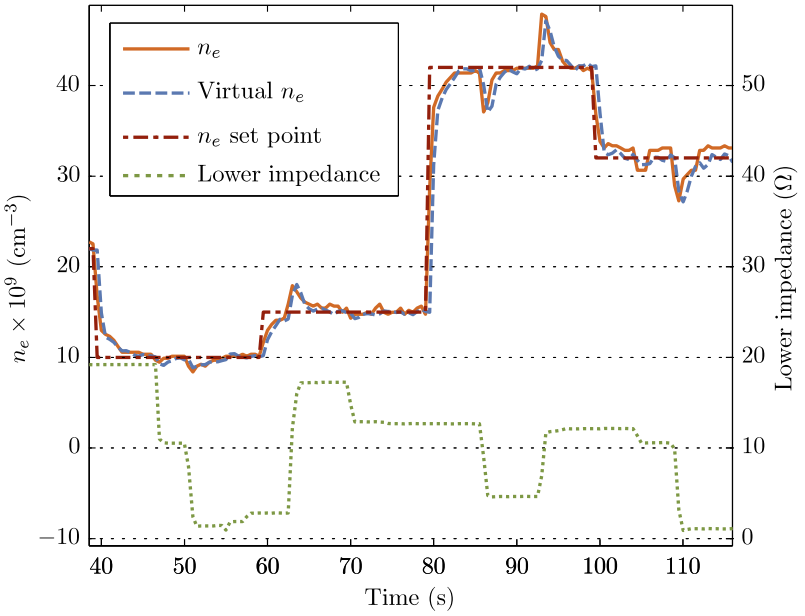



FIGURE 8.15: Control of electron density using VM measurements at 2 Hz with the PI controller parameters set to $K_{prop} = 1.69$, $K_{int} = 6.78$. Disturbances to the system are successfully followed by the VM model and compensated for by the control algorithm.

Tuning the PI controller with $K_{prop} = 3$, $K_{int} = 17$ to speed up the closed-loop response of the system causes large, slowly decaying oscillations to appear in the electron density output as shown in Figure 8.16. Such tuning served to reduce the response time and cause only small oscillations when actual electron density measurements were used in the control loop in Section 8.4.2, but introduces large oscillations when used with the VM system due to the VM measurement delay described in Section 8.5.3. The PI controller is incapable of compensating for the delay in the system without sacrificing response time. Increasing the sampling frequency of the PI control schemes will not improve this situation, as the measurement delay is fixed at 0.5 seconds. Further increases in $K_{prop}$ and $K_{int}$ to the levels that produced the response shown in Figure 8.10(b) causes an unstable electron density response.

To explicitly cater for the delay in the VM measurement of electron density, model predictive control (MPC) is employed, as described in Section 8.6.

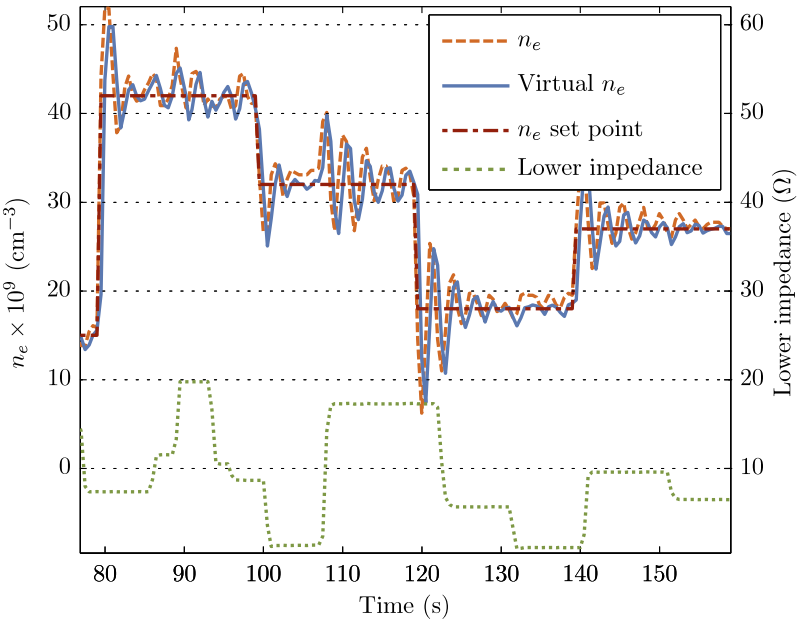



FIGURE 8.16: Control of electron density using VM measurements at 2 Hz with the PI controller parameters set to $K_{prop} = 3$, $K_{int} = 17$. Oscillations in the system response are observed because the PI controller, when tuned to a produce a fast response, cannot cater for the the measurement delay introduced by the VM model.

## 8.6 Predictive functional control

Model predictive control (MPC) or model-based predictive control (MBPC) was first used by the defence and petroleum industries in the 1970's. Since that time, MPC has been applied to thousands of control problems, spanning a plethora of industries [291]. MPC is the only control technique more advanced than standard PID control to have made a significant and widespread impact on industrial process control [292].

While classical control techniques such as PID control are suitable for many simple systems, the control of processes with considerable time delays, high-order dynamics, and/or constrained actuators is not easily accomplished. MPC provides a solution to these control challenges, and has become an attractive paradigm for industrial processes, offering a combination of simplicity and effectiveness. Originally, MPC was restricted to slower processes because it can require the solution of convex optimisation problems. However, improvements in computational capabilities in recent years have allowed MPC algorithms to be applied to high-bandwidth applications.

A wide range of MPC algorithms exist. A list of the main variants [291, 293] includes

- Dynamic matrix control (DMC)
- Extended prediction self-adaptive control (EPSAC)

- Model algorithmic control (MAC)
- Quadratic dynamic matrix control (QDMC)
- Generalised predictive control (GPC)
- Model predictive heuristic control (MPHC)
- Predictive functional control (PFC)

PFC is chosen as the MPC variant for the control of electron density in the plasma etch chamber. PFC is chosen for a number of reasons; it is easily implemented using a first order approximation to the system, it uses a single intuitively interpreted parameter during tuning, it is designed for single-input single-output systems, and it can control the etch system taking the VM delay into account. A closed form solution can be derived for a first-order PFC implementation, negating the requirement for online optimisation during control operations. An alternative option would be to use a conventional Smith predictor, but this is not considered in this research due to a high sensitivity to modelling errors [294].

### 8.6.1 Fundamental concepts

Three concepts that are central to all MPC algorithms are introduced in this section, and discussed from a PFC perspective. The treatment and use of these concepts differ from one MPC variant to another. The concepts are:

1. The internal model
2. The reference trajectory
3. The calculation of the manipulated variable.

PFC is differentiated from the other forms of MPC in that the internal models used by PFC are independent internal models that depend solely on the process input. Furthermore, the manipulated variable is constructed on a set of, typically polynomial, basis functions [295].

#### Internal model

The *internal model* is a model of the plant embedded in the predictive controller that is capable of predicting future process outputs. The internal model is not restricted to a

particular form and can be formulated as a transfer function, state-space, step-response, non-linear model etc. In the development of PFC control law described here, a first-order process is examined because many processes in production can be approximated by a first-order system model [289].

Consider a first-order process with a gain $K_p$ and a time constant $\tau_p$ subject to an input $u$. The process is represented by the Laplace transfer function

$$G_p(s) = \frac{Y_p(s)}{U_p(s)} = \frac{K_p}{1 + \tau_p s}. \tag{8.3}$$

The zero-order hold (ZOH) equivalent of $G_p(s)$ leads to the difference equation

$$y_p(k) = a_p y_p(k-1) + b_p K_p u(k-1), \tag{8.4}$$

where $y_p$ is the process output, $u$ is the process input, $k$ is the current sample number, $a_p = e^{\frac{-T_s}{\tau_p}}$, $b_p = 1 - a_p$, and $T_s$ is the sampling period. The ZOH represents the effect of using practical digital to analog convertor (DAC) for digital control. The process described by Equation (8.4) can be modelled by a first-order system model with a gain $K_m$ and a time constant $\tau_m$ as

$$y_m(k) = a_m y_*(k-1) + b_m K_m u(k-1), \tag{8.5}$$

where $y_m$ is the model output, $a_m = e^{\frac{-T_s}{\tau_m}}$, $b_m = 1 - a_m$. Two main categories of internal model arise, depending on the form of $y_*$ in Equation (8.5):

1. **Independent models.** Independent models calculate the model output $y_m(k)$ using only the known measured *process inputs* and past *model* outputs, as depicted in Figure 8.17 (a). Hence, for independent models, $y_*(k-1) = y_m(k-1)$. The process output and internal process variables are not used to calculate the future model outputs. Independent models are typically used in PFC. As the real process may be subjected to unknown disturbances and the process model will not be perfect, the process output $y_p(k)$ and the model output $y_m(k)$ may not be equal. However, $y_p(k)$ and $y_m(k)$ will evolve in parallel. Hence, during PFC, the process model is used to calculate a prediction of the increment of the process output rather than the absolute response of the process subjected to a particular input [295].

2. **Realigned models.** In a realigned model, the model output $y_m(k)$ is calculated using known and measured *process inputs* and the past measured (or estimated) *process* outputs as shown in Figure 8.17 (b). Hence, in a realigned model, $y_*(k-1) = y_p(k-1)$. Industrial vendors ADERSA claim that realigned models using noisy output measurements often suffer from offset errors [296]. Numerical imprecisions may arise for systems of order greater than one with unstable poles [295].

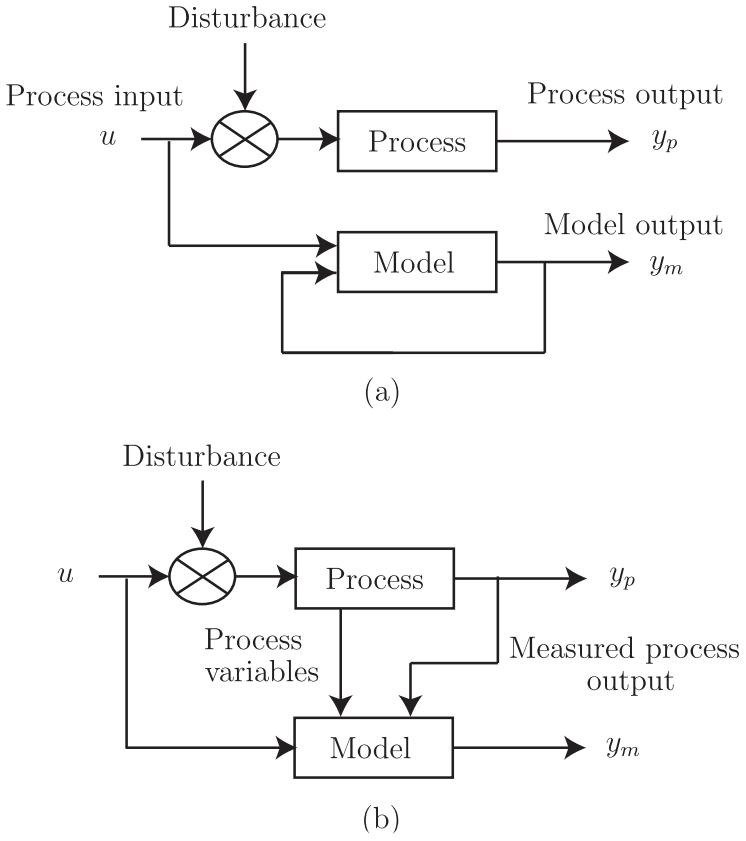



FIGURE 8.17: MPC internal model types. (a) Independent models calculate $y_m$ using only the process inputs and past *model* outputs. (b) Realigned models use *process* measurements and past *process* outputs to calculate the model output $y_m$.

For the remainder of this discussion, an independent model is assumed. The prediction of the process response using the process model described by Equation (8.5) is an equivalent problem to the solving of differential or difference equations. The solution of a finite difference equation from the instant $k = 0$ to a future time consists of two terms: the *free solution* and the *forced solution.*

The free, homogenous, initial condition response, or natural solution $y_A(k)$ is the model output that results when the future input signal is zero and the past input and output signals are non-zero. The free solution represents the process output when no

further external stimulus is applied and, in the case of asymptotically stable systems, the free solution tends towards zero according to the dynamics of the process. For the first-order linear model described by Equation (8.5), the free solution is given by

$$y_A(k+H) = y(k)a_m^H, \tag{8.6}$$

where $H$ is an integer number representing the future point for which the model output is to be found.

The forced, complementary, or inhomogeneous solution $y_F(k)$ is the process output when an external known stimulus is applied, assuming that all past signals, both input and output, are zero. For example, if a step input is applied with all past signals zero, the forced solution is the process step response. The future non-zero input signal is known and the future output is calculated by the model. For the first-order linear model described by Equation (8.5), the forced solution is given by

$$y_F(k+H) = K_m u(k)(1 - a_m^H), \tag{8.7}$$

assuming, for simplicity, that $u(k)$ is constant. For linear systems the full solution $y_m(k)$ is the sum of the free $y_A$ and the forced $y_F$ responses, such that

$$y_m(k+H) = y_A(k+H) + y_F(k+H) = y_m(k)a_m^H + K_m u(k)(1 - a_m^H). \tag{8.8}$$

**Reference trajectory**

In MPC algorithms, the *reference trajectory* is the desired future behaviour of the controlled variable. The reference trajectory is initialised on the current process output $y_p(k)$, and defines the desired path to be taken by the controlled variable towards the current set point $S(k)$. The reference trajectory can be directly interpreted as the desired closed-loop response when the process is subjected to a set point change and is re-initialised at each sampling instant using the measured or estimated process output.

The "coincidence horizon" is the set of points in the future where the process and the model outputs should be equal. For the sake of simplicity, only one coincidence point $H$ is considered in this explanation. Usually, an exponential reference trajectory is taken because such a trajectory takes only one point for initialisation, responds in a predictable manner without overshoot, and is easy to calculate in real time. For a constant set point $S(k) = S$, the exponential reference trajectory is defined such that

the error signal at a time $k+H$ is

$$S - y_p(k+H) = e(k+H) = e(k)\lambda^H, \tag{8.9}$$

where $\lambda = e^{\frac{-T_s}{\tau_r}}$, and $\tau_r$ is the required closed-loop time constant of the controlled system. One of the benefits of PFC is that the controller is tuned by simply adjusting the value of $\tau_r$, an easily interpretable variable with physical meaning.

Assuming a constant set point, the *desired* process output *increment* at the coincidence point (see Figure 8.18) $\Delta y_p(k+H)$ is such that

$$\Delta y_p(k+H) + e(k+H) = e(k) \tag{8.10}$$

Hence

$$\Delta y_p(k+H) = -e(k)\lambda^H + e(k) = e(k)(1-\lambda^H) = (S - y_p(k))(1-\lambda^H) \tag{8.11}$$

Note that rather than using only one coincidence point $H$, a *coincidence horizon* that consists of a number of points in the future can also be defined. The error between the reference trajectory and the predicted process output is minimised at all points on this horizon using more complex minimisation techniques. Multiple coincidence points are not considered in the work of this chapter.

### Calculation of the controlled variable

At each sample time $k$, the value for $\Delta y_p(k+H)$ is calculated from the process measurement, the reference trajectory, and the current set point value. Hence, it is required to calculate the future values of $u(k)$ that will produce a model output increment $\Delta y_m(K+H)$ that is equivalent to $\Delta y_p(K+H)$ at the sample time $k$.

To determine the required process inputs, the future model outputs $y_m(k+H)$ must be calculated, and as discussed earlier in this section, consist of two components:

- The effects of previous inputs that cannot be altered, but are known, i.e. the free output $y_A(k+H)$.
- The effects of future inputs that can be manipulated, i.e. the forced output $y_F(k+H)$ due to future values of $u$.

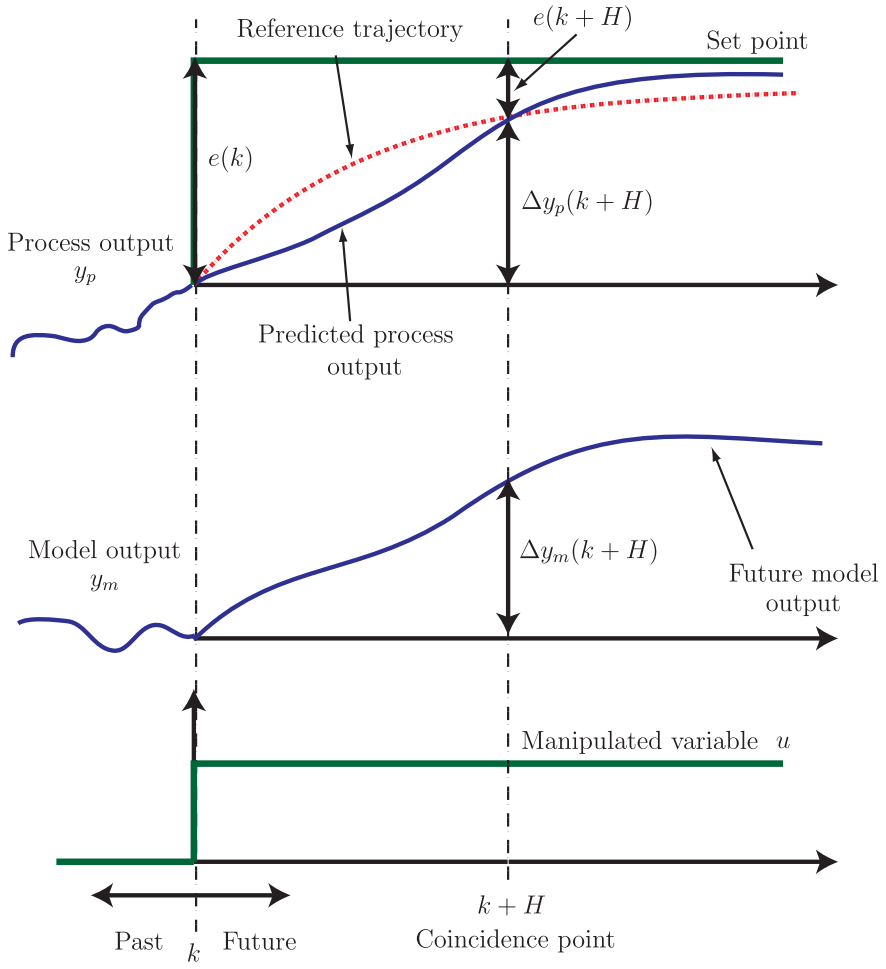



FIGURE 8.18: Reference trajectory, model increment, and process increment in PFC.

It is typical to impose a structure upon the future values of the manipulated variable $u(k)$. In PFC, the future manipulated variable is structured as a linear combination of basis functions $F_j$, which are chosen according to the nature of the process and the set point

$$u(k+i) = \sum_{j=0}^{N-1} \mu_j F_j(i) \tag{8.12}$$

where $j = 0, 1, \cdots, N-1$ and $0 \leq i \leq H$. It is required to determine $\mu_j$, the projections of $u(k)$ onto the finite set of basis functions. Thus the manipulated variable is expressed as a weighted sum of $N$ basis functions.

Normally, PFC uses a set of basis functions that consist of a polynomial basis, such that $F_j(i) = i^j$, where $j = 0$ for steps, $j = 1$ for ramps, and $j = 2$ for parabolas. The majority of set point reference paths can be expressed as combinations of these functions [297]. Using manipulated variables derived from basis functions leads to better control performance in terms of set point tracking, simpler calculations, and easier

implementation [295].

In the elementary case of a first-order process and step changes in the control set points (the case that applies here), the basis functions reduce to $N = 1$, $F_0(i) = i^0 = 1$. In this case $u(k+i) = u(k)$, i.e., a constant future manipulated variable is assumed such that it is the known forced output response of the internal model to a unit step that is multiplied by the value of $u(k)$.

At the next sample time, $k+1$, the procedure is repeated, resulting in a new reference trajectory, in essence creating a moving horizon.

As seen before

$$\Delta y_p(k+H) = (S - y_p(k))(1 - \lambda^H), \tag{8.13}$$

and $\Delta y_m(k+H)$ is given (see Figure 8.18) by

$$\Delta y_m(k+h) = y_m(k+H) - y_m(k), \tag{8.14}$$

$$\Delta y_m(k+h) = y_A(k+H) + y_F(k+H) - y_m(k), \tag{8.15}$$

$$\Delta y_m(k+h) = y_m(k)a_m^H + K_m u(k)(1 - a_m^H) - y_m(k). \tag{8.16}$$

The equation $\Delta y_p(k+H) = \Delta y_m(k+H)$ is fulfilled by

$$(S - y_p(k))(1 - \lambda^H) = y_m(k)a_m^H + K_m u(k)(1 - a_m^H) - y_m(k), \tag{8.17}$$

which can be solved for the manipulated variable $u(k)$

$$u(k) = \frac{(S - y_p(k))(1 - \lambda^H) - y_m(k)a_m^H + y_m(k)}{K_m(1 - a_m^H)}, \tag{8.18}$$

or more simply

$$u(k) = k_0 e(k) + k_1 y_m(k) \tag{8.19}$$

where $k_0 = \frac{1-\lambda^H}{K_m(1-a_m^H)}$, $k_1 = \frac{1}{K_m}$. This is the fundamental PFC control equation in its most elementary form [295]. Using a block diagram description, the control law in Equation (8.18) can be represented by Figure 8.19.

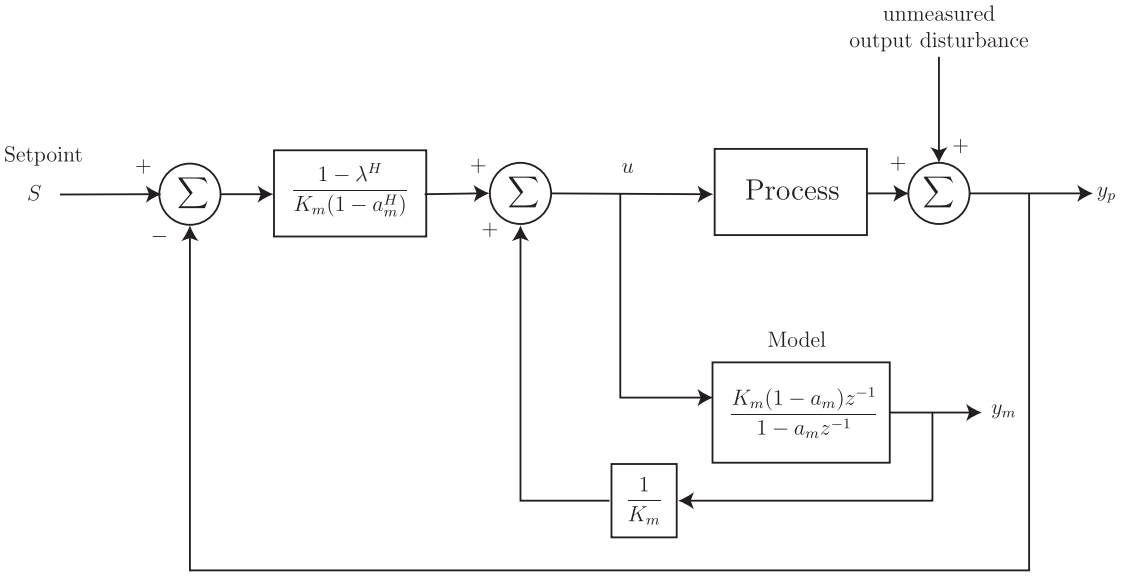



FIGURE 8.19: Block diagram implementation of PFC control law. The zero-order hold equivalent of the process model (Equation 8.5) is used in this implementation.

### 8.6.2 Systems with a pure time delay

Unlike PID controllers, predictive controllers can take pure time delay $\tau_d$ associated with the plant into account. Given a system with a time delay of $\tau_d = dT_s$ seconds, where $d$ is the number of sampling periods equivalent to the time delay, inputs applied at time $k$ will only affect the process output $\tau_d$ seconds later.

The delay is not included in the process model so that, ideally, $y_p(k) = y_m(k-d)$ and $\hat{y}_p(k+d) = y_p(k+d) = y_m(k)$, where $\hat{y}_p$ denotes a future prediction of $y_p$. Hence, due to the delay of $d$ samples, the change in the process output $y_p$ between times $k$ and $k+d$ is equal to the change of the model output $y_m$ between times $k-d$ and $k$, yielding

$$y_p(k+d) - y_p(k) = y_m(k) - y_m(k-d) \tag{8.20}$$

which rearranges to

$$\hat{y}_p(k+d) \simeq y_p(k+d) = y_p(k) + y_m(k) - y_m(k-d). \tag{8.21}$$

The reference trajectory is not initialised using the current of value $y_p(k)$, but on the predicted value of $y_p(k+d)$ in order to anticipate its response. $\hat{y}_p(k+d)$ is simply an extrapolation in time of the current value of $y_p(k)$ using the response of the model. The control equation given in Equation (8.18) is still valid by replacing $y_p$ with the expression for $\hat{y}_p(k+d)$ from Equation (8.21).

| | Precision | Transient response | Robustness |
|---|---|---|---|
| Basis Function | 1 | 0 | 0 |
| Reference Trajectory | 0 | 1 | 0.5 |
| Coincidence Horizon | 0 | 0.5 | 1 |

TABLE 8.4: Effect of PFC parameters on controller tuning [293]. The effect of each parameter on the corresponding controller performance is graded between 0 for minimum and 1 for maximum influence.

### 8.6.3 Controller Tuning

Three principal specifications are of interest during the tuning of a control system,

- the controlled variable precision in steady state,
- the dynamic response of the controlled variable to changes in set-point, and
- the robustness of the control system to model uncertainties and process structural variation.

Tuning of PFC controllers is a function of the order of the basis functions, the reference trajectory, the control horizon, and the closed-loop response time (CLRT). A general summary of the influence of the PFC parameters is given in Table 8.4, where the influence of PFC parameters on precision, transient response, and robustness are graded between 0 for minimum influence and 1 for maximum influence [293].

Many processes in production industries can be approximated by a first order system model [289], and in the majority PFC control applications, an exponential reference trajectory is used with a single coincidence horizon point and a zero-order polynomial basis function. In such cases, the main tuning parameter becomes the desired CLRT which is specified by $\tau_r$ in $\lambda = e^{\frac{-T_s}{\tau_r}}$. Because the uncontrolled process responds according to the open loop response time (OLRT), the ratio OLRT/CLRT plays an important role in the controller's tuning leading to an overshooting or non-overshooting manipulated variable, $u(k)$. If the CLRT is much smaller than the OLRT at the time of set point change, the manipulated variable will overshoot. On the other hand, if the CLRT is much larger than the OLRT, the manipulated variable will generally remain sufficiently constrained to avoid overshoot [295]. The dynamics of the manipulated variable can be an issue for physical systems where actuator speed and range is limited. For the plasma system investigated in this chapter, the OLRT is extremely fast, allowing considerable freedom in the specification of desired CLRT.

## 8.7 Predictive functional control of electron density

PFC is suitable for real-time control of plasma electron density using VM measurements because it can cater for the measurement delay introduced by the VM system and produce a first-order exponential electron density response. This section details the implementation of the PFC algorithm on the plasma etch chamber.

### 8.7.1 System Model

Figure 8.20 shows the relationship between power and electron density at three different chamber pressures. Over the range shown, the relationship between the power delivered to the chamber electrode and the plasma electron density is relatively linear for constant values of ground impedance and pressure. Significant changes in the system gain occur as the chamber pressure is changed. Smaller changes in gain are observed at each pressure set point as the ground impedance of the chamber is altered. To model the system using a consistent process model, the pressure is fixed at 250 mTorr for the experiments in this section.

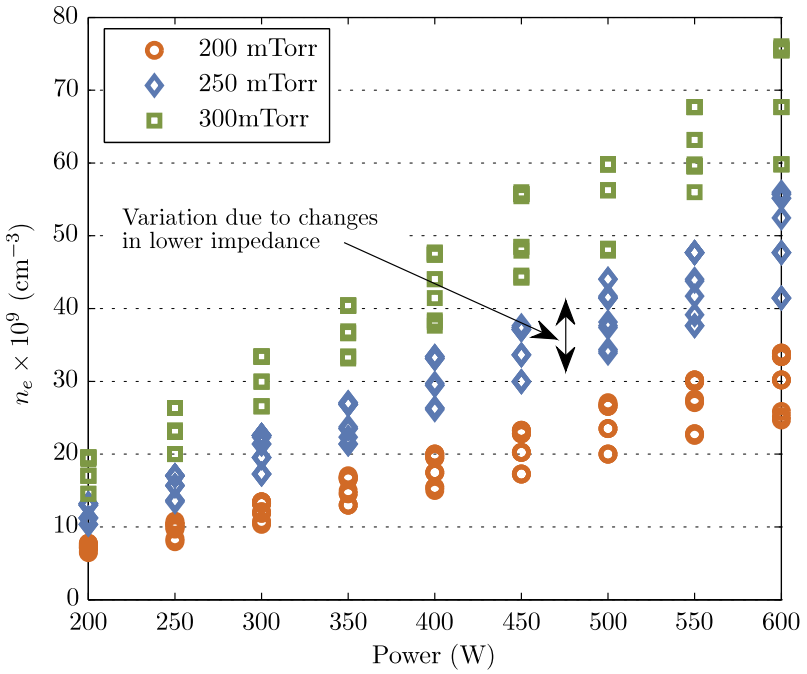



FIGURE 8.20: Electron density response to power at different pressures. Large changes in system gain occur as the the chamber pressure is changed. At each pressure set point, smaller changes in gain are observed as the ground impedance of the chamber is altered.

At a constant pressure, the system can be approximated as a pure gain $K_m$, with negligible dynamics and a delay term such that its transfer function is

$$G_m(s) = \frac{Y_m(s)}{U(s)} = K_m e^{-\tau_d s} \tag{8.22}$$

where, $Y_m(s)$ represents the system output, the plasma electron density, $U(s)$ represents the system input, the RF power, and $\tau_d = dT_s$ is the VM delay in seconds. No dynamics are used in this model because the relationship between power and electron density is effectively instantaneous at the sampling frequencies used in this thesis. The lack of dynamics in the system model simplifies the PFC control equations since $a_m = e^{\frac{-T_s}{\tau_m}} = 0$ and the system model equation without delay (from Equation 8.5) will consist only of the forced solution $y_F$,

$$y_m(k) = K_m u(k-1). \tag{8.23}$$

There will be no free solution $y_A$ as the system output does not depend on past outputs. Hence, the PFC control equation (Equation (8.18)) for the plasma system becomes

$$u(k) = \frac{(S - y_p(k))(1 - \lambda^H) + y_m(k)}{K_m}. \tag{8.24}$$

The electron density response at a fixed pressure with the VM delay included can be approximated to have a constant gain, and is represented by the linear model

$$y_m(k) = K_m u(k-d) - c_m. \tag{8.25}$$

Equation (8.25) has a non-zero vertical intercept. To allow the system to be treated as a simple gain term plus delay, the vertical intercept is included in the model as an input correction term of $-\frac{c_m}{K_m}$, as shown in Figure 8.21. Hence, to the PFC controller, the plant acts as a pure gain system with a time delay. As described in Section 8.6.2, the delay term is not included in the internal model during PFC operation.

As shown in Figure 8.20, using fixed values for $K_m$ and $c_m$ in Equation (8.25) is not a completely accurate representation of the process response for all values of the ground impedance at a fixed pressure. However, MPC-based controllers can maintain control of the system with zero steady state error even in the presence of model mismatch [295, 296].

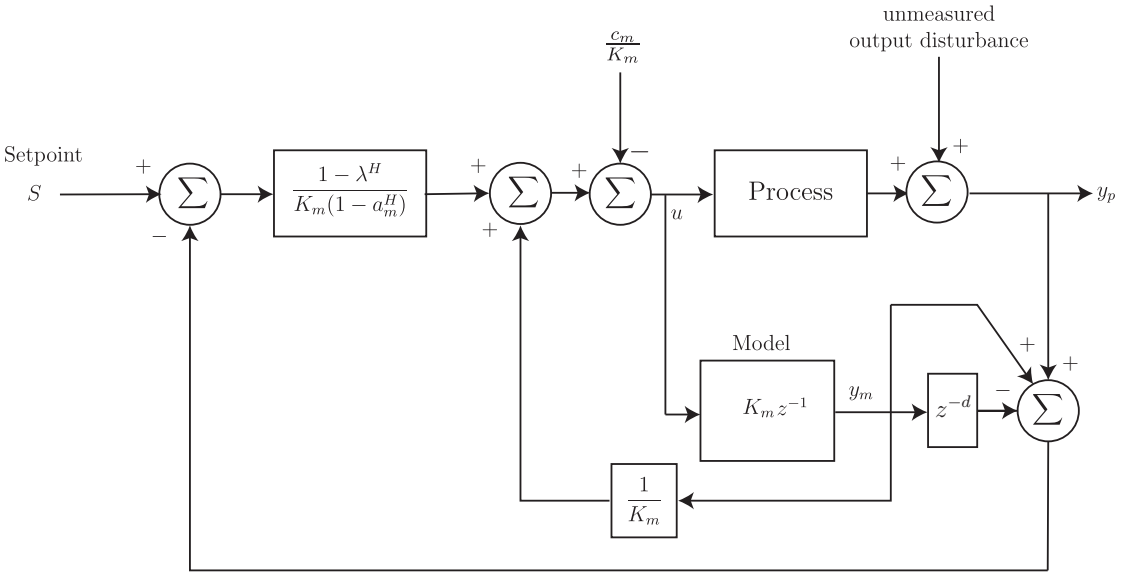



FIGURE 8.21: PFC block diagram with input correction term. The input correction term is introduced so that the plant and model appear as simple gain terms to the PFC controller. The model output is corrected for the system delay as described by Equation (8.21)

### 8.7.2 Results at $T_s = 0.5s$

As a first test of performance, and for direct comparison with the PI controller of Section 8.4, the PFC controller is implemented with a sampling frequency of 2 Hz and a required closed loop time constant of $\tau_r = 1$ s. The VM delay is equivalent to one sample time at this sampling rate. Successful set point tracking is achieved as shown in Figure 8.22(a), and the controller is capable of correcting for disturbances as shown in Figure 8.22(b). Small mismatches are seen between the virtual and the real measurements of electron density at different set points due to VM modelling errors. As expected, the electron density response is exponential with no overshoot due to the configuration of the PFC reference trajectory.

Because these experiments operated at a fixed pressure of 250 mTorr, the model gain term $K_m$ is fixed in Equation (8.25) to 0.1104 (determined from results shown in Figure 8.20) and hence it differs from the actual system gain $K_p$ depending on the ground impedance setting during experimentation. Although such model mismatch does not result in steady-state errors in the controlled variable, it does prevent the requested closed-loop time constant from being achieved. Hence, the time constant of the controlled electron density varies between 1 s and 2.5 s due to model mismatch at different operating points when $\tau_r = 1$. The degree of mismatch between the model and the process can be gauged by comparing the PFC model output $y_m$ with the actual process output $y_p$ (or VM measurement in this case) as shown in Figure 8.23.

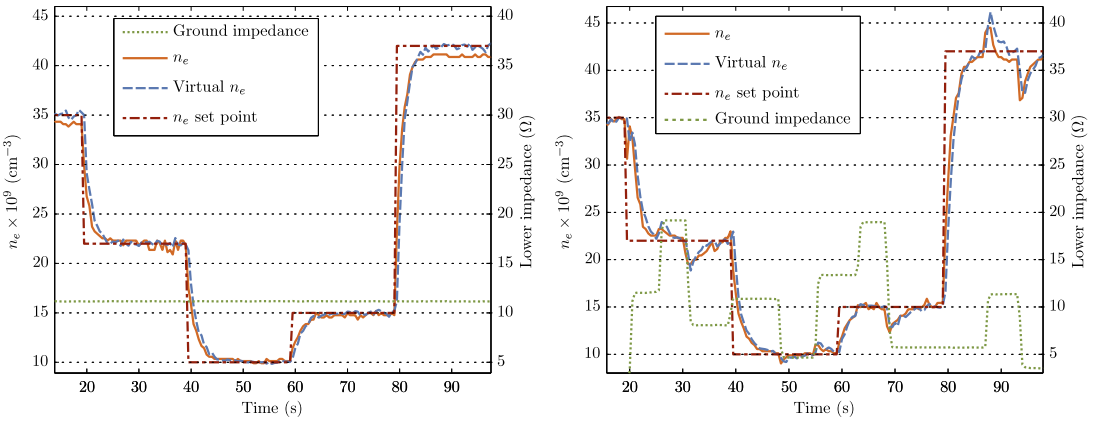



FIGURE 8.22: PFC control of electron density with $T_s = 0.5$ s. (a) shows the controlled response when the ground impedance remains constant, and (b) shows the controlled response when random disturbances in ground impedance are introduced. $\tau_r = 1$ s for both tests.

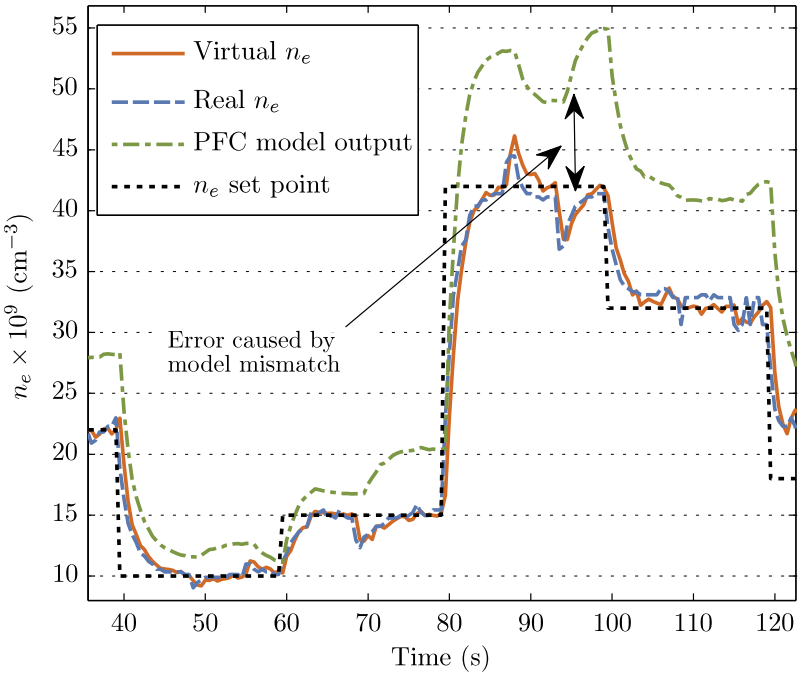



FIGURE 8.23: PFC model output compared to VM measurement of electron density. The PFC internal model output does not match the output of the process due to model mismatch, i.e. $K_p \neq K_m$. The degree of model mismatch depends on the operating point of the system because for the plasma system, $K_p$ varies with ground impedance.

In contrast to the tuning of the PI controller in Section 8.4, the PFC controller is intuitively tuned by adjusting a single parameter, $\tau_r$, to achieve faster or slower closed-loop response times. Figure 8.24 depicts the effect of varying $\tau_r$ to alter the time constant of the controlled variable. Unlike the PI control results of Section 8.4, overshoot and oscillations do not appear for relatively fast response times because the system delay

is explicitly catered for in the controller. Such fast response times cannot be achieved using the PI controller without taking the VM delay into account since the controller will continuously effect changes in the manipulated variables that are too late to compensate for current errors.

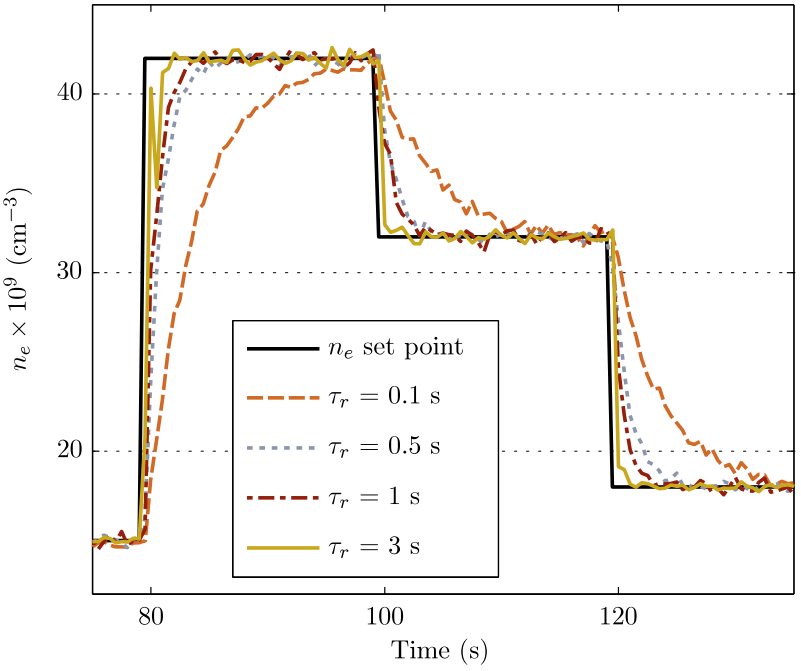



FIGURE 8.24: Effect of adjusting $\tau_r$ with $T_s = 0.5$ s. Adjusting the required closed-loop time constant $\tau_r$ increases the response speed of the controlled variable. Some inconsistencies in the electron density are seen for $\tau_r = 0.1$ seconds.

### 8.7.3 Results at $T_s = 0.1s$

Control at sampling speeds faster than 2 Hz are possible since the virtual measurement of electron density can be used in the control loop negating the necessity for slow communication with the electron density probe. The limiting factor on the sampling speed is determined by the computation time for each control output. The computation time for the VM ANN model output is the most demanding part of the calculations carried out at each sampling instant. However, the added accuracy of the ANN model over faster linear models justifies its use. The sampling frequency can be increased five-fold, to 10 Hz, without encountering computational difficulties. The VM and control algorithms are implemented on one computer using a 10 Hz sampling rate, while, for reference and validation purposes, electron density measurements are taken using the hairpin probe on a second computer at the slower sampling rate of 2 Hz. The results for $\tau_r = 1$ s with and without process disturbances are shown in Figure 8.25.

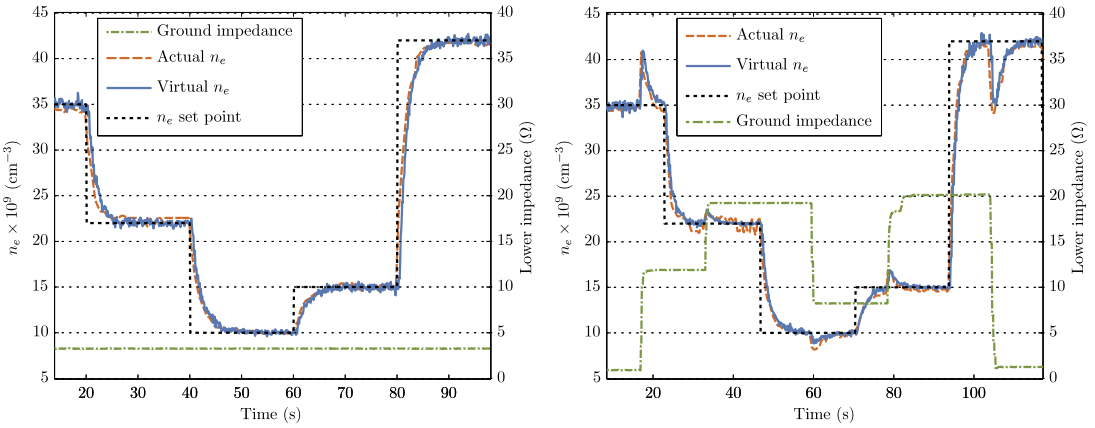



FIGURE 8.25: PFC control of electron density with $T_s = 0.1$ s. (a) shows the controlled response when the ground impedance remains constant, and (b) shows the controlled response when random disturbances in ground impedance are introduced. $\tau_r = 1$ s for both tests.

The electron density estimate responses for different closed-loop time constants are shown in Figure 8.26. Although the closed-loop settling time for $\tau_r = 0.1$ s should be approximately $3 \times \tau_r = 0.3$ s, this settling time is not achieved due to mismatch between the PFC model and process gains. Apart from this, the electron density responses are well-controlled to achieve exponential trajectories, and for $\tau_r = 0.5s$, 15 samples are taken per closed-loop settling time.

One disadvantage of increasing the sampling rate is that less time is available to average the values recorded from the PIM analog signals. As a result, the VM estimates are more susceptible to noise. When small $\tau_r$ values are used, the controller is more sensitive to this noise on the VM signal, as seen by irregularities in the trace for $\tau_r = 0.1$ s in Figure 8.26.

Step changes in ground impedance are unlikely to occur during wafer etching in a production process; it is more likely that the controller will only have to adjust the process power to counteract for changes in ground impedance caused by slow processes such as chamber drift and physical wear and tear. However, intermittent step changes in ground impedance may be brought about by preventative maintenance (PM) events. Step changes in electron density set point may occur depending on the requirements of particular processing recipes.

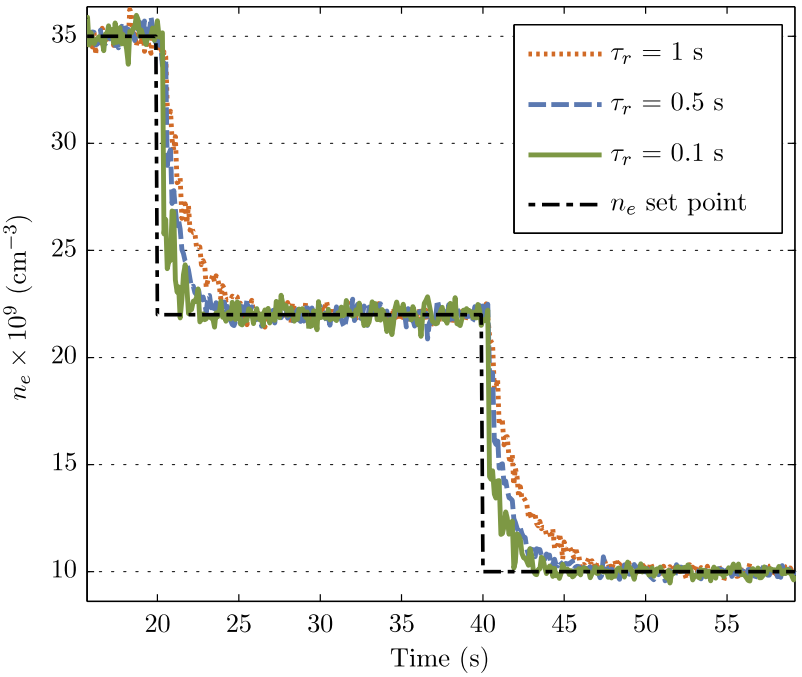



FIGURE 8.26: Effect of adjusting PFC controller $\tau_r$ with $T_s = 0.1$ s. Reducing the required closed-loop time constant $\tau_r$ increases the response speed of the controlled variable. The VM measurement of electron density is noisier at high sampling rates than at slower sampling rates and the controller is more sensitive to this noise when tuned with faster response times.

## 8.8 Adaptive PFC

This section investigates the use of an adaptive algorithm to update the internal model parameters of the PFC controller using measurements from the process. By correcting the gain term $K_m$ of the PFC model in response to changes in the system gain $K_p$, consistent electron density response times can be achieved over a broad range of operating conditions. Model updates are realised using recursive least squares (RLS).

As discussed in Section 8.7.1, in the plasma etch system, changes in $K_p$ can be caused by either changes in ground impedance or chamber pressure. In Section 8.8.1, the effect of model mismatch on controller performance is examined, where the model mismatch is effected by changes in chamber pressure. In Section 8.8.2, a brief explanation of the RLS algorithm and its application to PFC in this study is presented. Finally, in Section 8.8.3, experimental results using the adaptive control scheme are examined.

### 8.8.1 Effects of PFC model mismatch

#### Artificial simulation of model mismatch

Figure 8.27 shows PFC results for situations with gross mismatch between the PFC model gain and the actual process gain. The required time response $\tau_r$ for both sets of results in Figure 8.27 is fixed at $\tau_r = 0.3$ s. In Figure 8.27(a), the model gain $K_m$ is set higher than the actual system gain $K_p \sim 0.11$ such that $K_m = 0.25$. The resulting transient response is slow and the PFC model output is much higher than the actual process output. The controller consistently underestimates the input increments required to effect changes in the controlled variable. The opposite effect is seen in Figure 8.27(b) where $K_m = 0.06$. Here, the PFC model gain $K_m$ is lower than the process gain $K_p$ leading to a fast, overshooting response as the predictive controller overestimates the input increments required to effect changes in the controlled variable. The PFC model output is smaller than the real process output.

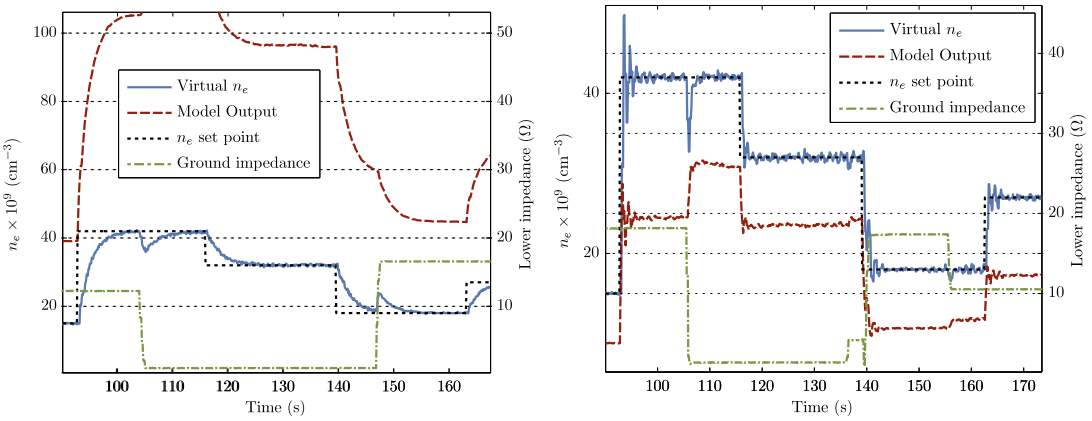



FIGURE 8.27: Effect of PFC model mismatch on the electron density transient response. In (a), $K_m = 0.25$ and in (b), $K_m = 0.06$. The actual value of $K_p$ is approximately 0.11. Because the system gain $K_p$ changes with the ground impedance, the size of the error between the model output and the controlled variable (virtual measurement of $n_e$) is affected by changes in the ground impedance value.

#### Model mismatch due to pressure changes

Figure 8.20 depicts the electron density recorded in the chamber for a range of power and ground impedances. The electron density system gain can be reasonably well approximated as a constant for each pressure individually and this fact allows a PFC controller with a constant internal model to be implemented for a fixed pressure in in Section 8.7. However, such a system may fail or perform poorly when changes are introduced

| | Chamber power (W) | Ground Impedance (Ω) | Pressure (mTorr) |
|---|---|---|---|
| Lowest value | 200 | 0 | 200 |
| Highest Value | 600 | 22 | 300 |

TABLE 8.5: Design of experiment inputs for VM model with varying pressure.

during system operation that cause substantial changes in the system gain, for example, changes in the chamber pressure.

The operating range of the VM model developed in Section 8.5 needs to be expanded to produce accurate estimates of electron density over a range of pressures. A new experiment is performed in which the chamber power, ground impedance, and pressure are varied over the ranges shown in Table 8.5. No direct measurement of pressure is available to the VM system in the particular equipment used during these experiments, but the VM model is capable of estimating the electron density over the explored pressure range using the PIM measurements alone. However, as a result of a greater reliance on the exact values of the PIM measurements, the electron density estimates become more sensitive to noise on the PIM signals, resulting in noisier electron estimates than those seen in Section 8.7.

The effect of chamber pressure variations on the PFC controller performance using a time-invariant PFC internal model is demonstrated in Figure 8.28. In the experimental data shown, the system conditions are changed at each "system change" point to the settings marked on the figure. The PFC internal model error, and hence the electron density time response, changes depending on the system conditions.

While gain scheduling could be implemented to change the PFC internal model to suit the particular operating regime in use at one particular time, a more elegant solution is to update the model parameters in real time in response to the system changes. This solution is explored in Section 8.8.2.

### 8.8.2 Recursive least squares

Due to the relatively simple form of the system model in use (see Section 8.7.1), the model update can be constructed as a linear regression problem. At each sample $k$, the system electron density estimate $\hat{n}_e$ is given by

$$\hat{n}_e(k) = K_m p(k-d) + c_m \tag{8.26}$$

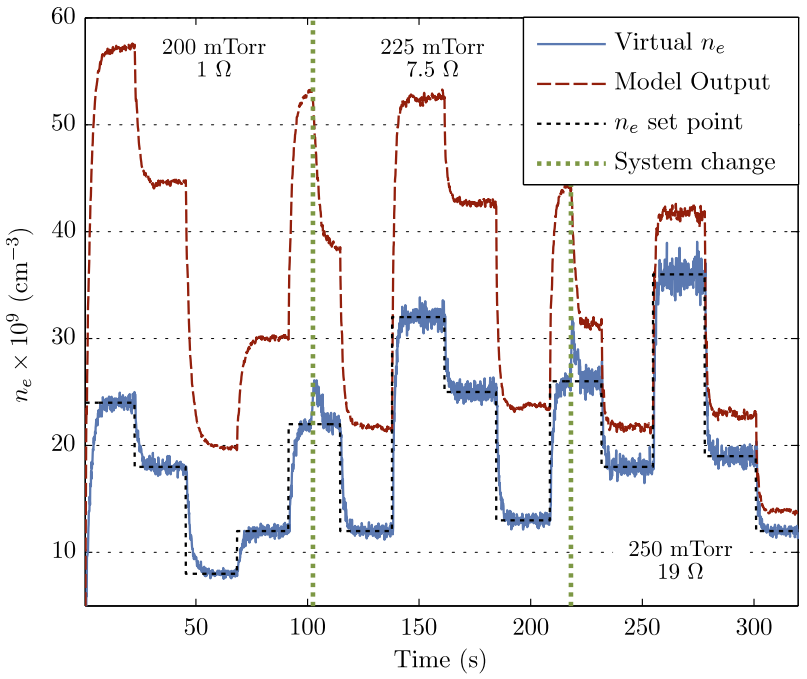



FIGURE 8.28: Effect of pressure disturbances on PFC controller performance. The pressure and ground impedance of the chamber are adjusted at each "system change" in the diagram. Pressure is adjusted from 200–225–250 mTorr. The PFC model gain $K_m$ is fixed at 0.1104, which best suits the 250 mTorr operating space.

where $K_m$ is the modelled process gain, $p$ is the power applied to the chamber, $d$ is the system delay in sample periods, and $c_m$ is an offset term. The electron density for samples $k$ to $k+N$ can be written as

$$\begin{bmatrix} \hat{n}_e(k) \\ \hat{n}_e(k+1) \\ \vdots \\ \hat{n}_e(k+N) \end{bmatrix} = \begin{bmatrix} p(k-d) & 1 \\ p(k-d+1) & 1 \\ \vdots & \vdots \\ p(k-d+N) & 1 \end{bmatrix} \begin{bmatrix} K_m \\ c_m \end{bmatrix}. \tag{8.27}$$

Equation (8.27) is of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ which can be solved using least squares regression techniques, as described in Section 3.1. However, rather than storing a fixed window of inputs and outputs and recalculating new values for $K_m$ and $c_m$ at each sample, recursive least squares (RLS) can be used to recursively update the $[K_m c_m]^T$ vector using each value of $p(k)$ and $\hat{n}_e(k)$ as they arise. The full derivation of the RLS algorithm is provided in Appendix C but the algorithm can be summarised [298], at sample $k+1$, by the following steps:

1. Collect $y(k+1)$ and form $\vec{\mathbf{x}}(k+1)$ from new data collected during the sample, where $\vec{\mathbf{x}}(k+1)$ is a row vector of variable measurements forming a new row of $\mathbf{X}$.

For the plasma system, $\vec{\mathbf{x}}(k+1)$ is simply the applied power from one delay time ago ($p(k+1-d)$) and a unit value, and $y(k+1)$, the process output measurement, is the virtual measurement of the electron density at sample $k+1$.

2. Calculate the current error $e(k+1)$ using

$$e(k+1) = y(k+1) - \vec{\mathbf{x}}(k+1)\hat{\boldsymbol{\beta}}(k), \tag{8.28}$$

where $\hat{\boldsymbol{\beta}}(k)$ are the model parameters from sample $k$, for this system a vector made up of $K_m$ and $c_m$.

3. Calculate the covariance matrix $\mathbf{P}(k+1)$ using

$$\mathbf{P}(k+1) = \mathbf{P}(k)\left[I - \frac{\vec{\mathbf{x}}(k+1)^T\vec{\mathbf{x}}(k+1)\mathbf{P}(k)}{\lambda_{RLS} + \vec{\mathbf{x}}(k+1)\mathbf{P}(k)\vec{\mathbf{x}}(k+1)^T},\right], \tag{8.29}$$

where $I$ is the identity matrix and $\lambda_{RLS}$ is the RLS forgetting factor.

4. Update the model parameters, $\hat{\boldsymbol{\beta}}(k+1)$ using

$$\hat{\boldsymbol{\beta}}(k+1) = \hat{\boldsymbol{\beta}}(k) + \mathbf{P}(k+1)\vec{\mathbf{x}}(k+1)^T e(k+1). \tag{8.30}$$

5. Return to step 1.

### 8.8.3 Application to PFC

RLS is included in the PFC control algorithm as shown in Figure 8.29.

The results for the PFC controller using the adaptive model are demonstrated in Figure 8.30. The same system excitations as for Figure 8.28 are used during the test. The PFC model output realigns to the VM estimation of the process output shortly after each system change, and with the PFC model correctly adapted, the PFC controller can control the process in line with the requested dynamics such that $\tau_r = 0.3$ s.

A forgetting factor of $\lambda = 0.995$ is used in the RLS algorithm to avoid rapid model updates in response to the noisy VM estimates of electron density. The evolution of the model parameters is shown in Figure 8.31. After each system change in the experiment, the RLS algorithm requires several different electron density set points before the model parameters settle on the correct values. The model estimates are inaccurate during these settling periods, and hence the system transient responses are not as desired. However, once appropriate values for the model parameters are ascertained, the system response is more accurate than the response achieved using a constant PFC internal model. To

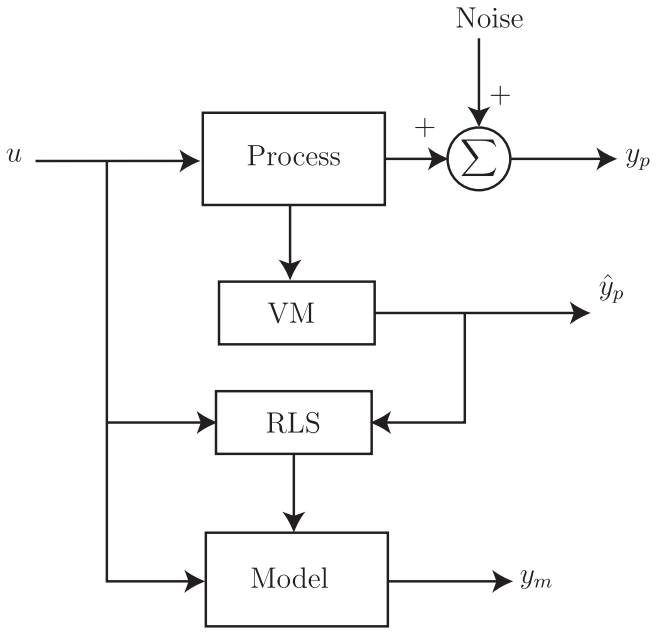



FIGURE 8.29: Recursive least squares (RLS) used to update PFC model parameters.

prevent unexpected or perhaps harmful behaviour in a production system, the calculated model parameters from the RLS algorithm would not be used in the online PFC internal model until they have converged.

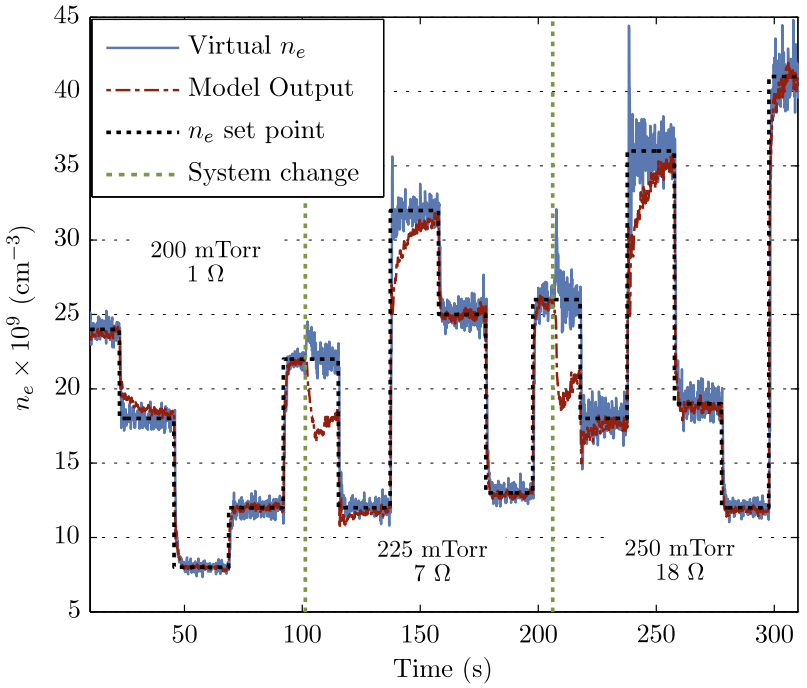



FIGURE 8.30: PFC with pressure changes using an internal model adapted with RLS. Once the model parameters are updated to match the process after each system change, the requested transient responses are observed in the controlled variable.

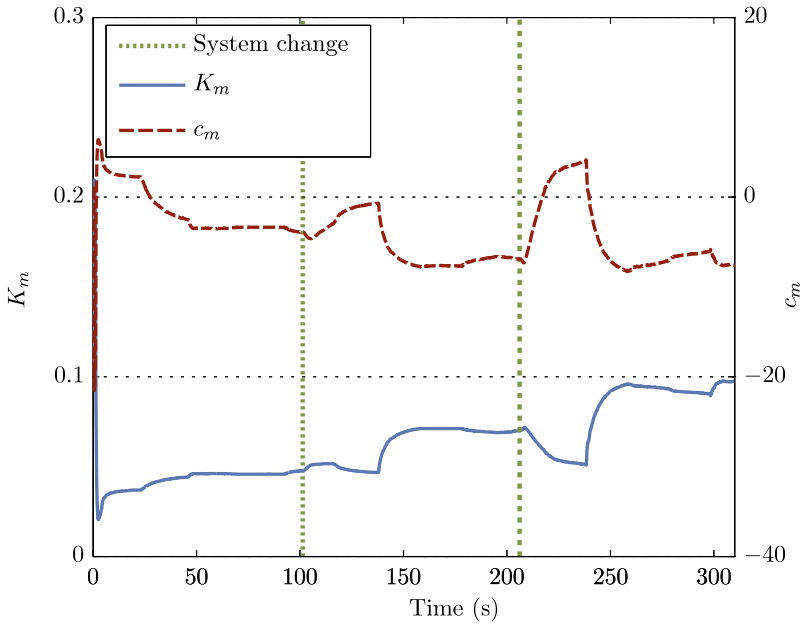



FIGURE 8.31: Evolution of model coefficients using RLS. The model is represented as a linear equation $n_e = K_m p + c_m$. Note that after every system change, some time is required before the coefficient values settle. During these transition periods, the model estimates are inaccurate.

## 8.9 Discussion

In this chapter, real-time control of plasma electron density using virtual measurements has been achieved in a production plasma etch chamber. Proportional-integral control is found to be unsatisfactory for control of electron density because of its inability to cater for the measurement delay introduced by the PIM signals used in the VM system. Predictive functional control (PFC) is chosen as the model predictive control (MPC) variant to implement the control. A first-order PFC implementation allows relatively fast set point tracking (settling times of approximately 1 second) without overshoot, explicitly deals with the time-delay introduced by the PIM-based VM system, and displays excellent disturbance rejection properties. The control system is expanded to operate across a larger operating range by adapting the PFC internal model parameters using recursive least squares (RLS). Closed-loop response times of less than one second can be reliably achieved over the full operating range using the adaptive system. In a production setting, the operating range is likely to be much smaller because process recipes typically operate around relatively fixed operating points.

Although the experiments in this chapter were completed using a non-etchant gas, helium, similar control performance is feasible using etchant gases with product wafers in the etch chamber, provided the control is implemented during the steady-state etching of wafer layers. More complex VM models would be required for control spanning multiple

etch layers. Such control will pave the way for more reliable etch performance, since the controller can compensate for disturbances introduced to the ground impedance as a result of component replacement or conditioning of the chamber in the production environment.

The results of this chapter have been accepted for publication in the International Federation of Advanced Control (IFAC) World congress in August of 2011.

# Chapter 9

# Conclusions and future directions

In this chapter, the main conclusions from the research in this thesis are provided in Section 9.1, and details on potential future directions for work are provided in Section 9.2.

## 9.1 Overall conclusions

From the literature review completed in Chapter 4, it is clear that accurate VM modelling of plasma etch processes is still a considerable challenge for industrial practitioners. VM research concentrates on many different processes and uses numerous techniques with no clearly advantageous approach emerging. While numerous studies exist that use data from designed experiments (such as factorial experiments) to develop virtual metrology (VM) models, research based on the analysis of production data sets is less common. Designed experiment data are useful for VM model development if the process variations introduced during the experiments are representative of the variations encountered during process operation. Examples of such situations are seen in the development of VM models for the VASIMR engine in Chapter 5, and, by a lesser amount, for the experimental electron density control system described in Chapter 8.

However, input variable variations introduced during designed experiments are typically not representative of process variations observed during production etch processes. During production, process input set points typically remain constant, and process variation is primarily caused by mechanisms such as chamber conditioning and preventative maintenance (PM) events. Such variations are not be captured in typical designed experiments and the measured dynamics and interrelationships between variables may not

be representative of the production environment. Also, factorial experiments for manufacturing processes can often be prohibitively expensive to carry out due to the high value nature of product wafers that may need to be sacrificed during experimentation.

To address this issue, Chapters 6 and 7 focussed on the development of wafer-level VM models for estimation of plasma etch rate using data collected from a plasma etch process during production. An extensive account of different VM approaches is provided for the benefit of industrial practitioners aiming to implement VM for etch processes.

The principal difficulty of using production data during the creation of VM models for time varying processes such as plasma etch is that VM estimates can be required for parts of the system operating space that have not been included in the model training data. As the system varies, new areas of the operating space are encountered, potentially rendering previously designed models obsolete. In general, estimation of process output variables for wafers dissimilar to those in the VM model training data are inaccurate, as shown by the cluster modelling results of Chapter 7. Hence, metrics and techniques such as the confidence intervals from Gaussian process regression (GPR) models, the $Q$ and $T^2$ statistics, Mahalanhobis distance, or clustering algorithms should be used to indicate when estimation from new regions of the system operating space are required, and care should be taken in such situations. When these metrics suggest, models may need to be updated or new models may be required to address the gap in operating space coverage, and the metrics could also be used to effect changes in the metrology rate as appropriate (dynamic sampling) to update or create the VM models. Estimation results from an interleaved data set are used in Chapter 6 as a metric for the potential accuracy of a VM system using a comprehensive data logging philosophy and an extensive data set. Increases in etch rate estimation accuracy are observed for the interleaved scheme.

A broad range of variable selection, data reduction, modelling, and data division schemes are examined throughout Chapters 6 and 7 for VM model development. Throughout the analysis, the maximum accuracy achieved by the VM models is approximately 1.14% MAPE, using a windowed GPR scheme. The accuracy levels achieved by other model types does not vary greatly from this level. The consistency of the results from different techniques lends confidence to the conclusion that the achieved accuracy represents the lower limit of achievable estimation accuracy for the etch rate of the process studied using data at the available frequency of metrology (approximately 4 % metrology rate). It is highly unlikely that further modelling techniques will substantially improve upon the accuracy reported. A MAPE of 1.14% is almost fully attributable to the cumulative effects of variations in incoming product material, variations in post-etch processes, and errors in measurement of trench depth.

For the process under study, the accuracy achieved is sufficient to successfully monitor whether the etch process is operating within specifications or not. The windowed-GPR modelling scheme is also accurate and fast enough to allow implementation of a wafer-to-wafer control scheme for etch rate. Currently, the time for the main etch step is varied on a manual basis, but the frequency and promptness of the manual updates are limited by a low etch depth measurement frequency and a metrology delay of the order of days. While small, high-frequency variations in etch rate are not accurately estimated by the VM models, the overall average etch rate is approximated accurately, and control of this average etch rate is achievable using the VM estimates with a wafer-to-wafer control scheme. Implementation of such a scheme would represent a considerable improvement upon the current standard practice for this process, and can be implemented without the addition of further sensors on the tools through the use of etch process (EP) data only.

Before the VM models are used for APC, further analysis with fully measured lots is required to ensure that models cater for intra-lot signal variations, for example, first wafer effects. However, apart from first-wafer effects, the etch process studied is known to operate relatively uniformly, and hence the VM scheme developed is conservatively conjectured to be capable of generating meaningful estimates for 70% of the wafers processed. From a manufacturing perspective, estimates for this proportion of processed wafer would represent a vast improvement on the existing infrastructure and would form a useful system for excursion detection and process monitoring before run-to-run control is considered.

With regard to the modelling techniques examined in this thesis, GPR modelling represents an advantageous new modelling technique for the semiconductor industry. With relatively few exceptions, GPR models perform consistently better for etch rate estimation than all other modelling techniques, including artificial neural network (ANN) models, the most common model type found in the semiconductor literature. This thesis documents the first application of GPR VM models to a semiconductor etch process. GPR models come with the additional advantage over ANN models of easily calculable confidence limits on each estimate, which are useful to gauge VM estimate reliability. The GPR models also perform relatively well when limited amounts of training data are available, which is often the case when data sets are disaggregated for local modelling. The confidence limits are found to be accurate for the data examined in Chapter 7. Models based on principal component analysis (PCA) yield consistently poor results. Least angled regression (LARS) models prove to provide more accurate results in general than the more popular stepwise regression models, with the exception of the windowed modelling investigations.

The local modelling results demonstrate the advantage of incorporating process knowledge during VM development to account for known process dynamics. Chapter 7 examines three modelling schemes that manipulate the input data in different ways to cater for the etch process dynamics: regional PM cycle modelling, clustered modelling, and windowed modelling. Regional PM cycle models were *not* found to aid VM estimates, demonstrating that inter-PM variation is greater than intra-PM variation. However, clustered models exhibit potential gains in etch rate estimation accuracy over comparable global models, and windowed models are proven to operate most effectively in the research. Cluster models were born from the knowledge that the etch process operates in distinct modes, with the potential that local models for each mode could prove more accurate than a single global model. Windowed modelling is used to maintain model currency in the knowledge that the etch process is time varying. The windowed and cluster models are similar in that they both provide local focus for etch rate estimates, but differ in that windowed modelling schemes discard historical data after every window movement, while cluster models retain all historical data. The weighted-window modelling scheme described in Chapter 7 incorporates knowledge of the chamber maintenance history to increase etch rate estimation accuracy for PLS models. However, for the data set studied, windowed-GPR models are found to be the most accurate method for estimation of etch rate over many thousands of wafers, estimating the actual etch rate of 493 unseen test wafers with an $R^2$ of 0.75 and MAPE of 1.14 %.

Assuming that no other sensors are available, the stronger the relationship between the VM model input variables and the process variables of interest, the more likely the particular VM implementation will be successful. In Chapter 5, optical emission spectroscopy (OES) is used to perform VM for engine component temperatures and in Chapter 8, plasma impedance monitor (PIM) variables are used to drive VM models for plasma electron density. The choice of VM model input variables, in both cases, was driven by prior knowledge of the systems being modelled. For the plasma etch data set, ultimately the relationship between the VM model input variables and etch rate was found to be relatively weak. In the case of the results presented in this thesis, the use of VM models with the etch process (EP) data set alone represents the minimum of investment from the manufacturer; all of the measurements used are already logged by each tool, and no further sensors are required. However, while there were no consistent differences in model accuracy between those models using PIM data and etch process (EP) variables, it is important to note that the models using the PIM sensor were the only models to follow some particular etch rate variations caused by PM operations during global modelling.

Although the data set used during etch rate modelling was restrictive in terms of

quantity and frequency of samples, based on the results of Chapters 6 and 7, some conjectures can be made on the potential effects of obtaining more data for VM model creation.

- Global model accuracy is likely to increase somewhat as more of the operating space is captured by larger training sets. However, global models tend to become overly general with large amounts of data, and it is likely that very large global models will lose local accuracy as the number of training samples increases greatly.
- Cluster models are expected to become more accurate as knowledge of new operating spaces allows new cluster models to be created and individual clusters become better defined. It is conjectured that a finite number of clusters will exist as the amount of data collected increases, as depicted in Figure 9.1. Hence, with a finite number of cluster models, local model accuracy can be achieved across the whole operating space.
- Windowed models, for a given window size, are unaffected by the number of training samples available in total, since old information, beyond the window size, is disregarded every time the window moves forward.
- ANN-based model accuracy is expected to increase as more samples are available to learn functional relationships, overcoming the difficulties of limited data sets. While this accuracy increase applies to global models, there is still a danger of models becoming overly general with additional data. ANN-based cluster models should also become more accurate, but there is the possibility of some clusters remaining sparsely populated and hindering ANN model accuracy.
- GPR models will perform equally well with large data sets as small data sets. However, the complexities of estimation calculations will increase since each estimate requires the inversion of a covariance matrix, the size of which is dictated by the number of training samples available. The operation of the GPR models will become a problem for very large data sets ($>\sim$ 10000 training points). In GPR modelling, essentially the training data forms the model, and so training data spread across larger regions of the operating space will aid estimation in these areas.

It is worthy of note that more modern etch processes tend to have higher measurement frequencies, and it can be reasonable assumed that the VM results will be more accurate for such processes (for example see [222]), since more measurements from each operating space are recorded before disturbances such as PM events occur.

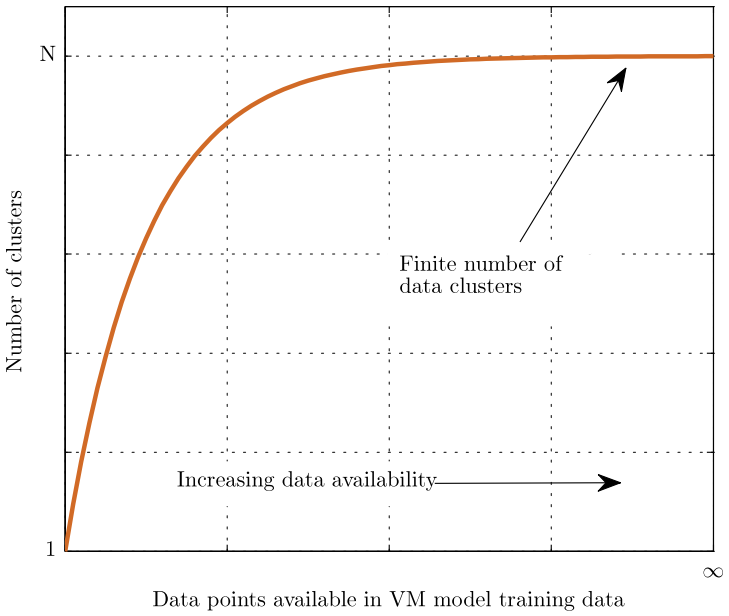



FIGURE 9.1: Conjectured limit of clusters existing in data.

In Chapter 8, a real-time VM and control system is developed for control of electron density in a production chamber. Over the operating ranges investigated, the response of the plasma electron density with respect to power is shown to be relatively linear, and closed-loop model predictive control with response times of less than one second, no overshoot, and good disturbance rejection properties is achieved using first-order predictive functional control (PFC). PIM variables are demonstrated to be suitable VM model input variables for estimation of electron density. Changes in the impedance of the ground path are used to simulate variations that are conjectured to be similar to PM event disturbances, and the VM system is designed to operate in the presence of such variations. The ground path impedance disturbances drive electron density variations and the control scheme is demonstrated to be capable of negating this effect. The implementation of a model update scheme for the PFC internal model allows the control scheme to operate effectively with changes in chamber pressure, and potentially, other variables that impact the relationship between the PIM variables and the plasma electron density.

The success of the non-invasive VM and real-time control scheme for electron density suggests that similar systems can be constructed for VM and control of other plasma variables. Example of such variables include ion densities or species concentrations, provided that measurements of such variables along with measurements of related ancillary variables can be attained for the purpose of VM model building. Such research paves the way for development of a multi-variable real-time control system such that process recipes can be specified in terms of plasma variables rather than process chamber input set points, without the requirement for expensive chamber alterations. From an

industrial perspective, non-invasive real-time control systems that can negate the effect of PM disturbances on plasma variables would be beneficial in terms of tool-matching applications and to achieve tightly controlled etch output variables.

As an overall concluding comment, although optimal VM implementation will never be achievable for plasma etch processes through the use of a standardised procedure, through an understanding of the limitations of the techniques employed along with the peculiarities of the etch process, the accuracy floor for the data available can be attained. Whether this accuracy is sufficient depends on the application. It is the opinion of the author that VM, given adequate investment of effort, represents an affordable technology for semiconductor manufacturers to enable wafer-to-wafer or real-time APC for many etch processes in the near future, potentially increasing fab-wide yields and saving millions of euro worldwide.

## 9.2 Future work

The research contained within this thesis highlights several areas for further research. This section examines some of the research topics for future consideration.

### 9.2.1 VASIMR state estimation

The temperature estimation system developed for the VASIMR engine in Chapter 5 can be expanded to operate in multiple helicon modes and across a larger operating regime than investigated through the use of local linear models and a mode detection algorithm. A potential improvement to the state estimation scheme would be the use of a Kalman filter in place of the the Luenberger observer used in this thesis. The Kalman filter algorithm may provide more accurate estimates of the engine temperatures because it takes the measurement and process noise variances into account during the determination of the estimator gain. However, estimates of the process and measurement noise for the VASIMR engine system are challenging to attain.

Ultimately, given access to the prototype system, the aim of the VM scheme is to provide feedback for an active cooling system to regulate the VASIMR engine temperature in final flight-ready designs. Flight-ready prototypes will be capable of running at 200 kW power, split between the helicon and ion-cyclotron resonance antennae. At such high powers, the engine heating will be more intensive than that seen in the experiments in Chapter 5, and active cooling systems may be required to maintain engine

temperatures within safe limits. Feedback mechanisms to implement closed loop control of the engine temperatures could use virtual measurements similar to the system developed in Chapter 5.

### 9.2.2 Plasma etch virtual metrology

The accuracy of the VM models examined in this thesis is limited by the available samples for modelling, and the relevance of the measurements taken from the process. Some further research should be completed to quantify the measurement frequency required to achieve the VM accuracy floor (the best accuracy possible), and to investigate the sensitivity of the VM estimation accuracy to the level of metrology available. For the data set explored in Chapters 6 and 7, etch rate metrology is available for approximately 4 % of wafers. An interesting expansion of the current results would be to examine the effect that reducing the number of metrology points would have on VM estimation accuracy. The frequency of metrology could be optimised to achieve the best VM estimation accuracy while minimising the risks of wafer scrap and maximising factory throughput. A hypothetical depiction of this effect is depicted in Figure 9.2 where the slope of the VM accuracy floor can be used as a metric for the sensitivity of the VM error to the metrology frequency. Such work was not completed in this thesis due to the restrictions of the data set available, a highly sampled data set would be required to comprehensively examine subsampling effects.

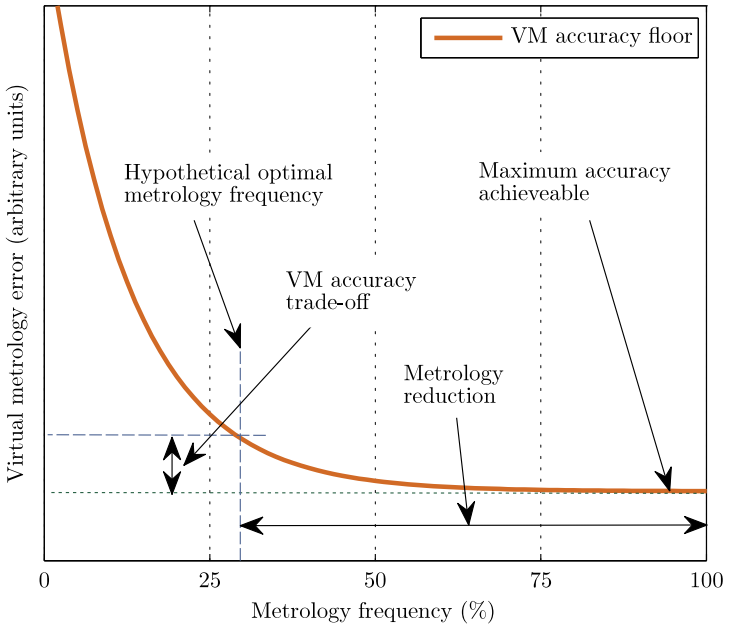



FIGURE 9.2: Hypothetical VM accuracy floor in relation to metrology frequency. An optimal measurement frequency can hypothetically be found that maximises factory throughput and VM accuracy while minimising metrology expenses and risk of wafer scrap.

Further research can be completed to compare further VM model input variables by collecting a data set containing concurrent measurements of EP, PIM, and OES data for the etch process. For a comprehensive investigation, data is needed from similar or greater amounts of wafers and PM cycles than that available in the data set used in this thesis. More modern processes compared to the one studied in this thesis are typically more frequently measured for etch rate, and hence, data sets from such processes are likely to be more suitable for VM investigations, assuming that ancillary variables from PIM and OES sensors can be obtained.

The modal behaviour of the plasma etch system is revealed during the cluster modelling results of Chapter 7. This modal behaviour has been observed in other research on etch processes [58, 229]. Further work is required in this area using larger data sets to investigate the potential applications and advantages of using specialised clustered models for etch rate VM. While it is conjectured in this thesis that a finite number of clustered operating points will appear as extensive process data is collected, this fact is yet to be confirmed through experiment.

The maintenance dependent weighted-window scheme developed in Chapter 7 improved the accuracy of the windowed partial least squares (PLS) models. Further research can be completed in the extension of the weighted scheme to the windowed GPR models examined. Such weighting could be implemented by adapting the marginal log-likelihood equation that is minimised during the optimisation of the covariance function hyperparameters. The implementation of weighted GPR models and the application of the weighted and non-weighted windowed modelling schemes to different industrial plasma etch data sets are areas of current research.

### 9.2.3 Real-time control of electron density

The real-time VM and electron density control scheme implemented in Chapter 8 yielded promising results that has spurred on further research into the area of real-time process control for plasma etch.

The next step in this research is the expansion of the VM models to allow for electron density estimation while wafers are being etched. The relationship between plasma electron density and polysilicon etch rate in $SF_6$ plasmas is currently being examined and characterised by a research group which includes the author, working in close collaboration with a semiconductor manufacturing company. The control system is being tested to examine whether regulation of the electron density stabilises the polysilicon etch rate in the presence of ground impedance disturbances. Ultimately, the relationship

between the plasma variables and the etch rate is to be characterised so that recipes can be specified in terms of etch variables. However, such abstraction requires extensive experimentation and analysis of the different interacting processes occurring during the etch of each wafer.

Further improvements to the control architecture include the inclusion of pressure as an explicit VM model input variable to allow more reliable estimation of the electron density variations over a range of chamber pressures (a pressure measurement was not available at the time of experimentation in this thesis). A Kalman filter could also be implemented to improve the current VM estimates across multiple chamber pressures, as the current VM estimates are relatively noisy. Spectral data should also be examined as potential VM model input variables to increase accuracy.

Other future work lies in the improvement of the adaptive internal model for the PFC controller. The PFC control performance can be unpredictable, in terms of settling time and overshoot, directly after changes in operating set point while the RLS algorithm settles on new values for the PFC internal model. To ensure predictable control performance and prevent potentially dangerous system operation, the internal model parameters could be scheduled from a lookup table, or else kept constant, in the interim period while the RLS algorithm converges. After convergence of the RLS parameter estimates, the RLS estimates can be used for the control, and the lookup table updated, if required, to reflect the newly calculated model parameter values.

Taking the results from Chapters 6 and 7 into account, it is likely that the VM models used for estimation of electron density will require periodic refreshing to maintain their currency. Techniques such as the clustered modelling may be employed if the system is found to be modal and the use of a GPR model for VM of electron density may be a useful addition because the confidence intervals on the VM estimates could be used as a metric to weight the control actions taken. Model refreshing can be achieved by collecting information using a retractable microwave hairpin probe during the etch of patterned test wafers that are routinely used to qualify chambers after PM operations. Some research is required to examine the possibilities of rapid characterisation of the chamber operating space during the etch of one patterned wafer to reduce the cost of model refreshing. Although the installation of a retractable probe on production chambers may incur substantial costs originally, the potential future gains may far outweigh this cost.

# Appendix A

# EP Data Variables

This appendix details the variables that are included in the EP data set for each of the five etch process steps.

| Step 1 | |
|---|---|
| **Variable Name** | **Description** |
| MEAN-POWER | Power delivered to chamber * |
| STD-DEV-TV_ANGLE | Std. dev. Throttle valve angle |
| MEAN-PHASE | Phase between voltage and current * |
| MEAN-IMPEDENCE | Plasma Impedance (V/I) * |
| MEAN-VOLTAGE | Voltage applied (V) * |
| MEAN-ENDPOINT_A | Monochromator signal A |
| MEAN-GAS_FLOW5 | MFC5 flow |
| STD-DEV-GAS_FLOW1 | Std. dev. MFC1 flow |
| STD-DEV-GAS_FLOW5 | Std. dev. MFC5 flow |
| STD-DEV-RF_FORWARD_GEN | Std. dev. RF power generated |
| STD-DEV-RF_REFLECTED | Std. dev. reflected RF power |
| STD-DEV-CHAMBER_PRESS | Std. dev. chamber pressure |
| MEAN-GAS_FLOW1 | Mean MFC1 flow |
| MEAN-RF_FORWARD_GEN | Mean RF power generated |
| MEAN-UP_ELECT_TEMP | Mean upper electrode temperature |
| MEAN-TV_ANGLE | Throttle valve angle |
| MEAN-RF_MATCH_1_TUNE | Matchbox tune inductor |
| MEAN-RF_MATCH_1_DC_BIAS | Induced DC bias on wafer |
| MEAN-RF_LOAD_MATCH_PH | Phase between pre and post match RF waveforms |
| MEAN-RF_LOAD_COIL_POS | Matchbox load coil position |
| MEAN-RF_LINE_IMP | RF line impedance |
| MEAN-LOW_ELECT_TEMP | Lower electrode temperature |
| MEAN-GAP | Gap distance between electrodes |
| MEAN-CHAMBER_PRESS | Chamber pressure |
| MEAN-RF_REFLECTED | RF power reflected from chamber |

Table A.1: Etch process variables recorded during Step 1 of the trench etch process. Mean value of time series variables are taken over the duration of the etch step. Standard deviation values are also recorded where noted. * denotes measurements derived from PIM sensor.

| Step 2 | |
|---|---|
| **Variable Name** | **Description** |
| MEAN-POWER | Power delivered to chamber * |
| STD-DEV-TV_ANGLE | Std. dev. Throttle valve angle |
| MEAN-CHAMBER_PRESS | Chamber pressure |
| MEAN-GAP | Gap distance between electrodes (8cm) |
| MEAN-LOW_ELECT_TEMP | Lower electrode temperature |
| MEAN-RF_LINE_IMP | RF line impedance |
| MEAN-RF_LOAD_COIL_POS | Matchbox load coil position |
| MEAN-RF_LOAD_MATCH_PH | Phase between pre and post match RF waveforms |
| MEAN-RF_MATCH_1_TUNE | Matchbox tune inductor |
| MEAN-RF_REFLECTED | RF power reflected from chamber |
| MEAN-UP_ELECT_TEMP | Mean upper electrode temperature |
| MEAN-TV_ANGLE | Throttle valve angle |
| MEAN-GAS_FLOW4 | MFC4 flow |
| MEAN-GAS_FLOW6 | MFC6 flow |
| MEAN-GAS_FLOW1 | MFC1 flow |
| MEAN-RF_FORWARD_GEN | Mean RF power generated |
| MEAN-RF_MATCH_1_DC_BIAS | Induced DC bias on wafer |
| STD-DEV-GAS_FLOW4 | Std. dev. MFC4 flow |
| STD-DEV-GAS_FLOW6 | Std. dev. MFC6 flow |
| STD-DEV-GAS_FLOW1 | Std. dev. MFC1 flow |
| STD-DEV-CHAMBER_PRESS | Std. dev. chamber pressure |
| STD-DEV-RF_FORWARD_GEN | Std. dev. RF power generated |
| STD-DEV-RF_REFLECTED | Std. dev. reflected RF power |
| MEAN-VOLTAGE | Voltage applied (V) * |
| MEAN-PHASE | Phase between voltage and current * |
| MEAN-IMPEDANCE | Plasma Impedance (V/I) * |

Table A.2: Etch process variables recorded during Step 2 of the trench etch process. Mean values of time series variables are taken over the duration of the etch step. Standard deviation values are also recorded where noted. This step is controlled using an endpoint signal and the length of the step is recorded in the step 3 parameters. * denotes measurements derived from PIM sensor.

| Step 3 | |
|---|---|
| **Variable Name** | **Description** |
| MEAN-ENDPT_STEP_TIME | Endpoint time of Step 2 |
| MEAN-POWER | Power delivered to chamber * |
| STD-DEV-TV_ANGLE | Std. dev. Throttle valve angle |
| MEAN-CHAMBER_PRESS | Chamber pressure |
| MEAN-GAP | Gap distance between electrodes |
| MEAN-LOW_ELECT_TEMP | Lower electrode temperature |
| MEAN-RF_LINE_IMP | RF line impedance |
| MEAN-RF_LOAD_COIL_POS | Matchbox load coil position |
| MEAN-RF_LOAD_MATCH_PH | Phase between pre and post match RF waveforms |
| MEAN-RF_MATCH_1_TUNE | Matchbox tune inductor |
| MEAN-RF_REFLECTED | RF power reflected from chamber |
| MEAN-UP_ELECT_TEMP | Mean upper electrode temperature |
| MEAN-TV_ANGLE | Throttle valve angle |
| MEAN-GAS_FLOW4 | MFC4 flow |
| MEAN-GAS_FLOW6 | MFC6 flow |
| MEAN-GAS_FLOW1 | MFC1 flow |
| MEAN-RF_FORWARD_GEN | Mean RF power generated |
| MEAN-RF_MATCH_1_DC_BIAS | Induced DC bias on wafer |
| STD-DEV-GAS_FLOW4 | Std. dev. MFC4 flow |
| STD-DEV-GAS_FLOW6 | Std. dev. MFC6 flow |
| STD-DEV-GAS_FLOW1 | Std. dev. MFC1 flow |
| STD-DEV-CHAMBER_PRESS | Std. dev. chamber pressure |
| STD-DEV-RF_FORWARD_GEN | Std. dev. RF power generated |
| STD-DEV-RF_REFLECTED | Std. dev. reflected RF power |
| MEAN-VOLTAGE | Voltage applied (V) * |
| MEAN-PHASE | Phase between voltage and current * |
| MEAN-IMPEDANCE | Plasma Impedance (V/I) * |

Table A.3: Etch process variables recorded during Step 3 of the trench etch process. Standard deviation values are also recorded where noted. * denotes measurements derived from PIM sensor.

| Step 4 | |
|---|---|
| **Variable Name** | **Description** |
| MEAN-POWER | Power delivered to chamber * |
| RANGE-ENDPOINT_A | Range of monochromator signal |
| STD-DEV-TV_ANGLE | Std. dev. Throttle valve angle |
| MEAN-PHASE | Phase between voltage and current * |
| MEAN-IMPEDANCE | Plasma Impedance (V/I) * |
| MEAN-VOLTAGE | Voltage applied (V) * |
| MEAN-ENDPOINT_A | Monochromator signal A |
| MEAN-GAS_FLOW2 | MFC2 flow |
| MEAN-GAS_FLOW7 | MFC7 flow |
| MEAN-GAS_FLOW6 | MFC6 flow |
| MEAN-RF_FORWARD_GEN | Mean RF power generated |
| MEAN-RF_MATCH_1_DC_BIAS | Induced DC bias on wafer |
| STD-DEV-GAS_FLOW2 | Std. dev. MFC2 flow |
| STD-DEV-GAS_FLOW7 | Std. dev. MFC7 flow |
| STD-DEV-CHAMBER_PRESS | Std. dev. Chamber pressure |
| STD-DEV-RF_FORWARD_GEN | Std. dev. RF power generated |
| STD-DEV-RF_REFLECTED | Std. dev. reflected RF power |
| STD-DEV-GAS_FLOW6 | Std. dev. MFC6 flow |
| MEAN-CHAMBER_PRESS | Chamber pressure |
| MEAN-GAP | Gap distance between electrodes |
| MEAN-LOW_ELECT_TEMP | Lower electrode temperature |
| MEAN-RF_LINE_IMP | RF line impedance |
| MEAN-RF_LOAD_COIL_POS | Matchbox load coil position |
| MEAN-RF_LOAD_MATCH_PH | Phase between pre and post match RF waveforms |
| MEAN-RF_MATCH_1_TUNE | Matchbox tune inductor |
| MEAN-RF_REFLECTED | RF power reflected from chamber |
| MEAN-TV_ANGLE | Throttle valve angle |
| MEAN-UP_ELECT_TEMP | Upper electrode temperature |

Table A.4: Etch process variables recorded during Step 4 of the trench etch process. Step 4 is the main trench etch step of the process. This step has the greatest influence on the overall trench depth. * denotes measurements derived from PIM sensor.

| Step 5 | |
|---|---|
| **Variable Name** | **Description** |
| MEAN-POWER | Power delivered to chamber * |
| RANGE-ENDPOINT_A | Range of monochromator signal |
| MEAN-CHAMBER_PRESS | Chamber pressure |
| MEAN-GAP | Gap distance between electrodes (8cm) |
| MEAN-LOW_ELECT_TEMP | Lower electrode temperature |
| MEAN-UP_ELECT_TEMP | Upper electrode temperature |
| MEAN-TV_ANGLE | Throttle valve angle |
| MEAN-RF_REFLECTED | RF power reflected from chamber |
| MEAN-RF_MATCH_1_TUNE | Matchbox tune inductor |
| MEAN-RF_MATCH_1_DC_BIAS | Induced DC bias on wafer |
| MEAN-RF_LOAD_MATCH_PH | Phase between pre and post match RF waveforms |
| MEAN-RF_LOAD_COIL_POS | Matchbox load coil position |
| MEAN-RF_LINE_IMP | RF line impedance |
| STD-DEV-GAS_FLOW1 | Std. dev. MFC1 flow |
| STD-DEV-CHAMBER_PRESS | Std. dev. Chamber pressure |
| STD-DEV-RF_FORWARD_GEN | Std. dev. RF power generated |
| STD-DEV-RF_REFLECTED | Std. dev. Of reflected RF power |
| STD-DEV-GAS_FLOW7 | Std. dev. MFC7 flow |
| MEAN-GAS_FLOW1 | MFC1 flow |
| MEAN-GAS_FLOW7 | MFC7 flow |
| MEAN-RF_FORWARD_GEN | Mean RF power generated |
| MEAN-VOLTAGE | Voltage applied (V) * |
| MEAN-PHASE | Phase between voltage and current * |
| MEAN-IMPEDANCE | Plasma Impedance (V/I) * |
| STD-DEV-TV_ANGLE | Std. dev. Throttle valve angle |

Table A.5: Etch process variables recorded during Step 5 of the trench etch process. Step 5 is the final step of the process where the bottom of the etched trenches are given the desired profiles. * denotes measurements derived from PIM sensor.

# Appendix B

# Variables removed from EP data

This appendix details the variables that are removed from the EP data set and not used during the development of the virtual metrology models for etch rate.

| Removed Variables | |
|---|---|
| **Step 3** | 4-STD-DEV-RF_FORWARD_GEN |
| 3-STD-DEV-TV_ANGLE | 4-STD-DEV-RF_REFLECTED |
| 3-MEAN-GAP | 4-STD-DEV-HBr_FLOW6 |
| 3-MEAN-RF_REFLECTED | 4-MEAN-CHAMBER_PRESS |
| 3-MEAN-TV_ANGLE | 4-MEAN-GAP |
| 3-MEAN-C2F6_FLOW4 | 4-MEAN-RF_REFLECTED |
| 3-MEAN-HBr_FLOW6 | 4-MEAN-TV_ANGLE |
| 3-MEAN-He_FLOW1 | **Step 5** |
| 3-STD-DEV-C2F6_FLOW4 | 5-RANGE-ENDPOINT_A |
| 3-STD-DEV-HBr_FLOW6 | 5-MEAN-CHAMBER_PRESS |
| 3-STD-DEV-He_FLOW1 | 5-MEAN-GAP |
| 3-STD-DEV-CHAMBER_PRESS | 5-MEAN-TV_ANGLE |
| 3-STD-DEV-RF_FORWARD_GEN | 5-MEAN-RF_REFLECTED |
| 3-STD-DEV-RF_REFLECTED | 5-STD-DEV-He_FLOW1 |
| **Step 4** | 5-STD-DEV-CHAMBER_PRESS |
| 4-RANGE-ENDPOINT_A | 5-STD-DEV-RF_FORWARD_GEN |
| 4-STD-DEV-TV_ANGLE | 5-STD-DEV-RF_REFLECTED |
| 4-MEAN-Ar_FLOW2 | 5-STD-DEV-Cl2_FLOW7 |
| 4-MEAN-Cl2_FLOW7 | 5-MEAN-He_FLOW1 |
| 4-MEAN-HBr_FLOW6 | 5-MEAN-Cl2_FLOW7 |
| 4-STD-DEV-Ar_FLOW2 | 5-STD-DEV-TV_ANGLE |
| 4-STD-DEV-Cl2_FLOW7 | |
| 4-STD-DEV-CHAMBER_PRESS | |

Table B.1: Variables removed from analysis due to low variance or on suspect of having no contribution to output modelling effort.

# Appendix C

# Derivation of recursive least squares

### RLS Algorithm details

Recall that the solution for equations of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ are given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a matrix of $n$ samples of $p$ variables, $\mathbf{y} \in \mathbb{R}^{n\times 1}$ is a vector of process outputs, and $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ is a vector of model coefficients. Define $\mathbf{X}(k)$ and $\mathbf{y}(k)$ to be

$$\mathbf{X}(k) = \begin{bmatrix} \vec{\mathbf{x}}_1 \\ \vec{\mathbf{x}}_2 \\ \vdots \\ \vec{\mathbf{x}}_k \end{bmatrix} \qquad \mathbf{y}(k) = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} \tag{C.1}$$

where $\vec{\mathbf{x}}_i$ represents the $i^{th}$ *row* of matrix $\mathbf{X}$, and in this case that corresponds to the row vector of variable measurements for sample $i$. $y_i$ represents the process output measurement for sample $i$. Hence $\mathbf{X}(k)$ and $\mathbf{y}(k)$ represent all information collected from samples 1 to $k$. The estimate for the process parameters at sample $k$ is given by

$$\hat{\boldsymbol{\beta}}(k) = \left[\mathbf{X}(k)^T\mathbf{X}(k)\right]^{-1}\mathbf{X}(k)^T\mathbf{y}(k). \tag{C.2}$$

At time $k+1$ further information from the process is gathered, and the matrices are increased in size to include the new information such that

$$\mathbf{X}(k+1) = \begin{bmatrix} \vec{\mathbf{x}}_1 \\ \vec{\mathbf{x}}_2 \\ \vdots \\ \vec{\mathbf{x}}_{k+1} \end{bmatrix} \qquad \mathbf{y}(k+1) = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{k+1} \end{bmatrix} \tag{C.3}$$

and

$$\hat{\boldsymbol{\beta}}(k+1) = \left[\mathbf{X}(k+1)^T\mathbf{X}(k+1)\right]^{-1}\mathbf{X}(k+1)^T\mathbf{y}(k+1). \tag{C.4}$$

We can express the terms on the right side of Equation (C.4) as

$$\mathbf{X}(k+1)^T\mathbf{X}(k+1) = \left[\mathbf{X}(k)^T\vec{\mathbf{x}}_{k+1}^T\right]\begin{bmatrix} \mathbf{X}(k) \\ \vec{\mathbf{x}}_{k+1} \end{bmatrix} = \mathbf{X}(k)^T\mathbf{X}(k) + \vec{\mathbf{x}}_{k+1}^T\vec{\mathbf{x}}_{k+1} \tag{C.5}$$

$$\mathbf{X}(k+1)^T\mathbf{y}(k+1) = \left[\mathbf{X}(k)^T\vec{\mathbf{x}}_{k+1}^T\right]\begin{bmatrix} \mathbf{y}(k) \\ y_{k+1} \end{bmatrix} = \mathbf{X}(k)^T\mathbf{y}(k) + \vec{\mathbf{x}}_{k+1}^T y_{k+1}^T. \tag{C.6}$$

Equations (C.5) and (C.6) allow Equation (C.4) to be updated at every sample. However, a method to directly update the inverse of Equation (C.5) is required. Define

$$\mathbf{P}(k) = [\mathbf{X}(k)^T\mathbf{X}(k)]^{-1} \tag{C.7}$$

$$\mathbf{B}(k) = \mathbf{X}(k)^T\mathbf{y}(k), \tag{C.8}$$

where $\mathbf{P}(k) \in \mathbb{R}^{p\times p}$, and $\mathbf{B}(k) \in \mathbb{R}^{p\times 1}$, such that

$$\hat{\boldsymbol{\beta}}(k) = \mathbf{P}(k)\mathbf{B}(k), \text{ and} \tag{C.9}$$

$$\hat{\boldsymbol{\beta}}(k+1) = \mathbf{P}(k+1)\mathbf{B}(k+1). \tag{C.10}$$

Substituting the definitions for $\mathbf{P}(k)$ and $\mathbf{B}(k)$ into Equations (C.5) and (C.6) yields

$$\mathbf{P}(k+1) = \mathbf{P}(k)^{-1} + \vec{\mathbf{x}}_{k+1}^T\vec{\mathbf{x}}_{k+1} \tag{C.11}$$

$$\mathbf{B}(k+1) = \mathbf{B}(k) + \vec{\mathbf{x}}_{k+1}^T y_{k+1}^T. \tag{C.12}$$

The *matrix inversion lemma* is used to find a direct update from $\mathbf{P}(k)$ to $\mathbf{P}(k+1)$. The matrix inversion lemma states that

$$(\mathbf{A}+\mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}. \tag{C.13}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C}$, and $\mathbf{D}$ are of the correct sizes (for example $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times k}$, $\mathbf{C} \in \mathbb{R}^{k\times k}$, and $\mathbf{D} \in \mathbb{R}^{k\times n}$). If we set $\mathbf{A} = \mathbf{P}(k)^{-1}$, $\mathbf{B} = \vec{\mathbf{x}}_{k+1}^T$, $\mathbf{C} = 1$, and $\mathbf{D} = \vec{\mathbf{x}}_{k+1}$, then a direct update for $\mathbf{P}(k+1)$ is obtained:

$$\mathbf{P}(k+1) = \mathbf{P}(k)\left[I - \frac{\vec{\mathbf{x}}_{k+1}^T\vec{\mathbf{x}}_{k+1}\mathbf{P}(k)}{1+\vec{\mathbf{x}}_{k+1}\mathbf{P}(k)\vec{\mathbf{x}}_{k+1}^T}\right]. \tag{C.14}$$

The only inversion in Equation (C.14) is that of a scalar value. Now, define the error variable $e(k)$ such that

$$e(k+1) = y_{k+1} - \vec{\mathbf{x}}_{k+1}\hat{\boldsymbol{\beta}}(k). \tag{C.15}$$

Using the expression for $y_{k+1}$ from Equation (C.15) in Equation (C.12) yields

$$\mathbf{B}(k+1) = \mathbf{B}(k) + \vec{\mathbf{x}}_{k+1}^T\vec{\mathbf{x}}_{k+1}\hat{\boldsymbol{\beta}}(k) + \vec{\mathbf{x}}_{k+1}^T e(k+1). \tag{C.16}$$

From Equations (C.9) and (C.10), $\mathbf{B}(k)$ and $\mathbf{B}(k+1)$ can be expressed as

$$\mathbf{B}(k) = \mathbf{P}(k)^{-1}\hat{\boldsymbol{\beta}}(k) \tag{C.17}$$

$$\mathbf{B}(k+1) = \mathbf{P}(k+1)^{-1}\hat{\boldsymbol{\beta}}(k+1). \tag{C.18}$$

Substituting the expressions from Equations (C.17) and (C.18) into Equation (C.16) yields the update for $\hat{\boldsymbol{\beta}}(k+1)$:

$$\mathbf{P}(k+1)^{-1}\hat{\boldsymbol{\beta}}(k+1) = \mathbf{P}(k)^{-1}\hat{\boldsymbol{\beta}}(k) + \vec{\mathbf{x}}_{k+1}^T\vec{\mathbf{x}}_{k+1}\hat{\boldsymbol{\beta}}(k) + \vec{\mathbf{x}}_{k+1}^T e(k+1) \tag{C.19}$$

$$\mathbf{P}(k+1)^{-1}\hat{\boldsymbol{\beta}}(k+1) = \mathbf{P}(k+1)^{-1}\hat{\boldsymbol{\beta}}(k) + \vec{\mathbf{x}}_{k+1}^T e(k+1) \tag{C.20}$$

$$\hat{\boldsymbol{\beta}}(k+1) = \hat{\boldsymbol{\beta}}(k) + \mathbf{P}(k+1)\vec{\mathbf{x}}_{k+1}^T e(k+1). \tag{C.21}$$

Hence, the full RLS algorithm can be implemented in five steps at each sample point. These are, at sample $k+1$,

1. Form $\vec{\mathbf{x}}_{k+1}$ from new data collected during the sample.
2. Calculate the current error using Equation (C.15).
3. Calculate the covariance matrix $\mathbf{P}(k+1)$ using Equation (C.14).
4. Update the model parameters, $\hat{\boldsymbol{\beta}}(k+1)$ using Equation (C.21).
5. Return to step 1.

A common adjustment to the RLS algorithm is the addition of a forgetting factor $0 \leq \lambda_{RLS} \leq 1$ which prevents the elements of $\mathbf{P}$ becoming too small and thus improves sensitivity of the algorithm. $\lambda_{RLS}$ adjusts Equation C.14 such that

$$\mathbf{P}(k+1) = \mathbf{P}(k)\left[I - \frac{\vec{\mathbf{x}}_{k+1}^T \vec{\mathbf{x}}_{k+1} \mathbf{P}(k)}{\lambda_{RLS} + \vec{\mathbf{x}}_{k+1}\mathbf{P}(k)\vec{\mathbf{x}}_{k+1}^T}\right]. \tag{C.22}$$

The above derivation, along with further discussion on self-tuning and adaptive systems, can be found in the work by Wellstead [298].

# Bibliography


[1] D. A. Gurnett and A. Bahattacharjee. *Introduction to plasma physics: with space and laboratory applications.* Cambridge University Press, Cambridge, UK, 2005.

[2] William Crooks. Radiant matter, a resum of the principal lectures and papers. Technical report, Royal Society of London and British Association for the Advancement of Science, James W. Queen & Co., 1879.

[3] Ad astra website, available online at http://www.adastrarocket.com.

[4] G. E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8):114–117, 1965.

[5] International technology roadmap for semiconductors. Executive summary 2009. Technical report, ITRS, 2009.

[6] T.F. Edgar, S.W. Butler, W.J. Campbell, C. Pfeiffer, C. Bode, S.B. Hwang, K.S. Balakrishnan, and J. Hahn. Automatic control in microelectronics manufacturing: Practices, challenges, and possibilities. *Automatica*, 36(11):1567–1603, Nov. 2000.

[7] S. Ruegsegger, A. Wagner, J.S. Freudenberg, and D.S. Grimard. Feedforward control for reduced run-to-run variation in microelectronics manufacturing. *IEEE T. Semiconduct. M.*, 12(4):493 –502, Nov. 1999.

[8] J. V. Ringwood, S. Lynn, G. Bacelli, B. Ma, E. Ragnoli, and S. McLoone. Estimation and control in semiconductor etch: Practice and possibilities. *IEEE T. Semiconduct. M.*, 23(1):87–98, Feb 2010.

[9] C.J. Spanos. Statistical process control in semiconductor manufacturing. *Proceedings of the IEEE*, 80(6):819–830, Jun. 1992.

[10] M. Sarfaty, S. Arulkumar, A. Schwarm, J. Paik, Z. Jimin, P. Rong, M.J. Seamons, H. LI, R. Hung, and S. Parikh. Advance process control solutions for semiconductor manufacturing. In *Advanced Semiconductor Manufacturing Conference*, pages 101–106, 2002.

[11] A. A. Khan, J.R. Moyne, and D.M. Tilbury. Virtual metrology and feedback control for semiconductor manufacturing processes using recursive partial least squares. *J. Process Control*, 18(10):961–974, 2008.

[12] K. Lensing and B. Stirton. Integrated metrology and wafer-level control. *Semiconductor International*, 29(6):44–54, June 2006.

[13] S. Joe Qin, G. Cherry, R. Good, J. Wang, and C.A. Harrison. Semiconductor manufacturing process control and monitoring: A fab-wide framework. *J. Process Control*, 16(3):179–191, Mar. 2006.

[14] Y-J. Chang, Y. Kang, C-L. Hsu, C-T. Chang, and T. Y. Chan. Virtual metrology technique for semiconductor manufacturing. In *International Joint Conference on Neural Networks*, pages 5289–5293, 2006.

[15] Yang Yang, Mingmei Wang, and Mark J. Kushner. Progress, opportunities and challenges in modeling of plasma etching. In *International Interconnect Technology Conference (IITC)*, pages 90–92, June 2008.

[16] A. Khan, D. Tilbury, and J. Moyne. Fab-wide virtual metrology and feedback control. In *Asian AEC/APC Symposium*. NSF Engineering Research Center for reconfigurable Manufacturing Systems, University of Michigan., 2006.

[17] A.A. Khan, J.R. Moyne, and D.M. Tilbury. An approach for factory-wide control utilizing virtual metrology. *IEEE T. Semiconduct. M.*, 20(4):364–375, Nov. 2007.

[18] M. A. Lieberman and A. J. Lichtenberg. *Priciples of Plasma Discharges and Materials Processing*. John Riley & Sons, Inc., Hoboken, NJ, USA, 2nd edition, 2005.

[19] J.J. Thompson. Cathode rays. *Philosophical Magazine*, 44(5):293–316, Oct. 1897.

[20] I. Langmuir. Oscillations in ionized gases. *Proceedings of the National Academy of Sciences of the United States of America*, 14(8):627–637, Aug. 1928.

[21] B. Chapman. *Glow Discharge Processes*. John Wiley & Sons Inc., 1980.

[22] F. F. Chen and J. P. Chang. *Lecture Notes on Principles of Plasma Processing*. Kluwer Academic/Plenum Publishers, 2003.

[23] N. Bohr. On the constitution of atoms and molecules part i. *Philosophical Magazine*, 26(151):1–26, Jul. 1913.

[24] M. Sugawara. *Plasma Etching, Fundamentals and Applications*. Oxford Science Publications, 1998.

[25] S. Lynn. An introduction to plasma and plasma etching. Technical Report EE/JVR/2/2007, Electronic Engineering Dept., National University of Ireland, Maynooth, Apr. 2007.

[26] J. W. Coburn and H. F. Winters. Ion and electron assisted gas-surface chemistry–an important effect in plasma etching. *J. Appl. Phys.*, 50(5):3189–3196, May. 1979.

[27] C. Y. Chang and S. M. Sze. *ULSI Technology*. McGraw-Hill International Editions, Singapore, 1st edition, 1996.

[28] G. S. May and C. J. Spanos. *Fundamentals of semiconductor manufacturing and process control*. Wiley-Interscience, 2006.

[29] J. W. Coburn and M. Chen. Optical emission spectroscopy of reactive plasmas: A method for correlating emission intensities to reactive particle density. *J. Appl. Phys.*, 51(6):3134–3136, June 1980.

[30] R. L. Stenzel. Microwave resonator probe for localized density measurements in weakly magnetized plasmas. *Rev. Sci. Instrum.*, 47(5):603–607, May. 1976.

[31] R. B. Piejak, J. Al-Kuzee, and N. St. J. Braithwaite. Hairpin resonator probe measurements in RF plasmas. *Plasma Sources Science and Technology*, 14(4):734–743, Nov. 2005.

[32] R. B. Piejak, V. A. Godyak, R. Garner, B. M. Alexandrovich, and N. Sternberg. The hairpin resonator: A plasma density measuring technique revisited. *J. Appl. Phys.*, 95(7):3785–3791, Apr. 2004.

[33] M. B. Hopkins and J. F. Lawler. Plasma diagnostics in industry. *Plasma Physics and Controlled Fusion*, 42(12B):B189–B197, Jun 2000.

[34] B. M. Wise, N. B. Gallagher, D. D. Butler, S. W.and White, and G. G. Barna. A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process. *J. Chemom.*, 13(3-4):379–396, 1999.

[35] H-F. Guo, C.J. Spanos, and A.J. Miller. Real time statistical process control for plasma etching. In *International Semiconductor Manufacturing Science Symposium (ISMSS)*, pages 113–118, 1991.

[36] M. N. A. Dewan, P. J. McNally, T. Perova, and P. A. F. Herbert. Use of plasma impedance monitoring for determination of SF6 reactive ion etching end point of the SiO2/Si system. *Microelectronic Engineering*, 65(1-2):25–46, Jan. 2003.

[37] M. Kanoh, M.I. Yamage, and H. Takada. End-point detection of reactive ion etching by plasma impedance monitoring. *Japanese Journal of Applied Physics*, 40(3A):1457–1462, March 2001.

[38] A. F. Bose. *Diagnostics and control of Plasma Etching Reactors for Semiconductor Manufacturing*. PhD thesis, Physical Electronics Laboratory, ETH Zurich, 1995.

[39] C. Almgren. Rf measurements and their role in the manufacturing environment. Article, Advanced energy / Symbios Logic, 2000.

[40] M. A. Sobolewski. Real-time, noninvasive monitoring of ion energy and ion current at a wafer surface during plasma etching. *J. Vac. Sci. Technol. A.*, 24(5):1892–1905, Sept. 2006.

[41] F. Galton. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15:246–263, 1885.

[42] A. M. Legendre. Nouvelles méthodes pour la détermination des orbites des cométes. Technical report, appears as an appendix, 1805.

[43] C.F. Gauss. Theoria motus corporum coelestium in sectionibus conicis solem ambientum, 1809.

[44] D. C. Montgomery, G. C. Runger, and N.F. Hubele. *Engineering Statistics*. John Riley & Sons, Inc., 2001.

[45] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, Inc., 2001.

[46] John A. Rice. *Mathematical Statistics and Data Analysis*. Thomson Brooks Cole, third edition edition, 2007.

[47] N.R. Draper and H. Smith. *Applied Regression Analysis*. John Wiley & Sons, Inc, 1998.

[48] G. M. Furnival and R. W. M. Wilson. Regression by leaps and bounds. *Technometrics*, 16(4):499–511, 1974.

[49] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, Apr. 2004.

[50] F.V. Berghen. LARS library: Least angled regression stagewise library. Technical report, IRIDIA, Universit Libre de Bruxelles, November 2005.

[51] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[52] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(7):417–441, 498–520, 1933.

[53] I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics, Aberdeen, UK, 2nd edition, 2004.

[54] A. Afifi, V.A. Clarke, and S. May. *Computer-Aided Multivariate Analysis*. Chapman & Hall/CRC, 4th edition, 2004.

[55] J. E. Jackson. *A User's guide to principal components*. Wiley Interscience, 1991.

[56] H. Wold. *Estimation of principal components and related models by iterative least squares*. Academic Press, New York, USA, 1966.

[57] P. Geladi and B. R. Kowalski. Partial least-squares regression: A tutorial. *Anal. Chim. Acta*, 185:1–17, 1986.

[58] Q.P. He and J. Wang. Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes. *IEEE T. Semiconduct. M.*, 20(4):345 –354, Nov. 2007.

[59] Theodora Kourti. Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of Adaptive Control and Signal Processing*, 19(4):213–246, January 2005.

[60] H.H. Yue, S.J. Qin, R.J. Markle, C. Nauert, and M. Gatto. Fault detection of plasma etchers using optical emission spectra. *IEEE T. Semiconduct. M.*, 13(3):374–385, Aug 2000.

[61] G. Spitzlsperger, C. Schmidt, G. Ernst, H. Strasser, and M. Speil. Fault detection for a via etch process using adaptive multivariate methods. *IEEE T. Semiconduct. M.*, 18(4):528–533, Nov 2005.

[62] D.A. White, B.E. Goodlin, A.E. Gower, D.S. Boning, H. Chen, H.H. Sawin, and T.J. Dalton. Low open-area endpoint detection using a PCA-based $T^2$ statistic and Q statistic on optical emission spectroscopy measurements. *IEEE T. Semiconduct. M.*, 13(2):193–207, May 2000.

[63] N.B. Gallagher and B.M. Wise. Development and benchmarking of multivariate statistical process control tools for a semiconductor etch process: Improving robustness through model updating. Technical report, Eigenvector Research Inc., 1997.

[64] H. Abdi. *Partial Least Squares (PLS) Regression*. Sage, 2003.

[65] S. Wold. Personal memories of the early PLS development. *Chemom. Intell. Lab. Syst.*, 58(2):83–84, 2001.

[66] R.l D. Tobias. An introduction to partial least squares regression. Technical report, SAS Institute Inc., 1997.

[67] S. de Jong. SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.*, 18(3):251–263, Mar. 1993.

[68] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5(5):115–133, 1943.

[69] D. O. Hebb. *The organisation of behavior*. Wiley, New York, USA, 1949.

[70] F. Rosenblatt. *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Spartan Books, New York, USA, 1962.

[71] M. L. Minsky and S. A. Papert. *Perceptrons*. MIT Press, Cambridge, MA, USA, 1969.

[72] D.S. Broomhead and D. Lowe. Multi-variable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355, 1988.

[73] J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the USA*, 79(8):2554–2558, Apr. 1982.

[74] T. Kohonen. *Self-organization and associative memory*. Springer, Berlin, Germany, 3rd edition, 1984.

[75] G. May. Manufacturing ICs the neural way. *IEEE Spectrum*, 9:47–51, 1994.

[76] K. Warwick, G.W. Irwin, and K.J. Hunt. *Neural Networks for Control and Systems*. Peter Peregrinus Ltd., 1992.

[77] B. Carse and T. C. Fogarty. Tackling the curse of dimensionality of radial basis functional neural networks using a genetic algorithm. *Lecture notes in Computer Science*, 1141:707–719, 1996.

[78] K. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3):183–192, 1989.

[79] B. Irie and S. Miyake. Capabilities of three-layered perceptrons. In *IEEE International Conference on Neural Networks*, pages 641–648 vol.1, 1988.

[80] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989.

[81] Haiyang Zhang, Michael Nikolaou, and Ying Peng. Development of a data-driven dynamic model for a plasma etching reactor. *J. Vac. Sci. Technol. B.*, 20(3):891–901, May 2002.

[82] C.D. Himmel and G.S. May. Advantages of plasma etch modeling using neural networks over statistical techniques. *IEEE T. Semiconduct. M.*, 6(2):103–111, 1993.

[83] S.F. Lee and C.J. Spanos. Prediction of wafer state after plasma processing using real-time tool data. *IEEE T. Semiconduct. M.*, 8(3):252–261, August 1995.

[84] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Letters to Nature*, 323:533–536, Oct. 1986.

[85] M. Fernndez-Redondo and C. Hernndez-Espinosa. Weight initilization methods for multilayer feedforward. In *European Symposium on Artificial Neural Networks, Bruges (Belguim)*, pages 119–124, April 2001.

[86] B. Kim, K-H. Kwon, S-K Kwon, J-M. Park, S.W. Yoo, K-S. Park, and B-W Kim. Modeling oxide etching in a magnetically enhanced reactive ion plasma using neural networks. *J. Vac. Sci. Technol. B.*, 20(5):2113–2119, 2002.

[87] R. Battiti and F. Masulli. BFGS optimization for faster and automated supervised learning. In *International Neural Network Conference*, volume 2, pages 757–760, 1990.

[88] M.T. Hagan and M.B. Menhaj. Training feedforward networks with the marquardt algorithm. *IEEE T. Neural. Networ.*, 5(6):989–993, Nov 1994.

[89] D. Marquardt. An algorithm for least squares estimation of non-linear parameters. *SIAM Journal on Applied Mathematics*, 11(2):431–441, June 1963.

[90] J.-H. Xia, Rusli, and A. S. Kumta. Feedforward neural network trained by BFGS algorithm for modeling plasma etching of silicon carbide. *IEEE T. Plasma. Sci.*, 38(2):142 –148, Feb. 2010.

[91] Portia A. Cerny. Data mining and neural networks from a commercial perspective. In *ORSNZ Conference Twenty Naught One*, 2001.

[92] C. K. I. Williams and C. E. Rasmussen. *Gaussian Processes for regression*, chapter 8, pages 514–520. MIT Press, 1996.

[93] C. K. I. Williams. Prediction with gaussian processes: from linear regression to linear prediction and beyond. Technical Report NCRG/97/012, Neural Computing Research Group, Aston University, Birmingham, October 1997.

[94] M. Ebden. Gaussian processes for regression: A quick introduction. Technical report, University of Oxford, August 2008.

[95] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[96] G. Gregorcic and G. Lightbody. Local model network identification with gaussian processes. *IEEE T. Neural. Networ.*, 18(5):1404–1423, Sept 2007.

[97] C. E. Rasmussen. *Evaluation of Guassian processes and other methods for non-linear regression*. PhD thesis, Department of Computer Science, University of Toronto, Toronto, ON, Canada, 1996.

[98] D. J. Leith, M. Heidl, and J. V. Ringwood. Gaussian process prior models for electrical load forecasting. In *8th International conference on probabilistic methods applied to power systems*, Iowa, U.S., Sept. 2004.

[99] E.T. Jaynes. *Probability Theory*. Cambridge University Press, 2003.

[100] P. J. Huber. *Robust Statistics*. Wiley Series in Probability and Statistics, 1981.

[101] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. Wiley Series in Probability and Statistics, 1987.

[102] B. Walczak and D. L. Massart. Robust principal components regression as a detection tool for outliers. *Chemometrics and Intelligent Laboratory Systems*, 27(1):41–54, 1995.

[103] P. Filzmoser. Robust principal component regression. In *Proc. 6th Int. Conf. Computer Data Analysis and Modeling*, volume 1, pages 132–137, Minsk, Belarus, 2001.

[104] Mia Hubert and S. Engelen. Robust PCA and classification in biosciences. *Bioinformatics*, 20(11):1728–1736, 2004.

[105] J. Gonzalez, D. Pena, and R. Romera. A robust partial least squares method with applications. Statistics and econometrics working papers, Universidad Carlos III, Departamento de Estadstica y Econometra, Mar. 2007.

[106] B. Walczak. Neural networks with robust backpropagation learning algorithm. *Analytica Chimica Acta*, 322(1-2):21–29, 1996.

[107] I. Klevecka and J. Lelis. Pre-processing of input data of neural networks: The case of forecasting telecommunication network traffic. *Telektronik*, 3/4:168–178, 2008.

[108] R-S Guh. Robustness of the neural network based control chart pattern recognition system. *Int. Jrnl. Quality and Reliability Management*, 19(1):97–112, 2002.

[109] G.E.P. Box and D.R. Cox. An analysis of transformations. *J. Roy. Statistical Society, B (Methodological)*, 26(2):211–252, 1964.

[110] Thomas Briegel and Volker Tresp. Robust neural network regression for offline and online learning. In *Advances in Neural Information Processing Systems*, 1999.

[111] M. Kuss, T. Pfingsten, L. Csat, and C.E. Rasmussen. Approximate inference for robust gaussian process regression. Technical Report 136, Max-Planck-Institut fur biologische Kybernetik, Mar. 2005.

[112] Q.V. Le and A.J. Smola. Heteroscedastic gaussian process regression. In *Int. Conf. on Machine Learning*, pages 489–496, 2005.

[113] E. Snelson, C.E. Rasmussen, and Z. Ghahramani. Warped Gaussian Processes. In *Advances in Neural Information Processing Systems 16*, pages 337–344, 2004.

[114] P. Sollich. Gaussian process regression with mismatched models. In *Neural Information Processing Systems*, pages 519–526, 2001.

[115] P. Sollich. Can gaussian process regression be made robust against model mismatch? In *Machine Learning Workshop*, number 3635, pages 199–210, 2004.

[116] T. A. Badgwell, T. Breedijk, S. G. Bushman, S. W. Butler, S. Chatterjee, T. F. Edgar, A. J. Toprac, and I. Trachtenberg. Modeling and control of microelectronics materials processing. *Comput. Chem. Eng.*, 19(1):1–41, Jan. 1995.

[117] M. J. Kushner. A kinetic study of the plasma-etching process. I. a model for the etching of Si and $SiO_2$ in $C_nF_m/H_2$ and $C_nF_m/O_2$ plasmas. *J. Appl. Phys.*, 53(4):2923–2938, Apr. 1982.

[118] M.A. Lieberman. Dynamics of a collisional, capacitive RF sheath. *IEEE T. Plasma. Sci.*, 17(2):338–341, Apr. 1989.

[119] J.M. Berg, A. Yezzi, and A Tannenbaum. Toward real-time estimation of surface evolution in plasma etching: isotropy, anisotropy, and self-calibration. In *Proc. 36th IEEE Conference on Decision and Control*, volume 1, pages 860–865. IEEE, Dec 1997.

[120] Shahram Abdollahi-Alibeik, James P. McVittie, Krishna C. Saraswat, Valeriy Sukharev, and Philippe Schoenborn. Analytical modeling of silicon etch process in high density plasma. *J. Vac. Sci. Technol. A.*, 17(5):2485–2491, Sept. 1999.

[121] Barbara Abraham-Shrauner. Plasma etch profiles of passivated open-area trenches. *J. Vac. Sci. Technol. B.*, 19(3):711–721, May 2001.

[122] C. Liu and B.. Abraham-Shrauner. Plasma-etching profile model for $SiO_2$ contact holes. *IEEE T. Plasma. Sci.*, 30(4):1579–1586, Aug. 2002.

[123] R. P. Bray and R. R. Rhinehart. A simplified model for the etch rate of novolac-based photoresist. *Plasma Chem. Plasma Process.*, 21(1):149–161, Mar. 2001.

[124] H.S. Folger. *Elements of Chemical Reaction Engineering.* Prentice Hall, 2nd edition, 1992.

[125] R. A. Gottscho, D. Cooperberg, and V. Vahedi. The black box illuminated. In *Frontiers in low temperature plasma diagnostics III.* Lam Research Corporation, Centre for Research in Plasma Physics, Feb. 1999.

[126] K. O. Abrokwah, P. R. Chidambaram, and D. S. Boning. Pattern based prediction for plasma etch. *IEEE T. Semiconduct. M.*, 20(2):77–86, May. 2007.

[127] A. J. Van Roosmalen. Plasma parameter estimation from RF impedance measurements in a dry etching system. *Appl. Phys. Lett.*, 42(5):416–418, Mar. 1983.

[128] S. Bushman, T. F. Edgar, and I. Trachtenberg. Radio frequency diagnostics for plasma etch systems. *J. Electrochem. Soc.*, 144(2):721–732, 1997.

[129] P. Colpo and R. Ernst. Determination of the equivalent circuit of inductively coupled plasma sources. *J. Appl. Phys.*, 85(3):1366–1371, Feb 1999.

[130] M. A. Sobolewski. Electrical characterization of radio-frequency discharges in the gaseous electronics conference reference cell. *J. Vac. Sci. Technol. A.*, 10(6):3550–3562, Dec. 1992.

[131] M. A. Sobolewski. Sheath model for radio-frequency-biased, high-density plasmas valid for all $\omega/\omega_i$. *Phys. Rev. E*, 62(6):8540–8553, Dec. 2000.

[132] M. A. Sobolewski. Noninvasive monitoring of ion energy drift in an inductively coupled plasma reactor. *J. Appl. Phys.*, 97(3):033301, Dec. 2005.

[133] D. C. Montgomery. *Design and Analysis of Experiments.* John Riley & Sons, Inc., 6th edition, 2005.

[134] P. Klimecky, C. Garvin, C. G. Galarza, B. S. Stutzman, P. P. Khargonekar, and F. L. Terry. Real-time reactive ion etch metrology techniques to enable in situ response surface process characterization. *J. Electrochem. Soc.*, 148(1):C34–C40, Jan 2001.

[135] T.L. Vincent, P.P. Khargonekar, and F.L. Terry. End-point and etch rate control using dual-wavelength laser reflectometry with a nonlinear estimator. *J. Electrochem. Soc.*, 144(7):2467–2472, 1997.

[136] G.S. May, J. Huang, and C.J. Spanos. Statistical experimental design in plasma etch modeling. *IEEE T. Semiconduct. M.*, 4(4):83–98, May 1991.

[137] Liang Tan, D. Cameron, and C. McCorkell. Steady-state regression analysis and optimization of multivariable plasma etching system. In *20th Int. Conf. on Industrial Electronics, Control, and Instrumentation*, volume 3, pages 1986–1991, Bologna, Italy, Sep. 1994.

[138] K. J. McLaughlin, S. W. Butler, T. F. Edgar, and I. Trachtenberg. Development of techniques for real-time monitoring and control in plasma etching - I. response surface modeling of $CF_4/O_2$ and $CF_4/H_2$ etching of silicon and silicon dioxide. *J. Electrochem. Soc.*, 138(3):789–799, Mar. 1991.

[139] S.W. Butler, K.J. McLaughlin, T.F. Edgar, and I. Trachtenberg. Development of techniques for real-time monitoring and control in plasma etching - II. multivariable control system analysis of manipulated, measured, and performance variables. *J. Electrochem. Soc.*, 138(9):2727–2735, Sept. 1991.

[140] P.K. Mozumder and G.G. Barna. Statistical feedback control of a plasma etch process. *IEEE T. Semiconduct. M.*, 7(1):1–11, Feb 1994.

[141] S.J. Hong, G.S. May, and D-C. Park. Neural network modeling of reactive ion etching using optical emission spectroscopy data. *IEEE T. Semiconduct. M.*, 16(4):598–608, 2003.

[142] B. Kim and B. T. Lee. Prediction of silicon oxynitride plasma etching using a generalized regression neural network. *J. Appl. Phys.*, 98(3):034912, Aug 2005.

[143] B. Kim and K. Park. Modeling plasma etching process using a radial basis function network. *Microelectron. Eng.*, 77(2):150–157, Oct 2005.

[144] E.A. Rietman, S.H. Patel, and E.R. Lory. Modeling and control of a semiconductor manufacturing process with an automata network: An example in plasma etch processing. *Comput. Oper. Res.*, 23(6):573–585, Jun. 1996.

[145] B. Kim, W. Choi, and H. Kim. Using neural networks with a linear output neuron to model plasma etch processes. In *IEEE International Symposium on Industrial Electronics*, volume 1, pages 441–445, 2001.

[146] B. Kim, K.H. Kwon, S.K. Kwon, J.M. Park, S.W. Yoo, K.S. Park, I.K. You, and B.W. Kim. Modeling etch rate and uniformity of oxide via etching in a CHF3/CF4 plasma using neural networks. *Thin Solid Films*, 426(1-2):8–15, Feb 2003.

[147] B. Kim and S. Kim. Partial diagnostic data to plasma etch modeling using neural network. *Microelectronics Engineering*, 75(4):397–404, Jul. 2004.

[148] B. Kim and B-T Lee. Effect of plasma and control parameters on SiC etching in a $C_2F_6$ plasma. *Plasma Chem. Plasma Process.*, 23(3):489–499, Sep 2003.

[149] Jing-Hua Xia, Rusli, and A. Kumta. Modeling of silicon carbide ECR etching by feed-forward neural network and its physical interpretations. *IEEE T. Plasma. Sci.*, 38(5):1091–1096, May. 2010.

[150] B. Kim and K. Kim. Prediction of profile surface roughness in CHF3/CF4 plasma using neural network. *Appl. Surf. Sci.*, 222(1-4):17–22, Jan 2004.

[151] B. Kim, D. Han, S. Moon, and K. K. Lee. Modeling of sidewall bottom etching using a neural network. *Journal of the Korean Physical Society*, 46(6):1365–1370, Jun. 2005.

[152] B. Kim, D. W. Kim, and G. T. Park. Prediction of plasma etching using a polynomial neural network. *Plasma Science, IEEE Transactions on*, 31(6):1330–1336, Dec. 2003.

[153] H.C. Kim, F. Iza, S.S. Yang, M. Radmilovic-Radjenovic, and J.K. Lee. Particle and fluid simulations of low-temperature plasma discharges: benchmarks and kinetic effects. *J. Phys. D: Appl. Phys.*, 38:R283–R301, Sept 2005.

[154] J. P. Verboncoeur. Particle simulation of plasmas: review and advances. *Plasma Physics and Controlled Fusion*, 47(5A):A231, Apr 2005.

[155] R. Krimke and H. M. Urbassek. Self-consistent simulation of a planar electron-cyclotron-wave-resonance discharge. *J. Appl. Phys.*, 81(11):7163–7169, Jun 1997.

[156] Yugo Osano and Kouichi Ono. Atomic-scale cellular model and profile simulation of poly-Si gate etching in high-density chlorine-based plasmas: Effects of passivation layer formation on evolution of feature profiles. *J. Vac. Sci. Technol. B.*, 26(4):1425–1439, Jul. 2008.

[157] S. H. Lee, F. Iza, and J. K. Lee. Particle-in-cell monte carlo and fluid simulations of argon-oxygen plasma: Comparisons with experiments and validations. *Phys. Plasmas*, 13(5):057102, May 2006.

[158] Mark Wilcoxson and Vasilios Manousiouthakis. Continuum fluid models for plasma etching reactor control. In *American Control Conference*, volume 30, pages 3013–3017, Los Angeles, CA, U.S.A., 1993.

[159] E Gogolides, M Stathakopoulos, and A Boudouvis. Modelling of radio frequency plasmas in tetrafluoromethane ($CF_4$): the gas phase physics and the role of negative ion detachment. *J. Phys. D: Appl. Phys.*, 27(9):1878–1886, Sept. 1994.

[160] E. Meeks and J. Won Shon. Modeling of plasma-etch processes using well stirred reactor approximations and including complex gas-phase and surfacereactions. *IEEE T. Plasma. Sci.*, 23(4):539–549, Aug. 1995.

[161] S. A. Sfikas, E. K. Amanatides, D. S. Mataras, and D. E. Rapakoulias. Fluid model of an electron cyclotron wave resonance discharge. *IEEE T. Plasma. Sci.*, 35(5):1420–1425, Oct 2007.

[162] P. A. Miller, G. A. Hebner, K. E. Greenberg, P.D. Pochan, and B.P. Aragon. An inductively coupled plasma source for the gaseous electronics conference RF reference cell. *J. Res. Natl. Inst. Stand. Technol.*, 100(4):427, Jul. 1995.

[163] T. J. Sommerer and M. J. Kushner. Numerical investigation of the kinetics and chemistry of RF glow discharge plasmas sustained in He, $N_2$, $O_2$, $He/N_2/O_2$, $He/CF_4/O_2$, and $SiH_4/NH_3$ using a monte carlo-fluid hybrid model. *J. Appl. Phys.*, 71(4):1654–1674, Feb. 1992.

[164] R.K. Porteous and D.B. Graves. Modeling and simulation of magnetically confined low-pressure plasmas in two dimensions. *IEEE T. Plasma. Sci.*, 19(2):204–213, Apr. 1991.

[165] A. V. Vasenkov and M. J. Kushner. Electron energy distributions and anomalous skin depth effects in high-plasma-density inductively coupled discharges. *Phys. Rev. E*, 66(6):066411, Dec 2002.

[166] M. J. Kushner. Modeling of magnetically enhanced capacitvely coupled plasma sources: Ar discharges. *J. Appl. Phys.*, 94(3):1436–1447, 2003.

[167] Y. Yang and M. J. Kushner. Modeling of magnetically enhanced capacitively coupled plasma sources: Two frequency discharges. *J. Vac. Sci. Technol. A.*, 25(5):1420–1432, Sept 2007.

[168] R. J. Hoekstra and M. J. Kushner. Modeling of finite 3d features in high density plasma etching tools. In *45th National Symposium of the American Vacuum Society*. University of Illinois, Semiconductor Research Corp and NSF., 1998.

[169] S.N. Sivanandam and S.N. Deepa. *Introduction to Genetic Algorithms*. Springer, 2008.

[170] K. Han, K. Park, H. Chae, and E. Yoon. Multi-way principal component analysis for the endpoint detection of the metal etch process using the whole optical emission spectra. *Korean J. Chem. Eng.*, 25(1):13–18, Jan. 2008.

[171] S. ThomasIII, H. H. Chen, C. K. Hanish, J. W. Grizzle, and S. W. Pang. Minimized response time of optical emission and mass spectrometric signals for optimized endpoint detection. *J. Vac. Sci. Technol. B.*, 14(4):2531–2536, Jul/Aug 1996.

[172] S.B. Dolins, A. Srivastava, and B.E. Flinchbaugh. Monitoring and diagnosis of plasma etch processes. *IEEE T. Semiconduct. M.*, 1(1):23–27, Feb 1988.

[173] H. E. Litvak. End point control via optical emission spectroscopy. *J. Vac. Sci. Technol. B.*, 14(1):516–520, Jan 1996.

[174] H. H. Yue, S. J. Qin, J. Wiseman, and A. Toprac. Plasma etching endpoint detection using multiple wavelengths for small open-area wafers. *J. Vac. Sci. Technol. A.*, 19(1):66–75, Jan/Feb 2001.

[175] K. Han, E.S. Yoon, J. Lee, H. Chae, K.H. Han, and K.J. Park. Real-time end-point detection using modified principal component analysis for small open area $SiO_2$ plasma etching. *Industrial & Engineering Chemistry Research*, 47(11):3907–3911, Jun. 2008.

[176] K. Han, S. Kim, K. J. Park, E.S. Yoon, and H. Chae. Principal component analysis based support vector machine for the end point detection of the metal etch process. In *Proc. 17th IFAC World Congress*, pages 4560–4565, Seoul, Korea, Jul. 2008.

[177] S. Rangan, C. Spanos, and K. Poolla. Modeling and filtering of optical emission spectroscopy data for plasma etching systems. In *IEEE International Symposium on Semiconductor Manufacturing*, volume 1, pages B41–B44, Albuquerque, NM , U.S.A., Jun. 1997.

[178] E. Ragnoli, S. McLoone, J. Ringwood, and N. Macgerailt. Matrix factorisation techniques for endpoint detection in plasma etching. In *Advanced Semiconductor Manufacturing Conference*, pages 156–161, May 2008.

[179] K. Ukai and K. Hanazawa. End-point determination of aluminum reactive ion etching by discharge impedance monitoring. *J. Vac. Sci. Technol.*, 16(2):385–387, Mar. 1979.

[180] G. Fortunato. End-point determination by reflected power monitoring. *J. Phys. E: Sci. Instr.*, 20(8):1051–1052, Aug. 1987.

[181] E. A. Rietman, J. T-C Lee, and N. Layadi. Dynamic images of plasma processes: Use of fourier blobs for endpoint detection during plasma etching of patterned wafers. *J. Vac. Sci. Technol. A.*, 16(3):1449–1453, May/Jun 1998.

[182] A.T.C Koh, N.F. Thronhill, and V.J. Law. Principal component analysis of plasma harmonics in end-point detection of photoresist stripping. *Electron. Lett.*, 35(16):1383–1385, Aug. 1999.

[183] V. Patel, B. Singh, and J. H. Thomas. Reactive ion etching end-point determination by plasma impedance monitoring. *Appl. Phys. Lett.*, 61(16):1912–1914, Oct. 1992.

[184] F. Bose, R. Patrick, and H. P. Baltes. Characterization of plasma etch processes using measurements of discharge impedance. *J. Vac. Sci. Technol. B.*, 12(4):2805–2809, Jul. 1994.

[185] H. L. Maynard, E. A. Rietman, J. T. C. Lee, and D. E. Ibbotson. Plasma etching endpointing by monitoring radio-frequency power systems with an artificial neural network. *J. Electrochem. Soc.*, 143(6):2029–2035, Jun. 1996.

[186] V.J. Law, A.J. Kenyon, N.F Thornhill, V. Srigengan, and I. Batty. Remote-coupled sensing of plasma harmonics and process end-point detection. *Surf. Eng., Surf. Instr. Vac. Tech.*, 57(4):351–364, Jun. 2000.

[187] M. Bonner and R. Clark. RF-based sensor technology improves cleaning efficiency on PECVD tools. *Semiconductor Fabtech*, 19:1–4, 2003.

[188] X. Li, M. Schaepkens, G. S. Oehrlein, R. E. Ellefson, L. C. Frees, N. Mueller, and N. Korner. Mass spectrometric measurements on inductively coupled fluorocarbon plasmas: Positive ions, radicals and endpoint detection. *J. Vac. Sci. Technol. A.*, 17(5):2438–2446, Sep/Oct 1999.

[189] B.B. Ma. Literature review: Endpoint detection in plasma etching. Technical Report EE/SMcL/1/2007, Dept. Elect. Eng., 2007.

[190] B. E. Goodlin. *Multivariate Endpoint Detection of Plasma Etching Processes*. PhD thesis, The University of Texas at Austin, April 2002.

[191] J. Zhang, E.B. Martin, and A.J. Morris. Fault-detection and diagnosis using multivariate statistical techniques. *Chem. Eng. Res. Des.*, 74(1):89–96, Jan. 1996.

[192] S.J. Hong and G.S. May. Neural-network-based sensor fusion of optical emission and mass spectroscopy data for real-time fault detection in reactive ion etching. *IEEE T. Ind. Electron.*, 52(4):1063–1072, Aug. 2005.

[193] R. Shadmehr, D. Angell, P. B. Chou, G. S. Oehrlein, and R. S. Jaffe. Principal component analysis of optical emission spectroscopy and mass spectrometry: Application to reactive ion etch process parameter estimation using neural networks. *J. Electrochem. Soc.*, 139(3):907–914, Mar. 1992.

[194] T. Sarmiento, S.J. Hong, and G.S. May. Fault detection in reactive ion etching systems using one-class support vector machines. In *Advanced Semiconductor Manufacturing Conference and Workshop*, pages 139–142, 2005.

[195] C.J. Spanos, H.F. Guo, A. Miller, and J. Levine-Parrill. Real-time statistical process control using tool data. *IEEE T. Semiconduct. M.*, 5(4.):308–318, Nov. 1992.

[196] K. A. Chamness. Diagnostics of plasma etch: PCA with adaptive centering and scaling. In *AEC/APC XV Symposium*, 2003.

[197] M-S. Chen, T.F. Yen, and B. Coonan. Controlling etch tools using real-time fault detection and classification. *MICRO*, 23(2):59–68, Mar. 2005.

[198] V. J. Law and N. Macgearailt. Visualization of a dual-frequency plasma etch process. *Measurement Science and Technology*, 18(3):645–649, Jan 2007.

[199] S.I. Imai. Virtual metrology for plasma particle in plasma etching equipment. In *Int. Symp. on Semiconductor Manufacturing.*, pages 1–4, Oct. 2007.

[200] Y-J. Chang. Fault detection for plasma etching processes using RBF neural networks. In *Int. Symp. Neural Networks*, pages 538–543, 2005.

[201] Q.P. He and Jin Wang. Large-scale semiconductor process fault detection using a fast pattern recognition-based method. *IEEE T. Semiconduct. M.*, 23(2):194–200, May. 2010.

[202] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley & Sons, 2001.

[203] J-M. Lee, S. J. Qin, and I-B. Lee. Fault detection and diagnosis based on modified independent component analysis. *AIChE Journal*, 52(10):3501–3514, Oct. 2006.

[204] Z. Ge and Z. Song. Semiconductor manufacturing process monitoring based on adaptive substatistical PCA. *IEEE T. Semiconduct. M.*, 23(1):99–108, Feb 2010.

[205] C.K. Hanish, J.W. Grizzle, and Jr. Teny, F.L. Estimating and controlling atomic chlorine concentration via actinometry. *IEEE T. Semiconduct. M.*, 12(3):323–331, Aug. 1999.

[206] M. V. Malyshev and V. M. Donnelly. Determination of electron temperatures in plasmas by multiple rare gas optical emission, and implications for advanced actinometry. *J. Vac. Sci. Technol. A.*, 15(3):550–558, May 1997.

[207] M.V. Malyshev and V.M. Donnelly. Trace rare gases optical emission spectroscopy: Nonintrusive method for measuring electron temperatures in low-pressure, low-temperature plasmas. *Physical Review E*, 60(5):6016–6029, Nov 1999.

[208] V.M. Donnelly, M.V. Malyshev, M. Schabel, A. Kornblit, W. Tai, I.P. Herman, and N.C.M. Fuller. Optical plasma emission spectroscopy of etching plasmas used in si-based semiconductor processing. *Plasma Sources Science and Technology*, 11(3A):A26, Aug. 2002.

[209] V.M. Donnelly and M.J. Schabel. Spatially resolved electron temperatures, species concentrations, and electron energy distributions in inductively coupled chlorine plasmas, measured by trace-rare gases optical emission spectroscopy. *J. Appl. Phys.*, 91(10):6288–6295, May 2002.

[210] M. A. Sobolewski. Measuring the ion current in high-density plasmas using radio-frequency current and voltage measurements. *J. Appl. Phys.*, 90(6):2660–2671, Sept. 2001.

[211] M. A. Sobolewski. Measuring the ion current in electrical discharges using radio-frequency current and voltage measurements. *Appl. Phys. Lett.*, 72(10):1146–1148, Mar. 1998.

[212] M. A. Sobolewski. Monitoring sheath voltages and ion energies in high-density plasmas using noninvasive radio-frequency current and voltage measurements. *J. Appl. Phys.*, 95(9):4593–4604, May. 2004.

[213] T. Yamashita, S. Hasaka, I. Natori, H. Fukui, and T. Ohmi. Minimizing damage and contamination in RIE processes by extracted-plasma-parameter analysis. *IEEE T. Semiconduct. M.*, 5(3):223–233, Aug. 1992.

[214] S. Wurm, W. Preis, and M. Klick. SEERS-based process control for plasma etching. *Solid. State. Technol.*, 42(6):103, Jun. 1999.

[215] A. Steinbach. Real time plasma etch diagnostics by plasma monitoring system hercules. In *Frontiers in low temperature plasma diagnostics III*. Centre for Research in Plasma Physics, Siemens, February 1999.

[216] J. S. Steinmetz, N. Rohn, T. Werner, M. Klick, W. Rehak, Kammeyer M., D. Suchland, S. Wurm, W. Preis, and Ch. Koelbl. Plasma diagnostic in an inductively coupled plasma using chlorine chemistry. In *Workshop on Self Excited Electron*

*Plasma Resonance Spectroscopy*. Lam Research, Adolf-Slaby-Institute, Infineon Technologies, 1999.

[217] V. Tegeder, R. Ronchi, S. Mueller, and M. Hofmann. The tremendous impact of APC for plasma etch. Solid State Technology Press, Oct. 2001.

[218] P.H Chen, S. Wu, J. Lin, F. Ko, H. Lo, J. Wang, C.H. Yu, and M.S.m Liang. Virtual metrology: a solution for wafer to wafer advanced process control. In *IEEE Int. Symp. on Semiconductor Manufacturing*, pages 155–157, 2005.

[219] N. Camara and K. Zekentes. Study of the reactive ion etching of 6H-SiC and 4H-SiC in $SF_6$/Ar plasmas by optical emission spectroscopy and laser interferometry. *Solid-State Electron.*, 46(11):1959–1963, Nov. 2002.

[220] E.A. Rietman, D.E. Ibbotson, and J.T.C Lee. Preliminary empirical results suggesting the mapping of dynamic in situ process signals to real-time wafer attributes in a plasma etch process. *J. Vac. Sci. Technol. B.*, 16(1):131–136, Jan. 1998.

[221] R. Chen, H. Huang, C. J. Spanos, and M. Gatto. Plasma etch modeling using optical emission spectroscopy. *J. Vac. Sci. Technol. A.*, 14(3):1901–1906, May 1996.

[222] E. Ragnoli, S. McLoone, S. Lynn, J. Ringwood, and N. Macgearailt. Identifying key process characteristics and predicting etch rate from high-dimension datasets. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference,*, pages 106–111, May. 2009.

[223] D.A. White, D. Boning, S.W. Butler, and G.G. Barna. Spatial characterization of wafer state using principal component analysis of optical emission spectra in plasma etch. *IEEE T. Semiconduct. M.*, 10(1):52–61, Feb 1997.

[224] D. Tsunami, J. McNames, B. Whitefield, P.Rudolph, and J. Zola. Oxide etch rate estimation using plasma impedance monitoring. In *Proc. SPIE*, volume 5755, pages 59–68, 2005.

[225] C. Garvin and J. W. Grizzle. Demonstration of broadband radio frequency sensing: Empirical polysilicon etch rate estimation in a Lam 9400 etch tool. *J. Vac. Sci. Technol. A.*, 18(4):1297–1302, Jul. 2000.

[226] S.F. Lee and C.J. Spanos. Equipment analysis and wafer parameter prediction using real-time tool data. In *Int. Symp. on Semiconductor Manufacturing*, pages 133–136, Jun. 1994.

[227] J.P. Card, M. Naimo, and W. Ziminsky. Run-to-run process control of a plasma etch process with neural network modelling. *Qual. Reliab. Eng. Int.*, 14(4):247–260, 1998.

[228] Dekong Zeng, Yajing Tan, and Costas J. Spanos. Dimensionality reduction methods in virtual metrology. In *Proc. of SPIE*, volume 6922, page 692238, Feb. 2008.

[229] D. Zeng and C. J. Spanos. Virtual metrology modeling for plasma etch operations. *IEEE T. Semiconduct. M.*, 22(4):419–431, Nov. 2009.

[230] P. Kang, H.J. Lee, S. Cho, D. Kim, J. Park, C.K. Park, and S. Doh. A virtual metrology system for semiconductor manufacturing. *Expert. Syst. Appl.*, 36(10):12554–12561, Dec. 2009.

[231] P. Kang, D. Kim, H.J. Lee, S. Doh, and S. Cho. Virtual metrology for run-to-run control in semiconductor manufacturing. *Expert Systems with Applications*, In Press:Corrected Proof, 2010.

[232] T.H. Lin, F.T. Cheng, W.M. Wu, C.A. Kao, A.J. Ye, and F.C. Chang. NN-based key-variable selection method for enhancing virtual metrology accuracy. *IEEE T. Semiconduct. M.*, 22(1):204–211, Feb. 2009.

[233] F.T. Cheng, Y.T. Chen, Y.C. Su, and D.L Zeng. Method for evaluating reliance level of a virtual metrology system. In *IEEE Int. Conf. Robotics and Automation*, pages 1590–1596, Apr. 2007.

[234] F-T. Cheng, Y-T. Chen, Y-C. Su, and D-L. Zeng. Evaluating reliance level of a virtual metrology system. *IEEE T. Semiconduct. M.*, 21(1):92 –103, Feb. 2008.

[235] F.T. Cheng, H-C. Huang, and C-A. Kao. Dual-phase virtual metrology scheme. *IEEE T. Semiconduct. M.*, 20(4):566–571, Nov. 2007.

[236] Y.C. Su, T.H. Lin, F.T. Cheng, and W.M. Wu. Accuracy and real-time considerations for implementing various virtual metrology algorithms. *IEEE T. Semiconduct. M.*, 21(3):426–434, Aug. 2008.

[237] A. Ferreira, A. Roussy, and L. Conde. Virtual metrology models for predicting physical measurement in semiconductor manufacturing. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pages 149–154, May 2009.

[238] A.J. Su, J.C. Jeng, H.P. Huang, C.C. Yu, S.Y. Hung, and C.K. Chao. Control relevant issues in semiconductor manufacturing: Overview with some new results. *Control. Eng. Pract.*, 15(10):1268–1279, Oct. 2007.

[239] A.J. Pasadyn and T.F. Edgar. Observability and state estimation for multiple product control in semiconductor manufacturing. *IEEE T. Semiconduct. M.*, 18(4):592–604, Nov. 2005.

[240] F.F. Chen. Plasma ionization by helicon waves. *Plasma Physics and Controlled Fusion*, 33(4):339–364, Apr. 1991.

[241] E.A. Bering and M. Brukardt. Ion acceleration by single pass ion cyclotron heating in the VASIMR engine. In *Proc. 29th International Electric Propulsion Conference*, number IEPC-2005-093, Princeton University, NJ, USA, Oct. 2005.

[242] T. Glover, G. F. Franklin, J.P. Squire, Jacobson V.P., Chavers D.G., and Carter M.D. Principal VASIMR results and present objects. In *Proc. Space Technology and Applications International Forum*, pages 976–982, Albuquerque, NM, Feb. 2005.

[243] J.P. Squire, F.R. Chang-Daz, T.W. Glover, M.D. Carter, L.D. Cassady, W.J. Chancery, V.T. Jacobson, G.E. McCaskill, C.S. Olsen, E.A. Bering, M.S. Brukardt, and B.W. Longmier. VASIMR performance measurements at powers exceeding 50–kW and lunar robotic mission applications. In *Proc. Int. Interdisciplinary Symp. on Gaseous and Liquid Plasmas, not paginated*, Akiu/Sendai, Japan, Sept. 2008.

[244] NASA. Mars exploration rover launches. Online at http://www.jpl.nasa.gov/news/press_kits/merlaunch.pdf, Jun. 2003.

[245] S. Upson. Rockets for the red planet. *IEEE Spectrum*, 46(6):42–47, Jun. 2009.

[246] R. W. Boswell and F.F. Chen. Helicons – the early years. *IEEE Trans. Plasma Sci.*, 25(6):1229–1244, Dec. 1997.

[247] R.W. Boswell and F.F. Chen. Helicons – the past decade. *IEEE Trans. Plasma Sci.*, 25(6):1245–1257, Dec. 1997.

[248] J. P. Rayner and A.D. Cheetham. Helicon modes in a cylindrical plasma source. *Plasma Sources Science and Technology*, 8(1):79–87, Feb. 1999.

[249] A.R. Ellingboe and R.W. Boswell. Capacitive, inductive and helicon-wave modes of operation of a helicon plasma source. *Physics of Plasmas*, 3(7):2797–2804, Jul. 1996.

[250] G.F. Franklin, J.D. Powell, and A. Emami-Naeini. *Feedback Control of Dynamic Systems*. Addison-Wesley Publishing Company, 3rd edition, 1993.

[251] D. Luenberger. An introduction to observers. *IEEE Trans. Auto. Cont.*, 16(6):596–602, Dec. 1971.

[252] S. Lynn, J.V. Ringwood, and J.I. Del Valle Gamboa. State estimation for the VASIMR plasma engine. In *Proc. 16$^{th}$ Irish Signals and Systems Conference*, pages 24–29, Galway, Ireland, 2008.

[253] S. Lynn, J.V. Ringwood, and J.I. Del Valle Gamboa. Temperature estimation for a plasma-propelled rocket engine - inferential measurement using optical emission spectrometer data. *IEEE Contr. Syst. Mag.*, 29(6):15–25, Dec, 2009.

[254] S. Lynn, J. Ringwood, and N. MacGearailt. Gaussian process regression for virtual metrology of plasma etch. In *IET Irish Signals and Systems Conference*, volume 2010, pages 42–47. IEE, 2010.

[255] R. Murray-Smith and T.A. Johansen. *Multiple Model Approaches to Modelling and Control*. Taylor and Francis, 1997.

[256] S. Lynn. Local modelling of a plasma etch data set. Technical Report EE/JVR/1/2010, Dept. of Elec. Eng., National University of Ireland, Maynooth, Feb. 2010.

[257] P-N Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.

[258] S. J. Qin. Recursive PLS algorithms for adaptive data modeling. *Comput. Chem. Eng.*, 22(4-5):503–514, 1998.

[259] E. R. Malinowski. Automatic window factor analysis-a more efficient method for determining concentration profiles from evolutionary spectra. *J. Chemom.*, 10(4):273–279, Jul. 1996.

[260] B.S. Dayal and J.F. Macgregor. Recursive exponentially weighted PLS and its applications to adaptive control and prediction. *J. Process Control*, 7(3):169–179, 1997.

[261] W. Li, H. H. Yue, S. Valle-Cervantes, and S.J. Qin. Recursive PCA for adaptive process monitoring. *J. Process Control*, 10:471–486(16), Oct 2000.

[262] L. Xu, J.H. Jiang, W.Q. Lin, Y.P. Zhou, H.L. Wu, G.L. Shen, and R.Q. Yu. Optimized sample-weighted partial least squares. *Talanta*, 71(2):561–566, Feb. 2007.

[263] S. Lynn, J. V. Ringwood, and N. MacGearailt. Weighted windowed PLS models for virtual metrology of an industrial plasma etch process. In *IEEE Int. Conf. on Industrial Technology*, pages 271–276, Valpariaso, Chile, Mar. 2010.

[264] S. Lynn, J. Ringwood, E. Ragnoli, S. McLoone, and N. MacGearailt. Virtual metrology for plasma etch using tool variables. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pages 143–148, May 2009.

[265] X.A. Wang and R.L. Mahajan. Artificial neural network model-based run-to-run process controller. *IEEE Trans. Components, Packaging, and Manufacturing Technology, Part C.*, 19(1):19–26, Jan. 1996.

[266] T.F. Edgar, W.J. Campbell, and C. Bode. Model-based control in microelectronics manufacturing. In *Proc. of the 38th IEEE Conf. on Decision and Control*, volume 4, pages 4185–4191, Phoenix, AZ , USA, Dec. 1999.

[267] J.P. Card, D.L. Sniderman, and C. Klimasauskas. Dynamic neural control for a plasma etch process. *IEEE T. Neural. Networ.*, 8(4):883–901, Jul. 1997.

[268] E.A. Rietman and S.H. Patel. A production demonstration of wafer-to-wafer plasma gate etch control by adaptive real-time computation of the over-etch time from in situ process signals. *IEEE T. Semiconduct. M.*, 8(3):304–308, Aug 1995.

[269] S. Limanond, J. Si, and Y.L. Tseng. Production data based optimal etch time control design for a reactive ion etching process. *IEEE T. Semiconduct. M.*, 12(1):139–147, Feb. 1999.

[270] G. Bacelli. Control in semiconductor manufacturing: A literature review. Technical Report EE/JVR/3/2007, Dept. of Elec. Eng., National University of Ireland, Maynooth, 2007.

[271] K.J. McLaughlin, T.F. Edgar, and I. Trachtenberg. Real-time monitoring and control in plasma etching. *IEEE Contr. Syst. Mag.*, 11(3):3–10, Apr. 1991.

[272] M. Sarfaty, C. Baum, M. Harper, N. Hershkowitz, and J.L. Shohet. Real-time monitoring and control of plasma etching. *Jpn. J. Appl. Phys.*, 37(4B):2381–2387, Apr. 1998.

[273] D. Stokes and G.S. May. Real-time control of reactive ion etching using neural networks. *IEEE T. Semiconduct. M.*, 13(4):469–480, Nov 2000.

[274] D. Stokes and G.S. May. Indirect adaptive control of reactive ion etching using neural networks. *IEEE T. Robotic. Autom.*, 17(5):650–657, Oct. 2001.

[275] I.G. Rosen, T. Parent, B. Fidan, C. Wang, and A. Madhukar. Design, development, and testing of real-time feedback controllers for semiconductor etching processes using in situ spectroscopic ellipsometry sensing. *IEEE T. Contr. Syst. T.*, 10(1):64–75, Jan. 2002.

[276] A. Armaou, J. Baker, and P.D. Christofides. Feedback control of plasma etching reactors for improved etching uniformity. *Chem. Eng. Sci.*, 56(4):1467–1475, Feb. 2001.

[277] B.R. Parkinson, Hyung Lee, M. Funk, D. Prager, A. Yamashita, R. Sundararajan, and T.F. Edgar. Addressing dynamic process changes in high volume plasma etch manufacturing by using multivariate process control. *IEEE T. Semiconduct. M.*, 23(2):185 –193, May. 2010.

[278] B.A. Rashap, P.P. Khargonekar, J.W. Grizzle, M.E. Elta, J.S. Freudenberg, and Jr. Terry, F.L. Real-time control of reactive ion etching: identification and disturbance rejection. In *Proc. 32nd IEEE Conference on Decision and Control*, volume 4, pages 3379–3385, Dec. 1993.

[279] B.A. Rashap, M.E. Elta, H. Etemad, J.P. Fournier, J.S. Freudenberg, M.D. Giles, J.W. Grizzle, P.T. Kabamba, P.P. Khargonekar, S. Lafortune, J.R. Moyne, D. Teneketzis, and Jr Terry, F.L. Control of semiconductor manufacturing equipment: real-time feedback control of a reactive ion etcher. *IEEE T. Semiconduct. M.*, 8(3):286–297, Aug. 1995.

[280] O.D. Patterson and P.P. Khargonekar. Reduction of loading effect in reactive ion etching using real-time closed-loop control. *J. Electrochem. Soc.*, 144(8):2865–2871, Aug. 1997.

[281] M. Hankinson, T. Vincent, K.B. Irani, and P.P. Khargonekar. Integrated real-time and run-to-run control of etch depth in reactive ion etching. *IEEE T. Semiconduct. M.*, 10(1):121–130, Feb. 1997.

[282] H. M. Park, C. Garvin, D. S. Grimard, and J. W. Grizzle. Control of ion energy in a capacitively coupled reactive ion etcher. *J. Electrochem. Soc.*, 145(12):4247–4252, 1998.

[283] V. Milosavljevic, A. R. Ellingboe, C. Gaman, and J. V. Ringwood. Real-time plasma control in a dual-frequency, confined plasma etcher. *J. Appl. Phys.*, 103(8):083302–083302–10, Apr. 2008.

[284] C. Lin, K.C. Leou, and K.M. Shiao. Feedback control of chlorine inductively coupled plasma etch processing. *J. Vac. Sci. Technol. A.*, 23(2):281–287, Mar 2005.

[285] C. Lin, K.C. Leou, H.M. Huang, and C.H. Hsieh. Feedback control of plasma electron density and ion energy in an inductively coupled plasma etcher. *J. Vac. Sci. Technol. A.*, 27(1):157–164, Jan. 2009.

[286] Pete I. Klimecky, J. W. Grizzle, and Fred L. Terry. Compensation for transient chamber wall condition using real-time plasma density feedback control in an inductively coupled plasma etcher. *J. Vac. Sci. Technol. A.*, 21(3):706–717, May. 2003.

[287] J.G. Ziegler and N.B. Nichols. Optimum settings for automatic controllers. *Transactions of the American Society of Mechanical Engineers*, 64(11):759–768, Nov. 1942.

[288] K.J Astrom and T. Haggland. *The Control Handbook*. IEEE Press, Piscataway, NJ, 1996.

[289] Aidan O'Dwyer. *Handbook of PI and PID Controller Tuning Rules*. Imperial College Press, 2009.

[290] K. H. Ang, G. Chong, and Yun Li. PID control system analysis, design, and technology. *IEEE T. Contr. Syst. T.*, 13(4):559–576, Jul. 2005.

[291] S.J. Qin and T.A. Badgwell. A survey of industrial model predictive control technology. *Control. Eng. Pract.*, 11(7):733–764, Jul. 2003.

[292] J. M. Maciejowski. *Predictive Control with Contraints*. Prentice Hall, 2002.

[293] M. T. Khadir. *Modelling and predictive control of a milk pasteurisation plant*. PhD thesis, Dept. of Elec. Eng., National University of Ireland, Maynooth, 2002.

[294] J. E. Marshall. *Control of time-delay systems*. P. Peregrinus, 1979.

[295] J. Richalet and D. O. Donovan. *Predictive Functional Control*. Springer, May. 2009.

[296] J.A. Rossiter and J. Richalet. Realigned models for prediction in MPC: a good thing or not? In *Proc. of the 6th Advanced process control Conf.*, pages 63–70, 2001.

[297] E.F. Camacho and C. Bordons. *Model Predictive Control*. Springer, 2004.

[298] P. E. Wellstead and M. B. Zarrop. *Self-tuning systems: Control and Signal Processing*. Wiley, 1991.