NATIONAL UNIVERSITY OF IRELAND, MAYNOOTH

NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

Prognostic Algorithms for Condition Monitoring and Remaining Useful Life Estimation

Shane Butler

A thesis submitted in partial fulfillment for the degree of

Doctor of Philosophy

in the
Faculty of Science and Engineering
Department of Electronic Engineering

Supervisor: Prof. John V. Ringwood
Head of Department: Dr. Seán McLoone

September 2012

Contents

Declaration of Authorship

I, Shane Butler, declare that this thesis titled, 'Prognostic Algorithms for Condition Monitoring and Remaining Useful Life Estimation' 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:

Abstract

To enable the benefits of a truly condition-based maintenance philosophy to be realised, robust, accurate and reliable algorithms, which provide maintenance personnel with the necessary information to make informed maintenance decisions, will be key. This thesis focuses on the development of such algorithms, with a focus on semiconductor manufacturing and wind turbines.

An introduction to condition-based maintenance is presented which reviews different types of maintenance philosophies and describes the potential benefits which a condition-based maintenance philosophy will deliver to operators of critical plant and machinery. The issues and challenges involved in developing condition-based maintenance solutions are discussed and a review of previous approaches and techniques in fault diagnostics and prognostics is presented.

The development of a condition monitoring system for dry vacuum pumps used in semiconductor manufacturing is presented. A notable feature is that upstream process measurements from the wafer processing chamber were incorporated in the development of a solution. In general, semiconductor manufacturers do not make such information available and this study identifies the benefits of information sharing in the development of condition monitoring solutions, within the semiconductor manufacturing domain. The developed solution provides maintenance personnel with the ability to identify, quantify, track and predict the remaining useful life of pumps suffering from degradation caused by pumping large volumes of corrosive fluorine gas.

A comprehensive condition monitoring solution for thermal abatement systems is also presented. As part of this work, a multiple model particle filtering algorithm for prognostics is developed and tested. The capabilities of the proposed prognostic solution for addressing the uncertainty challenges in predicting the remaining useful life of abatement systems, subject to uncertain future operating loads and conditions, is demonstrated.

Finally, a condition monitoring algorithm for the main bearing on large utility scale wind turbines is developed. The developed solution exploits data collected by onboard supervisory control and data acquisition (SCADA) systems in wind turbines. As a result, the developed solution can be integrated into existing monitoring systems, at no additional cost. The potential for the application of multiple model particle filtering algorithm to wind turbine prognostics is also demonstrated.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor, Professor John Ringwood, for all his help in carrying out the work presented in this thesis. Ever since supervising my final year project in my undergraduate years, John has been an incredible source of knowledge, advice, and encouragement and has provided me with countless opportunities. Beyond academic advice, John has also been a great mentor and I will miss our weekly meetings discussing all sorts of interesting topics and developments.

During the thesis, I have had the privilege of working with many companies who have provided the data and information, without which this document would not exist. Firstly, I would like to thank Niall MacGearailt at Intel Ireland. As "the man who can make things happen", Niall was always extremely generous with his knowledge, time and enthusiasm. At Edwards Vacuum, I would firstly like to thank Michael Mooney for all the help, interest and enthusiasm for this project. I would also like to thank Dave Kember, Nigel Gibbens and Matt McDonald, who were so generous with their time and knowledge in getting this project started. I would also like to thank Adrian Johnston at Seagate, who went out his way on many occasions to help me out. I'd also like to thank Frank O'Connor and Des Farren at ServusNet for providing me access to their data and all the help I needed. I would also like to thank Enterprise Ireland for providing the financial support for carrying out this work.

I am also grateful to all the staff in the Department of Electronic Engineering at NUIM, for providing such a great environment to work in. I'd also like to thank all my fellow postgraduates with whom I shared all the ups and downs along the road. In particular, I'd like to thank those with whom I shared an office for many years; Giorgio Bacelli, Shane Lynn, Niall Cahill, Violeta McCloone and Francesco Fusco.

I would also like to thank my parents, Kay and J.J., and my brother Stuart, who have supported me at every stage, and for that last-minute proof reading.

Finally, I would like to thank my wonderful and beautiful girlfiend Dee, whose unwavering love and support have helped me get to this stage. I couldn't wish for a more loving and fun partner in life.

List of Figures

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
BIT	Built In Test
BP	Booster Power
BT	Booster Temperature
CBM	Condition Based Maintenance
CT	Combustor Temperature
CVD	Chemical Vapour Deposition
DESP	Double Exponential Smoothing Prediction
DPT	Dry Pump Temperature
DRE	Destruction/Removal Efficiency
DWNN	Dynamic Wavelet Neural Network
EADS	Edwards Advanced Diagnostic System
EKF	Extended Kalman Filter
EM	Expectation Maximisation
EOD	End of Discharge
EoL	End of Life
EOP	End-of-Pipe
EP	Exhaust Pressure
EWMA	Exponentially Weighted Moving Average
FJ	Figueiredo-Jain
FMECA	Failure Modes and Criticality Analysis
FP	Foreline Pressure
GMM	Gaussian Mixture Model

GPR	Gaussian Process Regression
GWP	Global Warming Potential
HCDP	Harsh Chemical Dry Pump
HCMB	Harsh Chemical Mechanical Booster
HDP	High Density Plasma
HF	High-Fire
HFC	Hydrofluorocompound
HMI	Human Machine Interface
ITRS	International Technology Roadmap for Semiconductors
JIT	Just In Time
JITP	Just In Time Point
JSF	Joint Strike Fighter
LAN	Local Area Network
LF	Low-Fire
LSR	Least Squares Regression
LTI	Lead Time Interval
MLR	Multiple Linear Regression
MTBF	Mean Time Between Failure
O&S	Operation & Support
OEM	Original Equipment Manufacturer
PCA	Principal Component Analysis
PDF	Probability Density Function
PFC	Perfluorocompound
PHM	Prognostics and Health Management
PLS	Partial Least Squares
PoF	Probability of Failure
POU	Point-of-Use
PTFE	Polytetrafluoroethylene
QWT	Quench-Wall Temperature
RF	Radio Frequency
RNN	Recurrent Neural Network
RUL	Remaining Useful Life
RVM	Relevance Vector Machine

SCADA	Supervisory Control and Data Acquisition Systems
SIR	Sequential Importance Resampling
SIS	Sequential Importance Sampling
SMC	Sequential Monte Carlo
SMO	Sliding Mode Observer
SPC	Statistical Process Control
TOC	Total Ownership Cost
TPU	Thermal Processing Unit
WNN	Wavelet Neural Network
WSC	World Semiconductor Council
ZOH	Zero Order Hold

Dedicated to my fantastic parents, Kay and J.J., for providing me with every opportunity and encouragement in life

Chapter 1

Introduction

1.1 Background and Motivation

The maintenance of critical plant and machinery is a major expense for manufacturers and operators. Maintenance practices have traditionally employed one of two philosophies; preventative or corrective. Preventative maintenance involves performing regular scheduled maintenance to maintain equipment in good health and avoid in-service equipment failures. Corrective maintenance involves running equipment until it fails and then taking remedial action. Both approaches have drawbacks. Preventative maintenance is expensive to perform and the serviceable life of equipment and components is not maximised. Corrective maintenance maximises the serviceable life of equipment but risks damage to other equipment when failures occur. Regardless of which approach is taken, unexpected equipment failures result in equipment downtime, and thus the necessary maintenance will always be *reactive*. Consequently, the resulting equipment downtime will be prolonged while the necessary spare parts, personnel, and equipment, necessary to carry out the required maintenance, are organised.

Condition-based maintenance (CBM) is a new maintenance philosophy involving the real-time analysis of equipment sensor data to infer maintenance condition, or health. Maintenance activities are then performed on the basis of necessity, as identified by a condition-based maintenance system. In comparison with traditional maintenance philosophies, CBM offers the potential for minimising instances of equipment failures, a reduction in scheduled maintenance activities, maximisation of the serviceable life of life-limited components, and increased equipment availability. Critical to the success of implementing a condition-based maintenance philosophy however, are the necessary technical capabilities to infer equipment condition from real-time process measurements, so that informed maintenance decisions can be made.

The development of the technical capabilities to implement a condition-based maintenance philosophy, including the development of real predictive prognostics which estimate the remaining useful life of degrading equipment, are of major interest across almost all industrial environments in which the availability, reliability and performance of machinery is critical. However, developing such capabilities is a significant technical challenge. In this thesis, two industrial environments, namely semiconductor manufacturing and wind turbine power generation, are investigated in an effort to develop appropriate condition monitoring solutions.

1.1.1 CBM for Semiconductor Manufacturing

The International Technology Roadmap for Semiconductors (ITRS) is an industry grouping that identifies critical challenges and provides assessments of the semiconductor industrys future technology requirements. The future needs identified are used to direct research and development efforts among manufacturers, research facilities, universities, and national labs worldwide. In the latest ITRS annual report (2012), the importance of developing predictive maintenance capabilities for the semiconductor industry, and the associated challenges, is a topic which receives significant focus. This report describes how “a key challenge in the migration from reactive to predictive (maintenance) is the ability to establish accurate, robust, reconfigurable, real-time updateable and understandable models that are the basis for prediction. A key focus for prediction will be techniques for improving prediction accuracy and for utilising prediction accuracy information (along with the prediction itself) to optimise prediction systems.” [10]. Addressing exactly these challenges is the primary topic of this thesis.

1.1.2 CBM for Wind Turbines

The economic exploitation of wind energy is largely dependent upon the high reliability of wind turbines and their components. Wind turbines operate in harsh environments which generate large loads on wind turbine blades, which can lead to faults and failures in wind turbine components. In addition, with wind farms increasingly being located offshore, the costs of performing both scheduled and unscheduled maintenance are even greater. Studies have suggested that maintenance costs can consume up to 20 to 25% of the total income generated, and that a considerable percentage of these costs are due to unexpected equipment failure, which require corrective maintenance [11]. As a result, wind farm operators are keen to exploit condition-based maintenance in an effort to reduce overall maintenance costs.

1.2 Objectives

The main objectives of this thesis are to develop condition monitoring and prognostic capabilities for high-value critical equipment within the domains of semiconductor manufacturing and wind turbine power generation. The thesis aims to exploit system data which is already collected as part of regular data collection and monitoring activities within the semiconductor manufacturing environment or, in the case of wind turbines, data currently collected as part of supervisory control and data acquisition (SCADA) systems, which are installed as standard on most modern wind turbines. The primary theme in achieving these objectives is that all the data utilised was collected from real-world operational environments. As a result, the developed solutions have demonstrated capabilities and applicability to real world condition monitoring applications.

Within the semiconductor manufacturing environment, a wide variety of different equipment is used during each manufacturing process. Historically, for condition monitoring purposes, the data recorded by each piece of equipment is usually monitored in isolation. However, analysis of the data recorded by different equipment, such as semiconductor wafer processing tools and dry vacuum pumps, has identified potential benefits which might be realised by incorporating different signal data from different equipment in the development of condition monitoring solutions. As a result, the first objective in this thesis is to demonstrate the benefits of incorporating upstream process measurements in the development of a condition monitoring and fault prediction algorithm for dry vacuum pumps used in semiconductor manufacturing. The solution is developed exclusively on data collected from a large semiconductor manufacturing facility.

The second objective of the thesis is to develop a complete prognostic and health management (PHM) system for thermal abatement systems used in semiconductor manufacturing, in particular the ceramic liner component used within the thermal abatement system. Studies of available historical data has identified recurring issues with the thermal abatement ceramic liner. The development of a PHM system for thermal abatement systems has the potential to deliver major benefits including reduced maintenance costs, reduced in-service failures and increased equipment uptime. The dataset utilised in achieving this objective was again collected from a large semiconductor manufacturing facility.

The final objective of this thesis is to demonstrate how existing supervisory control and data acquisition (SCADA) systems can be exploited to detect and monitor the development of fault conditions within the main bearing of a large wind turbines. In addition, prognostic capabilities for wind turbines are developed which provide operators with an estimate of remaining useful life. The data utilised for this task was collected from a large operational windfarm.

1.3 Contributions of this Thesis

This thesis claims the following original contributions:

1. The development of a condition monitoring algorithm to detect, track, and predict the development of fluorine gas induced degradation in mechanical dry vacuum pumps used in semiconductor manufacturing. The notable feature of this algorithm is that, for the first time, upstream process measurement have been incorporated, in the form of foreline pressure measurement, which provide a means to quantify and model the relationships between pump process variable changes and the actual level of pump degradation. A comparison between multiple linear regression and artificial neural networks is demonstrated, and a double-exponential smoothing technique is applied to predicting the remaining useful life of a degrading pump.

2. The development of a multi-modal signal tracking algorithm for tracking changes in the signal distribution of the temperature signal generated by thermal abatement system used in semiconductor manufacturing. The proposed solution incorporates a novel approach to tracking changes in the non-Gaussian distribution of the generated signal. The non-Gaussian, multi-modal distribution of the monitored signal is tracked over time using a moving-window, whereby at each iteration a Gaussian mixture model is fit to the signal values within the bounds of the moving window. To the best of the authors knowledge, such an approach to tracking a non-Gaussian signal has not been demonstrated previously.

3. The development of a comprehensive multiple model particle filtering approach for predicting the remaining useful life of the ceramic liner used in thermal abatement systems for semiconductor manufacturing. This is the first demonstrated application of a multiple model particle filtering approach for prognostics.

4. A novel model weighting scheme for use within the multiple model particle filtering framework is developed which is demonstrated to improve both the accuracy and precision of long term predictions generated by the multiple model particle filtering framework for thermal abatement systems.

5. The development of a condition monitoring and prognostic capabilities for the main bearing on a large utility scale wind turbine. The main feature of this contribution is that the capabilities are developed exclusively using already installed onboard supervisory control and data acquisition systems. No previous work has considered the application of such data to main bearing condition monitoring. The demonstration of a prognostic solution for the main bearing on a wind turbine is also

presented. Prognostics for wind turbines has not previously been demonstrated in the relevant literature and the proposed solution, using the multiple model particle filtering framework, provides the necessary capabilities for addressing and representing the uncertainty regarding how a turbine fault will evolve, considering the uncertainty in future operating load.

A final notable feature of the work developed within this thesis is that the software and algorithms developed relating to dry vacuum pumps and thermal abatement systems was successfully commercialised and licensed to a large semiconductor equipment manufacturer, Edwards Vacuum (formerly BOC Edwards). Edwards are a major supplier of both vacuum and abatement systems to semiconductor, flat panel display, and solar panel manufacturers worldwide, with an annual turnover of almost $1 billion. The objective of the commercialisation agreement was that the technologies developed as part of this thesis will eventually be incorporated within Edwards existing condition monitoring software, FabWorks. As a result, there is the potential that the relevant algorithms developed within this thesis will, sometime in the future, be installed and operating in semiconductor manufacturing facilities worldwide. Another point to note is that the commercial agreements with Edwards Vacuum did not result in any major constraints regarding publishing the results of the work undertaken.

1.4 List of Publications

1. Butler, S., Ringwood, J.V., and MacGearailt, N., "Prediction of Vacuum Pump Degradation in Semiconductor Processing", *IFAC Symposium SAFEPROCESS: Fault Detection, Supervision and Safety for Technical Processes*, oral presentation, Barcelona, Spain, June/July 2009, pp. 1635–1640.

2. Butler, S. and Ringwood, J.V., "Particle filters for remaining useful life estimation of abatement equipment used in semiconductor manufacturing", *IEEE Conf. on Control and Fault-Tolerant Systems*, oral presentation, Nice, France, Oct. 2010, pp. 436-441.

3. Butler, S., Ringwood, J.V., "Particle Filtering for Prognostics of Thermal Abatement Systems", *Intel European Research & Innovation Conference*, oral presentation, Leixlip, Ireland, Oct. 2011.

4. Butler, S., Ringwood, J.V., and Mooney, M., "Multiple Model Particle Filtering for Prognostics of a Thermal Abatement System", *Reliability Engineering & System Safety*, submitted for publication Jan. 2012.

1.5 Thesis Layout

The layout of this thesis is as follows:

Chapter 2 provides an introduction to condition-based maintenance. The objective of this chapter is to provide the unfamiliar reader with a comprehensive overview of the main issues, benefits, challenges, and techniques involved in developing condition-based maintenance systems.

Chapter 3 introduces the mathematical techniques used in the development of condition monitoring and prognostic algorithms in this thesis.

Chapter 4 describes the development of a condition monitoring solution for dry vacuum pumps used in semiconductor manufacturing.

Chapter 5 describes the development of a Gaussian mixture model based condition monitoring solution for thermal abatement devices used in semiconductor manufacturing. The application of Gaussian mixture models for multi-mode pump tracking and condition monitoring in dry vacuum pumps is also presented.

Chapter 6 presents the development of a prognostic solution for thermal abatement devices. The developed solution uses a multiple model particle filtering approach to address the uncertainty challenges involved in predicting the remaining useful life of thermal abatement devices, where future operating loads and conditions are uncertain.

Chapter 7 describes the development of a condition monitoring solution for the main bearing on large utility scale wind turbines. The adaptability of the multiple model particle filtering approach developed in Chapter 6 is demonstrated for predicting the remaining useful life of a degrading main bearing.

Chapter 8 presents the general conclusions which can be drawn from the body of research presented in this thesis and briefly discusses potential future work arising from this research.

Chapter 2

Background

2.1 Background

Maintenance activities on critical equipment and systems have traditionally employed one of two maintenance philosophies; preventative or corrective. Preventative (or schedule-based) maintenance approaches use time-based intervals, or derived statistics such as mean-time-between-failures (MTBF), to schedule maintenance activities. An example of a preventative maintenance approach would be the performance of maintenance inspections and overhauls on aircraft engines, once the engines have been operated for a certain number of flight hours. Alternatively, the corrective maintenance approach entails operating equipment until it fails and then restoring it to good health. However, the corrective approach can only be applied to certain types of equipment, where system failure does not risk human safety.

Both preventative and corrective maintenance approaches have financial implications associated with them. The use of worst-case failure rate statistics to determine maintenance scheduling often results in conservative estimates regarding the likelihood of equipment failure. This can result in components regularly being replaced long before they have reached the end of their serviceable life. Alternatively, the use of a corrective maintenance approach ensures that the serviceable life of components is maximised. However, once a component does fail, it may cause damage to other parts of a system, resulting in significant repair costs with associated downtime and loss of revenue. The common factor to both of these approaches is that the “actual” condition of the equipment, before failure, is generally not taken into account when planning maintenance activities.

As systems and equipment become more complex and expensive, and increasing competition drives industries to become more lean and efficient, industrial and military communities are becoming increasingly concerned about system reliability and availability [3]. In many industries using complex machinery, the need to reduce maintenance costs, minimise the risk of catastrophic failures, and maximise system availability is leading a drive toward a new maintenance philosophy. Condition-based maintenance (CBM), or predictive maintenance, represents a new maintenance philosophy, whereby maintenance activities are only performed when there is objective evidence of an impending fault or failure condition, whilst also ensuring safety, reliability, and reducing overall total life costs [1],

The goal of a CBM approach is to optimise the availability of high-value critical assets, whilst also reducing overall maintenance and logistics costs. By performing maintenance only when there is evidence of abnormal behaviour, CBM programs aim to reduce maintenance costs by minimising the number of scheduled preventative maintenance actions, thus minimising the requirement, and cost, of maintaining a large inventory of spare parts, whilst also avoiding, potentially catastrophic, in-service equipment failures. The promise of reduced maintenance costs and increased availability is leading a change in maintenance philosophies, away from the traditional preventative and corrective approaches, toward a more condition-based approach.

2.2 Condition-Based Maintenance

Condition-based maintenance entails continuous monitoring of system data to provide an accurate assessment of the health, or status, of a component/system and performing maintenance based on its observed health. It involves using real-time system monitoring and data processing. Another capability that may form part of a CBM system is an ability to provide an estimate of the remaining useful life (RUL) of the system or component being monitored. This type of functionality is known as *prognostics*, as opposed to *diagnostics* which is used to assess the *current* condition of a monitored system.

A condition-based maintenance approach promises a range of improvements over existing approaches, with a potential reduction in overall maintenance costs being one of the primary drivers for developing such approaches. The cost associated with each of the various maintenance approaches is depicted in Figure 2.1. A corrective maintenance approach has a relatively low maintenance cost (minimal preventative actions), but high performance costs associated with the high cost of operational failures. In contrast, preventative maintenance generally has a low operating cost, associated with reduced instances of in-service failures, but often uses very conservative estimates regarding the probability of component failures and so has a high maintenance cost, associated with the removal of components before they have reached the end of their useful lives. It would seem, therefore, that the most cost efficient approach is to undertake maintenance when there is objective evidence of need, i.e. condition-based maintenance.

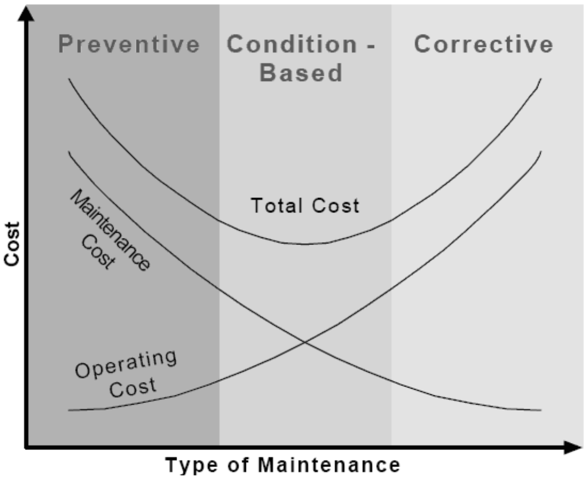


Figure 2.1: Costs associated with different maintenance approaches [1]

The development of CBM approaches has been enabled by developments and advancements in sensor technologies, data collection, storage and processing capabilities, and continuous improvements in algorithms and data analysis techniques. CBM systems are founded upon the ability to infer equipment condition using data collected from monitored systems. Ideally, a complete CBM system will incorporate both diagnostic and prognostic capabilities. The distinguishing factor between diagnostic and prognostic capabilities is the nature of the analysis. Diagnostics involves *posterior* event analysis (i.e. identifying the occurrence of an event which has already happened), while prognostics is concerned with *prior* event analysis (i.e. predicting the future behaviour of a system under observation) [12]. Sections 2.2.1 and 2.2.2 present an introduction to the principles of fault diagnostics and prognostics.

2.2.1 Diagnostics

The foundation of any CBM approach are robust and reliable fault diagnostic capabilities. Fault diagnostic algorithms are designed to detect system performance, monitor degradation levels, and identify faults (failures) based on physical property changes, through detectable phenomena [3]. Ideally, such systems will also identify the specific subsystem and/or component that is failing, as well as the specific failure mechanism that has occurred.

Fault diagnostic capabilities have been in development for over 50 years in various application domains. Some of the earliest fault diagnostic capabilities were developed in the form of built-in test (BIT) equipment for early generation aircraft [3]. In the intervening period, continuous developments and improvements in computer power and data storage capabilities have been reflected in the continuing development of more complex and capable fault diagnostic capabilities. Such capabilities have continued to improve such that, in many application domains, it is often possible to identify the presence of incipient fault conditions, which occur prior to equipment failure. Such capabilities enable maintenance personnel to potentially avoid catastrophic failures and reduce overall equipment downtime. In addition, such capabilities have driven efforts to develop capabilities beyond fault diagnostic capabilities, namely prognostic capabilities.

The term *fault diagnostics* is typically used to describe a broad range of tasks and capabilities. Well defined and accepted definitions for the various tasks covered by the term fault diagnostics have not yet become standardised, however, within the CBM community the following terms are becoming accepted definitions [3]:

- *Fault diagnosis* is concerned with detecting, isolating, and identifying an impending, or incipient, failure condition in a system. The term *fault* implies that the system under observation is still operational, but cannot continue operating indefinitely without maintenance intervention.
- *Failure diagnosis* is concerned with detecting, isolating, and identifying a system that has ceased to operate.

In the descriptions provided above, the terms fault *detection*, *isolation*, and *identification* generally imply the following meaning:

- Fault (failure) *detection* involves identifying the occurrence of a fault, or failure, in a monitored system, or the identification of abnormal behaviour which may be indicative of a fault condition.
- Fault (failure) *isolation* involves identifying which component/subsystem/system has a fault condition, or has failed
- Fault (failure) *identification* involves determining the nature and extent of a system fault condition or failure

2.2.2 Prognostics

To enable the benefits of a truly condition-based maintenance philosophy additional capabilities, beyond the realm of diagnostics, are required. Prognostics capabilities are designed to provide maintenance personnel with insight into the future health of a monitored system. To understand the realm of diagnostic and prognostic capabilities consider Figure 2.2, which illustrates the failure progression timeline of a typical system component. At the start of the components life, it is considered to be in proper working order and, after some time, an incipient fault condition develops in the component. As time progresses, the severity of the fault condition increases until the component eventually fails. If the system is permitted to continue operating, there is the potential that further damage may be caused to other secondary components or systems.

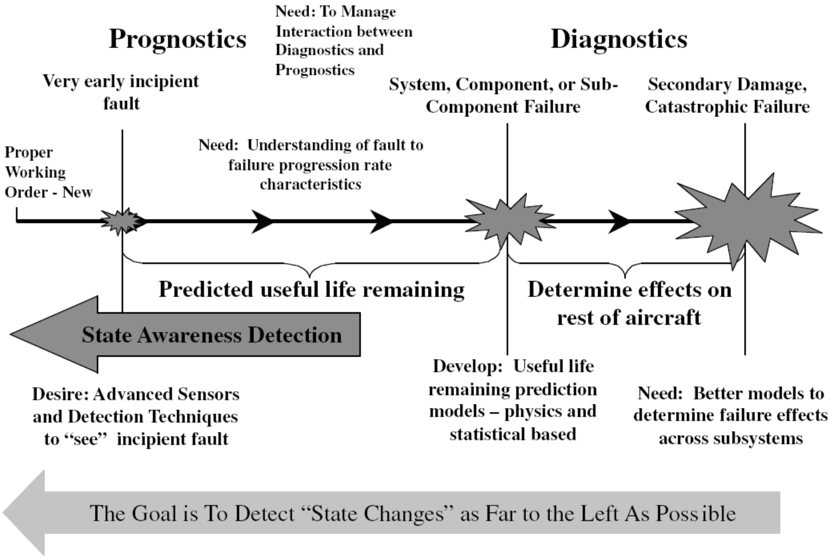

Prognostics
Need: To Manage Interaction between Diagnostics and Prognostics
Diagnostics
Very early incipient fault
System, Component, or Sub-Component Failure
Secondary Damage, Catastrophic Failure
Proper Working Order - New
Need: Understanding of fault to failure progression rate characteristics
Predicted useful life remaining
Determine effects on rest of aircraft
State Awareness Detection
Develop: Useful life remaining prediction models – physics and statistical based
Desire: Advanced Sensors and Detection Techniques to "see" incipient fault
Need: Better models to determine failure effects across subsystems
The Goal is To Detect "State Changes" as Far to the Left As Possible

FIGURE 2.2: Failure progression timeline [2]

The application domain of diagnostics has typically occurred at the point of component failure, or on the interval between component failure and eventual system-wide failure. However, if a fault condition can be detected at an early incipient stage, then maintenance actions can be delayed until the fault progresses to a more severe state, but before failure occurs. This interval, between the detection of an incipient fault condition and the occurrence of failure, defines the realm of prognostics. Provided a sufficient interval, commonly referred to as the lead-time interval (LTI) exists, between incipient fault detection and system failure, this enables a range of operational and maintenance benefits to be realised. With sufficient warning of upcoming maintenance events, remedial maintenance work can be planned in advance, with the necessary resources and personnel allocated as necessary. This capability is key to reaping the benefits of a truly condition based maintenance and delivering major costs savings.

From a high-level perspective, prognostics has the potential to deliver major improvements over more traditional maintenance approaches, including both reduced operational and support (O&S) costs and complete life-cyle total ownership costs (TOC), whilst also improving the safety of operating complex machinery and processes [3]. With the provision of a sufficient lead-time between the detection of an incipient fault condition and the occurrence of equipment failure, maintenance actions can become proactive instead of reactive, allowing necessary remedial maintenance work to be planned in advance. This is in stark contrast to more traditional maintenance approaches, in which equipment failure typically occurs without prior notice, leading to delays in organising the necessary personnel, spares, and tools, necessary to return the equipment to good health.

Furthermore, the costs associated with in-service equipment failure can often go far beyond the costs associated with repairing the failed piece of equipment. Consider the occurrence of equipment failure within a large manufacturing facility or on a commercial airliner. The failure of critical equipment in a manufacturing facility, such as a semiconductor fabrication (fab) plant, can potentially lead to major downtime on a specific manufacturing process. Such failures can potentially reduce the overall throughput of a manufacturing facility, resulting in incurred costs which can go far beyond the actual maintenance repair costs. In the case of a major fault on a commercial airliner, with passengers waiting at the gate, the costs can also quickly escalate beyond the actual repair costs. Furthermore, given the high utilisation rates of modern aircraft the effects of in-service failures can lead to delays for an extended period of time. Considering these issues, the potential benefits of accurate and reliable prognostic capabilities are obvious.

To enable the benefits of prognostic capabilities to be realised, maintenance staff need a reliable estimate of how long a system can continue to be operated safely, i.e. the remaining useful life (RUL) of the system, until a detected fault condition progresses to a failure condition. The generation of accurate predictions of RUL is the challenge presented in the development of prognostic algorithms. Since prognostics is associated with predicting the future, it inherently involves a large degree of uncertainty. Indeed, the task of prognostics is considered to be significantly more difficult task than diagnostics, since the evolution of equipment fault conditions is subject to stochastic processes which have not yet happened [13]. Later, in Section 2.5, a review of the main issues, challenges, and enabling technologies, used in the development of prognostic algorithms is presented.

2.3 Prognostics & Health Management

In recent times, the term prognostics & health management (PHM) has emerged, to describe systems developed to implement a CBM philosophy. The term PHM originated from the military applications and was the name given to the capability being developed for the new F-35 Joint-Strike Fighter (JSF) to enable the vision of autonomic logistics and to meet the overall affordability and supportability goals of the latest military fighter aircraft [14].

In the development of a PHM system, the term prognostics takes on a much wider definition than fault prediction and is used to describe a wide variety of activities including fault/failure detection, fault/failure isolation, enhanced diagnostics, material condition assessment, performance monitoring, and prognostics [14]. Within a PHM system, the capability to predict RUL is often termed as 'real predictive prognostics' to distinguish the capability from the more broader description of prognostics within a PHM system [2]. Figure 2.3 illustrates the typical stages within a CBM/PHM system, from appropriate signal pre-processing and feature extraction, fault detection and classification, to the prediction (prognostics) of RUL and finally, appropriate maintenance scheduling.

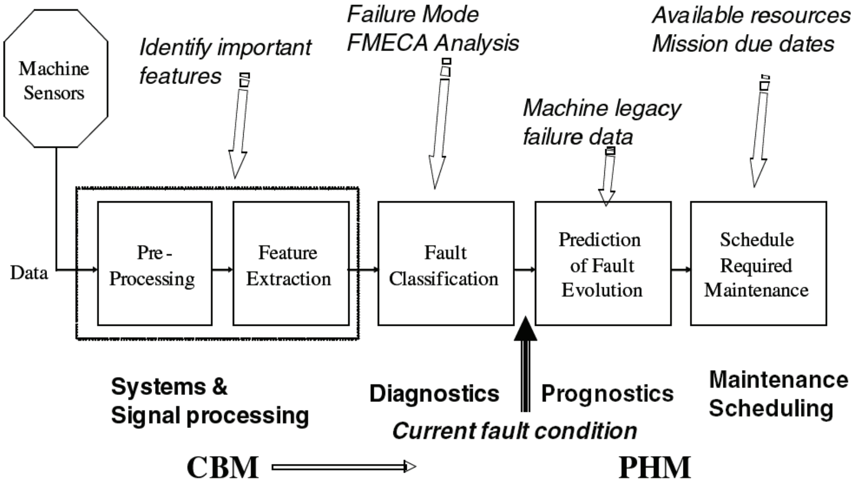

Machine Sensors
Identify important features
Failure Mode FMECA Analysis
Available resources Mission due dates
Machine legacy failure data
Data
Pre-Processing
Feature Extraction
Fault Classification
Prediction of Fault Evolution
Schedule Required Maintenance
Systems & Signal processing
Diagnostics
Prognostics
Maintenance Scheduling
Current fault condition
CBM
PHM

Figure 2.3: Stages within a typical PHM system [2]

The development of a comprehensive health management system goes beyond the ability to perform accurate diagnostics and real-predictive prognostics. Such systems will often include many types of functional capabilities which are designed to complement each other and deliver a greatly increased range of maintenance benefits, beyond what is possible in the application of any single capability. The following list describes some of the functionality and capabilities that might be contained within a modern PHM system, as described by Hess *et al.* [14].

- Fault detection / isolation
- Advanced diagnostics
- Predictive prognostics
- RUL and time-to-failure predictions
- Component life usage tracking
- Warranty guarantee tracking
- Health reporting and information management
- Utilisation tracking
- Decision support systems
- Fault accommodation
- Information fusion and reasoners

Beyond the obvious direct maintenance benefits of implementing a comprehensive PHM system, a range of other opportunities are presented from both a cost savings and business opportunity perspective, which reflect the original drivers for the development of PHM capabilities on the JSF. The provision of a lead-time, between detection of an incipient fault condition and actual system failure, is a key enabler of the concept of *autonomic logistics*. In most industries, where equipment uptime and availability is key, it is common for a large inventory of replacement components and parts to be maintained at all times. The procurement and logistics of managing a large inventory of spare parts is generally undertaken at significant cost. However, with the provision of a lead-time before failure, this presents an opportunity for improvements in how such logistic systems are managed. By maintaining a smaller inventory of replacement parts, PHM systems can be integrated into the logistics system, automatically ordering spare parts for those systems in which incipient failures are detected. In this way, the principle of just-in-time (JIT) manufacturing can be applied to the maintenance of critical equipment, reducing on-site inventory and the associated costs.

A further opportunity presented by the development of PHM systems are improvements to the implementation of *performance based contracting*. In many industries, it is common for manufacturers, or operators of complex systems, to outsource the maintenance of such systems to the original equipment manufacturer (OEM). In this way, operators can make major costs savings by not having to hire and train their own staff to maintain such equipment. Operators can also benefit from the knowledge, expertise, and worldwide support often offered by OEMs. A high profile example of such maintenance outsourcing is the "power-by-the-hour" concept offered by Rolls-Royce, in which Rolls-Royce take full responsibility for all maintenance, spares, and replacements, of aircraft engines installed on an airliners fleet. Rolls-Royce then receive payment for each hour an engine is operated [15].

The implementation of performance based contracting provides incentives for OEMs to make continuing investments in the development of PHM systems. Historically, OEMs and suppliers would have earned a large portion of their after-sales revenue from sales of replacement components to equipment operators. In such situations it was in the interests of OEMs that manufacturers would regularly swap out components at conservatively selected intervals as part of a preventative maintenance program. However, under the new performance-based maintenance contracts, it is now more beneficial for OEMs to maximise the service-life of systems and components. With the introduction of performance-based maintenance contracts, PHM technologies have the potential to drive a new business philosophy regarding after-sales service and revenue, whilst also ensuring no conflicts of interests between both parties. PHM systems, and, in particular, real predictive prognostics capabilities, have the potential to dramatically reduce the costs of providing maintenance contracts to equipment operators, whilst also improving the OEMs profit margins on such contracts. The sale of developed prognostic technologies could also provides a new and growing source of after-sales revenue. Equipment operators would also benefit from increased availability of equipment, ensuring a continuing good relationship between the two parties and potentially driving future business opportunities. Furthermore, the demonstrated benefits of such technologies can potentially motivate the continuing development of improved PHM technologies by OEMs, which can drive future after-sales revenues, and also provide OEMs with a competitive advantage over their competition, when bidding on future contracts.

Whilst the potential benefits of real predictive prognostics are obvious, and help explain the growth in related research in recent years [16], there are many reasons that such technologies have not yet become commonplace. The primary difficulty associated with the development of real predictive prognostics is the large level of uncertainty associated with the generation of long-term predictions of equipment health. How this uncertainty is represented and managed is is key to the success of any future prognostic technologies.

2.4 Fault Diagnostics

For several decades researchers and practitioners have been investigating and developing different techniques for failure detection, isolation and identification across a wide range of application domains in science, medicine and engineering. A comprehensive review of the techniques and methods used in fault diagnostics is beyond the scope of this document. Indeed, as described by Vachtenvanos *et al.* [3] “the diversity of application domains in fault diagnostics is matched only by plurality of enabling technologies that have surfaced over the years, in attempts to diagnose detrimental events”. For the interested reader, a series of review publications by Venkatasubramanian *et al.* [17–19] and Jardine *et al.* [12] provide an excellent introduction and reference source to the different approaches and techniques used in fault diagnostics, and the different applications to which such techniques have been applied.

In addition to the development of fault diagnostic capabilities for specific application domains, a number of fault diagnostic related issues are also often considered in the development of PHM solutions. Two such issues are failure modes and effects criticality analysis (FMECA) studies and feature extraction techniques. Section 2.4.1 discusses the issues and objectives in developing FMECA studies, which is followed in Section 2.4.2 by a brief overview of feature extraction methods. Finally, Section 2.4.3 presents a brief overview of the main approaches and techniques employed in the development of fault diagnostic capabilities.

2.4.1 FMECA Studies

A common first stage in the development of a PHM systems is a comprehensive FMECA study. The objective of FMECA studies is to relate failure events to root causes [3]. As part of this objective, FMECA studies investigate all relevant issues regarding potential failure modes of monitored systems including: the severity of different failure modes, their frequency of occurrence, their testability, the fault symptoms which are suggestive of a systems behaviour under different fault conditions, and the sensors and monitoring equipment required to monitor and track fault symptomatic behaviour.

In addition, advanced FMECA studies may also try to identify appropriate methods for identifying optimal features, or indicators, which can be used for detecting and isolating different fault conditions. FMECA studies are also often used as the starting template for the development of fault diagnostic capabilities, focusing efforts on addressing those issues which will provide the greatest maintenance benefits. The development of a FMECA study typically requires input from a diverse variety of sources including domain experts, maintenance personnel, equipment specialists, and designers. An excellent overview of the principle of FMECA studies and an example of a developed study is presented in [3].

2.4.2 Feature Extraction

The first stage in any PHM system typically involves appropriate preprocessing of equipment sensor data. This stage is often referred to as *feature extraction.* Feature extraction is the process of extracting useful information from raw signal data [12]. The feature extraction stage within a PHM system is designed to generate a vector of data features, which can be used to infer the current fault status of a monitored system.

The generation of an appropriate feature vector is typically application dependent and is one of the most important stages in a PHM system. For example, in the case of a vibration monitoring system, the feature extraction stage may be used to identify critical values relating to the magnitude of a vibration signal at critical values such as the gear-mesh frequency in a gearbox. Other examples of feature extraction could include identifying the parameters of a signal distribution for the purposes of tracking changes in such distribution which relate to developing fault conditions.

2.4.3 Fault Diagnosis Methods

From the highest level, fault diagnosis methods can be classified into one of two types of approaches, *model-based* and *data-based* [3]. Sections 2.4.3.1 and 2.4.3.2 present a brief review of both approaches.

2.4.3.1 Model-based approaches

Model-based fault diagnostic approaches employ a mathematical model of the system under observation. Using such a model, estimates of system/process outputs are generated which are then compared with the actual process outputs to generate a residual signal, or innovation. Based upon a comparison of the model outputs and the actual system outputs, potential fault conditions are identified on the basis of the values and properties of the generated residual signal. Figure 2.4 illustrates the basic concepts of a typical model-based approach to fault diagnostics.

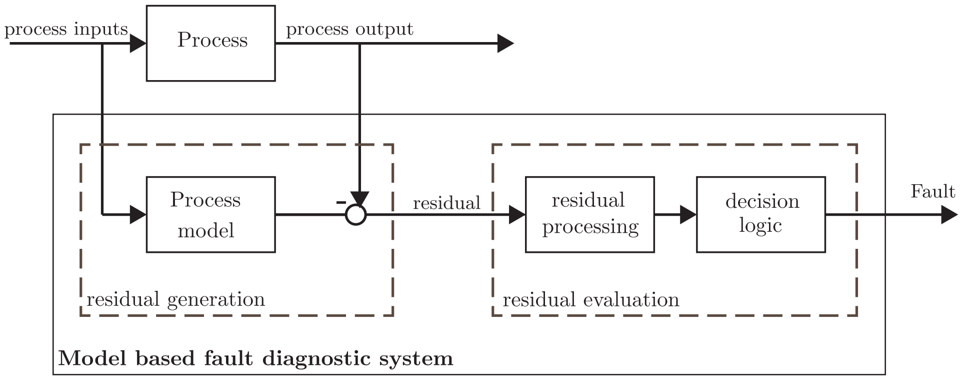


Figure 2.4: Model-based diagnostic approach

As indicated in Figure 2.4, a comparison between the actual system output and a model estimate are used to generate a residual signal. This process is known as *residual generation.* During fault-free operation, the value of residual signal should be approximately zero, indicating that the model, which describes fault-free behaviour, accurately describes the current behaviour of the system under observation. In the situation where the value of the residual signal deviates from zero, appropriate processing and analysis is applied to the residual signal. The appropriately processed residual signal is then forwarded to a decision logic routine which is used to map the behaviour of the residual signal onto a specific fault condition. This process is described as *residual evaluation.*

Whilst Figure 2.4 illustrates the general principle of model-based fault diagnostics, classical model-based fault diagnostic techniques can be further categorised. Isserman and Balle [20] categorise specific approaches into one of three types: *parameter identification*-based methods, *parity equation*-based methods and *observer*-based methods. Each of these approaches are briefly described.

Parameter identification based methods employ a dynamic model of the system under observation in which input/output data is used to estimate the values of the model parameters, using appropriate system identification techniques. Deviations in the values of the identified parameter values, over time, are used to identify the presence of a fault condition.

State and output observer based methods also employ a model of fault-free behaviour. In this case, system inputs are used to estimate system state variables, using state estimation techniques such as the Kalman filter. The estimated state variables are then used to reconstruct the system outputs which are then compared with the actual system outputs to generate a residual, which can be used to indicate faults.

Parity equation based methods compare the behaviour of an observed system with a process model which describes normal, non-faulty, behaviour. The basic principle is to check the parity (consistency) of the mathematical equations describing the observed system, also described as the analytical redundancy relations, using the actual system measurements. A fault is declared once the preassigned error bounds are surpassed. This approach has close similarities with the observer-based approaches [21].

The key feature of a model-based approach to fault diagnostics is the requirement for an accurate and robust mathematical model of the system under observation. Such models are usually derived from first principles, using ordinary differential equations, so that the different elements of the model relate to actual physical properties. Physical models of systems under observation are usually converted into state-space format, before applying the techniques described above.

A major benefit of model-based fault diagnostic approaches is the capability of detecting unanticipated faults [3], since the models employed are usually based upon the physics of failure. This is in contrast to data-driven approaches, which typically require historical examples of each fault condition they are designed to detect. However, in many real-world situations, it is not possible to apply model-based diagnostic approaches, since many processes are to complex to develop accurate mathematical models for.

2.4.3.2 Data-driven approaches

The general principle of data-driven approaches to fault diagnostics is to utilise pattern recognition techniques to map data in the measurement, or feature, space, to equipment faults within the fault space [12]. A wide variety of techniques have been applied to fault diagnostic problems. Jardine *et al.* [12] categorises data-driven approaches into two types; statistical approaches and artificial intelligence (AI) based approaches. Within these categorisations, a diverse range of techniques have been applied to a wide variety of fault diagnostic problems. In the following sections, a brief overview of methods and applications is presented.

Statistical Approaches A widely applied data-driven technique in fault diagnostics, taken from the domain of control theory, is statistical process control (SPC). SPC is used to measure deviations in signal behaviour about a predefined range or distribution. If a signal deviates outside the defined control limits this may be indicative of a fault condition. An example of multivariate SPC applied to fault detection in a semiconductor etch chamber is presented in [22].

Another commonly employed statistical approach is principal component analysis (PCA) and partial least squares (PLS). PCA is often applied to high-dimensional datasets to transform a number of related variables to a smaller set of uncorrelated variables, i.e. dimensionality reduction. The basic principle of PCA for fault diagnostics is to derive a PCA model using a dataset of normal fault-free behaviour. Future observations are then compared with this model using statistical measures such as the T^2 and Q statistics [19]. If the measured statistics exceed a defined limit, a potential fault condition is flagged. PLS is a multivariate regression algorithm based upon PCA. Whilst PCA is concerned with decomposing an input matrix X into its principal components, PLS is concerned with developing a linear regression model by first projecting the input matrix X and output matrix Y onto a lower dimensional space. A comprehensive overview of PCA/-PLS applied to fault diagnostic problems can be found in [23]. A more contemporary review of applications of PCA/PLS to fault diagnostics can be found in [24].

Classification techniques Data classification techniques, using both statistical methods and AI based methods, are some of the most widely applied techniques in data-driven fault diagnostics. The application of classification techniques relies upon the availability of a fault pattern library, or database, of historical failure examples, which relate extracted features from monitored systems to specific fault conditions. The objective in applying classification based techniques is to model the relationships between fault features, or fault indicator measurements, and fault classes [3]. Such approaches do not have the capability of model-based approaches which employ models, based upon the physics of failure, which are capable of detecting even unanticipated fault conditions. However, data-driven classification approaches can be constructed without the availability of a complex mathematical model of the monitored system or process. Some of the most common techniques applied for fault classification include artificial neural networks (ANNs), Bayesian networks, discriminant analysis, support vector classification, and fuzzy logic.

2.4.4 Novelty Detection

The development of fault diagnostic capabilities, particularly historical data-based approaches, rely upon the availability of historical failure examples from which the equipment behaviour, associated with fault (failure) conditions, can be "learned". However, in many real-world situations, the availability of sufficient historical failure data is often lacking, especially failure data which captures all the conceivable behaviour which might be observed in the presence of a fault condition. In contrast, data describing the fault-free behaviour of equipment is usually plentiful. Novelty detection methods, also known as *anomaly detection*, provide an approach for exploting such fault-free data in the development of fault diagnostic capabilities. The basic principle of novelty detection algorithms is to construct a model of a system, from collected system data, which describes its observed behaviour during normal, fault-free, operation. Once such models have been developed, future data collected is then compared with the model of fault-free operation and, using a distance metric and a specified threshold, abnormal or novel events are identified which might indicate the occurrence of a fault condition.

Novelty detection algorithms represent a type of classifier. However, instead of classifying the input data into a specific class, novelty detection algorithms act as a detector [25], in which the objective is to determine whether the input data is part of the data on which the classifier was trained, or if the input data is unknown. Novelty detection methods have been applied to a wide variety of applications including radar target detection, hand writing recognition and the detection of pathological features in mammograms.

In addition, novelty detection methods have seen wide application to fault detection problems. Similar to the development of fault diagnostic capabilities the variety of techniques applied in the development of novelty detection methods are vast. Some of the most commonly approaches involves constructing models which describe the distribution of fault-free signal behaviour. Such distributions are often multivariate and multi-modal. A range of parametric, non-parametric, and artificial intelligence based approaches are often employed to develop such models. An comprehensive review of novelty detection methods, and applications to which they have been applied, is available in a two part publication by Markou & Sing [25, 26].

2.5 Fault Prognostics

To enable the benefits of a truly condition-based maintenance philosophy, real predictive prognostic capabilities are required. Such capabilities are designed to provide maintenance staff with prior notice of pending equipment failure and ideally provide sufficient lead-time so that the necessary personnel, equipment and spare parts can be organised and deployed, thus minimising both equipment downtime and maintenance costs.

Real predictive prognostics is understood to be the generation of long-term predictions, describing the evolution of a signal of interest, or fault indicator, for the purpose of estimating the remaining useful life (RUL) of a failing system or component [27]. The primary difficulty encountered in the development of prognostic technologies is the significant uncertainty associated with the generation of long-term predictions of equipment health. For this reason, real predictive prognostics has been described as the Achilles' heel in implementing a comprehensive PHM system [3].

The development of prognostic capabilities is a difficult challenge owing to the inherent uncertainty associated with predicting the future behaviour of a degrading system. As a result, there are two key issues to consider when selecting appropriate techniques in the development of prognostics capabilities, *uncertainty representation* and *uncertainty management*. Uncertainty representation implies the ability to model various forms of uncertainty stemming from a variety of sources, whereas uncertainty management concerns the methodologies and tools needed to continuously "shrink" the uncertainty bounds as a fault evolves [3].

In this section, some of the issues involved in the development of prognostic algorithms are reviewed. Section 2.5.1 introduces the concept of the remaining useful life probability density function (PDF). Section 2.5.2 then reviews the different techniques which can be applied to predicting the RUL of systems under observation. The applicability and capabilities of the different methods are reviewed with reference to application examples of the different techniques. Finally, Section 2.5.3 presents a brief overview of the metrics which are available for evaluating and comparing the performance of different prognostic methods.

2.5.1 The Remaining Useful Life PDF

One of the key concepts within the prognostics framework is the RUL PDF. The RUL PDF is the output generated by a prognostic algorithm, describing the distribution in time of likely equipment failure times. Consider Figure 2.5, which illustrates the key concepts of a RUL PDF. At time t_P, a prediction is made and an estimate of the RUL PDF is generated. Once the RUL PDF has been generated, the next question is to decide when to carry out corrective maintenance actions. Ideally, the time chosen for maintenance action will both avoid equipment failure and maximise the useful-life of the equipment. However, these are conflicting requirements and, as a consequence, selecting when to perform maintenance is typically an exercise in risk management.

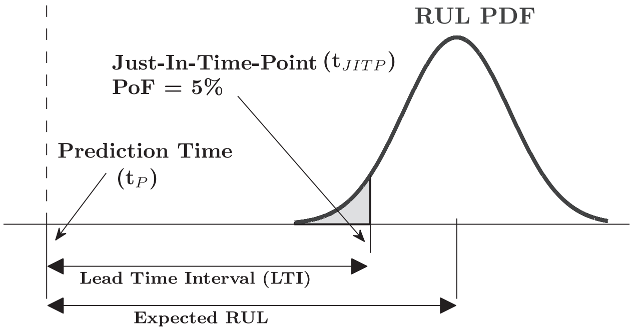


FIGURE 2.5: The remaining useful life PDF

In the development of a requirements specification for a prognostic algorithm, a key consideration will be the maximum allowable *probability of failure* (PoF). This value defines the maximum acceptable level of risk of equipment failure, beyond which equipment can no longer be operated as the risk of equipment failure is deemed excessive. Using the defined maximum allowable PoF and the estimated RUL PDF, an important value known as the just-in-time-point (JITP) can be identified. The JITP defines the latest point in time before which corrective maintenance actions must be carried out to avoid operating equipment beyond the maximum allowable PoF. In a real-life application, selecting the maximum allowable PoF would usually consider a number of factors.

The key factors include safety, criticality and economic considerations. In certain scenarios, where safety considerations are primary, the requirement might be to avoid as many in-service failures as possible, and thus a conservative value for the maximum allowable PoF might be chosen. Alternatively, a plant operator may accept a higher maximum allowable PoF value in situations where maximising the useful life of expensive equipment/components might be more economical than avoiding an occasional failure. An example of such a scenario might be the use of a diamond headed cutting tool.

In Figure 2.5, a maximum allowable PoF value of 5% is assumed for illustrative purposes. Once the JITP has been identified, another key measure can be computed, the lead-time interval (LTI). The LTI is defined as the time interval between the time the prediction is generated t_P, and the JITP t_{JITP}, so that

$$t_{LTI} = t_{JITP} - t_P \tag{2.1}$$

The LTI provides a real-time estimate of the remaining time before a system operates above the maximum allowable PoF. Maintenance actions must be performed before this time elapses, to avoid operating equipment beyond the maximum allowable PoF. The RUL PDF and the LTI value represent key information that should be presented to maintenance staff as part of the human-machine interface (HMI). This information allows for maintenance staff to make informed operational decisions, regarding when to perform maintenance and avoid instances of equipment failure. Additionally, the RUL PDF and LTI values can be used as inputs into an automated maintenance decision support system. As a failure approaches, the LTI value could be used to automatically generate maintenance tasks which must be performed to avoid system failure. In addition, both RUL and LTI estimates could potentially be incorporated into an autonomic logistics system, as described previously in Section 2.3, whereby replacement components and parts are ordered automatically when a predicted failure is approaching. This would allow for a smaller inventory of spare parts and components to be maintained onsite, further reducing total ownership costs (TOC).

2.5.2 Prognostic Techniques

In recent times, a wide variety of techniques have been applied to predicting the RUL of monitored systems, which is reflected in the diverse range of applications. Categorising the different approaches into different classes is difficult due to the wide variety of techniques which have been applied to prognostic problems. Vachtsevanos *et al.* [3] categorise prognostic approaches into one of three classes, *experience-based approaches*, *trending/data-driven* approaches and *model-based* approaches, as illustrated in Figure 2.6. The categorisation by Vachtsevanos *et al.* also relates different prognostic approaches to capability, applicability and complexity/cost. Figure 2.6 illustrates how, as we move from experience-based to model-based approaches, with increased capabilities and performance, there is a likewise decrease in the applicability of the different approaches. The reduction in applicability is a reflection of the increasing complexity/-cost of the different approaches, as the increased capabilities are achieved by adapting and tailoring solutions to specific prognostic applications.

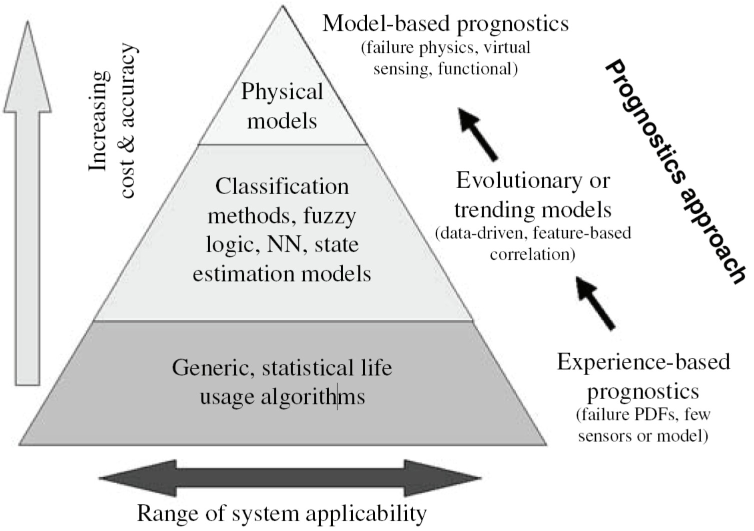

Increasing cost & accuracy
Physical models
Classification methods, fuzzy logic, NN, state estimation models
Generic, statistical life usage algorithms
Model-based prognostics
(failure physics, virtual sensing, functional)
Evolutionary or trending models
(data-driven, feature-based correlation)
Experience-based prognostics
(failure PDFs, few sensors or model)
Prognostics approach
Range of system applicability

Figure 2.6: Technical approaches to prognostics [3]

In the following sections, a review of the different technical approaches within these classes is presented, with reference to relevant applications of the different approaches.

2.5.2.1 Experience-Based Prognostic Approaches

The simplest prognostic approaches rely upon statistical information collected, which examines historical failure rates of systems or components. Such data can be used to develop life-usage models in terms of distributions of failure rates over time. Such approaches are used to develop preventative maintenance schedules in which maintenance is performed on the basis of mean time between failures (MBTF), which are derived from life-usage models. However, such approaches do not have any predictive capability and cannot be described as truly predictive prognostic techniques. However, such approaches have wide applicability in systems or components with low criticality or cost, or in situations where sensor data, which can be used to infer condition, is not available.

2.5.2.2 Model-Based Prognostic Approaches

The most capable prognostic approaches use *physics-of-failure* models of the system under observation, derived from first principles. The main application domain of such approaches, to date, have involved the use of fatigue models for modelling the initiation and propagation of cracks in structural components [28]. The main benefit of model-based approaches, using physics-of-failure models, is the ability to incorporate physical understanding of the process under observation, and additionally, the ability to predict degradation under different loads and operating conditions. However, model-based prognostic approaches are also limited by the ability to develop such high-fidelity models of often complex systems and processes. In many situations, where complex first principle models are not available, it is possible to assume a certain form for a dynamic model describing the evolution of a degradation process. Then, using observed inputs and outputs, the model parameters can be identified in a process known as *model identification* [3]. Prognostics approaches using such models are sometimes described as *hybrid* approaches, crossing the boundary between model-based and data-based prognostics.

With the availability of sufficiently descriptive models of a degradation process, either physics of failure models, or models derived to describe the behaviour of historical failure examples, the development of prognostic algorithms based upon the application of recursive Bayesian estimation techniques are possible. Dynamic model-based prognostic approaches based upon recursive Bayesian estimation techniques also provide a framework for addressing various sources of uncertainty in prognostics, and have been amongst the most successful approaches to prognostics. Recursive Bayesian estimation techniques combine information from the model, describing the degradation process, with measurements taken from the system under observation. In this way, the current level of degradation is modelled as a random variable, which allows for uncertainty in the current level of degradation to be quantified. In addition, the current level of uncertainty can then be propagated into the future to better describe the uncertainty in the predictions generated.

Various recursive Bayesian estimation techniques have been investigated for prognostic problems, such as the Kalman Filter, Extended Kalman filter, and Particle Filter. The Kalman filter is an optimal recursive Bayesian estimator in the situation where the process can be modelled by a linear state-space model with Gaussian noise processes. However, the dynamics of degradation process are typically non-linear, with noise processes that are often non-Gaussian [29]. In such situations, the Extended Kalman Filter, which linearises a non-linear state-space model around the current point, can be used [28]. However, as described by Saha *et al.* [30], if the initial estimate of the state is incorrect, or if the process is modelled incorrectly, the filter may quickly diverge, resulting in poor predictions being generated. In recent times, sequential Monte Carlo methods, more commonly known as particle filtering, have grown in popularity, due to their flexibility and ease of design in tackling *non-linear filtering* problems [29]. Indeed, particle filtering has been described as the de facto *state of the art* technique in failure prognostics [31]. The basic principle of particle filtering is to approximate the degradation state PDF with a set of *particles*, representing state values, and an associated set of particle *weights*, which represent the discrete probability masses of the individual particles. The particles can be generated and updated recursively using a non-linear state-transition model, describing the evolution of the process under observation.

The potential of particle filtering for prognostics was first demonstrated in a PhD thesis by Orchard [27], and since then particle filtering methods have been investigated and applied to a variety of prognostic applications including lithium ion battery prognostics [32], fatigue crack growth prediction [29, 33, 34] and pneumatic values [35]. In addition, particle filtering has been demonstrated to outperform alternative methods such as the EKF and ARIMA models in predicting the RUL of batteries [30].

2.5.2.3 Data-Based Prognostic Approaches

In many situations the complexity of the systems under observation makes it impossible to derive robust and accurate models which can be used for prognostic purposes. However, it is often the case that historical data, which captures the signal behaviour of measured signals or extracted features from the incipient fault stage to equipment failure, is available. In such cases, data-driven methods which model how such signals and features evolve can be utilised in generating predictions of RUL.

Data-driven prognostic approaches typically follow one of two strategies. The first strategy is a two-stage process. Firstly, appropriate dimensionality reduction, feature extraction, or pattern matching techniques are employed to map system signals or features onto a single dimension *damage*, *degradation*, or *health* index. Technically, this first step falls under the realm of fault diagnostics since it is concerned with posterior event analysis. Once the current level of degradation is identified, it is then extrapolated into the future until a predefined critical threshold limit is exceeded. A range of techniques can be applied in both of these steps. The second strategy is to directly model the relationship between monitored signals or features, and the remaining life of the system. In this situation, the remaining life of the system is the output generated by the models. In the following section, a brief overview of some of the data-based techniques which have been applied to prognostic problems are presented. More comprehensive reviews of data-based prognostic techniques can be found in [36–38]

Time Series Approaches The simplest data-driven approaches to prognostics rely upon projection methods, which project the current level of degradation into the future. This task is essentially a time series prediction problem and, within the realm of prognostics, has been addressed by a variety of approaches including autoregressive models [30, 39] and exponential smoothing techniques [40].

The autoregressive integrated moving average model (ARIMA) forms a general class of linear models that have historically seen wide use in modelling and forecasting time series [41]. Unsurprisingly, such approaches have also been applied to prognostic problems which are similar, in many respects, to forecasting problems. ARIMA models are derived from the more common autoregressive moving average (ARMA) model, which models a time series using two parts, an autoregressive (AR) part and a moving average (MA) part. However, since ARMA models can only be used to model stationary processes ARIMA models are often employed, which can be used to model non-stationary time series signals. Examples of ARIMA models applied to prognostic problems can be found in [30, 39].

Artificial Neural Networks Perhaps the most common data-driven technique applied to prognostic problems are artificial neural networks (ANNs). ANNs model relationships between input and output variables with a model structure inspired by the neural structure of the brain. The network weights and biases, which define the interconnections between the neurons, are adapted during a training process to maximise the fit between the input and output data on which the models are trained.

ANNs have been applied in a number of different ways for prognostics. The most common use of ANNs is in time series prediction, where the current degradation state is predicted into the future until it exceeds a threshold value. Typically, in a feed-forward ANN, previous values of the degradation index are used as the inputs to generate a one-step ahead prediction. The generated output is then fed back as an input to the next iteration, to generate long-term predictions. Examples of such applications on ANNs for prognostics can be found in [42, 43]. ANNs can also employed to estimate the current degradation index, using system features as inputs. The degradation index can then be predicted into the future using ANNs again, or via alternative prediction methods. Wang and Vachtsevanos [44] presented an application for bearing crack size prediction in which a wavelet neural network (WNN) is used to estimate the current crack size from extracted vibration signal features. A dynamic wavelet neural network (DWNN) is then used to predict the fault propagation process into the future and estimate the RUL. The first ANN used, the WNN, is a static feed-forward neural network used to derive a static relationship between inputs and ouputs. The second ANN used, the DWNN, is an example of a recurrent neural network (RNN), which incorporate feedback within the network structure to predict the timeseries evolution. Other examples of RNNs applied to prognostic problems can be found in [45, 46]. ANNs have also been used to directly model the relationship between system features and RUL [45]. ANNs learn by example and, as such, require sufficient instances of historical failure examples for training, and are typically data hungry. As a result, ANNs can generate poor prediction performance when future failure examples, on which the ANN was not trained, do not exhibit similar behaviour to the examples which formed part of the training set.

Other approaches Beyond time series and ANN based approaches, a range of other techniques have been applied to prognostic problems. Goebel [47] presented a comparison of ANNs against Gaussian process regression (GPR) and a relevance vector machine (RVM) approach. The intrinsic ability of GPR and RVM to generate confidence limits on generated predictions make such approaches attractive, whereas ANN approaches, in general, do not provide confidence limits associated with predictions. The availability of confidence limits associated with RUL predictions are highly desirable and provide a means for uncertainty management. Other data-based techniques which have been applied to prognostic problems include hidden Markov models [48], and Neuro-Fuzzy networks [49],

2.5.3 Prognostic performance metrics

The development of both online and offline metrics for evaluating and comparing the performance of different prognostic approaches has been an area of active research in recent years. Much of this work has been developed, and documented, in a series of publications by Saxena *et al.* [50–52]. These techniques were developed to address the perceived shortcomings in this area, particulary the unsuitability of standard forecasting metrics for prognostic problems. Much of the recent research on prognostic performance metrics has focused on the development of techniques for comparing the performance of different prognostic algorithms, when applied to the the same prognostic problem. A brief overview of some of the most useful metrics is provided below [52].

Prognostic Horizon evaluates how far, in advance of system failure, an algorithm can generate predictions within the desired accuracy range, around the actual failure time. The accuracy is described in terms of a percentage of the actual life of the system.

Relative Accuracy evaluates the percentage error of a prediction, relative to the actual RUL. In this way, a requirement for the algorithm to become more accurate as failure approaches is built into the metric.

Alpha-Lambda $(\alpha - \lambda)$ performance evaluates if an algorithms stays within desired performance levels, relative the actual RUL at a given time.

Chapter 3

Condition Monitoring and Prognostic Techniques

This chapter presents a review of the background and theory behind the mathematical techniques used in later chapters for the development of condition monitoring and prognostic algorithms. The layout of this chapter is as follows. Section 3.1 and Section 3.2 reviews the background and theory behind multiple linear regression models and artificial neural networks, respectively. These represent two techniques which can be used for modelling the relationship between input variables and an output variable. Both techniques are employed in Chapter 4, to model degradation in dry vacuum pumps. Section 3.3 introduces the background and theory behind Gaussian mixture models, which are the foundation of the condition monitoring algorithms developed in Chapter 5. Section 3.4 introduces particle filtering, which is a monte-carlo technique for implementing a recursive Bayesian filter in the presence of non-linear process models and non-linear noise processes. Particle filters are used in Chapter 6 to develop a prognostic solution for thermal abatement systems used in semiconductor manufacturing. Finally, in Section 3.5, sparse Bayesian learning for regression is introduced, which is a relatively recent, state of the art, technique for regression problems. This technique is used in Chapter 7 to develop a model which describes the fault-free behaviour of the main bearing temperature signal in a large utility scale wind turbine.

3.1 Multiple Linear Regression

A multiple linear regression (MLR) model is used to describe a linear relationship between a number of input variables x_i, and an output, or *response*, variable y. The subscripts i, where $i = 1, 2, ..., p$, refer to the individual input variables.

The form of a MLR model is given by

$$y(k) = \beta_1 + \beta_1 x_1(k) + \beta_2 x_2(k) + ... + \beta_p x_p(k) + \epsilon \tag{3.1}$$

where the model parameters β_i, $i = 0, 1, ...p$, represent the *regression coefficients* and ϵ represents the modelling error term, which should ideally be a random zero-mean Gaussian distributed variable. The form of the model in (3.1) assumes that a linear relationship exists between the input and output variables. Given a set of n input-output pairs of data from a system to be modelled, $\{\mathbf{x}, y\}$, where $n > p$, the most common approach to determining the values of β_i is via the method of *least squares*, resulting in a least squares regression (LSR) model. If y_i denotes the i-th observed response and x_{ij} denote the ith observation of input variable x_j, then for the i-th observation, y_i is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,\ i = 1, 2, \cdots, n. \tag{3.2}$$

$$= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i,\ i = 1, 2, \cdots, n. \tag{3.3}$$

The β_0 acts as a bias term which can be removed for notational simplicity if the inputs and output have been normalised to be zero-mean, prior to LSR modelling. With β_0 removed, Equation (3.2) can then be written in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.4}$$

where

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

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

where $\mathbf{y} \in \mathbb{R}^{n \times 1}$ 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.

The objective in LSR modelling is to find the vector of least-squares regression coefficient estimates $\hat{\boldsymbol{\beta}}$ that minimises the least-squares cost function, which is given by

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{3.5}$$

Equation (3.5) can be expanded as

$$S(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \tag{3.6}$$

Since $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}$ is a scalar, as is its transpose $(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y})^T$, such that $\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}$, Equation 3.6 can be written as

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

To identify the minimum solution to Equation (3.7), 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 as

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

leading to the *least-squares estimate* of $\boldsymbol{\beta}$ is

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

Provided $\mathbf{X}$ is invertible, then Equation (3.10) yields a unique solution for $\hat{\boldsymbol{\beta}}$ [53].

3.2 Artificial Neural Networks

Artificial neural networks (ANNs) represent a class of flexible nonlinear models, in which a network of interconnected processing elements, called nodes or neurons, are used for information processing. ANNs model relationships between input and output variables with a model structure inspired by the neural structure of the human brain. The original development of artificial neural networks is credited to McCulloch and Pitts [54], who developed the theory regarding how knowledge is learned by neurons within the brain. Since the original work by McCulloch and Pitts, the development of ANNs has continued. In recent decades, with the availability of increasing computing power, ANNs have emerged as a powerful tool for classification and regression problems, and have been applied to countless different applications across almost every relevant application domain. The primary feature of ANNs, which has made them so popular, is their ability to learn highly-complex nonlinear functional relationships between input and output training data. A wide variety of different types of ANNs have been developed including radial basis function (RBF) neural networks [55], Kohenen self-organising maps [56], and Hopfield networks [57]. However, by far the most commonly employed ANNs are multilayer perception (MLP) neural networks. A review of the theory of MLPs is presented in this section.

3.2.1 Multiple-Layered Perceptrons

A MLP neural network [58] is a type of ANN where the neurons are arranged in several layers. In all MLP networks, an *input* and *output* layer are present, along with one, or more, *hidden* layers between the input and output layers. The most common type of neuron used in MLPs is the McCulloch Pitts neuron. The form of a McCulloch Pitts neuron is illustrated in Figure 3.1. The neuron operates by calculating the weighted sum of the inputs in_i, with each input having an associated weight value, w_i. The neuron computes the weighted sum of the inputs and adds a bias value, b. The sum of the weighted inputs and the bias term are then passed to an activation function in the neuron $f(.)$, to yield the neuron output.

In an MLP neural network, the neurons in each layer receive weighted inputs from all neurons in the preceding layer, which are combined with a bias value, to calculate an output value. This output value is then passed to the neurons activation function and the neuron output is then passed as an input to the next layer. This type of ANN structure, in which data information processing flows in one direction, from the input layer to

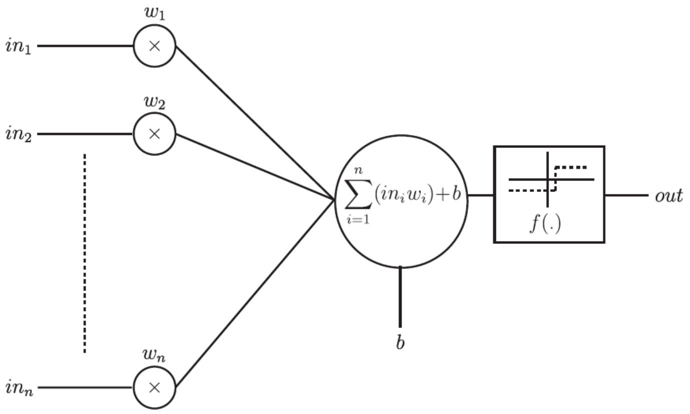


FIGURE 3.1: McCulloch Pitts neuron [4]

the output layer, is known as a *feed-forward neural network*. Figure 3.2 illustrates the structure of a typical feed-forward neural network.

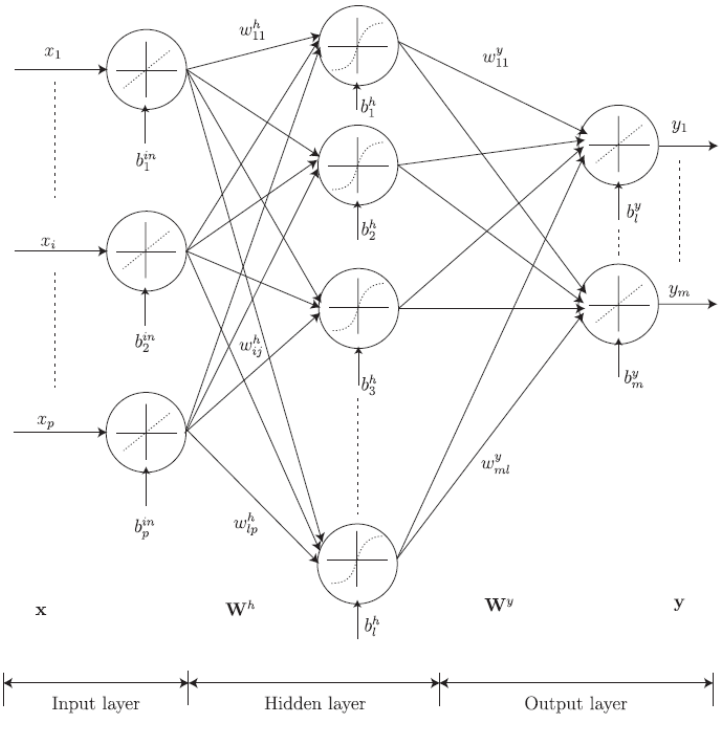


FIGURE 3.2: Structure of a feed-forward ANN [4]

The neurons in an MLP network can use a variety of linear and non-linear activation functions. Some of the most common activation functions include the step function, log-sigmoid, and tan-sigmoid functions. Log-sigmoid activation functions are often used in the output layer for classification problems, to limit the output to a small range, whilst linear activation functions are often used in the output layer for regression problems, to avoid limiting the output range. By allowing the number of neurons to increase indefinitely, it has been shown that a feed-forward ANN can be used to approximate any continuous function [58]. However, with noisy data, the ability to implement ANNs with a sufficiently large numbers of neurons required is a limiting factor.

A standard issue to be addressed in creating an MLP network is the selection of the network architecture, and the number of neurons to be used in the hidden layers. There does not exist a standardised method for choosing the optimal network structure, for a specific modelling problem, and the selection of the optimal network architecture is often a trial and error exercise. Selecting a network topology and the number of neurons in the hidden layer must consider a number of issues. The greater the number of neurons within the network, the greater the ability of the network to model more complex input-output relationships. However, during network training, as the number of neurons increases, the network will typically begin to model the noise and "over-fit" the training data, reducing the generalisation capabilities of the trained network. As a result, care must be taken in both network size selection and the network training process, to prevent over-fitting of the training data.

3.2.2 Network Training

To identify an optimal set of network weights and bias values, for a given input and output training data set, a network training procedure is required. For an MLP network, the training procedure is supervised and the training data used to identify the model parameters consists of input-output pairs, used by the network to learn the underlying functional relationship of the data. The most common approach for this task is the *backpropagation* algorithm, first introduced by Rumelhart *et al.* [59], which is described briefly below.

1. The training data, consisting of both the input data vectors, and output, or *target*, vector(s), are presented to the network.
2. Once the input data has been passed through the network, where the network weights and biases have usually been randomly initialised, the network output is computed. By comparing the model output with the target output, the sum squared error (SSE) is computed, where the SSE is defined as

$$\text{SSE} = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \hat{y}_{ij})^2 \tag{3.11}$$

where n is the number of input-output samples in the training data set, m is the number of model outputs, y_{ij} is the desired value, or target value, of the output j, and $\hat{y}_{ij}$ is the estimated network value for output j. The SEE is used as the *cost function*, to measure the training error at each iteration of the training procedure.

3. After evaluating the error signal on the first iteration, learning occurs by minimising the SSE, through modifications of the weights between each neuron. This is typically carried out using an iterative gradient descent algorithm, whereby the weights are adjusted at each iteration in the direction of decreasing values of the SSE. The weight optimisation routine is typically of the form

$$\mathbf{w}(m+1) = \mathbf{w}(m) - v\nabla(m) \tag{3.12}$$

   where $\mathbf{w}$ is vector containing the values of the network weight and biases, m is the current training iteration number, or *epoch*, and v is the *learning rate.* $\nabla(m)$ represents the gradient vector of the error function (3.11), with respect to the weight vector $\mathbf{w}$.

4. The training procedure, described by 1-3, is then repeated until the error reaches a desired threshold value, or until the required number of training iterations has been reached. At this point the network is deemed to have been trained, and can then be used to make predictions on previously unseen data.

3.2.3 Issues

A potential issue in training MLP networks is that the training procedure can begin to overfit the training data, resulting in poor generalisation capabilities when tested on previously unseen data. One technique for improving the generalisation capabilities of ANNs is to employ *early stopping* during network training. During typical network training, the training data set is generally split into three subsets known as the *training set*, *validation set*, and *test set.* The training set is employed for calculating the gradients, and update the network weights and biases. At each iteration, the error on both the training and validation sets is calculated. As training progresses, both the training set error and validations set error typically decrease. However, at some stage the network typically begins to over-fit the training data and the error on the validation set begins to rise. As the training algorithm continues to iterate, if the error on the validation set does not reduce further during the remaining epochs, then the network weights and biases which generated the minimum validation error are selected. The test set is not used during training but is instead used as a benchmark dataset to compare the performance of different trained networks.

Due to the random initialisation of the network weights and biases, it is common to train several networks, of the same structure, using a different random initialisation of the weights and biases. This is to address any potential sensitivity of the solution to the initial values of the weights and biases. The trained model with the best performance is then retained for future use.

In general, ANNs are a powerful tool for modelling complex input-output relationships. They provide the capability to model complex relationships where an understanding of the underlying physical relationships between the input and output data is unknown. However, this feature is also a drawback of ANNs, as the models created are completely black-box models and provide no direct insight into understanding the underlying relationship. In addition, ANNs typically require significant training times to identify the values of the weights and biases, and are relatively data-hungry, requiring a large number of samples to develop a useful model.

3.3 Gaussian Mixture Models

Mixture models are a probabilistic modelling technique which have found application in a wide range of engineering problems. To illustrate the principle of mixture modelling, consider the “old-faithful” data set [60], which is plotted in Figure 3.3. This data set comprises 272 measurements, which record the eruptions of the old-faithful geyser in the Yellowstone National Park in the US state of Wyoming. Each measurement comprises two values; the duration of an eruption in minutes and the time elapsed until the next eruption occurs.

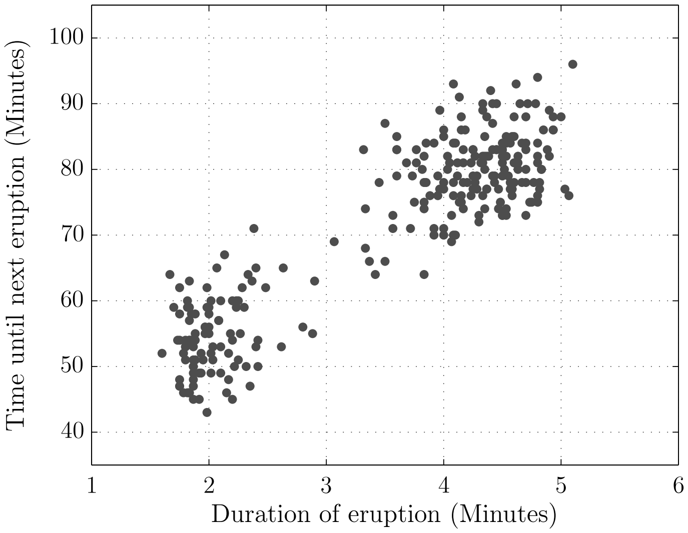


Figure 3.3: Old-Faithful data set

Looking at the plot of the old-faithful data set in figure 3.3, it can clearly be seen that there are two distinct clusters of data samples. If we wish to statistically model this data set using a parametric density function, it is immediately obvious that any single probability density function (PDF) will struggle to accurately represent the true underlying distribution of this data set. If instead, the set of observations is modelled using a combination of individual PDFs, it might significantly improve the overall accuracy of the model.

Figure 3.4 (a) illustrates the results of modelling the old-faithful data using a single 2-dimensional Gaussian PDF and Figure 3.4 (b) shows the results of modelling the old-faithful data set as a combination of two 2-dimensional Gaussian PDFs. The PDFs in each of the figures are presented as contour plots. The actual process of identifying the values of the parameters which define the PDFs illustrated in Figure 3.4 is discussed in Sections 3.3.1 and 3.3.2.

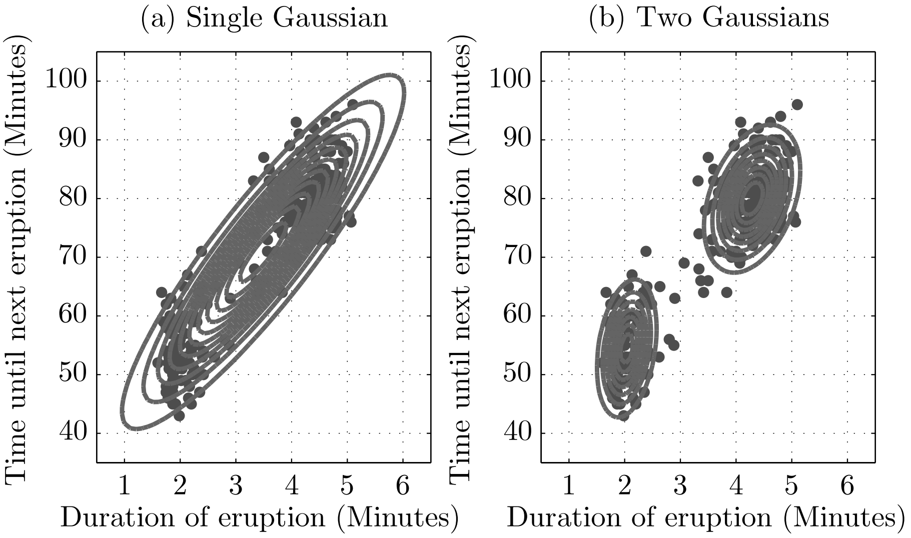


Figure 3.4: Old-Faithful data set: Model fitting

It is clear that by modelling the old-faithful data set using a combination of individual PDFs, the model can better capture the true underlying distribution of the data. These types of probabilistic models are formed via the linear superposition of individual PDFs, and are commonly known as mixture distributions or mixture models. These types of mixture distributions can be used to model very complex densities and, indeed, by employing a sufficient number of PDFs, and via appropriate choice of means, covariances, and mixture coefficient values, almost any continuous density can be approximated to arbitrary accuracy [60]. By far the most common probability density functions employed in real-world applications is the Gaussian PDF, although mixture models can be formed from mixtures of other parametric PDFs. The type of mixture models which are formed by taking linear superposition of individual Gaussian densities, and are more commonly known as Gaussian mixture models (GMMs).

A K-component Gaussian mixture model PDF is defined as

$$p(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k \, \mathcal{N}(\mathbf{x}|\mu_{\mathbf{k}}, \mathbf{\Sigma_k}) \tag{3.13}$$

where each Gaussian density $\mathcal{N}(\mathbf{x}|\mu_{\mathbf{k}}, \mathbf{\Sigma_k})$ is known as a component of the mixture model, and the parameters α_k are known as the *mixing coefficients*, which represent the weight of each component within the mixture model. The individual Gaussian components can be either univariate or multivariate depending on the dimensions of the data set being modelled. In the case of a single variable x, the Gaussian density is defined as

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{1}{2\sigma^2}(x-\mu^2)\right\} \tag{3.14}$$

where the Gaussian density is defined by two properties, the mean μ and the variance σ^2. In the case of a multivariate D-dimensional vector $\mathbf{x}$, a multivariate Gaussian density is defined as

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{3.15}$$

where the vector $\boldsymbol{\mu} \in \mathbb{R}^D$ is a D-dimensional vector containing the mean values of each variable, $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$ is a D x D covariance matrix and $|\boldsymbol{\Sigma}|$ is the determinant of the covariance matrix $\boldsymbol{\Sigma}$. To satisfy the requirements that the mixing coefficients α_k represent probabilities they are subject to a number of constraints, such that

$$\sum_{k=1}^{K} \alpha_k = 1 \quad \text{and} \quad 0 \leq \alpha_k \leq 1 \tag{3.16}$$

As a result, a K-component Gaussian mixture model is completely specified by parameter set $\boldsymbol{\theta}$, where

$$\boldsymbol{\theta} = \{\alpha_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \, \ldots \, , \alpha_D, \boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D\} \tag{3.17}$$

It follows then that the process of fitting a GMM to a set of observations involves optimising the values of the parameters in $\boldsymbol{\theta}$. There are two primary approaches within the literature for estimating the parameter values in $\boldsymbol{\theta}$, namely Bayesian estimation and maximum-likelihood estimation. By far the most commonly used approach is maximum-likelihood estimation, and within this thesis only maximum-likelihood approaches are considered.

3.3.1 Maximum-Likelihood Estimation

Consider a set of observations $\mathbf{X} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(N)}]$, where $\mathbf{x}^{(n)} = [x_1, ..., x_D]^T$ is a D-dimensional random vector representing a single observation in $\mathbf{X}$. The likelihood function for the parameter vector $\boldsymbol{\theta}$ is given by

$$L(\mathbf{X}|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\mathbf{x}^{(n)}|\boldsymbol{\theta}) \tag{3.18}$$

which quantifies the likelihood of observing the set of observations $\mathbf{X}$, conditioned on the values contained within the parameter set $\boldsymbol{\theta}$. The goal in using maximum-likelihood estimation techniques is to find $\hat{\boldsymbol{\theta}}$, which is the set of parameter values which maximises the likelihood function.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \, L(\mathbf{X}|\boldsymbol{\theta}) \tag{3.19}$$

A further issue to note is that it is common to maximise the logarithm of the likelihood function instead of maximising the likelihood function directly, as this is easier to handle analytically. This is called the log-likelihood function, and is equivalent to maximising the likelihood function due to the monotonicity of the logarithm function.

$$\log L(\mathbf{X}|\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}^{(n)}|\boldsymbol{\theta}) \tag{3.20}$$

Equation (3.20) describes the log-likelihood function in its generic format. In the case of a Gaussian mixture model, the log-likelihood function is described as follows

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \alpha_k \, \mathcal{N}(\mathbf{x}^{(n)}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \right\} \tag{3.21}$$

where K is the number of components in the mixture model, and N is the number of samples in the set of observations $\mathbf{X}$. For reasons which will become clear in Section 3.3.2, equation (3.21) is also commonly known as the *incomplete-data* log-likelihood function. In certain cases, it is possible to maximise the log-likelihood function directly by setting the partial derivatives of the log-likelihood function, with respect to each parameter in $\boldsymbol{\theta}$, to zero. For example, in the case of the log-likelihood function for a single Gaussian PDF, the maximum-likelihood solution leads directly to the standard equations for estimating the mean and variance of a set of observations [61].

In the case of the log-likelihood function for a Gaussian mixture model as in (3.21), there is no closed form solution possible once the derivative of the log-likelihood function is set to zero. This is due to the presence of the summation over k which appears inside the logarithm function [60]. As no closed form solution exists for the log-likelihood function of a Gaussian mixture model, the most common approach to identifying maximum-likelihood solutions for Gaussian mixture models is via the expectation-maximisation algorithm, which is discussed in the following section.

3.3.2 Expectation-Maximisation Algorithm

The expectation-maximisation (EM) algorithm is a two-stage iterative optimisation technique for finding maximum-likelihood estimates of the parameter values of a probability distribution from a set of observations. Inherent in the application of the EM algorithm is the interpretation of the set of observations being modelled, **X**, as an incomplete data set. In the case of the EM algorithm for mixture modelling, the missing part of the data set is a set of *latent* (hidden/unobserved) variables, **Z**, which indicate which component in the mixture model is responsible for generating each sample within the set of observations **X**. The missing set of latent variables **Z** has the form

$$\mathbf{Z} = \{\mathbf{z}^{(1)}, ..., \mathbf{z}^{(N)}\} \tag{3.22}$$

in which there are N labels associated with the N samples in the set of observations. Each individual label takes the form of a binary vector $\mathbf{z}^{(n)} = [z_1^{(n)}, \ ... \ , z_k^{(n)}]^T$ with $z_m^{(n)} = 1$ and $z_p^{(n)} = 0$, for $p \neq m$ if and only if $\mathbf{x}^{(n)}$ was produced by the m-th component of the mixture model. The values of $\mathbf{z}$ therefore satisfy $z_k^{(n)} \in \{0, 1\}$ and $\sum_k z_k^{(n)} = 1$. As a result, we can see that there are K possible states for the vector $\mathbf{z}^{(n)}$.

Having introduced the set of latent variables, it is now possible to define the *complete-data* log-likelihood function (3.23). Maximisation of this function with respect to $\boldsymbol{\theta}$ is simple if the missing data **Z** were observed, together with the actually observed set of observations **X**.

$$\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_k^{(n)} \log\,[\alpha_k \ p(\mathbf{x}^{(n)}|\boldsymbol{\theta_k})] \tag{3.23}$$

However, since the set of latent variables **Z** are not observable, the *complete-data* log-likelihood function cannot be maximised directly. To address this problem, the EM algorithm provides a method for indirect optimisation of the incomplete-data log-likelihood function via iterative optimisation of the conditional expectation of the *complete-data* log-likelihood function (3.23).

The EM algorithm operates by iteratively applying two-steps until some convergence criteria are satisfied. The two steps are known as the E-step (expectation) and the M-step (maximisation). In the E-step, the conditional expectation of the *complete-data* log-likelihood function is computed, conditioned on the set of available observations $\mathbf{X}$ and the current estimate of the parameter set $\boldsymbol{\theta}$ as

$$E\,[\,\log p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})\mid\mathbf{X},\boldsymbol{\theta}\,] \tag{3.24}$$

However, since the complete-data log-likelihood function (3.23) is linear with respect to the set of latent variables $\mathbf{Z}$, it is only necessary to compute the conditional expectation of the latent variables $\mathbf{W}$, as in (3.25), given the observed data $\mathbf{X}$, and the current estimate of the parameter set $\boldsymbol{\theta}$, which are available.

$$\mathbf{W} \equiv E\,[\,p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})\,] \tag{3.25}$$

As the elements of the set of latent variables $\mathbf{Z}$ are binary variables, their individual conditional expectations $w_k^{(n)}$ are given by Bayes' rule such that,

$$w_k^{(n)} \equiv E[\,p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})\,] = Pr[\,z_k^{(n)} = 1 \mid \mathbf{x}^{(n)},\boldsymbol{\theta}\,] \tag{3.26}$$

$$= \frac{\alpha_k\, p(\mathbf{x}^{(n)}|\boldsymbol{\theta}_k)}{\sum\limits_{j=1}^{K} \alpha_j\, p(\mathbf{x}^{(n)}|\boldsymbol{\theta}_j)} \tag{3.27}$$

where α_k is the *a priori* probability that $z_k^{(n)} = 1$ and $w_k^{(n)}$ is the *posterior* probability that $z_k^{(n)} = 1$ given the observation $\mathbf{x}^{(n)}$ and the current estimate of the parameter set $\boldsymbol{\theta}$. The $w_k^{(n)}$ values can also be viewed as the responsibility that component k takes for explaining the observation $\mathbf{x}^{(n)}$ [60]. Following the E-step, the M-step generates a revised estimate for the parameter values in $\boldsymbol{\theta}$ by maximising the complete-data log-likelihood function with respect to each of the elements in $\boldsymbol{\theta}$, where the $z_k^{(n)}$ in (3.23) are replaced with the $w_k^{(n)}$ estimated in the E-step.

The EM algorithm iterates with the property that each cycle or iteration of the EM algorithm will increase the *incomplete-data* log-likelihood function until some local or global maximum is reached [61]. This is typically determined using some type of convergence criteria. A common approach to identify convergence is to cease iterating the EM algorithm once the improvement in the *incomplete-data* log-likelihood function between successive iterations falls below some threshold. The following is a summary of the implementation of the EM algorithm applied to estimating the parameter values of a Gaussian mixture model:

1. Initialise the parameter set $\boldsymbol{\theta}$, which comprises the means $\boldsymbol{\mu}_k$, the covariances $\boldsymbol{\Sigma_k}$ and the mixing the coefficients α_k for the desired number of components K. Once these parameters have been set, an initial value of the *incomplete-data* log-likelihood function (3.21) is calculated.

2. **E-Step** Once an initial value for model parameters has been selected, the E-Step is performed in which the conditional expectations of the the set of latent variables is computed,

$$w_k^{(n)} = \frac{\alpha_k \, p(\mathbf{x}^{(n)}|\hat{\boldsymbol{\theta}}_k)}{\sum\limits_{j=1}^{K} \alpha_j \, p(\mathbf{x}^{(n)}|\hat{\boldsymbol{\theta}}_j)} \tag{3.28}$$

3. **M-Step** In the M-Step, the *complete-data* log-likelihood function (3.23) is maximised with respect to each of the parameters in $\boldsymbol{\theta}$, resulting in the following formulae for updating the parameter values,

$$\alpha_k^{t+1} = \frac{N_k}{N} \tag{3.29}$$

$$\boldsymbol{\mu}_k^{t+1} = \frac{1}{N_k} \sum_{n=1}^{N} w_k^{(n)} \, \boldsymbol{x}^{(n)} \tag{3.30}$$

$$\boldsymbol{\Sigma}_k^{t+1} = \frac{1}{N_k} \, w_k^{(n)} \, (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k^{t+1})(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k^{t+1})^{\mathrm{T}} \tag{3.31}$$

   where

$$N_k = \sum_{n=1}^{N} w_k^{(n)} \tag{3.32}$$

4. Following the E-step and M-step, the *incomplete-data* log-likelihood function is calculated using the latest parameter estimates in $\hat{\boldsymbol{\theta}}$

$$\log p(\boldsymbol{X}|\,\hat{\boldsymbol{\theta}}) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \alpha_k\, p(\mathbf{x}^{(n)}|\hat{\boldsymbol{\theta}}_k) \right\} \tag{3.33}$$

5. The final step is to check for convergence. If algorithm has not converged, return to step 2.

3.3.2.1 Shortcomings of EM Algorithm

In the application of the standard EM algorithm for mixture modelling, there are a number of issues and shortcomings of the algorithm to consider. These include

1. How to determine the number of components?
2. Sensitivity of solution to initial values chosen for parameters in $\boldsymbol{\theta}$
3. Convergence to the boundary of the parameter space

The first issue to consider is how to determine the optimal number of components with which to model a set of observations (*i.e. model structure*). Using the maximised likelihood function (3.21), for various values of k, might seem a reasonable approach; however, this method is ineffective for selecting the number of components since the maximised likelihood function is a non-decreasing function of K. Numerous researchers have considered this issue, with a wide variety of solutions proposed. EM-based approaches typically address this problem by adopting a "model-class/model" hierarchy approach. In this way the EM algorithm is used to obtain a sequence of estimates of $\boldsymbol{\theta}$ for a range of values of k, $\{\hat{\boldsymbol{\theta}}_{(\boldsymbol{k})}, k = k_{min}, ..., k_{max}\}$, with the final estimate $\hat{k}$ being defined [62] as that which minimises some specified cost function

$$\hat{k} = \arg\min_{k}\{C(\hat{\boldsymbol{\theta}}_{(k)}, k), k = k_{min}, ..., k_{max}\} \tag{3.34}$$

Typically, the cost function employed will be of the form

$$C(\hat{\theta}_{(\mathbf{k})}, k) = -\log p(\mathbf{X}|\boldsymbol{\theta}) + R(k) \tag{3.35}$$

where $R(k)$ is a complexity cost penalising larger values of k. Numerous criteria have been suggested to penalise larger values of k, such as Schwarz's Bayesian inference criteria

(BIC), Akaike's information criteria (AIC) and Rissanen's minimum description length (MDL) to name a few. A comprehensive study and performance review of a range of possible criteria was presented in [63].

The second issue which affects the performance of the EM algorithm is the initialisation of the parameter values in $\boldsymbol{\theta}$. Before applying the E-step, an initial estimate for the parameters in $\boldsymbol{\theta}$ is required. The selection of the initial values is critical as the EM algorithm will always converge to a local maximum of the likelihood function [64]. Thus, the final estimate of the parameter values will be dependent upon the initial estimate. Common approaches to addressing this issue include running the EM algorithm several times, with each run starting with a different initial estimate of the parameters in $\boldsymbol{\theta}$. The run which results in the highest value of the likelihood function is then selected as the final estimate. Other approaches use clustering techniques such as K-means, to generate an initial estimate of the component parameter values [65]. The common feature of any of these solutions, with regard to selecting the number of components and addressing the initialisation issues with the EM algorithm, is a significant increase in the computational burden of applying the EM algorithm.

The final issue to consider is that of the EM algorithm potentially converging to the boundary of the parameter space. This can result in an unbounded likelihood, where one of the component weights α_k approaches zero, resulting in the covariance matrix of the corresponding component becoming arbitrarily close to singular. This can occur frequently, when the assumed number of components is significantly larger than the true/optimal number [66].

In Chapter 5, Gaussian mixture models are utilised for a range of condition monitoring applications. One of the primary considerations in selecting an EM algorithm for use in these applications is that it must be robust to each of the issues described above. In particular, the issue of avoiding convergence to boundary of the parameter space was identified as the primary consideration. To address these issues, a variant of the EM algorithm known as the Figueiredo-Jain algorithm [66] was investigated and selected as an appropriate choice to address the shortcomings of the standard EM algorithm. This algorithm is described briefly in the following section.

3.3.3 Figueiredo-Jain Algorithm

The Figueiredo-Jain (FJ) algorithm [66] is a variant of the EM algorithm which addresses the three major shortcomings of the standard EM algorithm identified in Section 3.3.2.1.

The first issue which the FJ algorithm addresses is determining the number of components in the mixture model. As discussed in Section 3.3.2.1, the standard approach to selecting the number of components in a mixture model is to adopt a "model-class/model" hierarchy approach, in which candidate mixture models are computed for each model-class (i.e. number of components), from which the "best" model is selected. The FJ algorithm instead aims to find the "best" overall model directly from the set of possible models.

The FJ algorithm operates by starting with an initial large number of components k ($k \gg$ true/optimal number of components), and attempts to determine the correct number of mixture components by letting the value of some of the mixing coefficients be zero. Starting with a large number of components, the FJ algorithm iterates and automatically adjusts the number of components by annihilating and removing those components which are not supported by the data. In this fashion, the FJ algorithm also avoids the issue of convergence to the boundary of the parameter space. When one of the components becomes weak, meaning that it is no longer supported by the data, and the mixing coefficient value α_k approaches zero, then this component is annihilated by the algorithm.

The FJ algorithm also tackles the initialisation issue of the EM algorithm. By starting the algorithm with a large number of components, in which the initial estimates of the parameter values for each component are distributed throughout the parameter space, the initialisation issue is avoided. Instead, it is now only necessary to remove those components which are not supported by the data. The FJ algorithm uses a modified cost function, based upon the minimum message length (MML) criteria [66], which is given by

$$L(\boldsymbol{\theta}, \mathbf{X}) = \frac{V}{2} \sum_{m:\, \alpha_m > 0} \left(\frac{N\alpha_m}{12}\right) + \frac{k_{nz}}{2} \log \frac{N}{12} + \frac{k_{nz}(V+1)}{2} - \log p(\mathbf{X}|\boldsymbol{\theta}) \tag{3.36}$$

where V is the number of free parameters specifying each mixture component, N is the number of samples in the set of observations, k_{nz} is the number of components with non-zero weighting ($\alpha_m > 0$), and $\log p(\mathbf{X}|\boldsymbol{\theta})$ is the log-likelihood function, as in (3.21). A full description of the design and motivation for the cost function in (3.36) is presented in [66]. The EM algorithm can be then used to minimise (3.36) with respect to $\boldsymbol{\theta}$, which then represents the best mixture estimate. The FJ algorithm incorporates a modified M-step in which the mixing coefficient update formula (3.29) is modified such that

$$\alpha_k^{t+1} = \frac{\max\{0, (\sum_{n=1}^{N} w_k^{(n)}) - \frac{V}{2}\}}{\sum_{k=1}^{K} \max\{0, (\sum_{n=1}^{N} w_k^{(n)}) - \frac{V}{2}\}} \tag{3.37}$$

This modified update formula provides an explicit means for annihilating components by setting the mixing coefficient value to zero when a component is no longer supported by the data. In this way, the FJ algorithm automatically avoids the boundary of the parameter space and thus avoids this common pitfall of the standard EM algorithm. The M-step update formulae for the remaining mixture parameters, given in 3.30 and 3.31, are unchanged.

Another feature of the FJ algorithm addresses a potential failure mode which can occur through the use of the modified M-step (3.37). In a situation where the starting value of k is large, it is possible that none of the mixture components will obtain sufficient support from the data (i.e. $\sum_{n=1}^{N} w_k^{(n)} < \frac{V}{2}$, for $k = 1, 2, ..., K$), resulting in all of the mixing coefficients α_k having a value of zero. To address this issue, the FJ algorithm uses a component-wise EM algorithm (CEM) [67]. Instead of simultaneously updating all of the parameters in $\boldsymbol{\theta}$ as in the standard EM algorithm, the CEM algorithm updates the parameters in $\boldsymbol{\theta}$ sequentially for each component. Once the parameter values for a single component have been updated, the CEM algorithm recomputes the E-step (i.e. updates the $\mathbf{W}$ values, (3.28)). In this way, if any component is annihilated, its weighting is immediately redistributed amongst the remaining components. In this way, the algorithm can be started with an arbitrarily large value of k without any problems.

The final issue to discuss, regarding the convergence of the FJ algorithm, is the determination of the optimal number of mixture components. The component annihilation rule, described by (3.37), does not take into consideration the reduction in the cost function value $L(\boldsymbol{\theta}, \mathbf{X})$ that might be achieved by decreasing the value of k_{nz}. The FJ algorithm addresses this issues as follows; once the CEM algorithm converges, the component with the least weighting (i.e. smallest α_k) is annihilated, and the CEM algorithm is rerun until convergence. The procedure repeats until $k_{nz} = k_{min}$, where k_{min} is selected by the user. The estimate for each value of k, which leads to the minimum value of $L(\boldsymbol{\theta}, \mathbf{X})$, where $k = \{k_{min}, ..., k_{max}\}$, is then selected as the final estimate of the number of mixture components.

3.4 Particle Filtering for Prognostics

3.4.1 Introduction

Prognostics involves the generation of long-term predictions, describing the evolution in time of a signal of interest, or fault indicator, for the purposes of estimating the remaining useful life (RUL) of a degrading system [27]. Since prognostics involves projecting the current system state into the future, in the absence of future measurements, this inevitably introduces a significant degree of uncertainty into the generated predictions. To address this uncertainty, it is common to represent system state variables as random variables, whose probability distributions need to be estimated to derive confidence intervals on RUL predictions. In addition, the generation of long-term predictions requires a model describing the future evolution of a fault indicator. However, the dynamics of all damage processes are inherently nonliner processes, with uncertainties which are often non-Gaussian, therefore propagating these effects through time becomes a challenging and error-prone task [68].

To address these uncertainties, recursive Bayesian estimation techniques have emerged as a powerful technique for prognostics, which combine information from both dynamic models describing the evolution of a fault indicator, and on-line data collected from sensors monitoring key system variables [69]. In this way, long term predictions of a fault indicator are generated using dynamic models ,which describe the evolution of a fault indicator, and the initial conditions on those predictions, which are estimated using incoming measurement data. Within the domain of recursive Bayesian estimation techniques for prognostics, particle filters have emerged as the *de facto* state-of-the-art technique [69]. Particle filters are used for nonlinear Bayesian filtering and do not require the linearity and Gaussian noise assumptions of the Kalman filter. Particle filters also provide a robust framework for the generation of long term prognosis, while accounting effectively for uncertainties [33].

The basic principle of particle filtering is the approximation of the system state distribution using a set of particles (state values), and an associated set of particle weights which represent the discrete probability masses of the individual particles. The application of particle filtering to prognostics generally encompasses two separate steps, *state estimation* and *long term prediction*. In the first step, incoming measurement data is used to generate an estimate of the current system state [33]. Once an updated estimate of the state is available, the updated state estimate is used as a starting condition for the second step, which involves the generation of long term predictions of the evolution of the fault indicator. In Sections 3.4.2, 3.4.3, and 3.4.4, a technical review of the theory and principles of particle filtering is presented, with an emphasis on the application of particle filtering to prognostics.

3.4.2 Nonlinear Bayesian Filtering

Consider a dynamic system whose state at time t_k is represented by the vector $\mathbf{x}_k$. The evolution of the system state is described by a state-space model such that

$$\begin{aligned} \mathbf{x}_k &= \mathbf{f}_k(\mathbf{x}_{k-1}, \boldsymbol{\omega}_k) \leftrightarrow p(\mathbf{x}_k|\mathbf{x}_{k-1}) \quad (3.38) \\ \mathbf{z}_k &= \mathbf{h}_k(\mathbf{x}_k, \boldsymbol{\upsilon}_k) \quad \leftrightarrow p(\mathbf{z}_k|\mathbf{x}_{k-1}) \quad (3.39) \end{aligned}$$

where

- $\mathbf{f}_k : \mathbb{R}^{n_\mathbf{x}} \times \mathbb{R}^{n_{\boldsymbol{\omega}}} \rightarrow \mathbb{R}^{n_\mathbf{x}}$ is the state transition function (possibly nonlinear) which describes the evolution of the system state, where $n_\mathbf{x}$ and $n_{\boldsymbol{\omega}}$ are the dimensions of the state and process noise vectors respectively
- $\mathbf{h}_k : \mathbb{R}^{n_\mathbf{z}} \times \mathbb{R}^{n_{\boldsymbol{\upsilon}}} \rightarrow \mathbb{R}^{n_\mathbf{z}}$ is the measurement/observation function which describes the sequence of measurements $\mathbf{z}_k$ collected at successive time steps t_k
- $\boldsymbol{\omega}_k \in \mathbb{R}^{n_{\boldsymbol{\omega}}}$ is an independent identically distributed (i.i.d.) process noise sequence of known distribution
- $\boldsymbol{\upsilon}_k \in \mathbb{R}^{n_{\boldsymbol{\upsilon}}}$ is an i.i.d. measurement noise sequence of known distribution

Also shown in equations (3.38) and (3.39) are the equivalent probabilistic interpretations of the state-transition function and the measurement update function. The function $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ represents the state transition probability and the function $p(\mathbf{z}_k|\mathbf{x}_{k-1})$ is the *likelihood* function that defines the probability of observing the values in $\mathbf{z}_k$, given the current estimate of $\mathbf{x}_k$.

The objective of the filtering operation is to find filtered estimates of $\mathbf{x}_k$, based upon the set of all available measurements $\mathbf{z}_{1:k} = \{\mathbf{z}_j, j = 1, 2, ..., k\}$. Considering the filtering problem from a Bayesian perspective, the objective is to recursively calculate some degree of belief regarding the distribution of the state $\mathbf{x}_k$, given the set of observations $\mathbf{z}_{1:k}$ up to time t_k [70]. It is therefore required to construct the PDF $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ which is known as the filtered *posterior* state PDF. This PDF contains all the information about the state $\mathbf{x}_k$, which is inferred from the measurements $\mathbf{z}_{1:k}$ and the initial state PDF $p(\mathbf{x}_0)$, which is assumed known.

In principle, the *posterior* state PDF $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ can be estimated recursively by performing two sequential steps, *prediction* and *update* [71]. Given the probability distribution $p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})$ at time t_{k-1}, the *prediction* step uses the system model (3.38) to obtain the *a priori* state PDF $p(\mathbf{x}_k|\mathbf{z}_{1:k-1})$, at time t_k, via application of the Chapman-Kalmogorov equation

$$p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_{1:k-1})\, p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1}) d\mathbf{x}_{k-1} \tag{3.40}$$

Since the state transition equation in (3.38) describes a 1$^{\text{st}}$ order Markov process, in which the next system state $\mathbf{x}_k$ is a function of only the previous state $\mathbf{x}_{k-1}$, we have that

$$p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_{1:k-1}) = p(\mathbf{x}_k|\mathbf{x}_{k-1}) \tag{3.41}$$

and, hence, the prediction equation in (3.40) can be simplified as

$$p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1}) d\mathbf{x}_{k-1} \tag{3.42}$$

Following the *prediction* step, the *update* step incorporates the current measurement vector $\mathbf{z}_k$, the *a priori* state PDF $p(\mathbf{x}_k|\mathbf{z}_{1:k-1})$ calculated in the *prediction* step, the *likelihood* function $p(\mathbf{z}_k|\mathbf{x}_k)$, and Bayes rule, to estimate the *posterior* state PDF $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ as

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) = \frac{p(\mathbf{z}_k|\mathbf{x}_k)\, p(\mathbf{x}_k|\mathbf{z}_{1:k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})} \tag{3.43}$$

where the normalising constant

$$p(\mathbf{z}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{z}_k|\mathbf{x}_k)\, p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) d\mathbf{x}_k \tag{3.44}$$

is dependant upon the *likelihood* function $p(\mathbf{z}_k|\mathbf{x}_k)$ and the statistics of the measurement noise vector $\boldsymbol{\upsilon}_k$. In short, the objective of the *update* step is to incorporate the latest measurement vector $\mathbf{z}_k$, and modify the *a priori* state PDF $p(\mathbf{x}_k|\mathbf{z}_{1:k-1})$, to estimate the *posterior* state PDF $p(\mathbf{x}_k|\mathbf{z}_{1:k})$.

The procedure described above in (3.40 → 3.44) form the basis for the determination of the optimal Bayesian solution. However, the recursive computation of the *posterior* state PDF is more conceptual than practical and in general cannot be determined analytically. In a restricted set of cases, such as when using linear Gaussian state space models, the optimal solution leads to the well known Kalman filter [71]. However, in the presence of a nonlinear process model and/or non-Gaussian noise processes, an alternative approach must be considered. A common approach is to use particle filtering methods, which approximate the optimal Bayesian solution.

3.4.3 Particle Filtering (State Estimation)

Particle filtering, also known as Sequential Monte Carlo (SMC) methods [72], is a technique for implementing a recursive Bayesian filter via Monte Carlo simulations. The basic principle of particle filtering is to represent the *posterior* state PDF by a set of random samples or "particles", each with an associated weight, and to compute estimates based on these samples and weights. One of the most commonly used particle filter algorithms is the *sequential importance sampling* (SIS) particle filter. The SIS particle filter approximates the *posterior* state PDF $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$ by a set, S, of N_s weighted particles,

$$S = \{\mathbf{x}_{0:k}^{(i)}, w_k^{(i)}\}_{i=1}^{N_s} \tag{3.45}$$

where $\{\mathbf{x}_{0:k}^{(i)}, i = 1, ..., N_s\}$ is a set of particles representing state values, with an associated set of importance weights $\{w_k^{(i)}, i = 1, ..., N_s\}$ which are approximations to the relative posterior probabilities of the particles, and $\mathbf{x}_{0:k} = \{\mathbf{x}_j, j = 0, ..., k\}$ is the set of all states up to time k.

The weight values are also normalised such that,

$$\sum_i w_k^{(i)} = 1 \tag{3.46}$$

The *posterior* state PDF can then be approximated as

$$p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) \approx \sum_{i=1}^{N_s} w_k^{(i)} \, \delta(\mathbf{x}_{0:k} - \mathbf{x}_{0:k}^{(i)}) \tag{3.47}$$

This results in a discrete weighted approximation to the true *posterior* state distribution $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$. As the number of samples/particles becomes very large, the Monte Carlo characterisation becomes an equivalent representation to the usual functional description of the *posterior* state PDF, and the filter approaches the optimal Bayesian solution [71].

3.4.3.1 Importance Sampling

A key issue which must be addressed is how to determine the particle weight values $w_k^{(i)}$. Consider the *posterior* state PDF $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$. Without a closed form for describing this PDF it is often impossible to sample from this distribution. To address this issue, the principle of *importance sampling* is used [72].

Consider the situation where $p(x) \propto q(x)$, where $p(x)$ is a PDF from which it is difficult to draw samples, but $q(x)$ is a PDF from which samples can be easily drawn. If we let $x^{(i)} \sim q(x), i = 1, ..., N$ be a set of samples that are easily generated from a proposal distribution $q(\cdot)$, known as the *importance density*, then a weighted approximation to the density $p(\cdot)$ is then given [71] by

$$p(x) \approx \sum_{i=1}^{N_s} w^{(i)}\, \delta(x - x^{(i)}) \tag{3.48}$$

where the normalised weight of the ith particle is given by

$$w^{(i)} \approx \frac{p(x^{(i)})}{q(x^{(i)})} \tag{3.49}$$

Thus, if a set of samples $\mathbf{x}_{0:k}^{(i)}$ were drawn from an *importance density* $q(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$, then the weights in (6.3) are defined to be

$$w_k^{(i)} \propto \frac{p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})}{q(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})} \tag{3.50}$$

In an online application, the objective of the filtering operation is to recursively estimate the distribution $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$ at time t_k, from the distribution $p(\mathbf{x}_{0:k-1}|\mathbf{z}_{1:k-1})$ at time t_{k-1}. If at time t_k, we have a set of samples approximating $p(\mathbf{x}_{0:k-1}|\mathbf{z}_{1:k-1})$ and want to obtain a new set of samples approximating $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$, this can be achieved by placing a constraint on the *importance density* so that it can be factorised as

$$q(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) = q(\mathbf{x}_k|\mathbf{x}_{0:k-1}, \mathbf{z}_{1:k})q(\mathbf{x}_{0:k-1}|\mathbf{z}_{1:k-1}) \tag{3.51}$$

It is then possible to obtain a new set of samples $\mathbf{x}_{0:k}^{(i)} \sim q(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$, by appending onto each of the existing samples $\mathbf{x}_{0:k-1}^{(i)} \sim q(\mathbf{x}_{0:k-1}|\mathbf{z}_{1:k-1})$, the new state $\mathbf{x}_k^{(i)} \sim q(\mathbf{x}_k|\mathbf{x}_{0:k-1}, \mathbf{z}_{1:k})$. Following the update of the particle values at each iteration, it is also necessary update the particle weights. To derive the weight update equation, the first task is to factorise the *posterior* state PDF [71] as

$$p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) = p(\mathbf{x}_{0:k-1}|\mathbf{z}_{0:k-1})\frac{p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})} \tag{3.52}$$

Then, using (3.51) and (3.52), it is possible to derive a recursive weight update equation

$$\begin{aligned} w_k^{(i)} &\propto \frac{p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})}{q(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})} && (3.53) \\ &\propto \frac{p(\mathbf{x}_{0:k-1}|\mathbf{z}_{0:k-1})\ p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{k-1})}{q(\mathbf{x}_k|\mathbf{x}_{0:k}, \mathbf{z}_{1:k})q(\mathbf{x}_{0:k-1}|\mathbf{z}_{1:k-1})} && (3.54) \\ &= w_{k-1}^{(i)}\frac{p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{k-1})}{q(\mathbf{x}_k|\mathbf{x}_{0:k}, \mathbf{z}_{1:k})} && (3.55) \end{aligned}$$

Finally, if $q(\mathbf{x}_k^{(i)}|\mathbf{x}_{0:k-1}^{(i)}, \mathbf{z}_{1:k}) = q(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_k)$, then the *importance density* becomes dependent only upon $\mathbf{x}_{k-1}$ and $\mathbf{z}_k$. This is of particular benefit in an online application, where only a filtered estimate of $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ is required at each time step. In such a situation, only the current state estimate $\mathbf{x}_k^{(i)}$, and the set of observations $\mathbf{z}_{1:k}$ needs to be stored. The previous path of state $\mathbf{x}_{0:k-1}^{(i)}$ up to the current time can be discarded. This results in a modified update formula so that

$$w_k^{(i)} \propto w_{k-1}^{(i)}\frac{p(\mathbf{z}_k|\mathbf{x}_k^{(i)})p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})}{q(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_k)} \tag{3.56}$$

and the *posterior* filtered distribution $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ can be approximated as

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)}\delta(\mathbf{x}_{0:k} - \mathbf{x}_{0:k}^{(i)}) \tag{3.57}$$

where the weights are defined as in (3.56). The SIS particle filter computes the values of the particle weights $w_k^{(i)}$ by setting the *importance density* equal to the *a priori* PDF for the state [33], i.e. $q(\mathbf{x}_{0:k}|\mathbf{x}_{0:k-1}) = p(\mathbf{x}_k|\mathbf{x}_{k-1}) = f_k(\mathbf{x}_k|\mathbf{x}_{k-1})$. This means that the new set of particles $\mathbf{x}_k^{(i)}$ are generated from the previous set of particles $\mathbf{x}_{k-1}^{(i)}$, by projecting the previous set of particle values forward using the state transition function $f_k(\mathbf{x}_k|\mathbf{x}_{k-1})$. Substitution of the *a priori* state PDF $p(\mathbf{x}_k|\mathbf{x}_{k-1})$, for the *importance* *density* in (3.56), also leads to the simplification of weight update formula, such that the non-normalised weight update formula is now given by

$$w_k^{(i)} = w_{k-1}^{(i)} p(\mathbf{z}_k|\mathbf{x}_k^{(i)}) \tag{3.58}$$

In this way the algorithm is simplified, and the weights for the newly generated particle $\mathbf{x}_k^{(i)}$ are evaluated from the *likelihood* function for new observations $p(\mathbf{z}_k|\mathbf{x}_k^{(i)})$.

3.4.3.2 Particle Degeneracy

A common issue with the application of the SIS particle filter is particle degeneracy. It has be shown previously that as the SIS algorithm iterates, the variance of the particle weights can only increase until, after a few iterations, all but one of the particles will have negligible weight [72]. This can result in a poor approximation of the target *posterior* distribution $p(\mathbf{x}_k|\mathbf{z}_{1:k})$, and significant computational effort being spent updating particles whose contribution to the distribution is almost zero. To measure the level of degeneracy, a metric known as the effective sample size, N_{eff} is used [72], which is defined as

$$N_{eff} = \frac{N_s}{1 + \text{var}(w_k^{(i)})} \tag{3.59}$$

However, this value cannot be evaluated exactly [71] so, instead, an estimate $\hat{N}_{eff}$ is computed whereby

$$\hat{N}_{eff} = \frac{N_s}{\sum_{i=1}^{N_s} (w_k^{(i)})^2} \tag{3.60}$$

The level of particle degeneracy is determined at each iteration by comparing $\hat{N}_{eff}$ with a specified threshold value N_{thres}. Whenever $\hat{N}_{eff} \leq N_{thres}$, a particle resampling operation takes place. The idea behind the resampling operation is to eliminate those particles that have small weights and to focus on those particles with large weights. The resampling operation results in the generation of a new set of particles $\{\mathbf{x}_k^{(i*)}\}_{i=1}^{N_s}$ by resampling (without replacement) N_s times from the current discrete approximation of $p(\mathbf{x}_k|\mathbf{z}_{1:k})$

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)} \delta(\mathbf{x}_{0:k} - \mathbf{x}_{0:k}^{(i)}) \tag{3.61}$$

such that $Pr(\mathbf{x}_k^{(i*)} = \mathbf{x}_k^{(i)}) = w_k^{(i)}$. The new particle population represents a set of independent and identically distributed (i.i.d.) samples from the discrete density (3.61), and therefore the particle weights are reset such that $w_k^{(i)} = 1/N_s$. The incorporation of the resampling step leads to the SIS particle filter becoming known as the *sequential importance resampling* (SIR) particle filter.

3.4.4 Particle Filtering (Long-Term Prediction)

The second stage in the application of particle filtering for prognostics, long-term predictions, involves predicting the RUL of the system under observation. This is carried out in two steps. Firstly, long-term predictions of the system state are generated using the current state estimate as the starting condition. The generated predictions are then combined with a predefined critical threshold value of the system state, to derive a RUL PDF for the system. The details of this two-step process are presented below.

3.4.4.1 Step 1: Generation of Long-Term Predictions

To generate long-term predictions of the system state, the set of particles and weights $\{\mathbf{x}_k^{(i)}, w_k^{(i)}\}_{i=1}^{N_s}$ which define the current *posterior* TPU degradation state estimate are used as the initial conditions. Each particle is individually progagated into the future by recursively applying the state-transition model (3.38), until the value of each particle enters the predefined hazard zone. The hazard zone defines a range of critical values of the degradation state at which the system under observation is deemed to have reached the end of its serviceable life and there is a significant risk of equipment failure in continuing to operate the equipment beyond the range of values defined by the hazard zone. Typically, the lower and upper hazard zone bounds (H_{lb} and H_{ub} respectively) are determined statistically from historical failure data, or inferred using expert knowledge gathered from domain experts such as maintenance personnel.

Assuming that the current set of particles and weights $\{\mathbf{x}_k^{(i)}, w_k^{(i)}\}_{i=1}^{N_s}$ are a good representation of the system state at time t_k, then the predicted state PDF at time t_{k+p} can be approximated by using the law of total probabilities [33], whereby

$$\hat{p}(\hat{\mathbf{x}}_{k+p}|\hat{\mathbf{x}}_{k:k+p-1}) \approx \sum_{i=1}^{N_s} w_{k+p-1}^{(i)}\ \hat{p}(\hat{x}_{k+p}^{(i)}|\hat{x}_{k+p-1}^{(i)}) \tag{3.62}$$

To evaluate (3.62) it is necessary that the particle weights be modified at each prediction iteration, to account for the fact that noise and process nonlinearities may change the shape of the state PDF as time passes. However, at each prediction iteration, no measurement data is available, with the result that the weight update procedure, described previously (3.58), cannot be utilised. To address this problem, a number of solutions have been proposed [27]. The simplest approach is to consider that the error, associated with considering the particle weights invariant for future time instants to be negligible with respect to other sources of uncertainty, such as model inaccuracies and noise process assumptions [33]. Using this approach, the particle trajectories $x^{(i)}_{k:k+p}$ are simply propagated in time, using the state-transition model, while the current particle weights are propagated in time without any change. More accurate and robust methods for updating the particle weights at each prediction iteration are available (see [27]) for further details). However, it has been demonstrated previously that the approach described above provides satisfactory performance in describing how a system behaves in a practical application [27].

3.4.4.2 Step 2: Characterisation of Remaining Useful Life

Once the projected path for each particle, $\hat{x}^{(i)}_{k+p}$, has been generated, and the time at which each particle enters the hazard zone has been identified, this information is then combined with the weight of each particle $w^{(i)}_{k}$ to generate a RUL PDF for the system. The RUL PDF can be computed by applying the law of total probabilities [27], whereby

$$p_{ttf}(k+p) = \sum_{i=1}^{N_s} Pr(Failure|X = \hat{x}^{(i)}_{k+p}, H_{lb}, H_{ub})\,.w^{(i)}_{k+p} \tag{3.63}$$

where $p_{ttf}(k+p)$ is the probability of equipment failure at time t_{k+p}. The overall system RUL PDF is then approximated by the sum of the individual failure probabilities at each future time instant.

The application of particle filtering to prognostics, incorporating both state estimation and the generation of long-term predictions, is known as the *particle filtering framework* for prognostics [73]. The generation of long-term predictions is generally subject to a range of uncertainties, such as modelling errors and errors in updating the particle weights at future time horizons. The level of error generally increases with the length of the prediction horizon. To improve the prediction accuracy as more measurement data becomes available, a number of methods for correcting the predicted RUL are presented in [27]. These methods attempt to alter the RUL PDF on the basis of predictions errors, calculated from measurement data which subsequently becomes available as the fault progresses. A discussion of these methods, and techniques developed in this thesis to address issues of prognostic uncertainty, is presented later in Chapter 6, with reference to the application domain under investigation.

3.5 Sparse Bayesian Learning

Sparse Bayesian learning, of which the relevance vector machine (RVM) is a specific type of realisation, is a novel machine learning technique developed by Tipping [74], which uses Bayesian inference to obtain sparse models for regression and classification. Sparse Bayesian learning models have an identical functional form to support vector machines (SVMs), a state-of-the-art technique in regression and classification, but have a number of properties which make them more attractive. Predictions generated by sparse Bayesian learning models are probabilistic, incorporating uncertainty estimates which are not generated by SVMs. In addition, sparse Bayesian learning models have comparable generalisation capabilities to SVMs, but generally use dramatically fewer kernel functions as, during the training procedure, many of the model weights are set to zero, thus achieving sparsity. This section presents a brief review of the principles of sparse Bayesian learning for regression.

Given a set of input-target pairs $\{\mathbf{x}_n, t_n\}_{n=1}^{N}$, and considering a scalar valued target function t, a standard probabilistic formulation is to assume that the targets are samples from a model with additive noise, so that

$$t_n = y(\mathbf{x}_n, \mathbf{w}) + \epsilon_n \tag{3.64}$$

where the noise terms ϵ_n are independent samples from some noise process, which is assumed to be zero-mean and Gaussian distributed, with variance σ^2. Thus, the target values, t_n, follow a Gaussian distribution, with mean $y(\mathbf{x}_n)$ and variance σ^2.

Typically, a sparse model takes the form

$$y(\mathbf{x}) = \sum_{n=1}^{N} = w_n \phi(\mathbf{x}) = \mathbf{w}^T \mathbf{\Phi} \tag{3.65}$$

where $\mathbf{w} = (w_0, ..., w_N)^T$ is the weight vector, and $\mathbf{\Phi}$ is an $N \times (N+1)$ "design matrix", with $\mathbf{\Phi} = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), ..., \phi(\mathbf{x}_N)]$, where $\phi(\mathbf{x}_n) = [1, K(\mathbf{x}_n, \mathbf{x}_1), K(\mathbf{x}_n, \mathbf{x}_2), ..., K(\mathbf{x}_n, \mathbf{x}_N)]$, where $K(\mathbf{x}, \mathbf{x}_i)$ is a *kernel* function, defining a single basis function for each input-target pair. A common choice of Basis function is a Gaussian data-centred kernel function. Assuming independence of the t_n values, the likelihood function for the complete set of input-target pairs $\{\mathbf{x}, \mathbf{t}\}$ can can be defined as

$$p(\mathbf{t}|\mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-N/2} \text{exp} \left\{ -\frac{1}{2\sigma^2} \|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\| \right\} \tag{3.66}$$

The form of the model in (3.66) contains an equal number of model parameters as input-target pairs. As a result, it might be expected that maximisation of (3.66), with respect to $\mathbf{w}$ and σ^2, might lead to severe over-fitting of the data. To address this problem from a Bayesian perspective, a constraint is placed on the parameters by defining an explicit *prior* probability over them. A preference for smoother, less complex functions, can be made, by choosing a zero-mean Gaussian distributed prior distribution over $\mathbf{w}$, so that

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=0}^{N} \mathcal{N}(w_i|0, \alpha_i^{-1}) \tag{3.67}$$

where $\boldsymbol{\alpha}$ is a vector of $N+1$ hyperparameters (in Bayesian statistics, a hyperparameter is a parameter of a *prior* distribution). An individual, independent, hyperparameter is associated with each individual weight w_i, which moderates the strength of each weight. This is a key feature of the model, and is ultimately responsible for its sparsity properties.

The procedure for identifying the model weights involves first computing the posterior distribution over the weights, which is defined by

$$p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2) = \frac{p(\mathbf{t}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\boldsymbol{\alpha})}{p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2)} \tag{3.68}$$

To compute (3.68), it is first necessary to solve for its normalising constant, whereby

$$\begin{aligned} p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2) &= \int p(\mathbf{t}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\boldsymbol{\alpha}) \\ &= (2\pi)^{-N/2}|\sigma^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T|^{-1/2} \times \\ &\quad \exp\{-\frac{1}{2}\mathbf{t}^T(\sigma^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T)^{-1}\mathbf{t}\} \end{aligned} \tag{3.69}$$

Then, using equations (3.66),(3.67), and (3.69), it is possible to estimate the posterior distribution over the weights (3.68), as

$$\begin{aligned} p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2) &= \frac{p(\mathbf{t}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\boldsymbol{\alpha})}{p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2)} \\ &= (2\pi)^{-(N+1)/2}|\boldsymbol{\Sigma}|^{-1/2} \times \\ &\quad \exp\{-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu})\} \end{aligned} \tag{3.70}$$

where the posterior mean and covariance are given by

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{B}\mathbf{t} \tag{3.71}$$

$$\boldsymbol{\Sigma} = (\boldsymbol{\Phi}^T\mathbf{B}\boldsymbol{\Phi} + \mathbf{A})^{-1} \tag{3.72}$$

where $\mathbf{A} = \mathrm{diag}(\alpha_0, \alpha_1, ..., \alpha_N)$ and $\mathbf{B} = \sigma^{-2}\mathbf{I}_N$.

3.5.1 Hyperparameter Optimisation

Identifying values for the posterior mean (3.71) and covariance (3.72) of the model weights requires first identifying most-probable values for the hyperparameters, $\boldsymbol{\alpha}_{MP}$ and σ^2_{MP}, which can be computed by maximising (3.69) with respect to $\boldsymbol{\alpha}$ and σ^2. However, values of $\boldsymbol{\alpha}$ and σ^2 which maximise (3.69) cannot be obtained in closed form [75]. Instead, formulae for their iterative re-estimation can be derived.

For $\boldsymbol{\alpha}$, by deriving (3.69), equating to zero, and rearranging, it can be shown that [75]

$$\alpha_i^{new} = \frac{\gamma_i}{\mu_i^2} \tag{3.73}$$

where μ_i is the i-th posterior mean weight from (3.71) and the quantities γ_i are defined as

$$\gamma_i \equiv 1 - \alpha_i \Sigma_{ii} \tag{3.74}$$

where Σ_{ii} is the i-th diagonal element of the posterior weight covariance in (3.72), computed with the current estimates for $\boldsymbol{\alpha}$ and σ^2. For the current estimate of the noise variance σ, differentiation of (3.69) leads to the re-estimate

$$(\sigma^2)^{new} = \frac{\|\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{\mu}\|^2}{N - \Sigma_i \gamma_i} \tag{3.75}$$

The learning algorithm thus iterates by repeated application of (3.73) and (3.75), with concurrent updating of the posterior mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ from equations (3.71) and (3.72) respectively, until some convergence criteria are satisfied and $\boldsymbol{\alpha}_{MP}$ and σ^2_{MP} are identified. In practice, during re-estimation, it is generally the case that many of the α_i tend to infinity (or numerical equivalent, given machine accuracy). From (3.70), this implies that $p(w_i|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2)$ becomes highly (in principle, infinitely) peaked at zero, which implies that we are *a posteriori* certain that those w_i are zero. The corresponding basis functions can thus be ‘pruned’, and sparsity is realised.

3.5.2 Prediction Generation

Once the hyperparameter values have been identified, predictions can be generated by the model for new input data $\mathbf{x}_*$ in terms of the predictive distribution (3.76), based on the posterior distribution over the weights, conditioned on the maximising values $\boldsymbol{\alpha}_{MP}$ and σ^2_{MP}

$$p(t_*|\mathbf{t}, \boldsymbol{\alpha}_{MP}, \sigma^2_{MP}) = \int p(t_*|\mathbf{w}, \sigma^2_{MP}) p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}_{MP}, \sigma^2_{MP}) d\mathbf{w} \tag{3.76}$$

Since both terms in the integrand in (3.76) are Gaussian, the model output follows a Gaussian distribution, so that

$$p(t_*|\mathbf{t}, \boldsymbol{\alpha}_{MP}, \sigma^2_{MP}) = \mathcal{N}(t_*|y_*, \sigma^2_*) \tag{3.77}$$

with the predicted mean and variance estimate for the new input data $\mathbf{x}_*$ given by

$$y_* = \boldsymbol{\mu}^T \boldsymbol{\phi}(\mathbf{x}_*) \tag{3.78}$$

$$\sigma^2_* = \sigma^2_{MP} + \boldsymbol{\phi}(\mathbf{x}_*)^T \boldsymbol{\Sigma} \boldsymbol{\phi}(\mathbf{x}_*) \tag{3.79}$$

3.5.3 Fast Marginal Likelihood Maximisation

A major training performance improvement in the sparse Bayesian learning scheme was also introduced by Tipping and Faul [76], who presented a fast marginal likelihood maximisation scheme. The new highly accelerated algorithm exploits identified properties of the marginal likelihood function to enable maximisation via a principled and efficient sequential addition and deletion of candidate basis function. Details of the new algorithm can be found in [76].

Chapter 4

Dry Vacuum Pump Condition Monitoring

4.1 Introduction

The manufacturing of semiconductor devices often involves over 300 different processing steps to produce nanometer scale devices. Each step involves different manufacturing processes including deposition, ion implantation, etching and lithography. By repeatedly applying different processing steps, semiconductor devices are manufactured on silicon wafers. Semiconductor devices are manufactured in fabrication facilities, commonly referred to as *fabs*. Different processing tools within a fab are used for different processing steps and the semiconductor wafers are transported around the fab to each tool, as necessary. Processing tools are located in fab, which is a extremely clean, dust free, environment. Support equipment, such as chillers, pumps, and abatement equipment are usually located in a separate facility below the fab, commonly referred to as the *subfab*.

Most semiconductor manufacturing processes require some level of vacuum, or reduced pressure, to operate, typically in the range of slightly below atmospheric pressure, down to what is known as "ultra-high vacuum", where pressures are measured in the region of 10^{-9} Torr [77]. To achieve the necessary vacuum level required for a specific manufacturing processes, different types and combinations of vacuum pumps are used. Low to mid-levels of vacuum (a few mTorr) can usually be achieved by using a mechanical dry vacuum pump in isolation. To achieve ultra-high vacuum, a dry vacuum pump is often used in combination with a turbomolecular pump or cryogenic pump.

Vacuum pumps used in semiconductor manufacturing are generally very reliable. However, when applied to pumping on particularly harsh manufacturing processes, involving toxic gases and solid particulates, pumps can occasionally suffer from unexpected failures. At the same time, growing pressure on profit margins has lead semiconductor manufacturers to increase their focus on improving process yields, tool uptime, and wafer throughput. At the same time, the increase in the use of tools for 300mm wafers and the introduction of new material technologies and cleaning flows constantly raise the by-product challenge and increase the value of wafers [78]. The occurrence of a vacuum pump failure can cause irreparable damage to wafers, but also results in significant tool downtime and cleanup, which can can be a major expense. Furthermore, a vacuum pump failure results in unplanned maintenance of a pump which is significantly more expensive than planned maintenance, in terms of resources, planning, and manpower. As a result, semiconductor manufacturers are increasingly interested in the development of condition monitoring solutions for vacuum pumps, to reduce instances of pump failures and scrapped wafers, and also to reduce tool downtime and reductions in wafer throughput

In this chapter, the development of a condition monitoring solution for dry vacuum pumps used in semiconductor manufacturing is presented. The primary issue addressed is the pumping of large quantities of fluorine gas, which corrode the pumps and reduce their vacuum performance and remaining useful life. The layout of this chapter is as follows. Section 4.2 introduces dry vacuum pumps used in semiconductor manufacturing and describes the design of the specific type of dry vacuum pump investigated in this study. Section 4.2 also discusses the data collection system used to collect and store sensor data generated by the pumps. Section 4.3 describes the specific issue which this chapter addresses and the motivation for undertaking this study. Section 4.4 then discusses the signals available and preprocessing necessary for developing models to describe vacuum pump degradation. Section **??** then presents the results of the study in identifying and tracking the development of fluorine induced degradation in dry vacuum pumps. Section **??** then introduces an approach for predicting the remaining useful life (RUL) of pumps suffering from degradation. Finally, a brief conclusion of the chapter is presented in Section 4.6.

4.2 Dry Vacuum Pumps for Semiconductor Manufacturing

Dry vacuum pumps for semiconductor manufacturing were introduced in the late 1980's to replace oil-sealed rotary vane pumps which, until that time, had been the vacuum workhorse of the semiconductor industry. The term "dry" means that the pumps contain no oil or lubricating fluid within the swept pumping volume. The switch to dry vacuum pumps occurred due to the requirement for frequent oil changes in oil-sealed rotary vane pumps, and because of the risk of contamination caused by oil backstreaming into the processing chamber [79].

Several publications have addressed the issue of condition monitoring and predictive maintenance of dry vacuum pumps. Mooney and Shelley [78] presented an overview of new capabilities in pump predictive maintenance through the introduction of networked monitoring systems. Konishi [80] investigated the accumulation of process by-products within the running clearances of a dry vacuum pump, which causes friction resulting in the pump motor current exceeding current limits and causing the pump to shut down. An ARMAX model was developed to predict when the vacuum pump motor current would exceed acceptable limits. The use of fuzzy-logic based condition monitoring is considered in [81], where a fuzzy-model based diagnostic scheme to detect mechanical inefficiency and exhaust system blockage in a dry vacuum pump was designed. The system is based upon time and frequency analysis of the exhaust pressure signal. It is demonstrated that the power ratios of certain frequency components in the signal spectrum can be used to predict the gas load, motor current and, hence, mechanical efficiency.

In this chapter, the specific type of vacuum pump investigated are iH600 dry vacuum pumps manufactured by Edwards Vacuum, a major supplier of vacuum and abatement technology to the semiconductor industry. Section 4.2.1 provides a review of the design of iH series dry vacuum pumps.

4.2.1 iH Series Dry Vacuum Pumps

The Edwards iH series of dry vacuum pumps are an industry standard dry vacuum pump used in semiconductor fabs worldwide, on a wide variety of different manufacturing processes, where particulates, condensables and corrosive by-products are present. The iH system operates at pressures between atmospheric and ultimate vacuum with no lubricating fluid in the pumping chambers. The iH series of dry vacuum pump utilise two separate pumps mounted in series. The primary pump is a five-stage dry pump, known as the harsh chemical dry pump (HCDP). A second, single-stage, mechanical booster pump, known as the harsh chemical mechanical booster (HCMB) pump, is located at the inlet to the HCDP. Figure 4.1 illustrates the layout of a typical iH series dry vacuum pump.

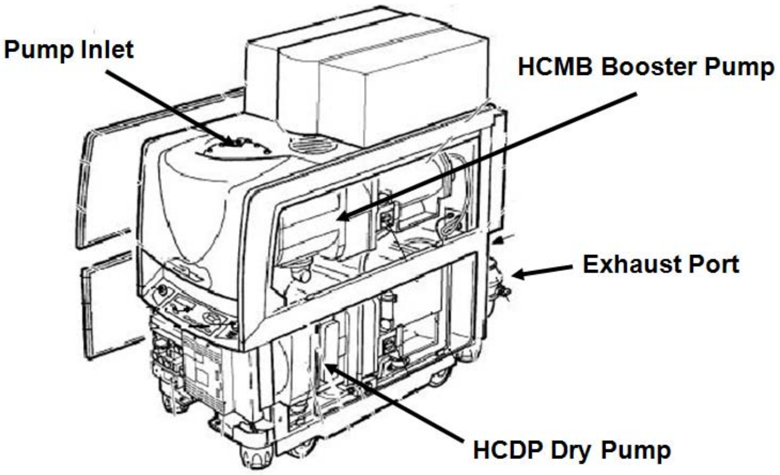


Figure 4.1: iH Series Dry Vacuum Pump

Other features of the iH series pumps include a nitrogen gas purge flow system to dilute the chemicals received from the upstream process tool and to keep corrosive chemicals away from critical components such as the pump shaft seals. A water cooling system is also employed to avoid pump overheating. A brief review of the design of the HCDP and HCMB pumps is presented below.

Harsh Chemical Mechanical Booster Pump

The harsh chemical mechanical booster (HCMB) pump on iH series pumps employ a roots mechanism, which is a valveless positive displacement device. The roots mechanism uses “figure-of-eight” rotors, also referred to as impellers or lobes, which are synchronised by timing gears. Figure 4.2 shows the cross section through a typical HCMB. The rotors rotate in opposite directions inside a stator and do not touch each other or the stator walls. The clearance is generally 0.1 to 0.5 mm when cold. The rotating lobes create an expanding volume which is then trapped between the rotor and the stator wall. As the rotors continue to rotate, the gas is compressed and moved around the stator wall of the pump. Once the rotor reaches the pump outlet, the compressed gas is then exhausted to the outlet. The HCMB operates at rotational speeds between 1400 and 4000 revolutions per minute (RPM) (without a motor inverter) [82].

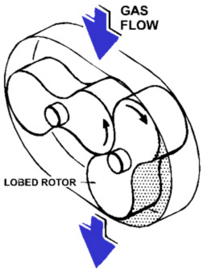


Figure 4.2: Roots Mechanism - Cross Section

Overheating is a source of trouble in roots pump designs. Heat generated by the compression process is transmitted to the rotors and the stator housing. Due to the clearance between pumping components, back-leakage of gas occurs at a rate governed by the pressure difference between the pump inlet and outlet (compression ratio) and the type of gas being pumped. The overheating issue is further aggravated by the back expansion of gas, at exhaust pressure, into the displacement volume where it undergoes further compression and heating. This inefficiency increases the rate of heat generation in the pump. Under high gas loads, the resulting temperature rise and expansion can potentially lead to the closure of the working clearances and consequent seizure. To address this issue, iH600 pumps incorporate a pressure relief valve to decrease the pressure differential across the booster pump. Larger variants of iH series pumps incorporate an electrical inverter, which increases the rotational speed of the booster pump up to approx 6000 RPM [83], but which is also used to reduce the HCMB rotational speed under high gas loads.

Harsh Chemical Dry Pump

The HCDP within each iH series variant is a five-stage positive displacement rotary pump. The iH series HCDP employs pairs of intermeshing rotors (of different types mounted on common shafts) which are held in correct phase relation by a pair of timing gears. The profiles/mechanisms employed consist of a roots stage followed by three claw stages. The final, fifth, stage is a 5-lobe roots rotor which is used to reduce the pump noise at the exhaust end, eliminating the need for a pump silencer. The combined roots/claw profile, used in all iH series HCDP pumps, is illustrated in Figure 4.3.

The main type of rotors employed on the dry pump are claw rotors. Unlike the roots mechanism used in the HCMB pump, the claw rotors mechanism is commonly described as a true compressor, as it can deliver to atmosphere without the need for exhaust

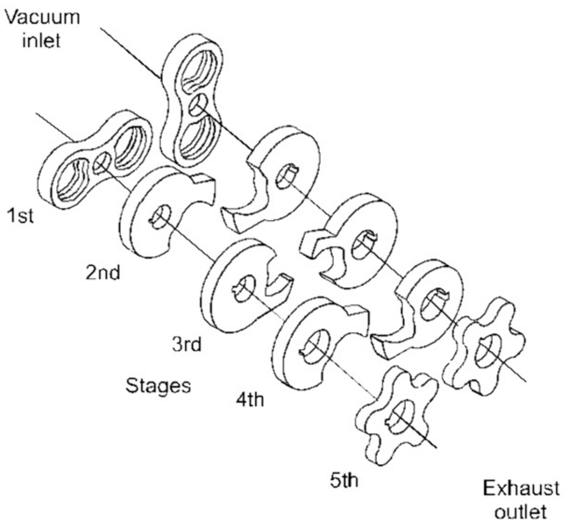


FIGURE 4.3: HCDP: combined roots/claw mechanism

cooling or pressure relief bypassing. The non-contact claw rotors are cylindrical for most of the circumference of the rotor, but have a deep depression followed by a protruding claw. Each claw combinations inlet and exhaust are arranged horizontally, rather than arranged vertically as in the roots mechanism used in the HCMB pump. As the two rotors rotate, gas is drawn in via an inlet slot, which matches the cavity in one of the rotors. Continued rotation closes the inlet while the "claws" compress the trapped volume of gas until the cavity in the second rotor exposes the outlet or exhaust slot (i.e. the mechanism is self valving). A small volume remains trapped and is "carried over" into the next rotational cycle.

The roots mechanism performs most efficiently at lower pressures, in contrast to the claw mechanism which performs more efficiently at higher pressures. A method of achieving optimum performance across a wide pressure range is to use both the roots and claw mechanisms together, as is the case in all iH series HCDP pumps. The use of a single roots stage at the HCDP pump inlet, followed by three claw stages, has been shown to be capable of developing 60% greater pumping speed than an all claw design, at 10^{-1} mbar [82]. This can be partly attributed to the better inlet conductance effect of the roots mechanism.

In the Edwards HCDP dry pumps, the claw stages are reversed compared to adjacent stages. This means that the outlet of the first claw stage is directly inline with the inlet of the second stage and provides the most direct gas path through the pump. This is illustrated in Figure 4.4. This arrangement of the claw stages contributes to good vacuum performance and improved power utilisation, enabling particles to pass easily through the pump. It also minimises the area available for the build up of corrosive or condensable residues.

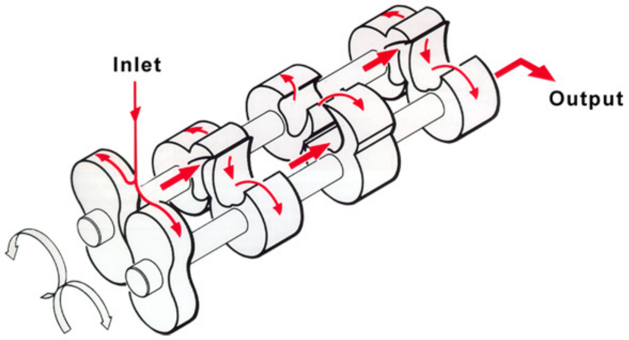


FIGURE 4.4: HCDP: reversed claw stages design

The use of the reversed claw arrangement also results in gas temperature increases through the pump, towards the exhaust side, with no cold spots present, which could arise if a reversed claw arrangement were not employed, due to the longer gas paths. The maintenance of elevated temperatures through the pumps substantially reduced the deposition of condensables. A further advantage of the claw mechanism is its ability to deal with precipitation on the internally exposed surfaces of the stator wall. This precipitation tends to reduce running clearances. The use of the rotating claws helps to "shave" the static surfaces as fast as deposition occurs. In normal use, few problems arise as the high gas throughputs are capable of transporting dust efficiently through the pump. Such debris is generally swept by centrifugal force around the periphery of the stage ahead of each claw, scouring the curved surfaces clean. However, rigid pump shafts and powerful motors are employed as occasional flakes of precipitate may sometimes detach and, in order to transport the debris towards the exhaust, the rotors must first pulverise the load via a milling action.

4.2.2 FabWorks Networked Monitoring System

In this study, all of the pump sensor data used was collected by the FabWorks networked monitoring system [84]. The FabWorks networked monitoring system, offered by Edwards Vacuum, allows real-time vacuum and exhaust gas management system monitoring over a local area network (LAN). The FabWorks system allows for data logging of up to 3000 items of equipment connected to the network, where each piece of equipment connected to the FabWorks network sends regular updates on sensor values and operating status information to a central server, where it is stored and analysed.

The FabWorks system was originally developed prior to the widespread proliferation of ethernet based network communications. As a result, network bandwidth in many older generations of FabWorks installations is limited. Newer generations of the FabWorks systems almost exclusively use ethernet based network communications and, as a result, network bandwidth issues are not as much of a concern. In the current study, the pump data was collected using a FabWorks network system installed before the availability of ethernet based communications. This means that network traffic across the FabWorks network needed be managed to avoid network saturation.

To address the issue of limited network bandwidth, the FabWorks system uses an event-based sampling scheme known as send-on-delta, or delta logging [78, 85]. The principle of send-on-delta sampling is to regularly sample the latest value for each sensor locally at each pump or abatement device. Once the latest value of a sensor variable $x(t_i)$ is sampled, it is compared to the value of the most recent sample sent across the network to the central database $x(t_{LastSent})$. If the difference between the two values is greater than a preset threshold δ, such that

$$|x(t_i) - x(t_{LastSent})| \geq \delta \tag{4.1}$$

then the latest value sampled, $x(t_i)$, is time stamped and sent to the central database. Otherwise, the latest sampled value is ignored. As a result, the major factor influencing the rate at which signal updates are sent across the network is the value of the δ chosen for each sensor variable. The δ value for each sensor essentially acts as a trade-off between signal tracking accuracy and network load. The smaller the value of the δ for a specific sensor, the greater the signal tracking accuracy and the greater the network traffic generated, and vice versa. Another feature of the FabWorks system is the timeout setting. In the event of the δ value for a sensor not being exceeded for a significant period of time, then an update containing the most recently sampled value is sent to the central database once a timer function elapses.

The FabWorks system allows for the user to select the data rates at which parameters are logged by selecting the δ values for each sensor variable. Another issue generated by the use of send-on-delta sampling is that the data recorded in the central database is on an irregular sampling interval. Overall, the send-on-delta data collected by the FabWorks monitoring system presents some challenges from a data analysis perspective, which are elaborated on further in Section 4.4.2 .

4.3 Problem Description

4.3.1 Fluorine Induced Pump Degradation

Vacuum pumps operating in the semiconductor manufacturing environment are exposed to a wide variety of often toxic and corrosive gas mixtures. On certain manufacturing processes, the types of gases pumped can result in a gradual deterioration in pump performance over time. One particular gas, commonly used in semiconductor manufacturing, which can cause significant and accelerated rates of pump degradation is Nitrogen trifluoride (NF_3).

In semiconductor manufacturing, one of the most common manufacturing processes involves the deposition of multiple thin-films, of various materials, onto the surface of a wafer. A common technique for performing this task is chemical vapour deposition (CVD). CVD facilitates the growth of thin-films of a material on the surface of a wafer, by exposing it to one or more volatile gases, which react on the surface of the wafer to produce the desired deposit. In addition to depositing material onto the surface of a wafer, CVD can also result in solid residues being deposited onto the internal surfaces of the wafer processing chamber. These unwanted residues can influence the electrical characteristics of the processing chamber, potentially leading to process drifts, particle defects, and yield loss [86]. In order to maintain production yields, and manufacturing quality, periodic cleaning of each chamber is necessary to remove these deposits from the chamber walls.

A common approach to maintaining processing chamber uniformity is to run a clean cycle between each processed wafer, to remove any deposited material. A current industry standard practice to remove deposited material involves the use of remote NF_3 plasma clean processes. The clean process operates by generating an NF_3 plasma upstream of the process chamber, using radio-frequency (RF), or microwave (MW) excitation. Once the plasma is generated, reactive fluorine neutrals drift into the process chamber where they react with the deposited residue, such as silicon dioxide (SiO_2) or silicon nitride (Si_3N_4). The reaction forms a volatile etch product such as (SiF_4) which is then pumped out of the chamber and down into the dry vacuum pump and onto a suitable abatement unit for processing [87]. Another major by-product produced by the remote plasma cleaning cycle is elemental fluorine (F_2). This is a highly reactive and corrosive gas which exits the processing chamber and enters the dry vacuum pump downstream. Depending on the manufacturing process, different quantities of NF_3 are employed in the cleaning cycle, as necessary. On those processes which use significant quantities of NF_3, the impact of generated fluorine gas on the operational lifetime of the dry pump can be significant.

Fluorine gas acts in a number of ways to degrade and reduce the performance and efficiency of dry vacuum pumps. One of the primary reasons for loss of vacuum performance is the degradation of the stator o-rings, which are gradually corroded by the fluorine gas. The stator o-rings seal each individual compression stage from those beside it and also provide a seal between the pump internals and the outside atmosphere. The degradation of the seals can result in leak paths between the pump stages and can eventually create leak paths from atmosphere into the pump. Figure 4.5 illustrates an example of an iH600 pump removed from service for poor vacuum performance and stripped down for failure analysis. Evidence of an external leak path can be seen at the bottom of picture where fluorine has corroded the o-ring and has caused whitening of the metal on the external side of the pump o-ring.

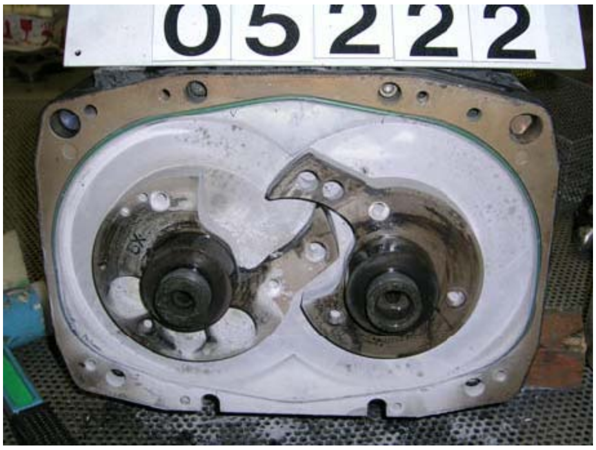


FIGURE 4.5: Evidence of o-ring damage resulting in an external pump leak-path

In addition to corroding the pump o-rings, the fluorine gas also corrodes and removes the polytetrafluoroethylene (PTFE), more commonly known as teflon, which coats many of the pump components. This results in an increase in the size of the pump running clearances as the teflon is removed. The increased running clearances further reduce the pumping performance and increases the rate of back-leakage of gas, causing the pump temperature to rise.

The degradation in vacuum pump performance, caused by the pumping of large quantities of NF_3 gas, results in a gradual increase in pressure in the upstream wafer processing chamber. Eventually the loss in vacuum performance causes the acceptable pressure limits in the wafer processing chamber to be exceeded, resulting in the automatic shutdown of a process tool, due to a high-pressure alarm. This can result in the loss of any wafers present in the chamber at the time, which can be of significant value. Furthermore, the shutdown of any tool in mid process often requires that the chamber undergoes a major clean, resulting in significant tool downtime and the unexpected removal of tool from the production line. This can have effects across the manufacturing facility, impacting on production scheduling and resulting in production bottlenecks, and lower wafer throughput rates. In addition, the costs associated with performing unplanned maintenance on a vacuum pump are far higher than the planned swap out of a pump.

Analysis of sensor data, collected from pumps suffering from fluorine induced degradation, clearly shows a gradual drift in pump sensor values over time, as the pumps begin to degrade. However, a number of issues presented in developing automated condition-monitoring solutions for this problem include; 1.) quantifying the level of pump degradation from analysis of the pump sensor data and 2.) estimating the remaining useful life (RUL) of a pump suffering from fluorine induced degradation. Addressing these two issues is the primary objective of this study.

4.3.2 Motivation

Semiconductor manufacturers operate in a highly competitive environment and are very protective of their process chemistries, recipes, and general operations. As a matter of routine they do not supply any information to outside parties regarding the types and quantities of different gases employed, or the rate of upstream wafer processing. These process related properties and operating characteristics are often proprietary and, as such, are often inaccessible to vacuum pump suppliers [88]. As a result, vacuum pump suppliers, who often operate and maintain pumps on-site on behalf of semiconductor manufacturers, are challenged to optimise pump maintenance, relative to the characteristics and operational profile of the different manufacturing processes.

The primary driver and motivation for this study was the interest, and willingness, of a major semiconductor manufacturer to provide assistance in carrying out this work. The manufacturer was experiencing excessive numbers of pump failures on a specific deposition process, which used significant volumes of NF_3 gas in the chamber cleaning cycle. As the fluorine induced degradation manifests itself in the form of increasing pressure in the upstream chamber, the semiconductor manufacturer was willing to provide access to relevant upstream process measurements.

The issue of fluorine induced pump degradation is also one which will become more important with the imminent introduction of 450mm wafers. The current standard size of wafers is 300mm. The increase in wafer sizes will require larger pump sizes to handle the increased gas loads, but will also require increased volumes of fluorine gas to clean the bigger process chambers. With the arrival of 450mm wafers, and the increasing reduction in the size of semiconductor components, the value of wafers will also increase significantly and, as a result, the costs associated with the in-service failure of a dry vacuum pump will also increase. Thus, the requirement for condition monitoring solutions to address fluorine induced vacuum pump degradation will only increase in future years.

A major driver for this study is to identify any benefits of incorporating such upstream process data in the development of condition monitoring algorithms for vacuum pumps. By having access to, and incorporating, such data in the development phase, it will be demonstrated how pump degradation can be identified, tracked, and predicted in real-time, using only the available pump data. In Section 4.4.1, a description of the specific upstream process measurement used this study, which were provided by the semiconductor manufacturer, is presented.

4.4 Preliminary Signal Analysis & Feature Extraction

4.4.1 Foreline Pressure Data

As described earlier in Section 4.3.2, a major semiconductor manufacturer was willing to provide access to upstream process data which might be useful in the development of algorithms for pump condition monitoring. The primary upstream process measurement which was identified as the signal most likely to provide visibility on the loss of vacuum performance is the foreline pressure (FP) signal. The FP sensor is located directly downstream of both the processing chamber and the turbo-pump, on the foreline between the turbo pump and dry vacuum pump. Figure 4.6 illustrates the equipment layout on the investigated manufacturing process and the location of the FP sensor.

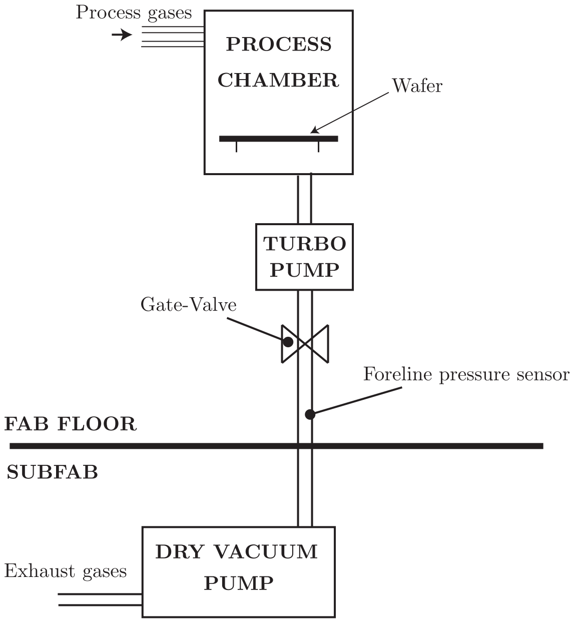


Figure 4.6: Equipment and sensor locations on investigated manufacturing process

The rational for investigating the foreline pressure signal as a suitable signal for quantifying the level of vacuum pump degradation is that, as a dry vacuum pump degrades, it was expected that the loss in vacuum performance in the dry vacuum pump would be reflected by increasing values of the FP signal. This sections describes how the foreline pressure signal, sampled at 1 second intervals by the processing tool, is preprocessed for use in describing vacuum pump degradation.

Semiconductor manufacturing processes are typically carried out using a specific "recipe" for each process. Each recipe step typically uses different mixtures of gases and requires different vacuum levels in the processing chamber. Control of vacuum in the chamber is achieved by using a gate-valve, which is automatically controlled to maintain the pressure at the required setpoint. Figure 4.6 ilustrates the location of the chamber gate-valve. As the gate-valve is controlled to try and achieve the desired pressure setpoint during each recipe step, the FP signal varies significantly in response to changing gas flows and gate-valve positions. As a result, some preprocessing of the FP signal was necessary.

The objective in using the FP signal is to generate a target signal which could then be used to train a model which quantifies the level of pump degradation, using the pump signals as inputs. To generate a target signal, the mean foreline pressure value over a specific recipe step was chosen to represent the FP for each individual processed wafer. In selecting a specific recipe step over which to average the FP values, a number of issues were considered:

1. The specific recipe step should be of significant duration relative to the overall wafer processing time so that the mean FP value over the chosen time-interval has least variance.
2. The recipe step should ideally be a high gas load recipe step so that the FP value rises significantly and that a high gas load is applied to the pump. Imparting a high gas load on the pump stresses the pump and, as the pump degrades, the loss of vacuum performance should be reflected in the FP value, during the specified recipe step, increasing over time.
3. The chamber gate valve should not be under control during the chosen recipe step and should ideally remain fully open for the duration of the recipe step. This ensures that the rate of gas load discharge from the chamber remains consistent for each wafer, and that the gate valve does not compensate for the loss of vacuum performance over time.

Figure 4.7 illustrates the FP signal, recorded during the processing of a single wafer. The corresponding recipe step number at each sample time is illustrated on the opposite axis. To protect the intellectual property of the semiconductor manufacturer, the FP data provided was mapped onto an arbitrary scale so that FP values have no physical meaning. In addition, the recipe step numbers have been altered and the actual duration of the recipe is not indicated on the time axis.

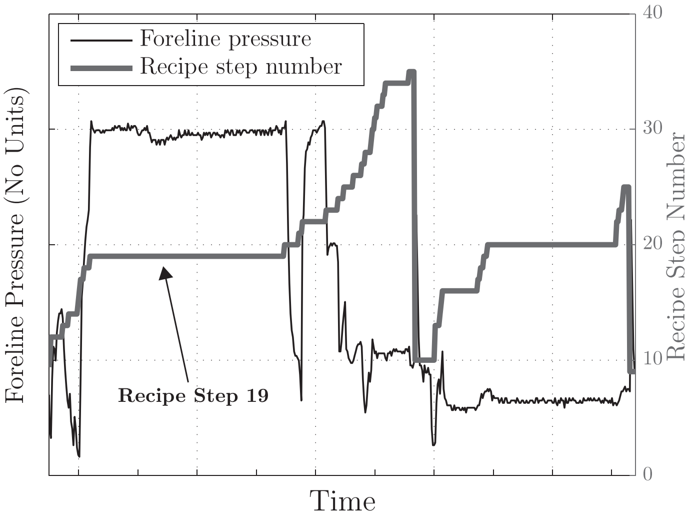


Figure 4.7: Foreline pressure (FP) by recipe step (single wafer)

As illustrated in Figure 4.7, the variability in the value of the FP signal over the course of the recipe is significant. The variability in FP is due to changes in gas flows. At each recipe step, different gases flow at different rates causing fluctuations in the FP value. The gate-value is controlled to maintain the chamber pressure at the desired setpoint for each recipe step.

From analysis of the FP signal illustrated in Figure 4.7 and the chamber gate-valve signal, the mean FP value during recipe step 19 was selected to represent the FP value for each individual processed wafer. The FP values for each processed wafer were then combined to form a vector of FP values over time. This vector of FP values over the lifetime of a pump was then employed to represent, and quantify, the level of vacuum pump degradation over time. Figure 4.8 below plots the extracted FP value for each processed wafer over the lifetime of a single pump which failed in-service due to poor vacuum performance. As illustrated in Figure 4.8, the FP remains relatively constant for a period of time. As the pumped fluorine begins to corrode the pump, the FP signal begins to rise in response to increasing levels of pump degradation. The values of the FP signal have also been mapped onto the interval $[0, 1]$. The values on the interval $[0, 1]$ are thus used to represent the level of vacuum pump degradation.

The rational for using the FP signal to describe the evolution of fluorine induced vacuum pump degradation is clearly motivated by the observed behaviour of the FP signal in Figure 4.8. This behaviour, involving monotonically increasing values of FP signal

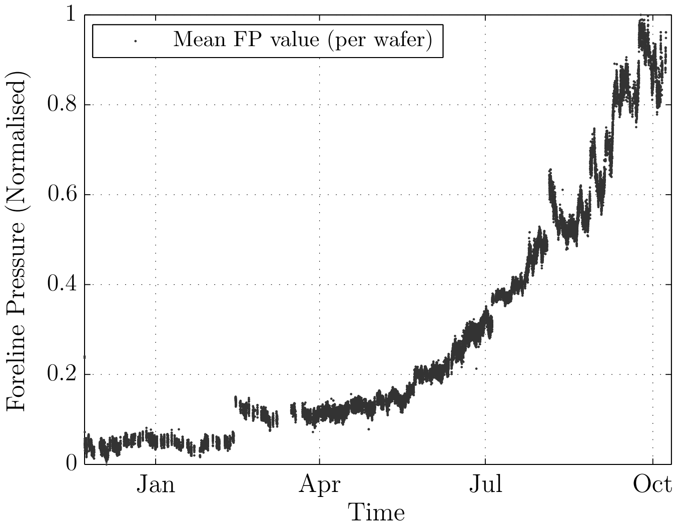


FIGURE 4.8: Foreline pressure by recipe step

which reflect the loss in vacuum performance leading up to pump failure, was observed in all the available historical pump failure examples.

4.4.2 Pump Signal Data

For the period of this study, pump sensor data collected from three high-density-plasma (HDP) CVD processing tools was available for developing a condition monitoring algorithm for detecting, tracking, and predicting fluorine induced pump degradation. Pump data covering the period January 2007 to December 2008 was available for the study and, during this period, a total of 4 pumps failed in-service or were removed from service for poor vacuum performance.

The dry vacuum pumps employed on each processing tool were Edwards iH600 mechanical dry vacuum pumps. A description of the design of iH series pumps was presented previously in Section 4.2.1. Each HDP-CVD processing tool has two individual wafer processing chambers, with each chamber having its own iH600 mechanical backing pump located in the subfab. The sensor data from each of the iH600 vacuum pumps was collected using the FabWorks networked monitoring system. Note, in the following sections, the HCDP and HCMB pumps are referred to as the dry pump and booster pump respectively. The FabWorks system recorded a number of pump variables for the period of this study, including

- Booster power

- Booster temperature
- Drypump power
- Drypump temperature
- Exhaust pressure
- Shaft-seal pressure
- Nitrogen purge flow

The FabWorks networked monitoring system records pump sensor data using a send-on-delta based logging scheme, as described previously in Section 4.2.2. This logging scheme presented a number of challenges. The logging profile for each pump variable is defined by two parameters; the δ value and the timeout value. The original logging profile used excessively high values for the δ values for each sensor variable. As a result, most sensor updates occurred due to the timer function elapsing and an automatic update being sent to the central server. However, the timer setting was originally set to 1-hour, resulting in significant periods of time between signal updates. The issue of poor signal resolution was identified at an early stage and the logging profiles on the FabWorks system were adapted to try and improve the signal resolution. The timeout setting on each sensor variable was reduced to 5 minutes, to guarantee a signal update in the event of the delta value not being exceeded during this interval. This significantly improved both the time and amplitude resolution in the temperature sensor values, though issues with the resolution of the power signals remained.

In addition to difficulties presented by the delta-logging profiles, the quantisation levels on the power sensors employed on the iH600 pumps also presented a number of issues. The Edwards iH pumps series are manufactured in a range of sizes, with the iH600 being the smallest sized pump within the iH series range. However, the sensors employed on each iH variants are the same and, in the case of the iH600, the quantisation levels on the power sensors are relatively large when compared to the typical operating range of the power signals. Consider Figure 4.9 which illustrates the booster power (BP), booster temperature (BT), drypump power (DPP), and dry pump temperature (DPT) signals recorded from an iH600 pump which was removed from service for poor vacuum performance. The pump signals in Figure 4.9 illustrate a number of issues presented by the delta-logging profile, and the sensor quantisation levels.

The quantisation levels on the power sensors are 0.1kW, however the typical operating range of the BP signal is between approximately 0.6kW and 0.9kW, and the operating

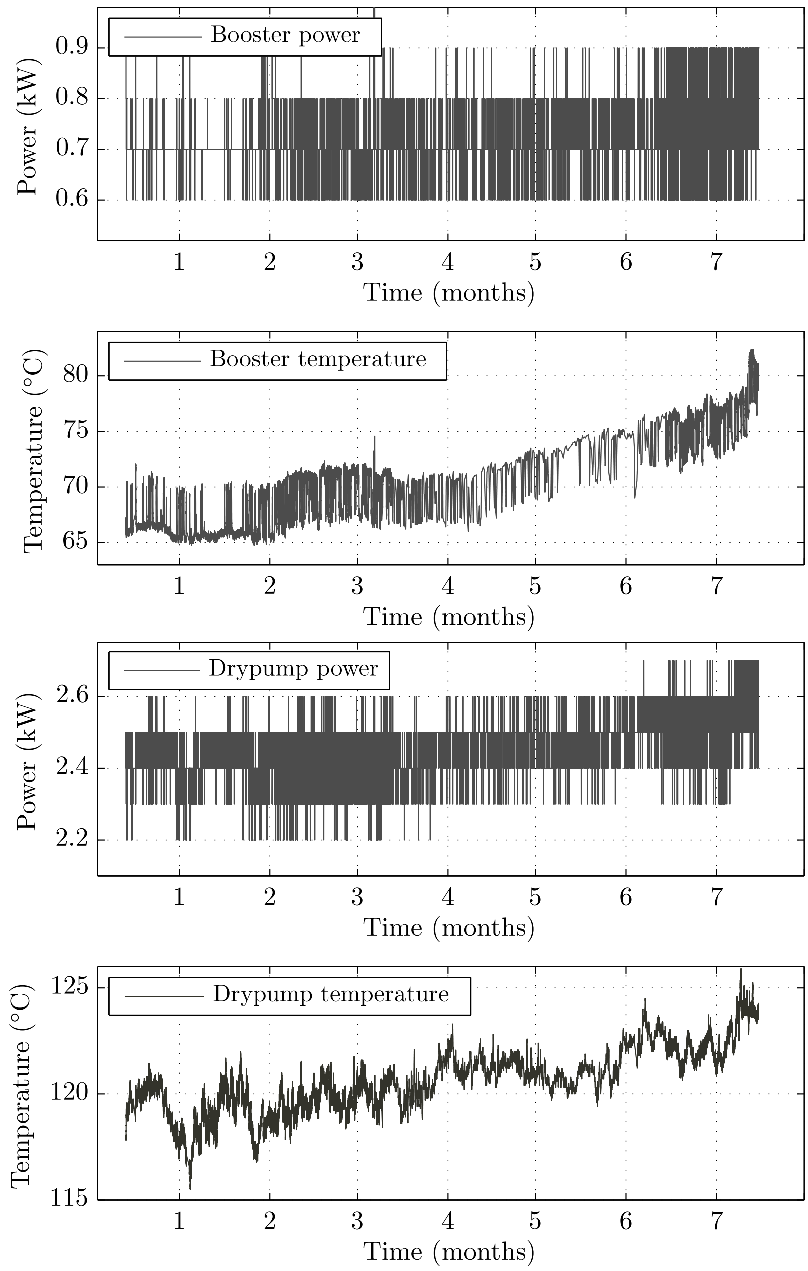


Figure 4.9: Evidence of o-ring damage resulting in external pump leak-path

range of the DPP signal is between 2.2 and 2.6kW. As a result, the signal resolution is quite poor in the amplitude domain. In addition, the delta values on the profile were set to 0.3kW, which means the power signal must change by this quantity in order for the pump to send an update to the central database. Otherwise, a new update is not sent until the sensor times out and a new update is sent. As a result, the quality of the signal resolution for the BP and DPP signals is very poor.

A further issue requiring attention is the bi-modal nature of the booster temperature signal. The short-term rise and fall of the BT signal is a result of the intermittent processing of wafers, which generate a gas load on the pump which, when compressed in the booster pump, generates heat, resulting in a rise in observed temperature values. If a signal indicating the operational state of the processing chamber is available then the booster temperature signal can be separated by pump mode in a supervised manner. However in the absence of such a signal, the pump status must be inferred from the available pump signals. A simple method for tracking the BT signal, in each of the pump operating modes, is presented in Section 4.4.4.

4.4.3 Signal Resampling & Filtering

To extract signals/features suitable for detecting and tracking pump degradation it was necessary to preprocess the power and temperature signals using smoothing techniques, to generate suitable signals which correlate with the loss of vacuum performance. However, the use of a send-on-delta based sampling approach for collecting pump sensor data presents a number of challenges. Most applications in digital control and signal processing have traditionally used periodic sampling, as there exists a well established systems theory for periodically sampled data. The lack of such a systems theory for dealing with event-driven data suggests that in order to apply classical systems theory, such as system identification or regression, some form of reconstruction of the event-based sampled data onto a periodic sampling interval is necessary.

In the current application, signal updates occur either due to the delta values being exceeded, or the five-minute timeout interval being triggered. This results in the data for each signal being recorded on an irregular sampling interval, which is different for each pump signal. To address this issue each of the pump signals were resampled onto a regular 1-minute sampling interval. The sampling interval was chosen on the basis of the observed signal update rate characteristics. In general, for both the temperature and power signals, the vast majority of signal updates are triggered by the timeout setting and a 1-minute interval was deemed appropriate to sample each of the signals onto a single regular sampling interval.

Once the periodic sampling interval was chosen, the resampling method had to be selected. The most common approaches to resampling use either a zero-order hold approach whereby, in the absence of a signal update, the most recently available signal update is used at the current sample time. The alternative approach is to use linear interpolation between available signal update values. In this study, both ZOH and linear interpolation were employed. The power signals from both the booster and drypump were resampled using a zero-order hold (ZOH) approach and the temperature signals from the booster and dry pump were resampled using linear interpolation. The selection was based upon the relative dynamics of the signals in question. Temperature is typically a relatively slow changing process and so it is reasonable to assume that the temperature values between available updates moved somewhat linearly between the recorded values. In the case of the power signals, these can change almost instantaneously and to assume that the values between updates change linearly has no basis and so, given the available signal resolution, the power signals were resampled using a ZOH approach.

Once all of the available pump data was resampled onto a periodic sampling interval of 1-minute, it was then necessary to filter the data in order to extract the relevant trends from both the power and temperature signals. The primary objective of the filtering operation was to smooth the data, particularly in the case of the power signals which have poor amplitude resolution, and to extract relevant trends which, over time, correlate with the loss of vacuum performance.

The resampled pump signals were smoothed using an exponentially weighted moving average (EWMA) filter. The principle of an EWMA filter is similar to that of a simple moving average filter except that equal weighting is not applied to each sample within the window. Instead, the weighting for older data points decreases exponentially, giving more importance to recent observations whilst not discarding older observations entirely. Equation 4.2 describes the operation of an EWMA filter

$$s_t = \alpha x_t + (1 - \alpha) s_{t-1} \tag{4.2}$$

where s_t is the output of the filter and x_t is the original signal value at time t. The filter constant α is the *smoothing factor* which controls the degree of smoothing applied to the input signal. The lower the value of α the greater the level of smoothing applied to the data. The smoothing constant α is analogous to the cut-off frequency in a low-pass filter. Figure 4.10 illustrates how the smoothing filter operation is used to extract a signal for tracking changes in the power signals from both the booster and drypump from the original, poorly resolved, data. The value of α used to generate the filtered signals in Figure 4.10 was 0.001. The value of α was chosen so as to try and extract a long-term trend from the available, poorly resolved, power signal.

The filtering operations described above were also applied to both the booster and dry-pump temperature signals to smooth the short-term fluctuations in these signals. In

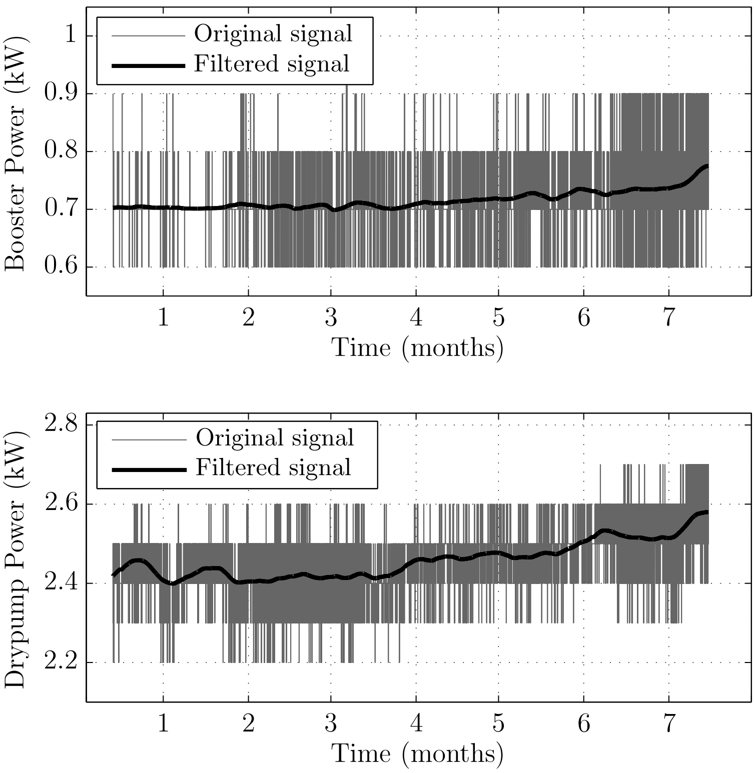


Figure 4.10: Smoothed pump motor power signals for pump lifetime

addition to the smoothing operation, additional preprocessing on the booster temperature signal was necessary due to the multi-modal nature of this signal. The details of this process are described in then following section.

4.4.4 Multi-Mode Booster Temperature Signal Tracking

The load imparted on a dry vacuum pump is a function of the gas load generated in the upstream processing chamber. Once this gas load reaches the pump it is compressed, which generates a response in the pump variables. Typically, a response is seen in the booster pump variables, as the booster pump is located at the pump inlet. However, due to the poor booster power signal resolution, the only pump variable, in the investigated data set, which illustrates a response to change in gas load imparted on the pump is the booster temperature (BT) signal. Figure 4.11 illustrates a typical BT signal profile over a period of 5 days. The BT signal profile exhibits bi-modal behaviour which reflects the operating modes of the pump.

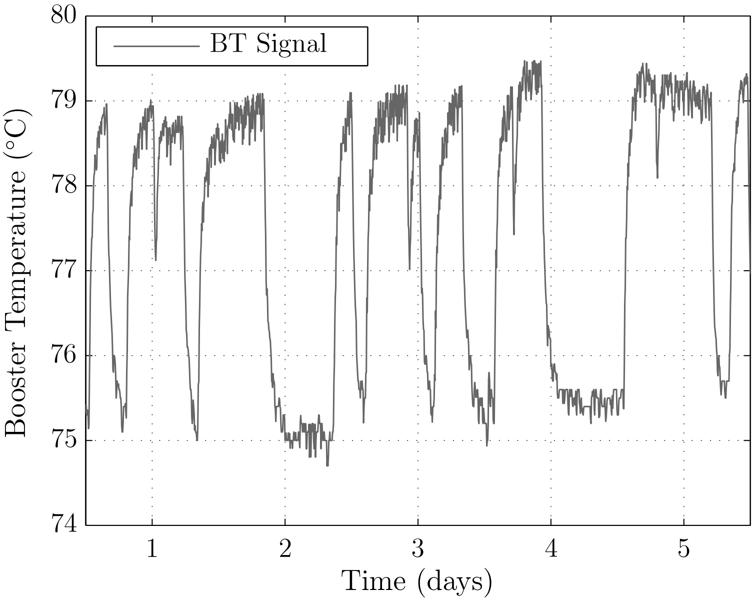


FIGURE 4.11: Typical BT signal profile

The BT signal is characterised by two different responses. During periods of no wafer processing in the upstream chamber, described henceforth as *idling mode*, the BT signal remains steady about a lower range of operating values. During periods of wafer processing in the upstream chamber, the BT signal responds and increases to a higher range of values. This higher range of values is associated with the compression of the process gases used in the upstream manufacturing process, which generates heat. Individual wafers are processed sequentially and usually arrive at the processing tool in batches of 25 wafers. During periods of upstream wafer processing the pump mode is described henceforth as *processing mode*. Once wafer processing is completed, the BT signal decays back to the idling mode range of values, until a new batch of wafers arrives.

Once a gas load is imparted on the pump, the booster pump power signal also generally increases to maintain a constant rotational speed. However, the quantisation levels on the iH600 power sensors are too small to identify and track this process. In Section 5.5, an algorithm for inferring the current pump mode from the booster power signal is presented. In the application presented in Section 5.5, the booster power signal is recorded from a larger iH1800 pump, in which the operating range of the BP signal is much larger, allowing the pump mode to be inferred. For the purpose of tracking the changes in the characteristics of the BT signal, in response to increasing levels of pump degradation, it is necessary to identify some method for tracking the BT signal in each of the pump operating modes. From observations of a typical BT signal profile, it is possible to classify the BT signal *status* as being in one of four states, as illustrated in Figure 4.12

- HIGH
- LOW
- RISING
- FALLING

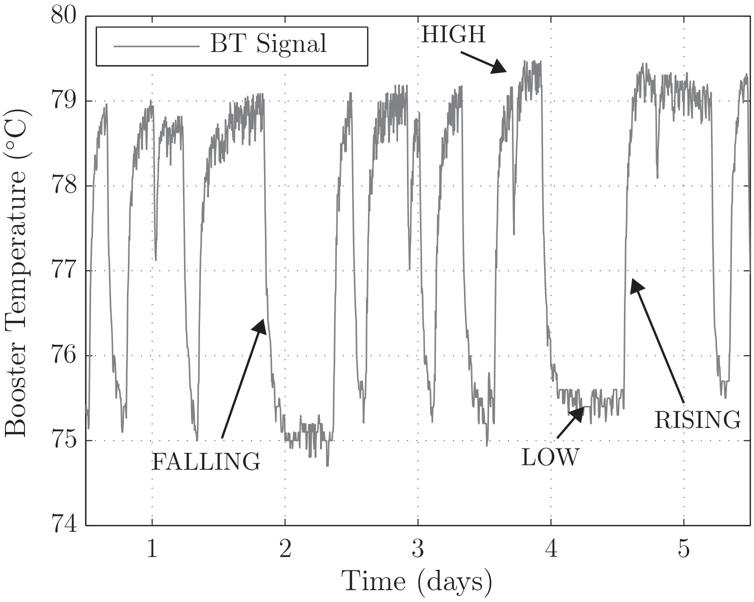


FIGURE 4.12: BT signal: signal status classification

The BT signal HIGH and RISING states are associated with periods of upstream wafer processing, whilst the FALLING and LOW states are associated with periods of no wafer processing in the upstream chamber. To classify the BT signal state at each sample time, a simple moving-window classifier was developed, which infers the BT signal state on the basis of the slope of the BT signal. Using the BT signal, which was first resampled onto a regular 1-minute sampling interval, a moving window of 20-minutes length was applied. At each iteration of the moving window, the slope of the line of best-fit was computed from the samples within the bounds of the window.

Initially, the algorithm identifies those sample times at which the BT state is either RISING and FALLING. Threshold slope values were determined for both the RISING and FALLING states on the basis of the historically observed slope profiles of the BT signal. Those samples which exceeded the relevant thresholds were classified as either RISING or FALLING. The remaining samples were left unclassified at this stage. Figure 4.13 illustrates the output of the algorithm in identifying those BT signal samples at which times the BT signal state was either RISING or FALLING.

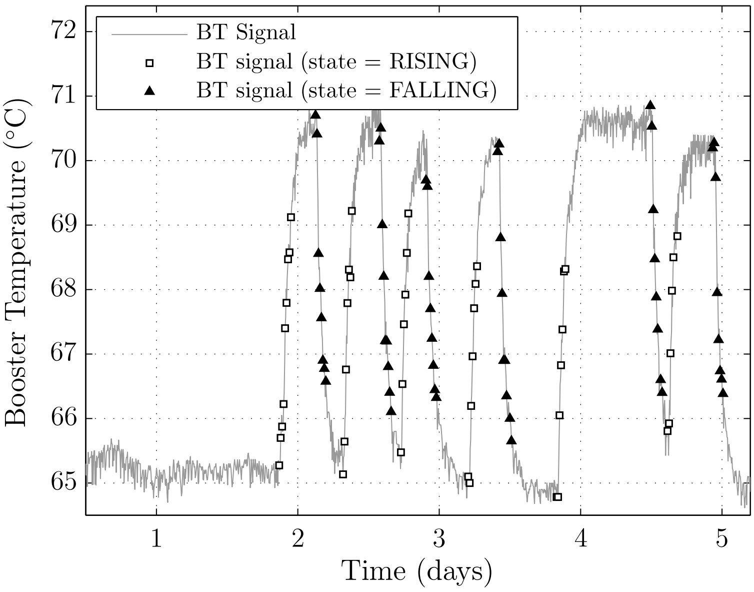


FIGURE 4.13: Pump mode identification using BT signal

Using the values of the BT signal samples, classified as either RISING or FALLING, a moving average signal was generated from these samples. On the basis of whether the remaining BT signal samples were above or below this moving average signal, the remaining BT samples states were classified as either HIGH or LOW. Once the BT signal state was identified at each sample time, two moving average signals were then generated, to describe the BT signal in each of the pump operating modes; idling and processing. Those BT samples recorded when the BT signal state was either LOW or HIGH were each arranged into two individual timeseries vectors. Both of these timeseries vectors occur on an irregular sampling interval, which reflect the different times at which the relevant BT signal states occurred, and so each of these timeseries vectors was then resampled onto the original 1-minute sampling interval and smoothed using an EWMA filter, as described previously in Section 4.4.3. The final output of the algorithm is two timeseries vectors describing the evolution of the BT signal in both the idling and processing pump modes, which is illustrated in Figure 4.14.

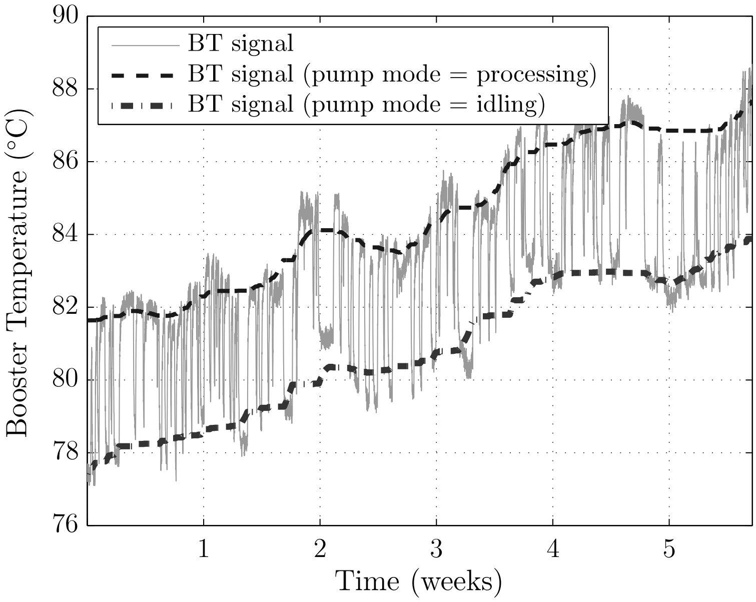


FIGURE 4.14: Evolution of BT signal under different operating modes

The separation of the BT signal into two signals, representing the operating modes of the pump, was necessary for a number of reasons. The primary reason was that the BT signal, in its original form, was not suitable for analysing pump condition over time. If the signal were simply smoothed to produce a signal useful for correlating with pump degradation, then the value of the smoothed signal would be a function of the pump utilisation rate within each short period of time. This would generate significant fluctuations in the BT signal which would not be a function of pump condition but instead be a function of the pump utilisation rate. By separating the BT signal into two separate modes it is now possible to describe the evolution of the BT signal regardless of the rate of pump utilisation.

In addition, the approach described above provides a crude method for tracking the utilisation rates of these types of pumps, which corresponds to those times when the BT signal state is identified as RISING or HIGH. This information can be useful for identifying relationships between pump failure rates and utilisation rates, as opposed to purely time based relationships which do not take into consideration actual pump usage rates, which can vary significantly.

4.5 Dry Pump Condition Monitoring Algorithm

In this section, the proposed algorithm for identifying, tracking and predicting fluorine induced vacuum pump degradation is described. A flowchart describing the proposed algorithm is illustrated in Figure 4.15. Firstly, raw sensor data recorded in real time by the FabWorks monitoring system is passed to the preprocessing stage. During this stage, various resampling, filtering, and feature extraction routines, as described in Section 4.4, are applied to the FabWorks data to generate a set of inputs for use in estimating the current level of pump degradation. The list of inputs generated by the preprocessing routines described in Section 4.4, is presented below. The preprocessing routine generates a value for each input on a regular 1-minute sampling interval.

- u_1: Booster Power
- u_2: Booster Temperature (Idling Mode)
- u_3: Booster Temperature (Processing Mode)
- u_4: Dry Pump Power
- u_5: Dry Pump Temperature

Once the set of inputs are generated by the preprocessing stage, the inputs are passed to the modelling stage, where the current level of pump degradation is estimated. Two approaches to modelling the level of pump degradation are considered, namely multiple linear regression (MLR) and artificial neural networks (ANNs). The processing of training and testing the models to estimate pump degradation on historical pump failure data is described in Section 4.5.1.

Once an estimate of the current level of pump degradation is estimated, the level of estimated degradation is compared to a threshold value. If the current level of pump degradation is below the defined threshold, the algorithm continues to iterate and estimate the level of pump degradation at 1-minute intervals. Alternatively, if the current level of pump degradation exceeds the defined threshold, a prediction of the remaining useful life (RUL) of the pump is generated. The algorithm then continues to iterate and update the current degradation estimate and the RUL prediction. The RUL predictions are generated using a double exponential smoothing prediction (DESP) technique, which is described in Section 4.5.5.

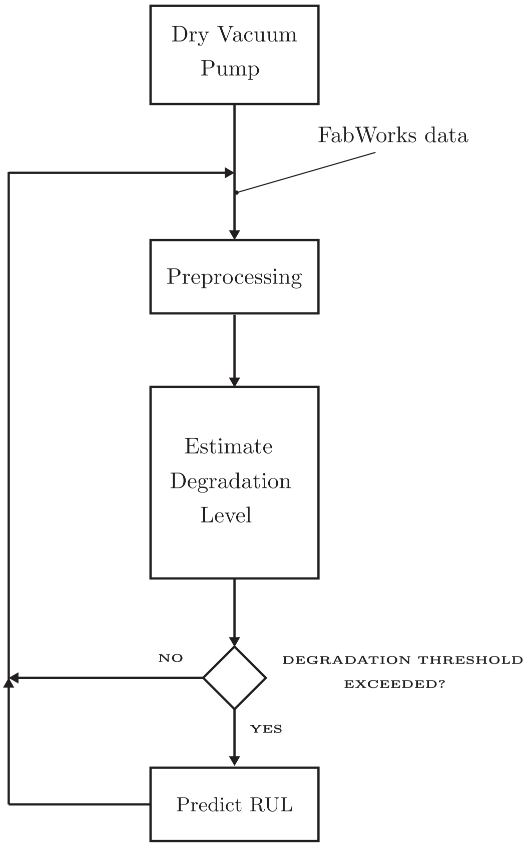


Figure 4.15: Dry pump condition monitoring algorithm

4.5.1 Modelling Pump Degradation

Having applied suitable preprocessing to the pump signal data and the upstream foreline pressure signal, the next stage is to develop a model to estimate the current level of pump degradation, using the pump variables as inputs. To achieve this task, static models relating the value of pump variables to upstream pressure values were developed. The models developed to estimate the level of pump degradation are of the form

$$\hat{r}(k) = f(u_1(k), u_2(k), ..., u_m(k)) \tag{4.3}$$

where $\hat{r}(k)$ is the estimated level of pump degradation at time k, and $u_i(k)$, $i = 1, ..., m$ are the model inputs, which comprise the values of the preprocessed pump sensor variables at time k. The input variables u_i represent the following pump variables, which were all preprocessed as described in Sections 4.4.3 and 4.4.4.

- u_1: Booster Power
- u_2: Booster Temperature (Idling Mode)
- u_3: Booster Temperature (Processing Mode)
- u_4: Dry Pump Power
- u_5: Dry Pump Temperature

The relationship between the pump sensor variables and the upstream foreline pressure is assumed time-invariant. To model the relationship between the level of pump degradation and the pump sensor variables, two different modelling approaches are considered, multiple linear regression (MLR) and artificial neural networks (ANNs). A review of the theory and backgrounds of MLR and ANNs is available in Sections 3.1 and 3.2 respectively, which should be referred to as necessary.

The motivation for investigating MLR was to determine if the relationship between the pump variables and upstream foreline pressure could be described by a linear model. In addition, MLR models are computationally inexpensive, which is a desirable property when monitoring potentially hundreds of pumps in a semiconductor manufacturing facility. In contrast, ANNs are capable of modelling complex non-linear relationships between input and output data. ANNs were investigated to determine if improved performance could be achieved in modelling pump degradation, versus MLR-based approaches.

The input and outputs signals, following the preprocessing steps discussed in Section 4.4.3, are all on a regular 1-minute sampling interval. However, since vacuum pumps generally operate continuously, this results in significant volumes of data being generated over periods of weeks and months. To address this issue, the input and output signal were downsampled to a five-minute interval. This was done simply by selecting every fifth sample. Due to the filtering applied to smooth the input and output signals, no loss of signal fidelity arose from this operation. As described previously in Section 4.4.2, a total of four historical pump failure examples are available. Of these, three failed in service and another was removed from service when poor vacuum performance was suspected.

4.5.2 Model Training using MLR

A multiple linear regression (MLR) model is used to describe a linear relationship between the preprocessed pump sensor variables $u_i(k)$, and the preprocessed upstream foreline pressure value $r(k)$. The subscripts i, where $i = 1, 2, ..., p$, refer to the individual input variables. The form of the MLR model is given by

$$r(k) = \beta_0 + \beta_1 u_1(k) + \beta_2 u_2(k) + ... + \beta_p u_p(k) + \epsilon \tag{4.4}$$

where the model parameters β_i, $i = 0, 1, ...p$, represent the *regression coefficients* and ϵ represents the modelling error term, which should ideally be a random zero-mean Gaussian distributed variable. Given a set of n observations to be modelled, representing input-output pairs of preprocessed pump sensor values and foreline pressure values, the values of β_i are identified via the method of *least squares*, as described previously in Section 3.1. Identifying the β_i values is a batch process, and the pump failure examples need only be separated into training and test data. Thus, for each available pump failure case, the β_i parameters were identified using the input-output data describing three of the failure examples, and the identified model was then tested on the single remaining failure example.

4.5.3 Model Training using ANNs

In this study, a standard feed-forward neural network was employed to model pump degradation. Various network architectures were considered and tested. The selected architecture comprised a single hidden layer with 10 neurons, using tan-sigmoid activation functions, and a single output layer using a log-sigmoid activation function, to limit the output to the range [0,1]. The range [0,1] defines the range to which the upstream foreline pressure signal were mapped to in Section 4.4.1, to represent the level of pump degradation. A single hidden layer was chosen as no improvement in estimation accuracy was observed when tested using two hidden layers. Additionally, a single hidden layer reduces the complexity of the ANN, potentially improving the generalisation capabilities of the network. For training and testing purposes, the 4 historical failure examples were separated into training, validation and test data. For network training and testing purposes, two failures were selected for use as training data, a single failure was selected for validation data, and the remaining failure was retained for testing data. The network was trained using the Levenberg-Marquardt backpropagation algorithm [58].

4.5.4 Results

This section presents the results of the models developed for identifying and tracking the development of fluorine induced vacuum pump degradation. In each case presented, the performance of the developed models was evaluated on previously unseen data, i.e. the test cases presented were not used in training either the MLR or ANN models.

Figure 4.16 presents the first examples of the ANN and MLR models tested on an historical pump failure example. As illustrated in Figure 4.16, the ANN model tracked the more general trend of increasing degradation, but the ANN output suggests that failure is occurring at a faster rate than is actually the case. The MLR model tracks the increasing degradation quite accurately though, in the later stages, the MLR models fails to describe the observed behaviour and underestimates the level of pump degradation.

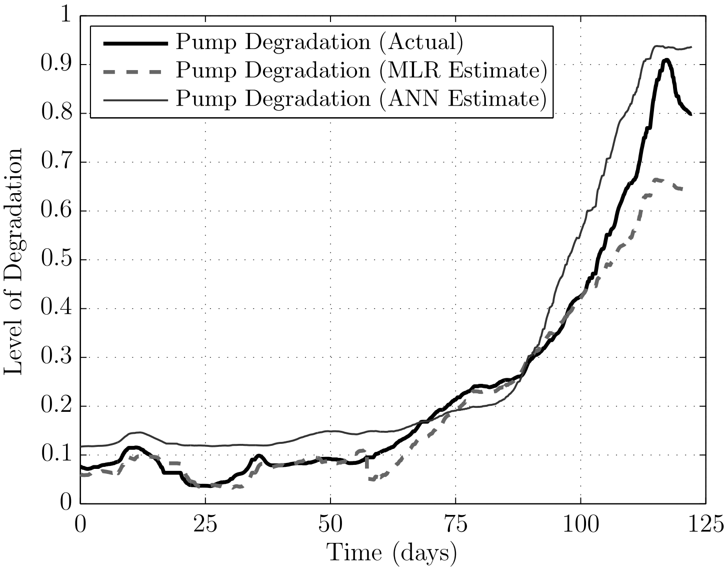


Figure 4.16: Pump degradation tracking performance (Pump A)

Figure 4.17 presents the case of the pump which was removed from service prior to failure. In this example, the MLR model overestimates the increase in the level of pump degradation, compared to the degradation tracking performance of the ANN. However, the accuracy of the MLR estimates were greater at the time the pump was removed from service.

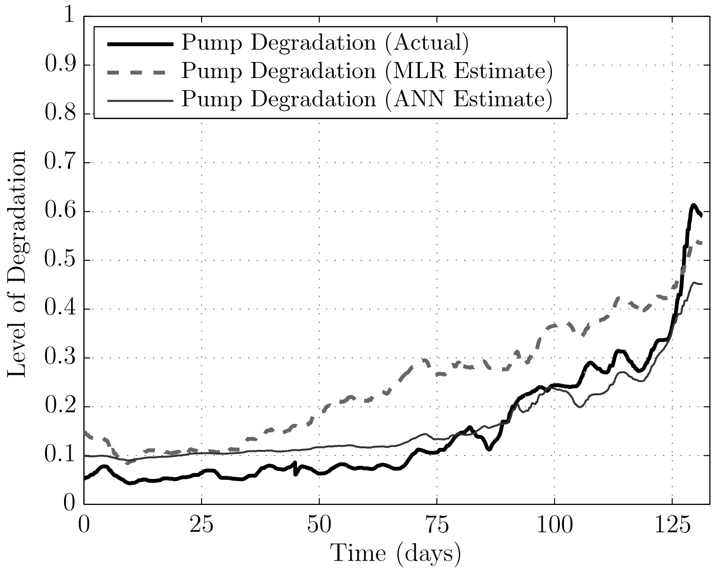


Figure 4.17: Pump degradation tracking performance (Pump B)

The final example illustrated in Figure 4.18 demonstrates good degradation tracking performance for the both the ANN and MLR models. However, the accuracy of the MLR model in the later stages of degradation fails to describe the observed behaviour.

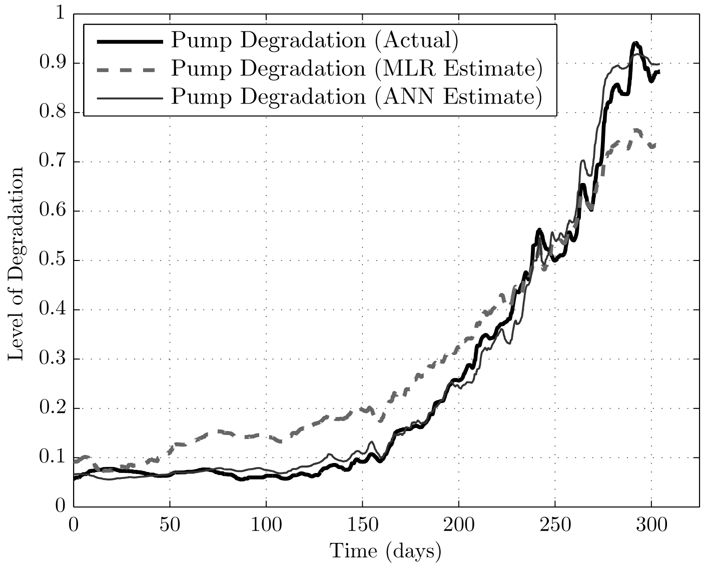


Figure 4.18: Pump degradation tracking performance (Pump C)

To compare the performance of the ANN and MLR models, Table 4.1 shows the root mean squared error (RMSE), evaluated for each failure cases by comparing the estimated and actual levels of vacuum pump degradation, where the RMSE is defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2}. \tag{4.5}$$

where ϵ_i^2 is the error between the estimated and actual level of pump degradation at time i, and n is the number of model estimates over which the RMSE is computed. The RMSE values in Table 4.1 were computed only using the model estimates generated when the actual level of pump degradation was above 0.2, as the performance of the models belows this level of degradation is less important.

Failure Case	MLR	ANN
Pump A	0.0654	0.0769
Pump B	0.0983	0.0546
Pump C	0.0803	0.0371
Pump D	0.1104	0.0624

Table 4.1: RMSE for degradation level greater than 0.2

As Table 4.1 illustrates, the performance of the ANN models is generally superior to that of the MLR models, with the exception of pump failure A. However, as illustrated in Figure 4.16, the behaviour of this failure examples was better described by the ANN, only that the ANN estimated a faster rate of degradation.

4.5.5 Predicting Pump Remaining Useful Life

Section 4.5.4 demonstrated the ability to estimate the current level of pump degradation using pump sensor variables. While such information is of benefit to pump operators, by far the most useful information for operators is an estimate of the RUL of the pump. This allows for planning of pump replacements at a convenient time, with minimal disruption to manufacturing. Predicting the RUL of a pump is a difficult task as there are a number of factors which cannot be determined *a priori*, such as the future utilisation rate of the pump (wafer throughput rate) and the number of times the pump is shutdown and restarted, which can influence the RUL.

Looking at the failure examples illustrated in Figures 4.16 and 4.18, as the level of pump degradation approaches 70% (0.7), the time series observations of the pump degradation levels can be “roughly” approximated by sloped line, of the form in Equation (4.6).

Double exponential smoothing-based prediction models a given time series using a simple linear regression equation, where the y-intercept β_0, and the slope β_1, are slowly changing over time. In such cases, double exponential smoothing can be used to apply unequal weighting to the individual elements of the time series.

$$y_t = \beta_0 + \beta_1 t + \epsilon_t \tag{4.6}$$

In order to obtain updated estimates of the time series, double exponential smoothing uses what is called the single and double smoothed statistics, S_t and $S_t^{[2]}$. These values are computed using two smoothing equations, (4.7) and (4.8), where both equations use the same smoothing constant α, which lies within the range [0,1]. This value determine the degree of smoothing applied to the data. The first equation smoothes the original time series and the second filters the S_t values.

$$S_t = \alpha y_t + (1-\alpha)S_{t-1} \tag{4.7}$$

$$S_t^{[2]} = \alpha S_t + (1-\alpha)S_{t-1}^{[2]} \tag{4.8}$$

Using S_t and $S_t^{[2]}$, updated estimates of β_0 and β_1 are determined as:

$$\beta_0(t) = 2S_t - S_t^{[2]} - t\beta_1(t) \tag{4.9}$$

$$\beta_1(t) = (\frac{\alpha}{1-\alpha})(S_t - S_t^{[2]}) \tag{4.10}$$

Having estimated β_0 and β_1, the forecast made at time t for the future value of $y_{t+\tau}$ is given by,

$$\hat{y}_{t+\tau}(t) = \beta_0(t) + \beta_1(t)(t+\tau) \tag{4.11}$$

With some manipulation, see [89], it can be shown that the forecast $\hat{y}_{t+\tau}(t)$ is given by,

$$\hat{y}_{t+\tau}(T) = (2 + \frac{\alpha\tau}{(1-\alpha)})S_t - (1 + \frac{\alpha\tau}{(1-\alpha)})S_t^{[2]} \tag{4.12}$$

Such an approach to estimating the RUL of a degrading system has previously been applied to determining time-to-wash intervals for shipboard gas turbine engines which experience gradual performance degradation caused by the ingestion of salt [90].

The choice of a number of parameters must be carefully considered in the application of the double exponential smoothing prediction. The smoothing constant α is determined by simulated forecasting of an historical dataset. Using a section of available historical data, a regression line is fitted to the data and the initial least-squares estimates of β_0 and β_1 are determined. Using (4.9) and (4.10), initial values of S_t and $S_t^{[2]}$ can be calculated. Using these values, the historical dataset is used to update S_t and $S_t^{[2]}$ and in each time period a forecast is computed using the current values of S_t and $S_t^{[2]}$. The procedure is repeated for a range of α values, and the value which minimises the sum of the squared forecast errors in selected for use in forecasting future values.

The time series generated by the output of the ANN and MLR models generally comprises thousands of values, as the input variables are on a 5-minute sampling interval, and the difference between each consecutive value is quite small. In order to extract the relevant linear trend from the data over a smaller dataset, a new dataset comprising every 12th sample of the time series (i.e. 1 hour) was complied. This dataset was used for the prediction of the estimated RUL of the pump, and the generation of approximate 95% confidence limits.

Figure 4.19 illustrates the performance of the DESP method applied to the prediction of the RUL of a pump from 70% observed degradation, as estimated by the ANN model. Also shown is the actual rate of degradation in the pump, and the approximate 95% confidence limits of the prediction.

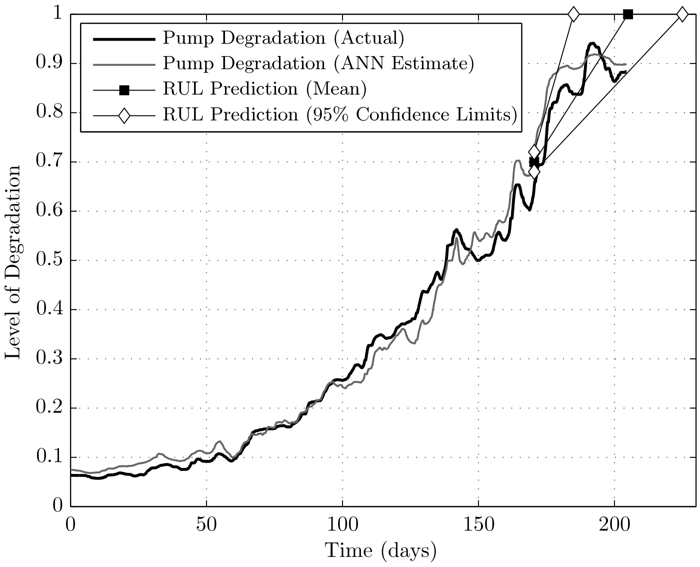


FIGURE 4.19: Pump degradation: RUL prediction at 70% degradation

Assuming a value of 90% degradation is used to define the point at which at pump is deemed to have exceeded its serviceable life, then the predictions generated fall well within the generated confidence limits. In this example, the mean estimate of the RUL was approximately 22 days, with pump failure occurring (in the sense of exceeding the acceptable level of degradation) approximately 20 days after the RUL prediction was generated.

4.6 Conclusions

In this chapter, a method for identifying and tracking the development of fluorine induced degradation in a dry vacuum pump was developed. The developed solution has the potential to reduce instances of unexpected pump failures caused by vacuum pump degradation. This could have significant monetary benefit in terms of reduced pump maintenance costs but also reduce the costs associated with scrapped wafers, tool downtime and chamber cleaning, following an unexpected pump failure.

To model pump degradation, both MLR and ANNs were investigated. The performance of the ANN models was significantly better than the MLR models. This is due to the ANNs ability to model non-linear functional relationships between the input and output variables. This chapter also demonstrated the potential benefit to semiconductor manufacturers in providing access to upstream process measurements in the development of condition monitoring solutions for dry vacuum pumps. By having access to such data in the development phase, it is now possible to identify and, more importantly, quantify the current level of vacuum pump degradation using only pump sensor variables. In this way, potential pump failures can be avoided and the serviceable life of dry vacuum pumps operating on harsh processes can be maximised.

RUL prediction capabilities for dry vacuum pumps were also developed using a double exponential smoothing prediction technique. This capability will provide maintenance personnel with an estimate of the RUL of a pump so that corrective maintenance can be arranged at a suitable time, thus minimising equipment downtime and the associated costs.

Chapter 5

TPU Condition Monitoring

5.1 Introduction

The primary focus of this chapter is the development of a signal tracking and condition monitoring algorithm for thermal abatement systems used in semiconductor manufacturing. The specific type of thermal abatement system investigated is known as a thermal processing unit (TPU). The main challenge presented in developing condition monitoring solutions for TPU systems is the multiple operating modes, which generate multi-modal, non-Gaussian, distributed signals. To address this challenge, a moving-window based solution is developed which addresses the multi-modal and non-Gaussian distributed signals by modelling the underlying signal distribution as a mixture of Gaussian probability density functions (PDFs), i.e. a Gaussian mixture model. The proposed condition monitoring algorithm for TPU systems is developed and tested exclusively using data collected within a large semiconductor manufacturing facility. The signal tracking algorithm allows for changes in the maintenance condition of a TPU system, which generate changes in the underlying distribution of the signals generated, to be tracked, allowing for accurate real-time condition monitoring. In addition, the extracted features also provide the foundation for the development of prognostic capabilities for TPU systems, which will be presented in Chapter 6.

In addition to the signal tracking algorithm for TPU systems, Gaussian mixture model-based solutions are also developed for a number of other applications. In Section 5.4.4, a Gaussian mixture model representation is used to develop a novelty detection algorithm for TPU systems, which can detect the occurrence of unanticipated fault conditions. This is demonstrated on an historical TPU failure example. Finally, in Section 5.5, a Gaussian mixture model-based solution for multi-mode signal tracking within a mechanical dry vacuum pump is developed. The proposed solution identifies the underlying distribution of the booster power signal as a mixture of Gaussian components, where each component describes a different operating mode of the pump. Bayesian inference is then applied to identify the pump mode at any instant in time. This approach allows for changes in the statistical characteristics of different pump variables, within the different pump operating modes, to be tracked over time. This permits greater insight and ability to identify and track the development of pump degradation. This capability is demonstrated on historical pump failure data.

5.2 Semiconductor Abatement Technology

5.2.1 Introduction

Semiconductor manufacturing utilises a wide variety of often toxic, corrosive and reactive gas mixtures, which pose a significant health, safety, and environmental, hazard. As a result, a key component of many semiconductor manufacturing processes is the exhaust gas management system. The exhaust gas management system is responsible for the safe handling and disposal of the effluent gas streams generated by semiconductor manufacturing processes. The exhaust gas management system is generally composed of two separate systems. Vacuum systems are employed in the form of mechanical dry vacuum pumps and turbomolecular (turbo) pumps which provide the necessary vacuum conditions to permit processing to occur, and also serve to remove hazardous by-products and gases from the processing chamber. The vacuum systems then pump these often toxic and reactive by-products downstream to devices known as abatement systems, which break these substances down for safe disposal. Abatement systems are designed to reduce the emission levels of dangerous gases from the effluent gas stream of semiconductor manufacturing processes to such a level that they can be safely exhausted to atmosphere, or discharged in liquid form into the waste-water treatment system.

In a large modern semiconductor fabrication facility, the abatement systems will typically utilise a range of different processes, each of which is designed to handle the different types of effluent gas streams, generated by different manufacturing processes. The principles of operation of different abatement devices generally falls into one of the following categories:

- **Wet scrubbing** of semiconductor manufacturing effluent streams involves the contacting of the effluent gas with a scrubbing liquid, which results in the absorption of undesired effluent components by the scrubbing liquid. Alternatively, the gas effluent may react with the scrubbing liquid; for example a caustic solution may be used to react with an acidic gas effluent to neutralise the undesired components of the effluent stream.

- **Dry scrubbing** involves contacting the effluent stream with a solid material, often in the form of small pellets designed to maximise the overall surface area for contact, which treats the effluent stream by chemisorption, in which a chemical reaction occurs at the exposed surface of the material to remove the undesired components of the effluent stream.

- **Thermal destruction** involves combustion of the effluent gas stream to oxidise and destroy the hazardous elements, at high temperatures.

- **Non-thermal plasma abatement** involves the destruction of specific components within the effluent stream using a plasma discharge. The development and in-service application of non-thermal plasma devices is at a much less mature stage when compared to the techniques descried above, but is becoming much more widespread. One of the primary drivers of plasma abatement is the lack of available and reliable natural gas sources at certain manufacturing locations around the world.

Abatement devices are also commonly described by their physical location relative to the processing tool which they serve. Local point-of-use (POU) devices are generally located immediately downstream of the vacuum systems to control emissions from a single or several processing tools. In contrast, end-of-pipe (EOP) devices are generally installed further downstream where they operate to abate emissions from multiple tools, or possibly the entire fab [91]. Figure 5.1 illustrates the layout of the processing tool and the exhaust gas management system on a typical semiconductor manufacturing process. The processing tool, located within the fabrication (fab) facility clean room, performs a specific manufacturing process on a single, or batch of, silicon wafers. The process effluent gases, generated by the manufacturing process, then flow down through the fab floor into the sub-fab. Located in the sub-fab is the primary dry vacuum pump which feeds the process gases into the point-of-use abatement device, located directly downstream of the dry vacuum pump. The treated gases then flow into the fab-wide exhaust system for further treatment downstream or discharge to the atmosphere.

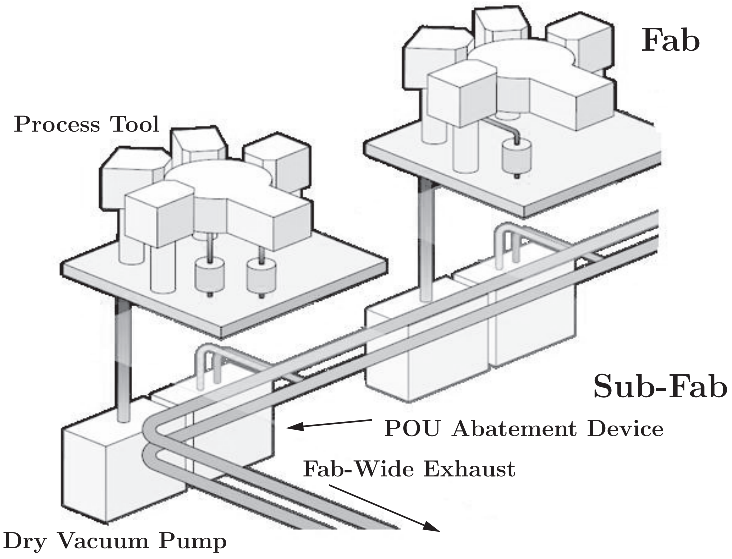


Figure 5.1: Typical equipment layout

One of the primary drivers for the development of new abatement technologies and particularly thermal and plasma abatement devices, is the challenge presented by a range of gases known as Perfluorocompounds (PFCs). These types of gases are commonly used in semiconductor manufacturing due to many desirable properties they exhibit; however, they also represent a significant environmental hazard when released to the atmosphere due to their high infra-red absorption capabilities.

5.2.2 PFC Gas Abatement

Perfluorocompounds (PFCs), including CF_4,C_2, F_6, NF_3, SF_6, and hydrofluorocarbons (HFCs) such as CHF_3, are used extensively in semiconductor manufacturing for plasma etching, chemical vapour deposition (CVD), and process chamber cleaning because of their low impact on employee safety, and because of the unique chemical properties they possess [92].

In particular, these types of gases provide uniquely effective performance when etching high aspect ratio features on semiconductor wafers. These gases also provide a safe and reliable source of fluorine gas used for remote plasma cleaning of deposition chambers [93]. Finding suitable replacements for these PFC gases has proven difficult due to the specific chemical properties of these gases.

PFC gases are mostly chemically inert and noncorrosive; however, these types of gases are known to intensely absorb infrared radiation and survive for long periods in the atmosphere. As as result, PFC gases are classed as high global-warming potential (GWP) gases due to their high infrared absorption capacity when compared to CO_2. For example, one metric tonne of SF_6, a gas commonly used in plasma etching, released into the atmosphere, is the equivalent of 23,900 metric tons of CO_2 in terms of its potential effect on global warming over a time horizon of 100 years [91]. A further concern with the use of PFC gases is the remaining volume of such gases that are not consumed by the manufacturing process, and thus requiring abatement. This is due to the relatively low utilisation efficiencies of these gases within fabrication processes [92]. To address the issue of PFC gas emissions, the World Semiconductor Council (WSC), in 1999, declared an industry-wide goal of 10% or greater reduction of aggregate PFC gas emissions over 1995 levels, from semiconductor fabrication facilities, by the year 2010 [87]. This represented a ambitious goal considering the average annual growth rates of 15% which the semiconductor industry has experienced almost every year since the 1970s. According to a statement released by the WSC in 2010, the industry was on target to achieve this goal [94].

To address the issue of PFC gas abatement, the semiconductor industry has investigated different approaches to reducing the volume of PFC gases discharged into the atmosphere. The range of approaches have included process optimisation, the introduction of alternative chemistries, capture/recovery of emissions and the abatement of PFC gases from process gases [93]. Each of these approaches have achieved varying levels of success, and it is understood that a complete solution will incorporate aspects from each of the proposed approaches. One of the most successful approaches to date has been the use of thermal abatement systems which have become commonplace in large semiconductor manufacturing facilities worldwide, as a method for reducing the levels of PFC gas emissions into the atmosphere.

5.3 Thermal Abatement Systems

5.3.1 Introduction

Amongst the range of approaches to reducing PFC emissions, one of the most commonly used in semiconductor manufacturing facilities worldwide is thermal abatement of effluent gas streams. Thermal abatement operates on the principle of oxidisation of the PFC gas components of effluent gas streams, via combustion at temperatures in excess of 900°C. Thermal abatement devices have become the technology of choice for the abatement of PFC gases due to their high performance, measured in terms of their destruction/removal efficiency (DRE). The DRE measures the percentage of PFC gas components which are destroyed or removed by a thermal abatement device. One of the highest scoring thermal abatement devices, in terms of its DRE, is the thermal processing unit (TPU) [95], which is manufactured by Edwards Vacuum, who are a major supplier of both vacuum and abatement solutions to the semiconductor industry worldwide. The TPU has become one of the most prevalent thermal abatement devices in semiconductor, solar, and flat-panel manufacturing facilities worldwide.

5.3.2 Edwards Thermal Processing Unit (TPU)

The TPU system, manufactured by Edwards Vacuum, is a point-of-use abatement device which is generally located in the subfab of a manufacturing facility, directly downstream of the exhaust of the mechanical dry vacuum pump employed on a particular manufacturing process, as shown in Figure 5.1. The TPU system has become an industry standard device for abatement on both CVD and etch processes, and has been shown to achieve greater than 99% DRE in the treatment of PFC gases [95].

A TPU utilises an inwardly fired combustor to oxidise the effluent gas stream, followed by a three-stage wet scrubber. The combustion chamber utilises inward-fired combustor technology developed by the Alzeta$^{\mathrm{TM}}$ corporation, and employs four gas injector nozzles to control the level of combustion, by injecting natural gas and oxygen into the combustion chamber.

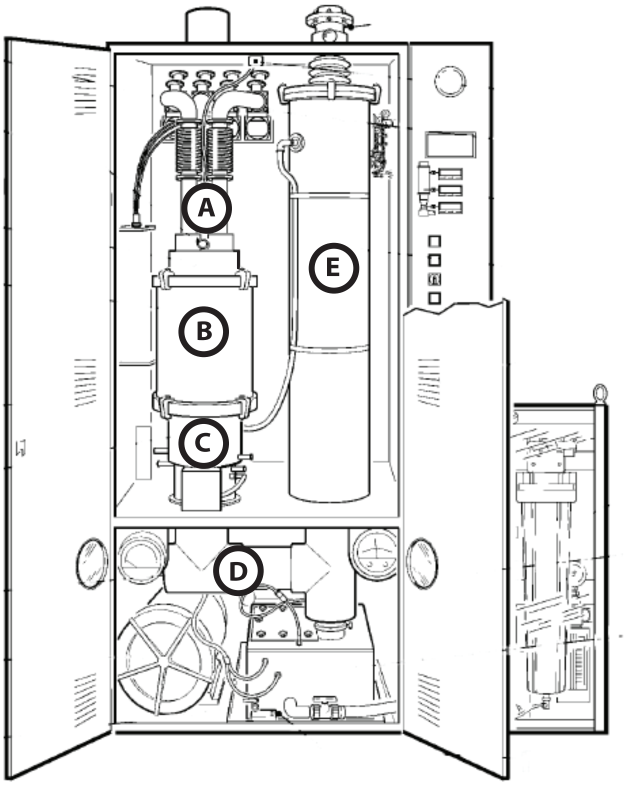


Figure 5.2: TPU system layout [5]

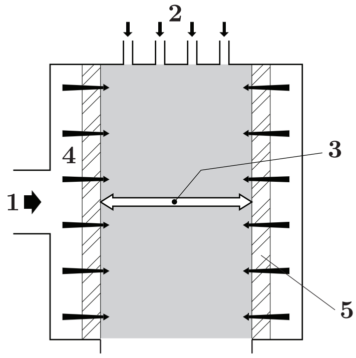
5.3.2.1 System Design

In this section, a general overview of the design and operation of a TPU is presented with reference to Figure 5.2, in which each of the major functional units within a TPU, labelled A-E, are described [6].

A. Head Unit The head of the TPU unit contains four inlets, each of which can be connected to an individual processing tool chamber, which feed the process gases into the TPU combustion chamber directly below.

B. Combustion Chamber Within the TPU combustion chamber the process gases undergo oxidisation at very high temperatures. Figure 5.3 illustrates the design of the TPU combustion chamber.

The combustion chamber contains a porous ceramic liner (5), which separates the inner surface walls of the combustion chamber from the high temperature reaction zone (3) inside the ceramic liner. Oxygen and natural gas are injected (1) into the inner surface walls of the combustion chamber which serves to flush the inner



1	Natural Gas and Oxygen
2	PFC Containing Process Effluent
3	Radiation Exhange
4	Annular Supply Plenum
5	Ceramic Liner

FIGURE 5.3: TPU combustion chamber design [6]

walls, keeping reaction materials away from the walls and preventing corrosion and the build up of solids. The area between the outer walls and the ceramic liner also serves as an annular supply plenum (4) for the delivery of the oxygen and natural gas mixture to the cylindrical ceramic liner.

The injected natural gas and oxygen, known as the fuel-air mixture, passes through the porous ceramic linear, where combustion and oxidation of process gases occurs on the inner surface of the ceramic liner. Temperatures at the surface of the liner can exceed 900°C. Sustained exposure of the process gas effluent to these high temperatures ensures complete destruction of the PFC gas components of the effluent stream. The principle of operation of the TPU combustion chamber is shown in Figure 5.4, where the arrows illustrate the annular supply of the fuel-air mixture around the ceramic liner, which then passes through the ceramic liner and into the inner combustion zone.

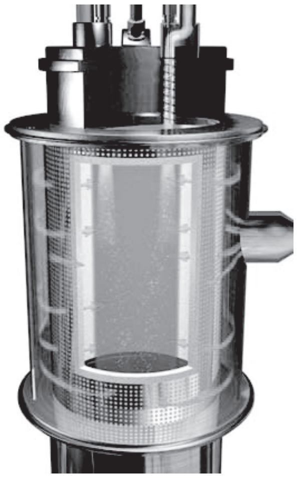


FIGURE 5.4: TPU combustion chamber illustration [7]

Also located within the combustion chamber is a temperature sensor which monitors the temperature within the combustion zone of the chamber. The temperature at this location is one of two TPU sensor variables monitored and recorded by the FabWorks networked monitoring system. This variable is named "combustor temperature" (CT).

C. Flux Force Condensation Scrubber (Quench-Unit) After the combustion chamber, the hot gases pass through to the first stage of a three-stage water scrubber. This first stage is known as the "quench" unit. Two water spray jets generate a fine mist which is sprayed across the hot gas stream to both; cool the gas, and condense on the surface of any particulates in the gas, causing the particulates to drop to the bottom of the quench-unit, removing those components from the exhaust gas stream.

Additionally, water is continuously flushed across the walls of the quench unit, both to keep the walls cool, and to prevent corrosion from the hot gases. A temperature sensor located in the wall of the quench unit, monitors the temperature of the walls. The temperature at this location is the second TPU sensor variable monitored by the FabWorks system and is named quench-wall temperature (QWT).

D. Cyclone Scrubber Following the quench unit, the cooled gas stream passes through the cyclone scrubber. During this stage, the gases and the water vapour are spun at high rotational speeds to ensure that any particulates, not removed from the gas stream in the previous scrubber stage, become entrained in the water which forms on the walls of the cyclone scrubber.

E. Packed Tower Scrubber The final water scrubber stage is the packed tower. The remaining components of the effluent gas stream rise up through the packed tower, which contains high surface area packing, where water which flows down from the top of the tower and scrubs the remaining acidic components from the gas. The cleaned gas stream then passes through two mist filter elements which remove any remaining water droplets, and the gas then exits through the exhaust gas outlet and into the fab-wide exhaust extraction system.

During manufacturing operations, TPU systems operates in two distinct modes; low-fire (LF) and high-fire (HF). In HF mode, increased flows of natural gas and oxygen are injected into the combustion chamber via the gas inject nozzles, to increase the combustion temperature, and ensure the destruction of all PFC gases within the effluent stream. The starting and stopping of the HF mode is usually controlled via communication between the semiconductor processing tool, and the TPU device. Typically, the HF mode is triggered by the radio frequency (RF) generator (i.e. the RF power signal) in the processing chamber, indicating the generation of a plasma in the chamber, and the pending arrival of an effluent gas stream containing PFC gas components, which requires higher temperature within the combustion zone to ensure their complete destruction.

5.3.2.2 Data Collection & Resampling

As part of this study, a dataset collected over a number of years was made available for analysis, and development of condition-monitoring and prognostic algorithms for TPU systems. The data was collected from a number of TPU systems installed on chemical vapour deposition (CVD) processes, within a large semiconductor manufacturing facility. Within the manufacturing facility, each TPU was connected to the FabWorks monitoring system. An overview of the FabWorks networked monitoring system was presented previously in Section 4.2.2.

To briefly recap, each device connected to the FabWorks networked monitoring system sends regular updates on sensor values and its current operating status (i.e. starting,running,off,etc) to a central server, where they are stored and analysed. The network typically carries data from each mechanical dry vacuum pump, turbomolecular pump, and TPU within a manufacturing facility. To ensure stability and robustness, message traffic across the network must be managed so that the network does not become saturated, and that regular communication between devices and the central database is maintained.

The primary method by which the volume of network traffic is controlled, is via the use of a send-on-delta sampling technique [85]. Using this sampling technique, sensor value updates are only sent to the central database when the difference between the current sensor value and the value of the most recent update sent to the central database exceeds a preset threshold, δ. The value of δ, for each individual sensor variable, acts as a trade-off between signal resolution and signal update rates, directly influencing the network bandwidth consumed. As the value of δ increases, the rate of signal updates reduces, with a corresponding reduction in signal resolution. Likewise, as the values of δ are decreased, there is an increase in signal updates rates, and hence network bandwidth utilisation, alongside an increase in signal tracking resolution. For the period of this study,, the FabWorks monitoring system was setup to record two sensor variables on each TPU monitored

1. Combustor temperature

2. Quench-wall temperature

The location of each of these sensors is detailed in Section 5.3.2.1. In addition to the sensor variables monitored, status information from each TPU system is also recorded and updates sent to the central database whenever the TPU status changes. The TPU status at any time is described by one of the following states

- Off
- On
- Processing
- Stopping
- Shutdown
- No Communication
- Communications Fail

During normal operation, the TPU status signal indicates "On" during periods of low-fire mode and indicates "Processing" during periods of high-fire mode. The status signal provides a means to track the system switching between operating modes, in a supervised manner.

5.3.2.3 Preliminary Signal Analysis

In Section 5.3.2, a brief description of the operation of a TPU system was presented, which reviewed the variables and status information that are recorded by the FabWorks networked monitoring system. In Figure 5.5, a plot of the combustor temperature (CT) and quench-wall temperature (QWT) signals from a normally functioning TPU system is shown over a period of approximately two days. The large fluctuations observed in both the CT and QWT signal values are a function of the TPU system switching between LF and HF operating mode, as changing flows of oxygen and natural gas are injected into the TPU combustion chamber.

Another factor to note is the operating ranges of both the CT and QWT signals. In Figure 5.5, it can be seen that the operating range of the CT signal is generally between 850° and 1000° C. The operating range of the QWT signal is generally between 35° and

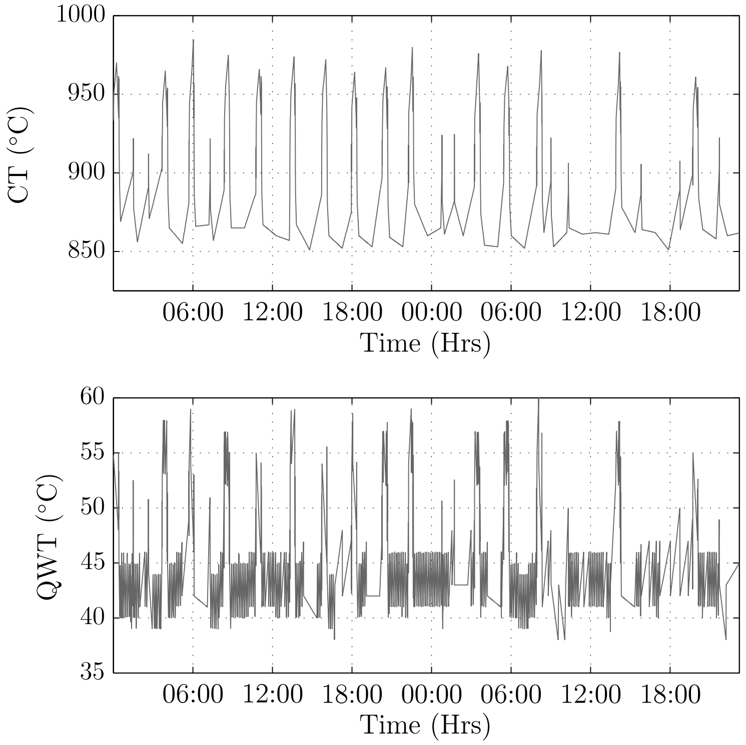


Figure 5.5: TPU data example

60° C. It is also clear that the behaviour of both the CT and QWT signals is similar, in response to a change in the operating mode of the TPU, with both signals rising and falling in response to periods of HF mode operation.

Visual analysis of the CT and QWT signals illustrated in Figure 5.5 clearly shows that both signals exhibit multi-modal behaviour, which reflect the switching of the TPU system between the LF and HF operating modes. This multi-modal behaviour of both the CT and QWT signals is a significant issue and a major driver in the selection of appropriate techniques for the development of condition monitoring algorithms for TPU systems, which are presented in this chapter. A final issue to note is that, unless otherwise stated, both the CT and QWT signals were resampled onto a regular 1-minute sampling interval, using linear interpolation. This resampling operation is necessary as the data collected by the FabWorks networked monitoring system, using a send-on-delta sampling technique, results in the data being collected on an irregular sampling interval.

5.3.2.4 TPU Failure Modes

TPU devices suffer from range of maintenance issues due to the harsh environment in which they operate. The result of an in-service failure of any major component of a TPU will often result in downtime for the entire manufacturing process as wafer processing cannot continue if the downstream abatement system is not operational. For the duration of this study, the FabWorks monitoring system only recorded the values of the CT and QWT signals from each TPU system. As a result, the work presented in this chapter only considers maintenance issues within either the combustion chamber or the quench-unit, which generate a response in the temperature sensors monitoring these stages of the TPU system. In future versions of the FabWorks networked monitoring system, and in the next generation of TPU system designs, it is likely that many more sensor variables will be monitored and recorded as ethernet-based communications become more prevalent, providing greater network bandwidth for communications. This may potentially enable the development of condition monitoring solutions for other components of a TPU system.

Combustion Chamber Operating Issues

One of the primary maintenance concerns when operating TPUs on certain manufacturing processes are silica particles (SiO_2), which can be generated as a by-product of the upstream manufacturing process. These particles, which often form as a fine white powder, can become deposited on the walls of the ceramic liner and crystalise to form a hard "glaze", as illustrated in Figure 5.6 [8].

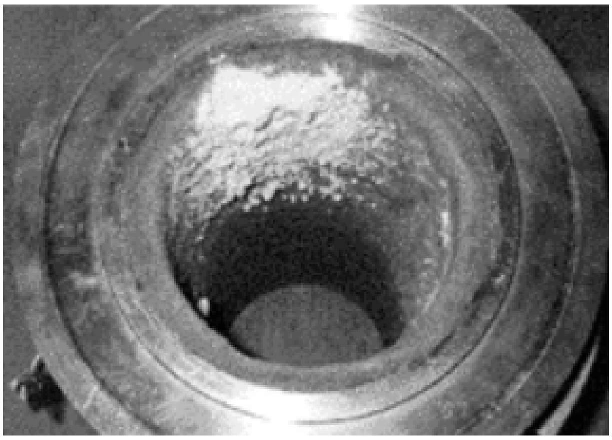


Figure 5.6: Example of deposit build-up in a TPU combustion chamber [8]

As the level of deposits increase, the pores of the ceramic liner become blocked, reducing the overall porosity of the ceramic liner. The reduced porosity causes a reduction in the volume of the fuel-air mixture which passes through the ceramic liner, from the annular supply plenum, into the inner combustion zone. This affects the temperature profile within the inner combustion zone. Figure 5.7 illustrates the response seen in a TPU CT signal (which monitors the temperature in the inner combustion zone) to a build-up of silica deposits on the walls of the ceramic liner. As seen in Figure 5.7, the values of CT signal, at the lower limits, begin to decay in response to a increase in the level of deposits on the walls of the ceramic liner.

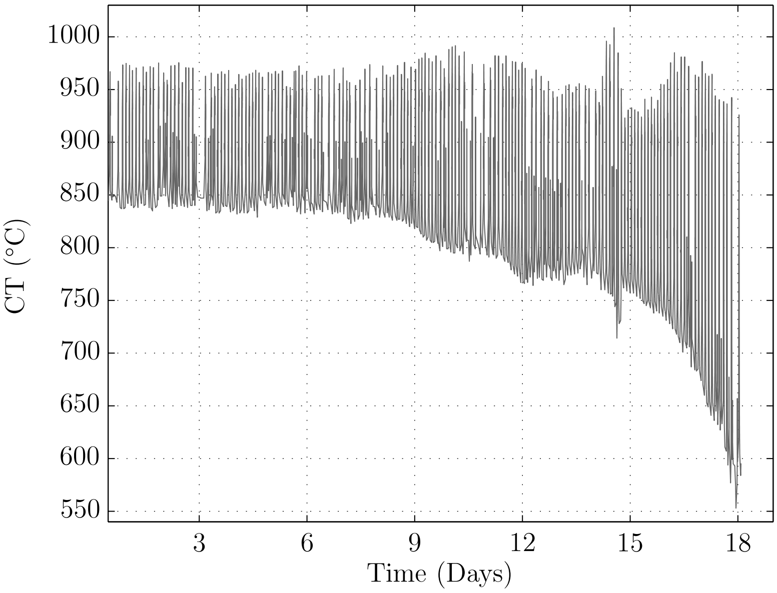


Figure 5.7: Typical TPU combustor temperature (CT) signal response to a build-up of silica deposits on the walls of the ceramic liner

The gradual decay in the CT values is reflective of the reducing porosity of the ceramic liner, resulting in less fuel-air mixture reaching the inner combustion zone and, hence, a reduction in the temperature within the inner combustion zone. It is also noticeable that the effect of deposit build-up on the observed CT signal values is less significant during periods of high-fire mode operation, particularly during the early stages of deposit build-up. It seems that the increased flows of fuel-air mixture during high-fire mode are sufficient to overcome the reduced porosity of the ceramic liner. This may be due the the additional fuel-air mixture being injected, raising the pressure on the annular supply plenum side of the ceramic liner and increasing the volumetric flow through the liner. However, it has also been observed that the time necessary for the temperature within the inner combustion zone to transition from low-fire mode values to high-fire mode values increases, as the CT signal decays in response to a build-up of deposits.

From a maintenance perspective, the build-up of deposits causes a number of undesirable consequences, which must be considered in choosing when to perform maintenance on TPUs. These include;

Combustion Effects: The build-up of deposits affects the temperature profile within the inner combustion zone of a TPU, as illustrated in Figure 5.7. As the level of deposits becomes excessive, it is possible that the destruction/removal efficiency (DRE) of the TPU may be reduced, potentially affecting the volume of PFC gases released into the atmosphere.

Ceramic Liner Damage: The air-flow through the pores of the ceramic liner is designed to transport heat build-up in the ceramic liner into the inner combustion zone. However, as the pores of the ceramic liner become blocked, this air-flow reduces, meaning that the ability to dissipate heat build-up within the liner is reduced. This results in hardening of the ceramic liner (which is usually quite soft to touch) which significantly reduces its remaining serviceable life [8].

Fault Propagation: As the level of deposits become excessive, there is an increased risk of large silica deposits becoming dislodged and falling into the water scrubber below, potentially damaging the cyclone scrubber unit which rotates at high speeds.

The build-up of deposits within a TPU combustion chamber essentially represents a process of gradual *degradation* of the maintenance condition of a TPU. Considering each of the issues described above, it is imperative that TPU systems are maintained in peak condition to both; ensure maximum DRE at all times, and to reduce the potential for in-service failures and subsequent unplanned tool downtime. The build-up of deposits on the ceramic liner within the TPU combustion chamber, and the gradual degradation in TPU condition, are the primary TPU maintenance issues for which condition monitoring and prognostic algorithms are developed in this chapter and the following chapter.

Quench Unit Operating Issues

The quench unit is responsible for cooling the hot gas flows which emerge from the TPU combustion chamber, and for removing any particulates in the effluent gas stream. Within the quench unit, it is imperative that the cooling water temperature and flows are maintained at required levels. This is important, as otherwise the quench unit walls risk becoming damaged by both the high heat, and the potentially corrosive elements within the effluent gas stream. Occasionally, problems with cooling water flows can arise which risk damaging the quench-unit. In Section 5.4.4, a simple, generic, technique, for monitoring and detecting potential problems within the quench-unit is presented, which can identify the occurrence of uncharacteristic or “novel” events within the multi-modal QWT signal.

5.3.2.5 TPU Maintenance Practices: Current & Future

The current method for maintaining TPUs is for regular preventative maintenance (PM) to be performed at fixed intervals. The intervals can be based upon elapsed time, or upon the volume of wafers processed since the last PM was performed, if such information is available. Performing regular PM on TPUs has a number of benefits and shortcomings from a maintenance perspective. By performing regular PM, at sufficiently conservative intervals, the likelihood of in-service failure is reduced and the DRE of the TPU device is maintained at high levels. However, regular PM is expensive to perform; manpower must be allocated to perform the maintenance tasks as scheduled and a significant inventory of spare parts must be maintained to perform the regular swap out of components. In a manufacturing facility operating hundreds of process tools, the regular PM requirements for TPUs, and the cost and logistics of maintaining a large inventory of spares, can be significant. Furthermore, in those periods between regular PM overhauls, and in the absence of any condition monitoring system, all in-service failures of TPUs will occur unexpectedly, resulting in major disruption to process tool availability and potentially impacting on production scheduling across a manufacturing facility. The consequences of an in-service TPU failure are even more significant when you consider that a single TPU might be connected to up to four individual wafer processing chambers.

Semiconductor manufacturing is becoming an increasingly lean manufacturing operation due to increased competition [96]. As a result, maximising equipment and production uptime is becoming ever more critical. From a TPU perspective, the importance of overall production uptime is clearly illustrated by the fact that some manufacturers are now operating two TPU systems in parallel on certain manufacturing processes, to try and achieve 100% abatement system uptime. This is achieved by having one TPU system operating normally, with a second TPU system operating in a standby role. A bypass valve is incorporated into the design, allowing one TPU to treat process gases whilst the other system remains in standby, or undergoes maintenance. In the event of one of the TPU systems failing, the process gases can be automatically diverted into the other TPU system via the bypass valve, resulting in no impact on wafer processing in the upstream chamber(s). This approach to achieving 100% abatement uptime illustrates the importance attached to maintaining high levels of equipment uptime within the semiconductor manufacturing industry. Running two TPU systems effectively doubles both capital and operating costs of the thermal abatement system. However, for those manufacturers operating TPU systems in parallel, the costs associated with unplanned production downtime must, presumably, far exceed the costs associated with running two TPU systems in parallel.

Within the semiconductor manufacturing environment, the increasing focus on maximising equipment uptime is serving as a driver for the development and deployment of new maintenance philosophies and practices on TPU systems. As time and technology progresses, maintenance activities on TPU systems are expected to move away from regular PM activities, toward more condition-based, or as-necessary, maintenance. Currently, TPU maintenance personnel often spend significant amount of time viewing and analysing the signals recorded on TPU systems to try and identify upcoming maintenance events. This practice is often time consuming and the analysis will always be somewhat subjective.

The overall objective in developing new condition-based maintenance practices on TPU systems is to minimise preventative maintenance activities, whilst also avoiding in-service failures. A certain level of preventative maintenance will always be necessary to swap out life-limited components, but minimising such activities will be key to minimising maintenance and total ownership costs. Key to the success of new condition-based maintenance approaches for TPU systems, is the development of new signal tracking and condition monitoring algorithms for TPUs. Such systems will provide an accurate assessment of overall TPU health in real-time, and provide maintenance personnel with warnings of current and approaching maintenance issues. The rest of this chapter describes the development of a condition monitoring algorithms for TPUs. In Chapter 6, the features generated by the TPU condition monitoring algorithms developed in this chapter are used to infer the current level of TPU degradation, and predict the remaining useful life of a TPU system suffering a build-up of deposits.

5.4 TPU Condition Monitoring & Feature Extraction

This section describes the development of condition monitoring solutions for TPU systems. Firstly, Section 5.4.1 describes existing condition monitoring approaches for TPU systems. Sections 5.4.2 and 5.4.3 then introduce a proposed signal tracking algorithm which uses Gaussian mixture models to track changes in the underlying distribution of the combustor temperature signal. The combustor temperature sensor which is located inside the TPU ceramic liner. The changes in the underlying distribution of the CT signal can be used to detect the buildup of deposits on the ceramic liner or other maintenance issues which may generate a response in the CT signal. Finally, Section 5.4.4 describes how the Gaussian mixture model approach, described in the prior sections, can also be used to detect the occurrence of novel events within the TPU quench-unit.

5.4.1 Existing TPU Condition Monitoring Approaches

This section provides a brief review of existing condition monitoring techniques for TPU systems. These techniques have been developed both as part of the FabWorks networked monitoring system, and also as an optional FabWorks add-on application known as Edwards Advanced Diagnostic Systems (EADS) [97]. Whilst much of the specifics of how these existing condition monitoring techniques operate remain proprietary, the basic principles are understood and are presented here for comparison with solutions developed as part of this thesis.

5.4.1.1 Alarm Thresholds

The most basic condition monitoring techniques employed on TPU systems are univariate alarm thresholds. These operate by defining threshold limits on individual sensor values which, when exceeded, generate a warning or alarm, notifying maintenance personnel of a maintenance issue. In some cases, excessively high sensor value readings can result in the automatic shutdown of a TPU system. The use of univariate alarm thresholds is one of the most common condition monitoring techniques across all types of machinery and equipment, due to the ease of implementation. However, such approaches suffer from a number of shortcomings. The primary shortcoming is the excessive generation of false alarms, which can sometimes result in unnecessary maintenance being performed, or unnecessary downtime to investigate the root cause of an alarm being triggered. As a result, maintenance personnel can become distrustful and lose confidence in the usefulness of such approaches.

In the case of TPU systems, a further concern with the use of univariate alarms is the multi-modal signal characteristics of the CT and QWT signals. This multi-modal signal behaviour is caused by a TPU system switching between LF and HF operating modes. For example, consider setting a threshold limit on a temperature signal, which is triggered when excessively high values are detected. This approach may fail to detect a potential fault condition if the TPU system is operating in LF mode. This is because the range of temperature values observed in LF mode are significantly less than those observed in the HF mode. Whilst a significant rise in temperatures may occur relative to normal operating ranges, they may not be sufficient to exceed the alarm threshold which may be set above the typical operating range of the HF mode. Furthermore, setting threshold limits for each operating mode will also result in numerous false alarms as the temperature values rise and fall when switching between TPU operating modes.

Another consideration with the setting of alarm threshold limits on TPU systems, particularly within the semiconductor manufacturing environment, is that the temperature profiles observed on each TPU system may vary significantly across different manufacturing processes. This is due to the different process chemistries employed on each manufacturing process, which can influence the temperatures which are generated within the combustion chamber. In developing new condition monitoring techniques for fault detection in TPU systems, it is important that such techniques be robust to the variability in TPU operating temperature profiles, which can vary across different manufacturing processes.

To address the shortcomings of alarm thresholding on TPU systems, Section 5.4.4 introduces a simple univariate condition monitoring algorithm capable of detecting potential fault conditions in either the combustion chamber or quench-unit of a TPU system. The proposed solution comes from the domain of condition monitoring techniques known as "Novelty Detection" algorithms, previously discussed in Section 2.4.4. The new solution presented in Section 5.4.4 provides a simple solution to the issue of multiple operating modes, allowing for the detection of fault conditions independent of the TPU operating mode. It also avoids the requirement of having to select specific threshold values for each process. The proposed solution is also insensitive to the different temperature profiles observed from TPU systems operating on different processes as the distributions of signal values is identified online, using data collected from each individual TPU.

5.4.1.2 Existing Moving Window-Based TPU Signal Monitoring

Section 5.3.2.4 discussed the issue of deposit build-up in TPU combustion chambers. Current condition monitoring algorithms for TPU systems, offered as part of the EADS system, employ moving window based methods to detect and track the build-up of deposits in a TPU combustion chamber from the CT signal. Moving windows of different lengths are used and, within each window, characteristics such as max and min signal values of the CT signal are identified, and moving average values are calculated. Gradient based techniques, using the features generated by the windowing methods, are used to track and trend the loss of temperature in the combustion chamber, which reflect the build-up of deposits. These features are then analysed in a proprietary rule-based framework to identify the build-up of deposits and generate notifications for maintenance personnel.

Alternatively, in the absence of any automated TPU condition monitoring for TPU systems, it is common for maintenance personnel to spend significant lengths of time viewing and analysing the plots of CT values for each TPU in a facility, on a daily, or sometimes more frequent, basis. The purpose of this visual analysis is to identify those TPUs experiencing a build-up of deposits within the combustion chamber and to track this build-up over time. In addition to being quite time consuming, the visual analysis of signal plots can also be somewhat subjective.

A shortcoming of existing moving-window based methods, is that they do not account for the different operating modes which a TPU may operate in, over any window period. As a result, fluctuations in the max and min values of the CT signal will be a function of the different operating modes in which a TPU operated over that period. In Section 5.4.2, the development of a new signal tracking algorithm for TPU systems is presented. The new technique expands upon the existing moving-window based approaches, described above, by employing Gaussian mixture models to track the changes in the non-Gaussian distribution of CT signal, in each of the TPU operating modes, over time. This new approach allows for the identification and generation of many more system features which can be incorporated into the existing rule-based framework for TPU monitoring. The new signal tracking approach employs a moving window, similar to existing TPU monitoring methods described above, but provides a more robust, accurate, and adaptable approach, to tracking deposit build-up within the combustion chamber. Furthermore, the approach allows for separate tracking of the CT signal distribution in each of the TPU operating modes.

It will be illustrated in later sections how the new signal tracking approach provides a range of improvements over existing methods including, freeing maintenance personnel from manual analysis of CT plots, allowing them more time for maintenance actions and planning, and additionally removing any subjectivity from the analysis. The new signal tracking approach is designed to be flexible and robust with regard to the quality of signal resolution available in the CT signal. Additionally, the proposed signal tracking approach can easily be expanded to multivariate signal tracking, which will be necessary as the next generation of TPU systems will have many more signals available for monitoring. This capability will be illustrated briefly in Section 5.4.4.3.

Finally, with the introduction of new TPU operating modes, designed to reduce the consumption of the fuel-air mixture whilst no processing is occurring upstream, the new operating modes will be easily accommodated by the proposed signal tracking approach, and will provide the ability to monitor and track TPU condition regardless of the operating mode.

5.4.2 TPU Condition Monitoring using Gaussian Mixture Models

Previously, in Section 5.3.2.3, a brief introduction to the characteristics of the CT and QWT signals, generated by a typical TPU device, was presented. The primary characteristic observed in both signals was the large fluctuations in temperature values as a TPU system switches between operating modes. This switching betweens operating modes results in the generation of a multi-modal signal, which presents some issues from a signal tracking perspective. Figure 5.8 illustrates an example of deposit build-up within a TPU combustion chamber. Highlighted in Figure 5.8 are two periods of data, labelled 1 and 2, which each cover a 24-hour period of data. The data in period 1 was recorded when the TPU was operating normally and the data in period 2 was recorded when a build-up of deposits in the TPU combustion chamber had commenced. This build-up of deposits within the combustion chamber results in the gradual reduction in the "lower" range of CT values observed.

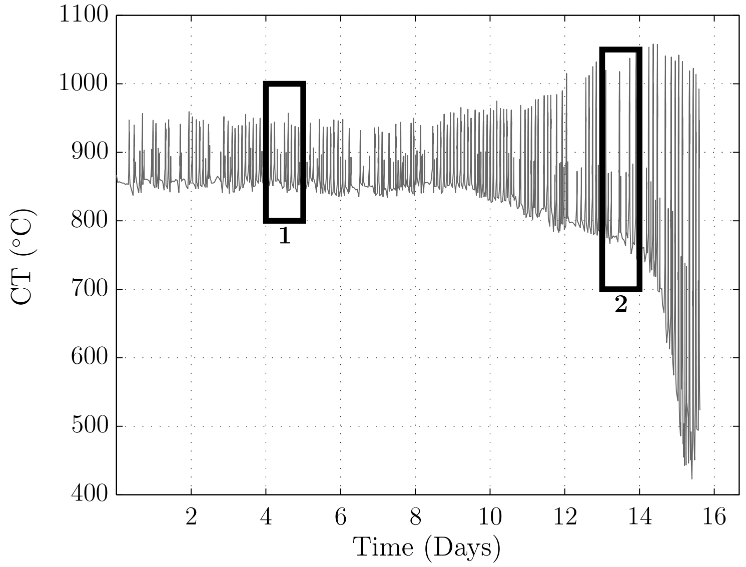


FIGURE 5.8: Evolution of CT signal as level of deposits increase

In Figure 5.9, the distribution of the CT signal during each of the data periods labelled 1 and 2 in Figure 5.8 is presented in histogram format. Looking at the upper illustration in Figure 5.9, which represents the distribution of CT signal values during normal (fault-free) operation, we can clearly see two distinct regions, or clusters, of CT signal values. These two regions reflect the distribution of CT values generated from the TPU operating in both LF and HF operating modes.

In the lower illustration of Figure 5.9, the distribution of the CT signal, following a build-up of deposits in the combustion chamber, is shown. Comparing the two distributions in

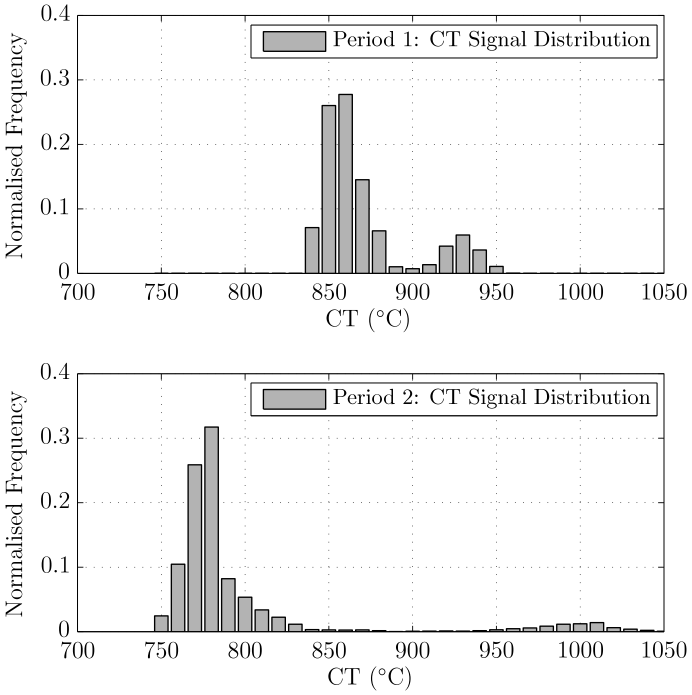


FIGURE 5.9: Combustor temperature signal: Changing distribution

Figure 5.9, it can be seen that the signal distribution at both times maintains the same general shape. However, once a build-up of deposits occurs, the two distinct clusters of CT values both diverge from each other, and spread out across a wider range of temperature values. Since the general shape of the CT signal distribution is maintained as deposit build-up occurs, this suggests a possible approach to tracking the multi-modal CT signal. By parameterising the underlying multi-modal distribution of a typical CT signal, this would provide a means to accurately track the development of deposit build-up, by analysing the changes in the parameter values defining the underlying signal distribution.

One potential solution to the problem of parameterising the underlying CT signal distribution is through the use of Gaussian mixture models (GMMs). GMMs allow for the parameterisation of multi-modal densities, by modelling the underlying distribution as a sum of individual Gaussian probability density functions (PDFs). Furthermore, mixture models can be used to describe situations where each observation is modelled as having been produced by one of a set of alternative mechanisms [64]. In the case of equipment monitoring applications, where equipment operating modes define the set of alternative mechanisms, GMMs can provide a powerful tool for signal tracking and fault detection applications in a multi-modal signal environment.

GMMs have previously seen application to a number of condition monitoring and fault detection problems. Yu and Qin [98] applied GMMs to modelling complex industrial processes, in which each of the operating modes of the process is described by a component of the GMM. Then, using a derived global probabilistic index, process measurements indicative of a potential fault condition are identified. Choi *et. al* [99] combined GMMs and principal components analysis (PCA) for monitoring of multi-modal processes.

5.4.2.1 Gaussian Mixture Models

A mixture model is a type of probabilistic model used for density estimation, in which a combination of individual probability density functions are used to model the underlying distribution of a set of observations. Mixture models are often employed for unsupervised learning, or clustering of data [62, 63]. More specifically, a Gaussian mixture model is a parametric probability density function, which represents an observed distribution as a weighted sum of individual Gaussian PDFs. A review of the theory and application of mixture modelling was presented previously in Section 3.3, which should be referred to as necessary.

From a set of observations, the parameters of a GMM are usually identified using the expectation maximisation (EM) algorithms. However, as discussed previously in Section 3.3.2.1, the standard EM algorithm suffers from a number of shortcomings. To address these shortcomings, the Figueiredo-Jain (F-J) version of the EM algorithm [66] was chosen to estimate GMM model parameters in all examples shown. The primary motivation for the use of the F-J version of the EM algorithm is its robustness in addressing the issues of initialisation sensitivity and convergence to boundary of the parameter space. These issues can result in an unbounded likelihood estimate for one of the components, with the covariance matrix of the that component becoming arbitrarily close to singular.

The F-J algorithm also has the capability to identify the optimal number of mixture components with which to model a set of observations, $K_{optimal}$. To identify $K_{optimal}$, the user selects a range of possible values for K, by setting values for K_{min} and K_{max}. The value of K which minimises the cost function, defined within the F-J algorithm, is then chosen as the optimal number of components with which to model the set of observations. In many of the applications presented in the following sections, the number of mixture model components, K, is known *a priori.* In such situations, where K is known, the F-J algorithm was modified slightly. Instead of returning the optimal number of components, as identified by the cost function, the algorithm instead returns a GMM of size K_{min}

where, in this case, K_{min} is the number of components to be identified from the set of observations, and is known *a priori*.

By setting K_{min} to the desired number of components, and by setting K_{max} to a significantly larger value, the algorithm addresses the issue of initialisation sensitivity. By starting with a large value for K_{max}, each of the components are initially distributed randomly throughout the parameter space. Then, by iteratively killing and removing those components not supported by the data, and at each iteration removing the component with the smallest weighting until the desired number of components K_{min} is identified, the algorithm exhibited robust performance in identifying GMM parameter values. Indeed, using the standard EM algorithm, it was not possible to generate the results presented in the following sections for the moving-window applications. This is because the standard EM algorithm frequently converged to the boundary of the parameter space, resulting in unbounded likelihood estimates and, thus, failing to identify the parameter values of the specified GMM from the presented set of observations.

5.4.2.2 Modelling the CT Signal Distribution

Figure 5.10 shows an example of a TPU system which suffered from a build-up of silica particles on the ceramic liner. In the upper plot of Figure 5.10, a region of fault-free data is highlighted. The distribution of CT signal within this highlighted region is shown in the lower plot of Figure 5.10. To model the distribution of the CT signal within the highlighted region using a GMM, the first issue is to determine how many individual Gaussian PDFs to use to model the underlying multi-modal distribution. The F-J version of the EM algorithm was used for this task, and the optimal number of components identified, with which to model the underlying distribution of the signal, was three.

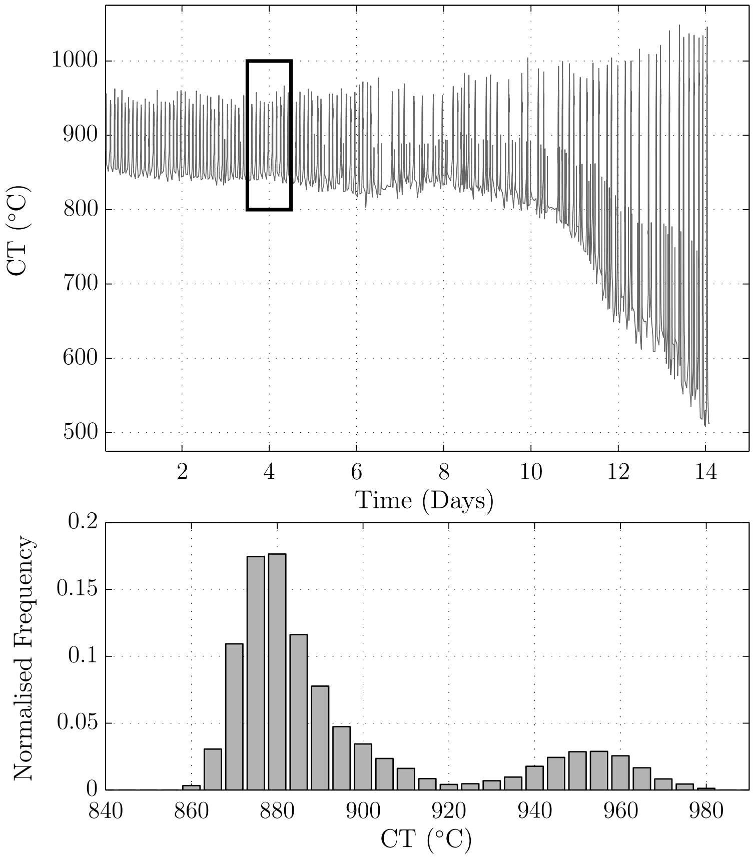


FIGURE 5.10: Fault-free CT signal distribution

Figure 5.11 illustrates the CT signal distribution in Figure 5.10 modelled as a superposition of three Gaussian PDFs. The upper plot in Figure 5.11 shows the distribution of CT values, with the identified 3-component GMM PDF overlayed. Shown in the lower plot are the individual Gaussian PDFs, the weighted superposition of which make up the overall Gaussian mixture model density shown in the upper plot. As illustrated in

Figure 5.11, the 3-component GMM accurately models the underlying distribution of the CT signal.

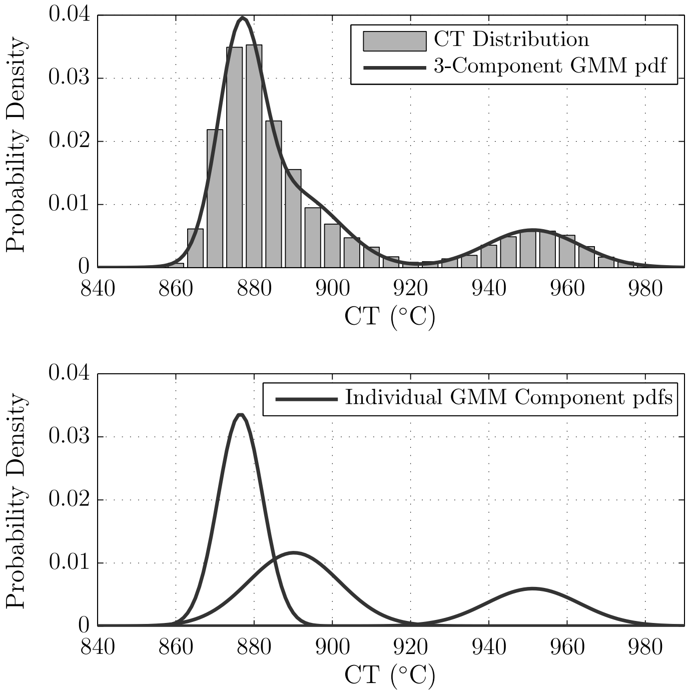


Figure 5.11: GMM representation of CT signal distribution

5.4.3 Multi-Mode TPU Condition Monitoring Algorithm

This section describes a proposed algorithm for TPU condition monitoring, specifically for detecting deposit buildup on the TPU ceramic liner. Figure 5.12 shows a flow chart illustrating the proposed multi-mode TPU condition monitoring algorithm. Firstly, raw sensor data is collected by the FabWorks monitoring system. This data is then passed to the resampling stage where the combustor temperature (CT) signal is resampled to a regular 1-minute sampling interval, as described in Section 5.3.2.3. The resampled CT signal is then passed to the multi-mode signal tracking stage, which is described in detail later in this section. The general purpose of this stage is described below.

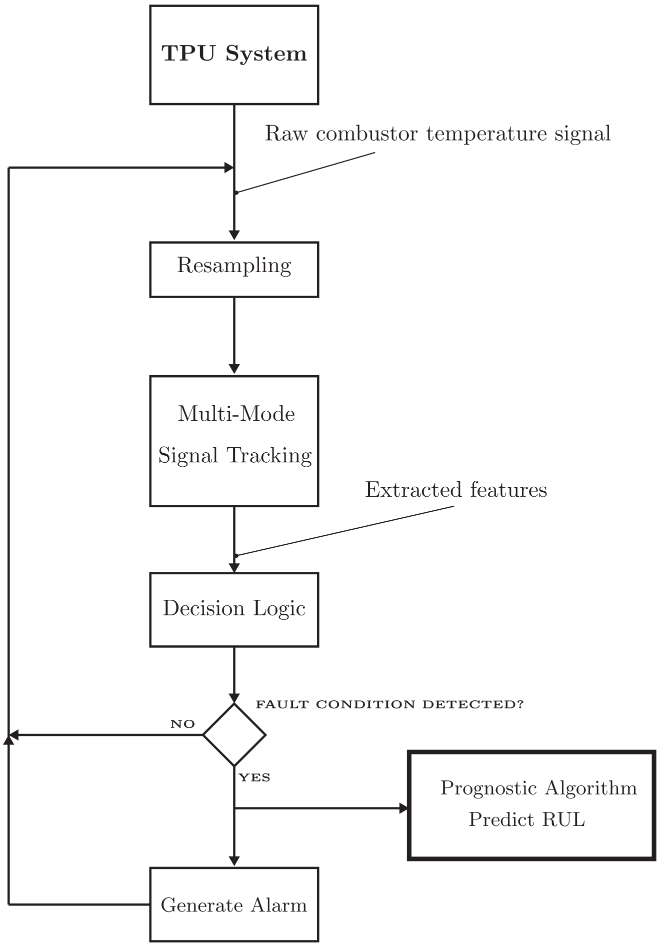

TPU System
Raw combustor temperature signal
Resampling
Multi-Mode Signal Tracking
Extracted features
Decision Logic
FAULT CONDITION DETECTED?
NO
YES
Prognostic Algorithm Predict RUL
Generate Alarm

Figure 5.12: Multi-mode CT tracking algorithm

Section 5.4.2.2 illustrated how GMMs can be used to model the distribution of a typical TPU CT signal. However, to track how the distribution of the CT signal changes in response to deposit build-up within the TPU combustion chamber, it is necessary to employ some form of moving-window approach. A moving-window based solution is proposed to track changes in the distribution of the multi-modal TPU CT signal. The proposed solution employs a moving window where, at each iteration, the parameters of a specified GMM are identified from the set of observations within the current window. In this way, the changes in the parameter values of the GMM identified at each iteration can be used to track the changes in the underlying distribution of the CT signal.

The identified values, comprising the means and variances of each of the GMM components, are then passed as features to the decision logic stage. The decision logic stage identifies any parameters deviating from normal operating ranges and identifies potential fault conditions. If no fault condition is identified, the algorithm continues to iterate. If a fault condition is identified, an alarm condition is generated and the algorithm continues to iterate. Furthermore, when a fault condition is identified, which is indicative of deposit buildup occurring on the ceramic liner, the prognostic stage is initialised. The prognostic stage generates a prediction of the remaining useful life (RUL) of the TPU. The algorithm then continues to iterate and a specific feature (described in Section 5.4.3.4) generated by the multi-mode signal tracking stage is continuously passed to the prognostic stage to recursively update the RUL predictions as the fault condition continues to develop. The design of the prognostic stage of the algorithm is described in Chapter 6.

The remainder of this section describes the specifics of how the multi-mode signal tracking stage is implemented. Section 5.4.3.1 describes how the underlying distribution of the CT signal is modelled using GMMs. The challenges presented by the different operating modes of the TPU systems in modelling the underlying distribution of the CT signal are presented and a proposed solution involving separating the CT signal by operating mode is discussed. The design of the moving window, in terms of window length and the step size between iterations of the algorithm, is discussed in Section 5.4.3.2. Section 5.4.3.3 then presents some results of the signal tracking algorithm applied to historical TPU failure examples and illustrates how changes in the values of the features generated can be used to indicate the presence of a fault condition or deposit buildup. Finally, Section 5.4.3.4 describes how one the features generated by the multi-mode signal tracking stage can be used to track deposit buildup and for predicting the RUL of a TPU.

5.4.3.1 CT Signal Distribution Identification

Figure 5.11 illustrated how a 3-component GMM can be used to model the underlying distribution of a typical TPU CT signal. This suggests that in applying a moving-window approach to tracking the changes in CT signal distribution, the parameters of a 3-component GMM should be identified from the set of CT signal values within each window. In this way, the changes in the mean and variance values of each of the components could be used to track the changes in the underlying distribution of the CT signal, in each of the TPU operating modes. However, such approach raises some issues which must be considered.

Using a moving-window approach, with a specified fixed window length, assumes that the distribution of CT signal values will be similar within successive windows. In this way, changes in the parameters describing the underlying distribution of the CT signal could then be used to infer the condition of a TPU system, within each of the TPU operating modes. However, wafer processing within a manufacturing facility does not typically occur at regular, specified intervals. This means that within a specified window length, a TPU may not have operated in the HF mode and, instead, may have operated solely in the LF mode for the period of the window. As a result, any attempt to identify the parameters of a 3-component GMM from the data samples within a fixed-length data window, where, if it is assumed that the component with the highest mean value describes the distribution of the CT signal within the HF mode, then this assumption will be incorrect.

A solution to this issues is provided by the TPU status signal, which is recorded by FabWorks. The TPU status signal records the times when the TPU switches between LF and HF operating modes, as described previously in Section 5.3.2.2. By using this status signal, it is possible to separate the original CT signal into two separate signals. The first signal, the LF mode CT signal, is composed of only those samples recorded when the TPU was operating in LF mode. The second signal, the HF mode CT signal, is composed of only those samples recorded when the TPU was operating in HF mode. By separating the original CT signal on the basis of the TPU operating mode, it is now possible to employ a moving-window approach to track changes in the signal distribution, in each of the TPU operating modes, independently. In this way, the problem of irregular wafer processing is addressed. Once the original CT signal is separated into two separate signals, there remain a number of design issues to consider. The first issue is to identify the underlying distribution of both the LF mode CT signal and the HF mode CT signal, which is discussed in this section.

Once the CT signal is separated into two signals, using the TPU status signal, the next issue is to investigate the distribution of these two signals. Figure 5.13 shows the distribution of the LF mode CT values from the same dataset that generated the distribution shown in Figure 5.11, except that only those samples recorded during periods of LF mode operation are shown. The F-J algorithm [66] was used to identify the optimal number of GMM components necessary to model the underlying distribution of the LF mode CT signal. The optimal number of components identified was two. The upper plot in Figure 5.13 illustrates the identified 2-component GMM PDF overlayed on the LF mode CT signal distribution, with the individual GMM component PDFs are shown in the lower plot.

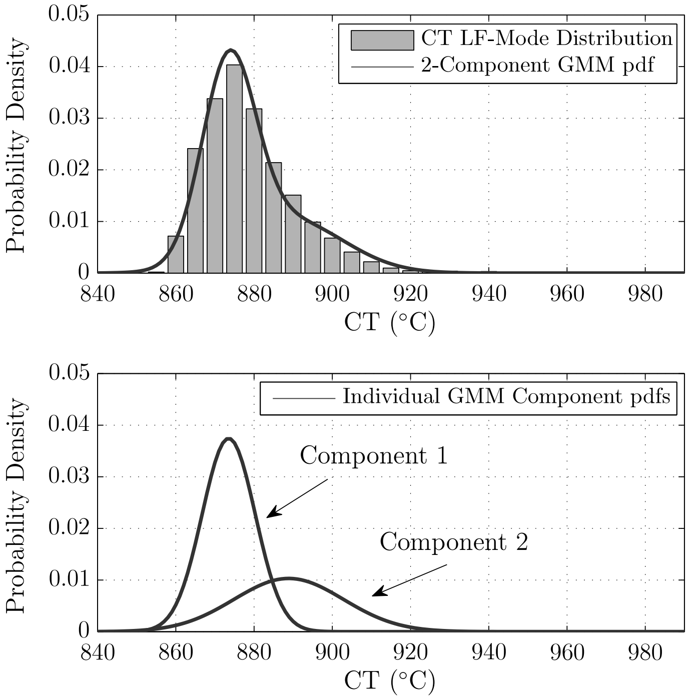


FIGURE 5.13: GMM representation of LF mode CT signal distribution

Considering that TPU systems operate in only two modes, it might be expected that the signal distribution in each mode might be represented by a single Gaussian density. However, it has been clearly illustrated that the underlying distribution is best represented by a 2-component GMM. The reason for the presence of the additional component, labelled Component 2 in Figure 5.13, is due to the TPU switching between operating modes. Once a TPU switches from HF mode into LF mode, the CT values must decay from the high values, associated with the TPU operating in HF mode, back to lower values associated with the TPU operating in LF mode. Another reason for the non-Gaussian distribution of the LF mode CT signal is due to the TPU switching into HF mode for short periods, whilst abating PFC gases from an upstream chamber clean cycle. These clean cycles, which generally last for shorter durations than wafer processing cycles, cause short duration spikes in CT values and, once the TPU switches back into LF mode, the CT signal must again decay back to normal range.

Having selected a 2-component GMM to model the distribution of the LF mode CT signal, the distribution of the HF mode CT signal must also be identified. Figure 5.14 shows the distribution of the HF mode CT signal taken from the same data set as used previously in Figure 5.11. However, in this example, only those samples recorded during periods of HF mode operation are selected.

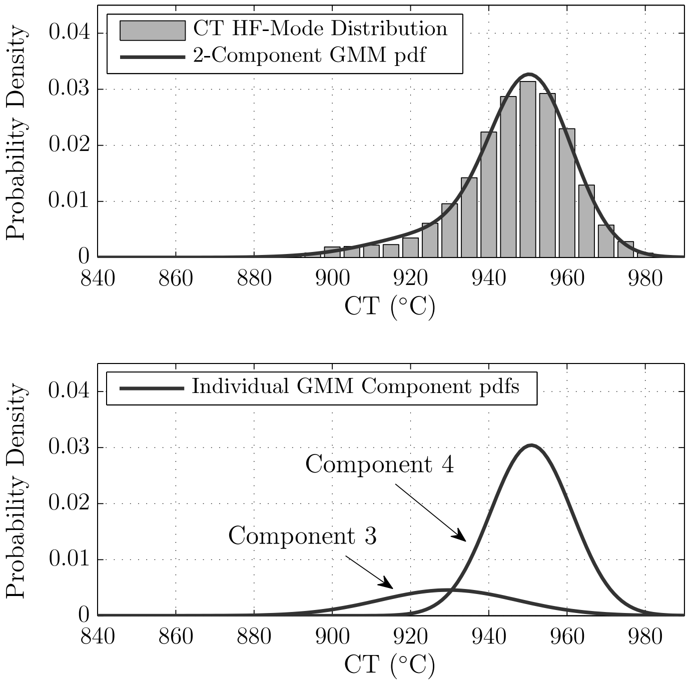


FIGURE 5.14: GMM representation of HF mode CT signal distribution

Using the F-J algorithm, a 2-component GMM was identified as the optimal number of components with which to model the underlying distribution of the HF mode CT signal. As with modelling the LF mode signal distribution, an additional component is necessary to model the HF mode signal distribution. This additional component, labelled Component 3 in Figure 5.14, models the periods of transition between LF and HF operating modes, as the temperature values rise in response to additional flows of oxygen and natural gas.

5.4.3.2 Moving Window Design

The proposed multi-modal CT tracking algorithms employs two separate moving windows, which operate independently to the track the CT signal distribution within each of the TPU operating modes. This means that appropriate window lengths and step sizes must be chosen for tracking the CT signal in each operating mode. In selecting a window length, another issue to consider is how to specify the bounds of the moving-window at each iteration. Since the two signals to be tracked are comprised solely of samples recorded during the specified operating mode, and that a TPU regularly switches between operating modes, then the upper and lower bounds of the moving-window cannot be defined in terms of absolute time. Instead, for tracking within each of the TPU operating modes, the moving-window is filled with a specified number/window of the most recent samples recorded, whilst the TPU was operating within the relevant mode. For example, the window length for tracking the LF mode CT signal might comprise the most recent two-hours of data recorded when the TPU system was operating in the LF mode.

Another issue to consider is the performance/accuracy trade-off in choosing different window lengths. Increasing the window length means a greater number of samples from the underlying distribution are available, increasing the likelihood that an accurate representation of the true underlying distribution of the CT signal will be identified. However, increasing the window length also introduces a lag in the estimates of the mixture model parameters. Much like a moving-average (low-pass) filter, the greater the length of the window, the greater the lag introduced. Given that the objective is to track changes in the underlying distribution, it is preferable to introduce as small a lag as possible so that the GMM parameters values respond quickly to changes in the underlying signal distribution. This is achieved by reducing the length of the window. However, as the window length becomes smaller, this results in greater fluctuations in the estimates of the mixture model parameters between successive windows. This is due to having an insufficient number of samples to identify the true underlying distribution.

Considering the issues described above, a wide range of window lengths were considered for tracking the CT signal distribution in each of the TPU operating modes. Employing a quantitative approach to selecting the optimal window length in each operating mode proved difficult. This was due to the absence of a metric against which to optimise the window length versus the signal tracking accuracy and the different phase lags introduced by the different window lengths. Following exhaustive visual analysis of the signal tracking performance, across a range of possible window lengths, a window length of 300-minutes was selected for tracking the LF mode CT signal, and a window-length of 75-minutes was selected for tracking the HF mode CT signal.

A final issue to consider in the design of the moving-window solution is the step-size to be taken between successive windows. This determines how often the parameters of the GMM must be calculated for each TPU. Due to the relatively slow changing characteristics of the CT signal distribution, a step-size of 15 minutes was selected.

5.4.3.3 Application Examples - CT Signal Tracking

Sections 5.4.3.1 and 5.4.3.2 have described the development of a multi-modal CT tracking algorithm, to track the changes in the response of the TPU CT signal to deposit build-up within a TPU combustion chamber. In this section, a number of examples of the algorithm applied to historical TPU failure examples are presented.

Low-Fire Mode Tracking

Figure 5.15 illustrates the performance of the algorithm in tracking the changes in the CT signal distribution within the LF operating mode, as the level of deposits within the TPU combustion chamber increases. As described previously in Section 5.4.3.1, the LF mode CT signal distribution is tracked using a mixture of two Gaussian PDFs. The labelling of the components in Figure 5.15 is the same as used previously in Section 5.4.3.1.

The lower valued component, labelled Component 1 in Figure 5.15, tracks the "steady-state" operating range of the CT signal during continued periods of LF mode operation. The second component, labelled Component 2, tracks the changing distribution of the CT signal during periods when the CT signal transitions from HF mode operation to LF mode operation. The upper plot in Figure 5.15 shows the evolution of the mean value (μ_1) of Component 1, and the 95% confidence limits on the estimation. The 95% confidence limits are computed from the variance value (σ_1^2) identified for Component 1 of the mixture model at each iteration, and are defined as $\mu_1 \pm 1.96\sigma_1$, where σ_1 is the standard deviation of Component 1. The lower plot in Figure 5.15 shows the evolution of the mean value (μ_2) of Component 2, and the 95% confidence limits on the estimation at each iteration.

While Figure 5.15 illustrates the evolution of the GMM components, which are overlayed on the original CT signal to illustrate the changing distributions of the two components, Figure 5.16 illustrates the evolution in time of the actual mean and standard deviation values identified for each component, at each iteration. As can be seen in Figure 5.16, the parameter estimates, generated by the F-J algorithm at each iteration of the moving-window, exhibit a significant level of fluctuation between successive iterations. To smooth these fluctuations, each of the signals were filtered using a low-pass filter. Forward-backward filtering was utilised to ensure a zero-phase delay as illustrated. Note that the smoothed parameter values were also used for the illustrations shown in Figure 5.15.

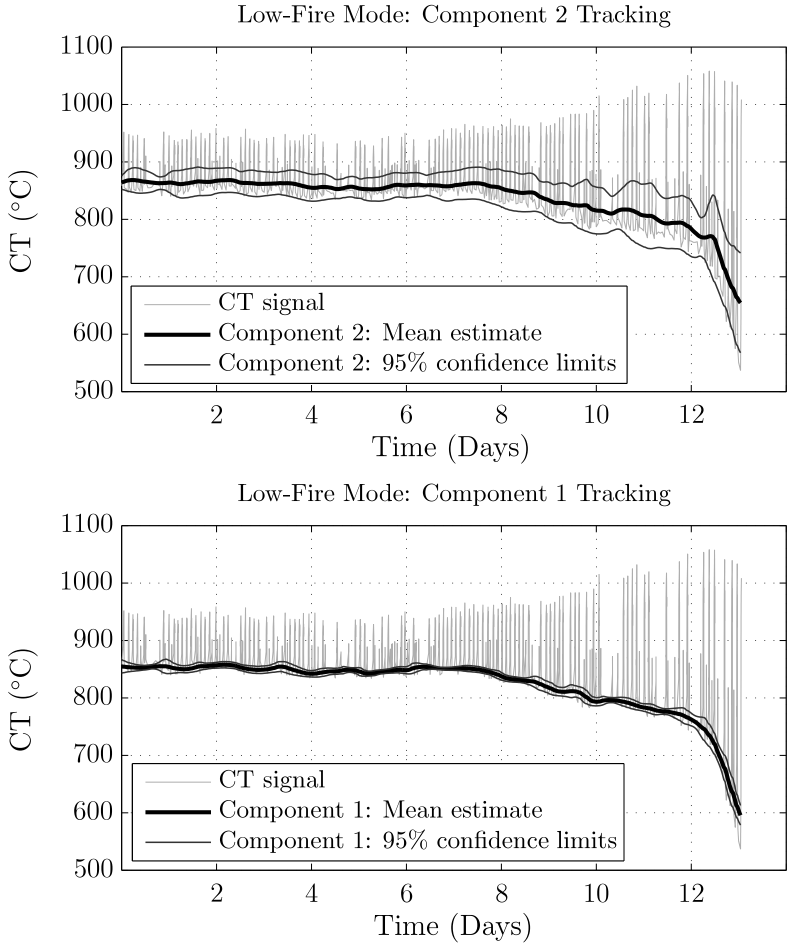


Figure 5.15: TPU Low-Fire mode CT signal tracking

From an automated condition monitoring perspective, the most useful feature from the set of parameter values tracked each iteration is the mean value of Component 1 (μ_1). This parameter, which is inferred from the original multi-modal CT signal, tracks the decay in the values of the CT signal at the lower limits of the CT signal, and provides a means to infer the degree to which the porosity of the ceramic liner has been reduced, in response to an increasing build-up of deposits on the walls of the ceramic liner.

From a stand-alone perspective, this feature can be used to trigger warnings and alarms as its value gradually decays and passes through threshold limits. However, this feature can be used for a far more useful application, as presented in the following section where the use of this feature in the development of prognostic algorithms for TPU systems is discussed. The evolution of the mean value of Component 2 (μ_2) is very similar to that of Component 1, thus making this feature somewhat redundant. However, the standard deviation of Component 2 (σ_2) exhibits a somewhat monotonic increase, which could be

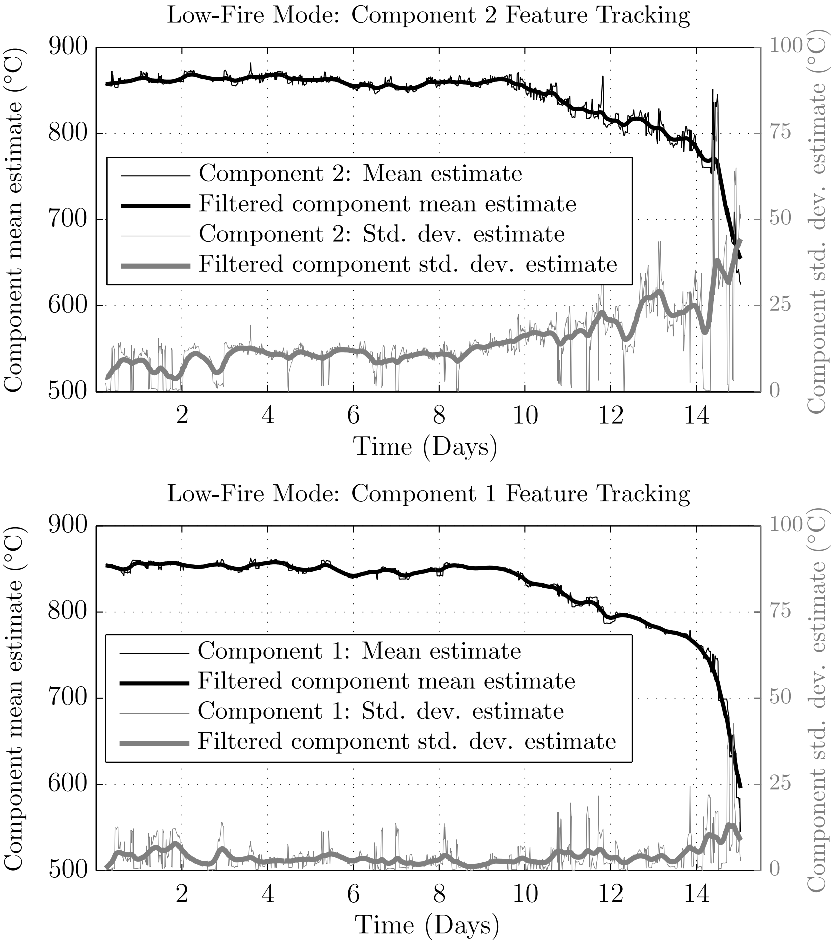
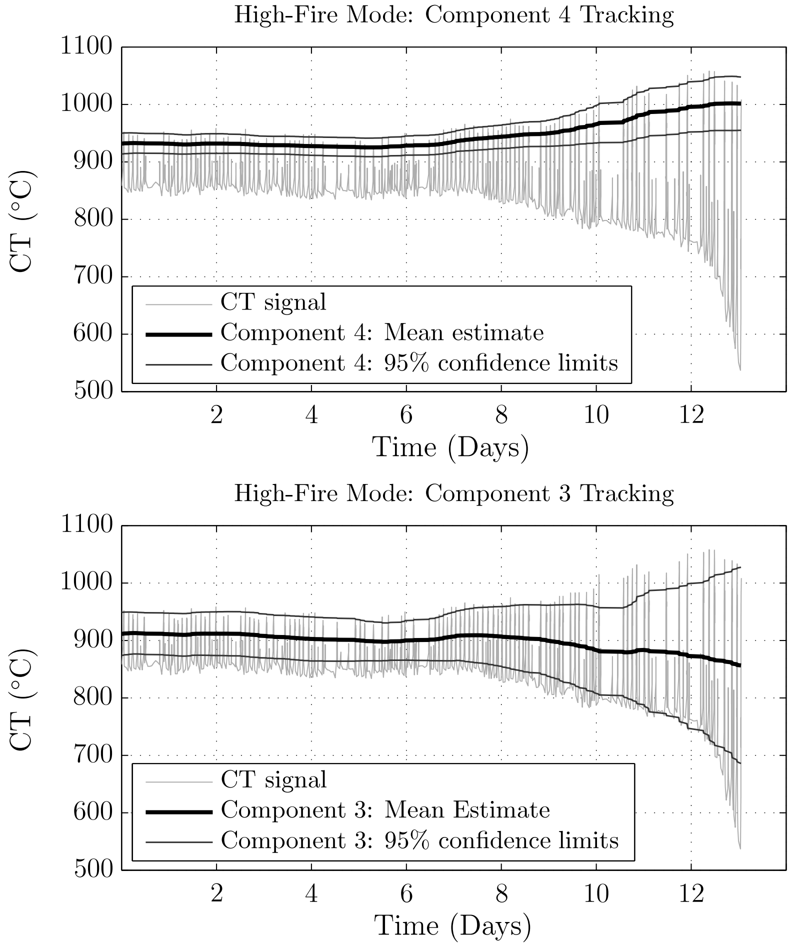


Figure 5.16: TPU Low-Fire mode CT signal tracking (component means and standard deviations tracking)

incorporated within a multivariate rule-based framework for the generation of warnings and alarms.

High-Fire Mode Tracking

Figure 5.17 illustrates the performance of the CT tracking algorithm in monitoring the changes in the CT signal distribution within the HF operating mode. As described previously in Section 5.4.3.1, the HF mode CT signal is tracked using a mixture of two Gaussian distributions. The labelling of the components in Figure 5.17 is the same as used previously in Section 5.4.3.1.



High-Fire Mode: Component 4 Tracking

High-Fire Mode: Component 3 Tracking

Figure 5.17: TPU High-Fire mode CT signal tracking

Component 3, in Figure 5.17, tracks the changing distribution of the CT signal during those periods when the CT signal rises in response to additional flows of natural gas and oxygen, when the TPU system first switches into HF mode. In contrast, Component 4, in Figure 5.17, tracks the "steady-state" range of operating values of the CT signal during continued periods of HF mode operation.

Figure 5.18 shows the evolution of the actual parameter values identified by the F-J algorithm at each iteration. A notable observation is the reduced level of fluctuations in the identified parameter values between successive iterations, compared to the tracking performance within the LF mode. This is due, in part, to the comparatively better signal resolution with the HF mode of operation, as the rise and fall of CT signal results in the δ values for the CT signal being exceeded more frequently.

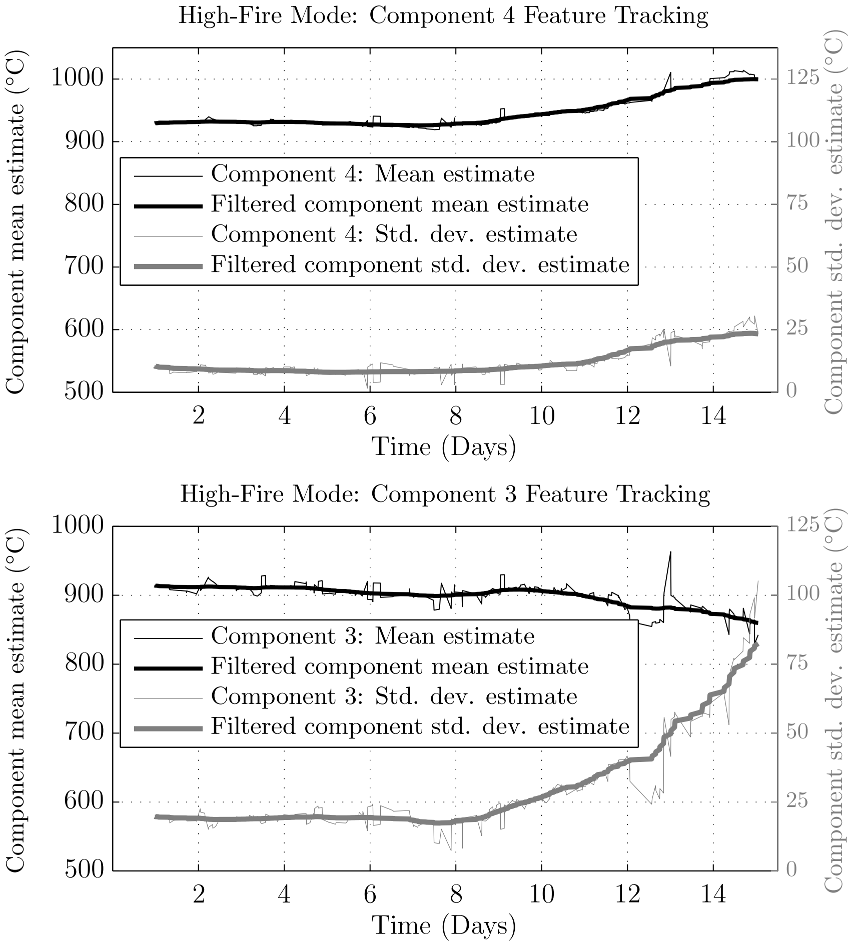


Figure 5.18: TPU High-Fire mode CT signal tracking (component means and standard deviations tracking)

From an automated condition monitoring perspective, two of the tracked parameters within the HF mode provide the most useful features for inferring TPU condition. In the case of Component 4, the most useful feature is the tracking of the mean value of this component (μ_4). As can be seen in Figure 5.18, there was a significant rise in the mean value of Component 4 (μ_4), as the peak values of the CT signal increased as TPU failure approached. This process of increasing peak values of the CT signal has been identified previously by TPU maintenance personnel as a precursor to a pending TPU failure. The mean value of Component 4 tracks this steady rise in CT values beyond the normal operating range.

In the case of Component 3, the most useful feature is the tracking of the standard deviation of this component (σ_3). It can clearly be seen that the standard deviation value increases significantly as TPU failure approaches. The increase in (σ_3) value is reflective of the increasing temperature differential between the steady-state LF mode temperature values and the steady-state HF mode temperature values. Considering that the (σ_3) parameter value increases in value, by several orders of magnitude as the fault evolves (in a relatively monotonic fashion), both the value and gradient of this parameter could potentially be incorporated within a rule-based framework, alongside other features, to quantify the current level of degradation and generate alarms and warnings as the level of degradation passes through certain thresholds.

5.4.3.4 Fault Indicator Measurement for Prognostics

Section 5.4.3.3 illustrated how the parameters tracked by the multimode CT tracking algorithm can be used as features to infer the current condition of a TPU. This allows maintenance staff be be aware of developing problems and to make inferences regarding when corrective maintenance action should be performed. In this section, the development of prognostics for TPU systems is introduced. Prognostics is understood to be the generation of long-term predictions describing the evolution in time of a particular signal of interest, or *fault indicator*, for the purpose of estimating the remaining useful life (RUL) of a failing system, subsystem, or component [27].

Prognostics is designed to provides maintenance personnel with sufficient visibility of approaching maintenance events, so that informed decisions, regarding when to perform corrective maintenance actions, can be made. Prognostics has the potential to deliver major improvements in both equipment uptime and overall maintenance costs for TPU systems. In the case of prognostics for TPU systems, the useful life of a TPU is exceeded once the level of deposits on the ceramic liner becomes excessive, such that the porosity of the ceramic liner is reduced below an acceptable limit. To avoid the potential in-service failure of a TPU, and also to maximise the useful life of a TPU ceramic liner, accurate estimates of the RUL of TPU systems suffering from deposit buildup are required. To develop this capability, it is first necessary to identify a signal of interest, or *fault indicator measurement*, which can be predicted into the future to estimate the RUL of a TPU.

Figure 5.15 illustrates an example of the proposed fault indicator measurement to be used in the development of prognostic capabilities for TPU systems. The fault indicator measurement is shown overlayed on the original CT signal. The fault indicator measurement comprises the mean values of Component 1, identified at each iteration of the LF mode CT tracking algorithm. This signal was chosen to represent the fault indicator measurement for a number of reasons.

The maintenance condition of a TPU is characterised by the level of deposits which have formed on the walls of the ceramic liner. The only method available for inferring the level of deposits is the response seen in the CT signal. As the level of deposits increases, the porosity of the ceramic liner is impacted, which reduces the volume of fuel-air mixture which passes through the ceramic liner into the inner combustion zone. As discussed previously in Section 5.3.2.4, the effect of reduced porosity is most evident during periods of LF mode operation, when there are no upstream process gases which are undergoing oxidisation, and hence a smaller volume of fuel-air mixture is being injected into the

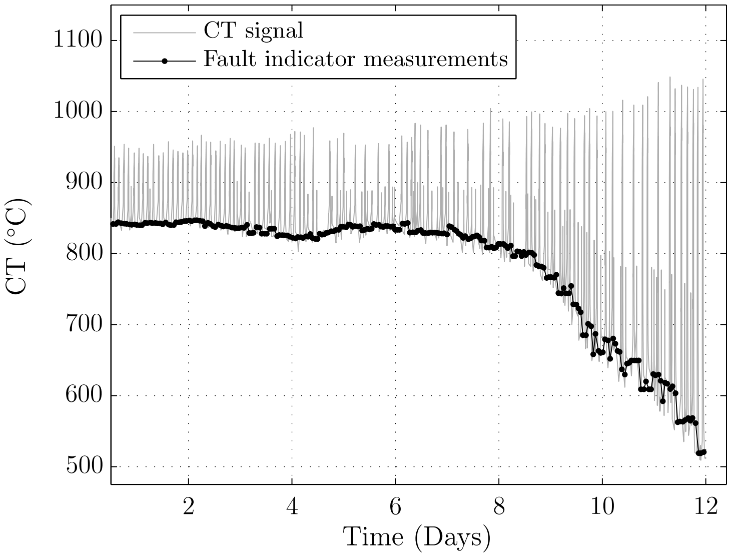


Figure 5.19: TPU Low-Fire mode CT signal tracking

TPU combustion chamber. The decay in the CT signal, during LF mode operation, directly reflects the reducing porosity of the ceramic liner and provides the best method for inferring the level of deposit buildup.

By generating long-term predictions of the evolution of the fault indicator measurement, and predicting when it will decay below a predefined threshold, accurate and actionable estimates of TPU RUL can be generated. However, as this problem involves predicting the future, there is a significant level of uncertainty associated with generating such predictions which must be accurately represented and managed, to predict TPU RUL. The development of prognostic capabilities for TPU systems, which address these uncertainty challenges presented, is presented in Chapter 6.

5.4.4 Quench-Unit Fault Detection

In this section, a simple application, again using Gaussian mixture models, is presented. The purpose of the application is to identify the occurrence of an unexpected, or "novel" event, from monitored sensor data, which might indicate the occurrence of a fault condition. The specific solution presented here falls within the realm of condition monitoring algorithms known as novelty-detection algorithms. The basic principle of novelty detection algorithms is to build a model of a system, which describes the observed behaviour of the system during normal, fault-free, operation. Future data collected is then compared with the model of fault-free operation and, using a distance metric and a specified threshold, abnormal or novel events are identified, which might indicate the occurrence of a fault condition. Novelty detection algorithms have found widespread application in a number of fields, such as fault detection, hand writing recognition, and radar target detection. A brief review of the principles of novelty detection algorithms was presented previously in Section 2.4.4.

5.4.4.1 Problem Description

Figure 5.20 illustrates an example of a fault which occurred in the quench-unit of a TPU. It is understood that this fault occurred due to a problem with the delivery of cooling water to the quench-weir. The reduction in cooling water resulted in a rise in the temperature recorded at the wall of the quench-weir. If the walls of the quench-unit do not remain covered by a sufficient protective water stream at all times, there is a risk of the process gases corroding the walls of the quench unit. The longer such issues remain uncorrected, the greater is the likelihood that the system may suffer further damage, resulting in more costly maintenance requirements. Additionally, if such a problem remains unchecked, the temperature values may continue to rise and potentially result in the automatic shutdown of a TPU system. As a result, it is imperative that the time between fault occurrence and fault detection is minimised.

In the example shown in Figure 5.20, a condition monitoring systems was in operation in the form of univariate alarm thresholds, which are used to indicate the presence of a potential fault condition. Upon the QWT signal exceeding a fixed threshold value, an alarm is generated which notifies maintenance personnel of an issue requiring immediate attention. As shown in Figure 5.20, an alarm was generated at 17:11:03 on the day in question, when the QWT signal exceeded the alarm threshold value of 71° C. The entire TPU system was shutdown by maintenance personnel shortly after, and all planned

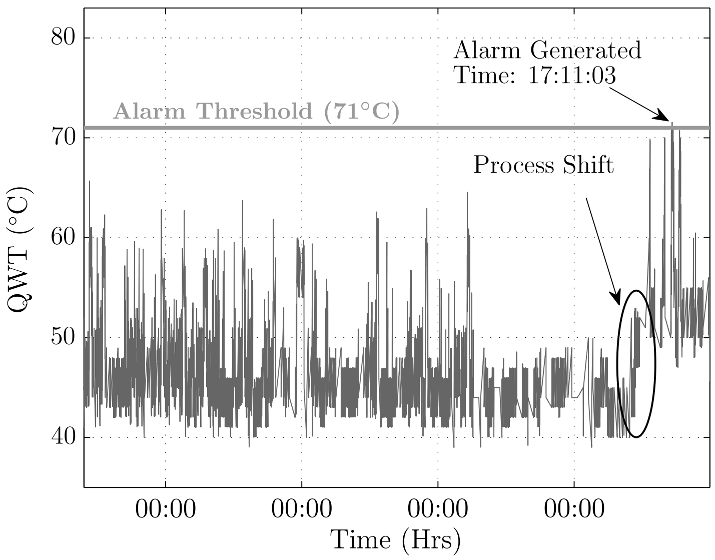


FIGURE 5.20: QWT Signal: Crack in Quench Wall

processing on the upstream tool(s) connected to this TPU had to be cancelled until the maintenance issue was resolved.

Whilst the use of a univariate alarm threshold detected the occurrence of an abnormal event, the time of notification of the event occurred after a significant period of time had elapsed since the actual event occurred. Within Figure 5.20, an enclosed region, labelled as "Process Shift" is shown. During this period, there was a significant shift in the mean operating values of the QWT signal, some 5-6 hours prior to generation of the alarm condition. This delay in identifying the presence of a fault condition is of significant concern for a number of reasons:

- The initial fault condition may subsequently propagate, potentially resulting in damage to other components. This can have a twofold impact
  - an increase in maintenance costs with potentially more components to be replaced
  - an increase in the time necessary to carry out corrective maintenance, resulting in increased manufacturing downtime

The primary reason for the delay between fault occurrence and alarm generation is that the fault monitoring system, in the form of a univariate alarm threshold, does not take into account the different operating modes of the TPU system. Whilst the alarm threshold will detect excessively high QWT values, such high QWT values will generally only occur when the TPU is operating in HF mode. During periods of HF mode, additional oxygen and natural gas are injected into the TPU, raising the gas temperature and, subsequently, the QWT signal values. However, TPU systems will typically only operate in HF mode between 6% and 9% of the time. The actual percentage of time spent in HF mode is determined by the rate of wafer processing in the upstream chambers. As a result, the current univariate condition monitoring approach will not notice any process shift in QWT values, if the system is operating in LF mode.

In the following section, a simple novelty detection algorithm is presented, which takes into consideration the different operating modes of the TPU system. The solution uses a GMM to model the underlying distribution of the QWT signal and, on the basis of comparison of subsequent QWT values and the current TPU mode, an alarm is generated when the process is detected as having deviated from normal operation, where normal operation is defined by the parameters of the GMM.

5.4.4.2 Data Modelling

Figure 5.21 illustrates the distribution of the QWT signal and the 2-Component GMM used to model the underlying QWT signal distribution. The parameters of the GMM were identified using the F-J algorithm. The underlying distribution of the QWT signal is modelled as 2-component GMM, which reflect the two operating modes of the TPU; LF and HF. Unlike the situation with the CT signal, in which 3 Gaussian components are required to model the distribution of the CT signal, a third mixture component is not necessary to model the QWT signal distribution. This is due to the significantly smaller operating range of a typical QWT signal compared to the CT signal. The transition in temperature values in response to a change in TPU operating mode is much smaller, at approximately 10° C and, thus, no additional component is required to model these transition periods.

The parameters of the GMM were identified from 3 days of fault-free data, collected following the most recent maintenance overhaul on the TPU. It is necessary to refit the parameters of the GMM after each major maintenance overhaul since, following maintenance, the normal operating values of the system often change slightly. This can be due to a number of reasons, including the replacement of different sensors/components at each interval, manufacturing variations in replaced components, and the specific actions of different maintenance personnel in returning the system to fully operational status.

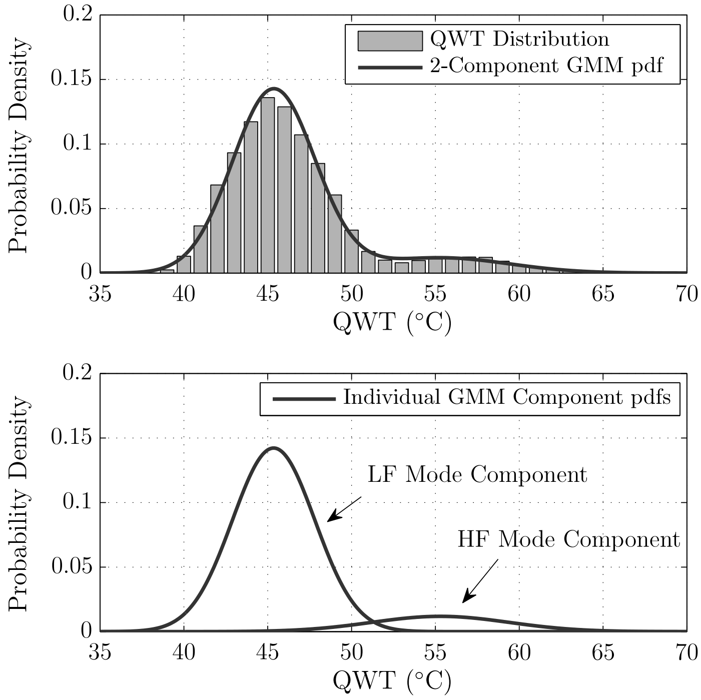


FIGURE 5.21: QWT Signal: 2 Component GMM

Once the model describing the distribution of the QWT signal during fault-free operation has been identified, each subsequent QWT value can be compared with the GMM and, using a distance measure and a specified threshold, future samples which appear unlikely to have been generated by the distribution describing normal fault-free operation can be labelled as "novel", indicating the potential existence of a fault condition. For this application, the distance metric used was simply the likelihood function, which describes the probability that the current sample was generated from the GMM component describing the distribution of the fault-free QWT signal, in the LF operating mode (θ_{LF}).

The likelihood function for a QWT signal value, recorded during LF mode operation, is given by

$$p(x_k|\theta_{LF}) = \frac{1}{\sqrt{2\pi\sigma_{LF}^2}} \exp\left\{\frac{-(x_k - \mu_{LF})^2}{2\sigma_{LF}^2}\right\} \tag{5.1}$$

where x_k is the value of the QWT signal at time t_k, θ_{LF} is the component of the GMM which describes the distribution of the QWT signal in the LF mode, and μ_{LF} and σ_{LF}^2 which are the mean and variance values of the LF mode component, contained in θ_{LF}. Figure 5.22 illustrates the output generated by the simple novelty detection algorithm described above. In the upper plot, the original QWT signal value is illustrated. In the lower plot, the likelihood function (5.1), computed for each QWT signal value x_k, recorded during LF mode operation, is shown. A novelty threshold limit for the likelihood function, below which an x_k value is declared as novel, was set at a value of 0.03. This value was chosen based on testing of the algorithm on historical data.

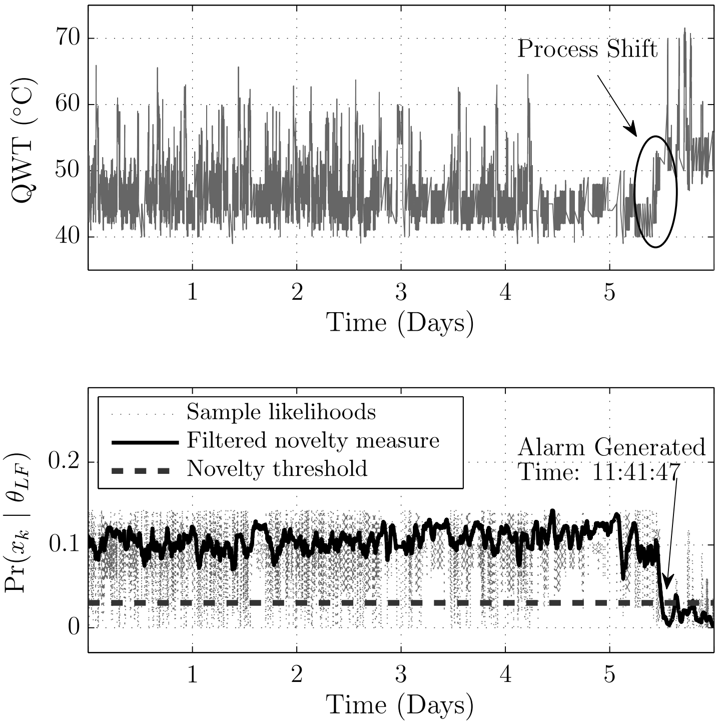


Figure 5.22: Evolution of QWT signal and inferred novelty measure, illustrating fast detection of process shift in QWT signal

As can be seen in Figure 5.22, the novelty output for the observed x_k values in LF mode regularly falls below the novelty threshold. This is caused by the TPU switching from HF to LF mode. Before the QWT signal decays back to normal LF mode operating values, the threshold limit is commonly exceeded. To address this issue, the output of the likelihood function if filtered using a low-pass filter to remove the short-term fluctuations caused by the TPU system switching from HF to LF mode. The resulting filtered novelty outputs are shown in the lower plot of Figure 5.22. Following the filtering operation, the novelty threshold is no longer exceeded regularly. As can be seen, the new filtered novelty output does not fall below the novelty threshold until the time the process shift occurs in the QWT signal.

In this example, an alarm is now generated at 11:41:47, some five and a half hours before an alarm was generated by the original unimodal fault detection approach, as previously shown in Figure 5.20. This provides maintenance staff with earlier warning of a maintenance issue, potentially reducing the overall impact of the fault condition on equipment availability.

5.4.4.3 Discussion

Each of the examples so far have illustrated the applications of GMMs for univariate modelling. The lack of multivariate modelling presented to date has been due the low number of signals monitored by the FabWorks network and the irregularly sampled nature of the data. Using the send-on-delta sampling approach, each of the signals are recorded on an irregular sampling interval. This means each sample must first be resampled onto the same, periodic, sampling interval. However, this can result in significant uncertainty regarding true signal values when no updates on signal values are received for an extended period of time.

In the development of the algorithms presented here, the issue of extending the described approaches into the multivariate domain was always considered. With the move toward ethernet based communication, more signals, recorded at fast sampling rates, will become available. This will permit the extension of the approaches described here into the multivariate domain. For example, consider Figure 5.23, which illustrates the joint distribution of the CT and QWT signals modelled as a two component GMM, using the F-J algorithm to identify the GMM parameters. Using the GMM approach to modelling multi-modal multivariate signal data, both signal tracking and novelty detection applications will be possible using the approaches described in the previous sections.

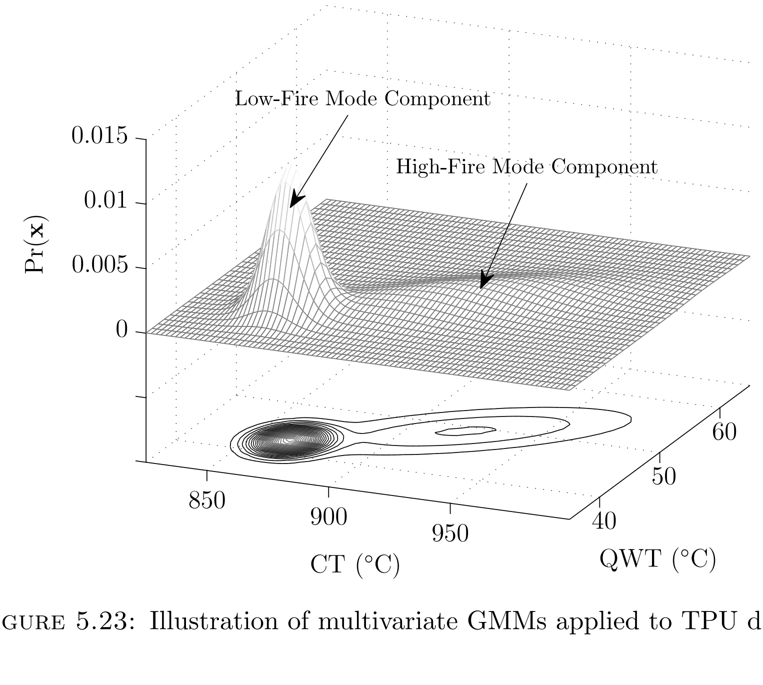


Figure 5.23: Illustration of multivariate GMMs applied to TPU data

5.5 Multi-Mode Vacuum Pump Signal Tracking

In this final section, an example of how the Gaussian mixture model representation of monitored signals can also be used within the dry-vacuum pump environment, for multi-mode signal tracking and condition monitoring. Figure 5.24 illustrates the booster power (BP) and exhaust pressure (EP) signals recorded from an iH1800 dry vacuum pump operating on a CVD manufacturing process. The upper plot in Figure 5.24 shows the BP signal, which exhibits multi-modal behaviour. This behaviour is reflective of the varying gas loads on the vacuum pump. The varying gas loads are generated by the different recipe steps in the manufacturing processes, when different combinations of gas types and flow rates are injected into the process chamber.

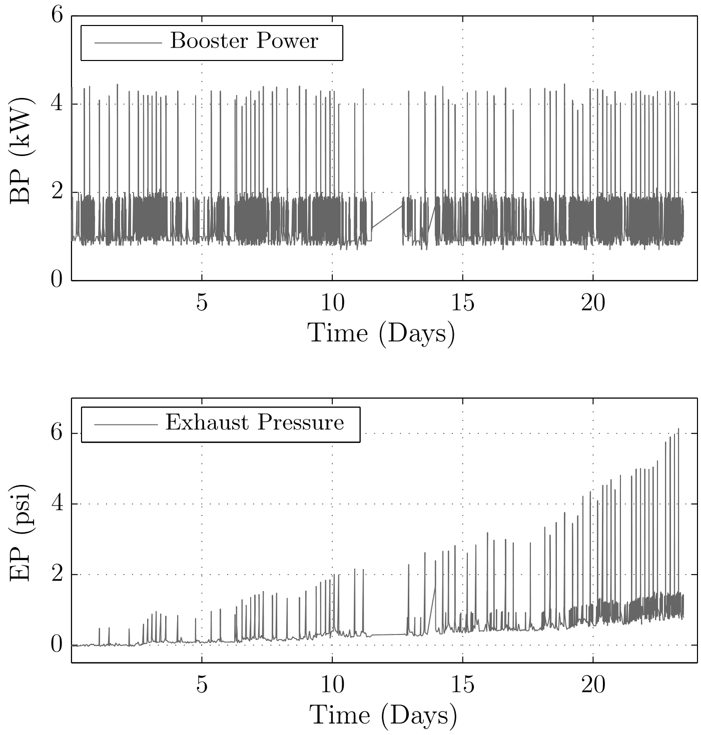


Figure 5.24: Booster Power & Exhaust Pressure signals example

The lower plot of Figure 5.24 shows the EP signal. During the initial stages of the recorded EP signal, there is no visible response in the EP signal to the peak values of the BP signal. However, as time evolves, the EP signal begins to respond with increasing peak values, at the same time as the peak values in the BP signal occur. The increasing peak values of the EP signal are caused by a build-up of deposits within the pump exhaust line. The deposits are similar to those which form on the walls of a TPU ceramic liner. As the volume of deposits increases, the pump exhaust can eventually become blocked, leading to an automatic in-service pump shutdown. This can lead to a loss of any wafers undergoing processing in the upstream chamber, and significant manufacturing downtime. Additionally, the loss of vacuum in the process chamber may result in the process chamber having to be stripped down for a total clean, at a cost which is often far in excess of the pump replacement costs.

The issue of exhaust pressure blockages has been investigated previously, resulting in a number of relevant publications. Twiddle *et. al* [81] developed a fuzzy-model based diagnostic scheme to monitor two fault conditions in a dry-vacuum pump: Mechanical inefficiency and exhaust system blockage. The approach described involves time and frequency analysis of the exhaust pressure signal. Power ratios of certain frequency components in the EP signal are used to monitor the pump condition, and changes in the periodic features of the signal, symptomatic of fault conditions, are detected using a fuzzy reference model. The approach described was successful in detecting both mechanical inefficiency and exhaust blockage. Twiddle *et. al* [100] also describe the development of a sliding mode observer (SMO) for on-line condition monitoring of dry vacuum pumps. The exhaust pressure signal is described using an auto-regressive (AR) model, and a discrete-time SMO was designed to estimate the AR model coefficients based on a short data set, sampled from the EP signal, and a nominal set of model coefficients estimated from fault-free data. The results showed that the build-up of deposits in the pump exhaust can be detected by monitoring the injection signal of the SMO.

The two publications by Twiddle *et al.* [81, 100], which were both developed in an experimental test-bed environment, detail promising solutions to the issue of exhaust pressure blockage within dry-vacuum pumps. However, the data requirements of these approaches, in terms of necessary signal resolution and sampling frequency, are not yet achievable within the manufacturing environment. This is due to sampling limitations which arise from having hundreds of pumps and TPUs, which run continuously, connected to a networked monitoring system. This section details the development of a exhaust pressure tracking solution, which addresses the multi-modal operating characteristics of typical vacuum pump signals, and which can be implemented with the currently achievable signal tracking resolution.

5.5.1 Pump Mode Tracking

Figure. 5.24 illustrated how the process of deposit buildup in a pump exhaust line results in a gradual change in the underlying distribution of the EP signal. The magnitude of the change in the underlying EP signal distribution varies according to the gas load on the pump. This is due to the the relatively low net gas flow rates which are typical of dry-vacuum pumps. Under low gas loads, a reduction in the effective diameter of the exhaust line can be difficult to detect in the exhaust pressure signal. Under higher gas loads, it is more likely that a reduction in the effective diameter of the exhaust line will generate a response in the EP signal. These principles are reflected in the observed EP signal behaviour in Figure 5.24.

To track how the distribution of the EP signal changes under different gas loads, it is first necessary to determine the gas load, or *pump mode*, at any instant in time. However, unlike the case of TPU monitoring, in which changes in the TPU status are recorded, there is no automatic method for tracking the dry-vacuum pump mode. Instead, the pump mode must be inferred from the available pump signals. The BP signal in a dry-vacuum pump provides the best method for tracking changes in the gas load on the pump. This is because the pump controller manipulates the BP signal value to maintain a fixed rotational speed of the booster pump, in response to varying loads on the pump. Figure 5.25 shows a zoomed section of the BP signal shown previously in Figure 5.24, which illustrates the response of the BP signal to changing gas loads.

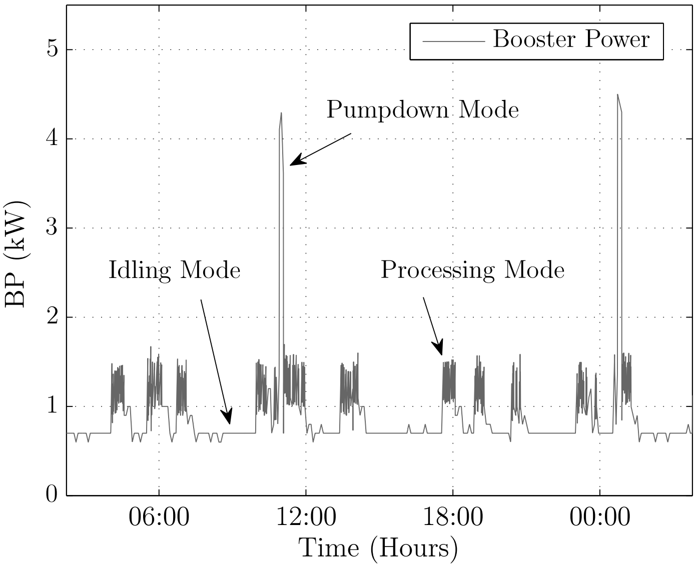


FIGURE 5.25: Booster Power signal

The quality of signal resolution, seen in the BP signal in Figure 5.25, was made possible by changes to the pump logging profile. Previously, all pump data recorded by FabWorks used a 1-hour timeout on the pump logging profiles. This meant that if the relevant δ value for a pump signal was not exceed for 1 hour, then a new update, containing the latest signal value is automatically sent to the central database. The use of a 1 hour timeout interval often results in extended periods of time, up to 1 hour, when no updates are received. This results in a significant level of uncertainty regarding the true signal values during these periods, which makes it extremely difficult to to generate statistical characterisations of the recorded pump signals.

The new logging profile, using a 30 second timeout interval, was implemented on a single pump, to investigate how improved signal resolution in dry-vacuum pumps may improve condition monitoring capabilities. As the FabWorks monitoring system moves toward ethernet based communications, such logging profiles will become common, resulting in improved signal tracking accuracy. The logging profile on this specific pump was maintained for a period of 18 months, during which a number of maintenance events were captured. The recorded pump signals were then resampled onto a regular 10 second sampling interval, using linear interpolation.

Figure. 5.25 highlighted three distinct pump modes which could be identified in the BP signal, which were labelled as follows

Idling Mode This mode corresponds to periods of no processing in the upstream chamber resulting in the booster power remaining at steady-state operating values.

Processing Mode This mode appears to reflect a period of sustained gas load being imparted on the pump, resulting in the BP value increasing to maintain the booster pump rotational speed at a fixed value, in response to the increases in the gas load.

Pumpdown Mode This mode refers to a stage within most CVD manufacturing processes in which the wafers are exposed to a sustained period of very high gas flows to achieve the desired level of deposition on the wafers. This results in a large sustained increase in BP values as seen in Figure 5.25

Figure 5.26 shows the distribution of the BP signal, shown in Figure 5.25, in histogram format. Once again it is possible to distinguish the three pump operating modes within the BP signal. Using a GMM, it is possible to model the distribution of the BP signal in Figure 5.26 using a mixture of individual Gaussian PDFs, in which each of the individual Gaussian PDFs characterises the distribution of the BP signal in one of the pump operating modes. Once the model has been developed, simple Bayesian inference can be used to infer the pump mode at any future instant using the BP signal. Figure 5.27 shows the BP signal in Figure 5.26, modelled as a three component GMM. The overall GMM PDF density identified using the F-J algorithm is shown in the upper plot, and the individual PDF densities are shown in the lower plot.

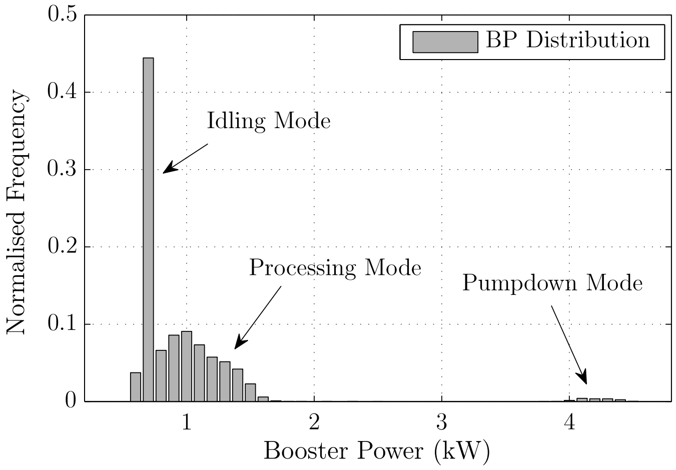


FIGURE 5.26: Booster Power signal distribution

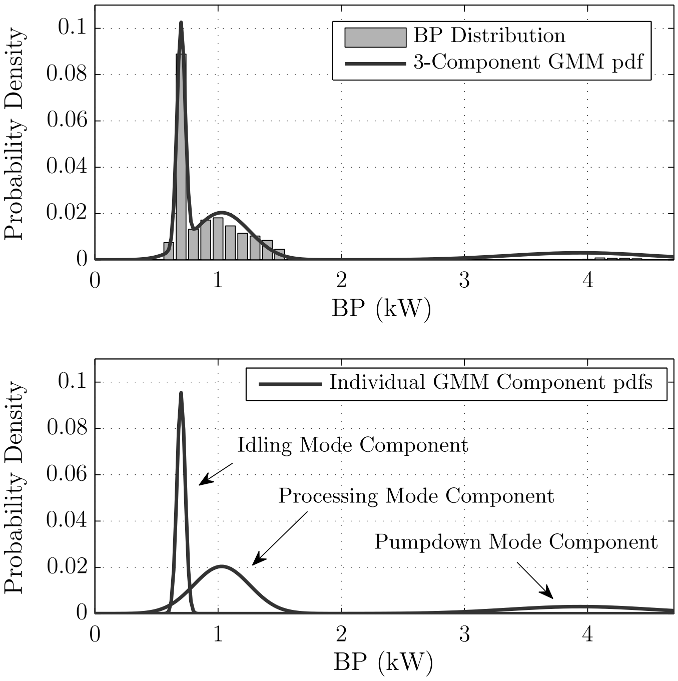


FIGURE 5.27: Booster Power signal distribution modelled using a GMM

Once the GMM describing the distribution of the BP signal has been generated, it is then possible to infer the pump mode at any time in the future using the BP signal value at that time. To determine the pump mode, the conditional probability that each of the GMM components was responsible for generating the observed BP value, x_i at time t_i, is computed (5.2). The GMM component which generates the largest conditional probability value is selected as the pump mode at that time instant

$$p(\text{Mode}_k|x_i) = \frac{\alpha_k \mathcal{N}(x_i|\mu_k, \sigma_k)}{\sum_{k=1}^{K} \alpha_k \mathcal{N}(x_i|\mu_k, \sigma_k)} \tag{5.2}$$

where $p(\text{Mode}_k|x_i)$ describes the probability that the observed value of the BP signal x_i, at time t_i, was generated by component k, which describes the distribution of the BP signal in a specific pump operating mode. Using (5.2), it is possible to infer the current pump mode at each sample time. Then, similar to the multi-mode CT tracking algorithm in which the TPU status signal was used to separate out the original CT signal by TPU mode, the inferred pump mode signal can be used to track each of the other monitored pump signals, in each of the pump operating modes. This process is illustrated in Section 5.5.2

Aside from the use of the inferred pump mode signal for tracking the EP signal, the inferred pump mode can be used for a range of more general applications. Firstly, pump maintenance personnel are often unaware of the volume of wafers which are processed in an upstream processing chamber, over any period of time. As a result, pump swap out times, which are often chosen on time-based intervals, may not not correlate well with the actual load imparted on a pump over a specific time interval. By inferring the pump mode, the load on the pump can be tracked in real-time, which can be used to develop more accurate mean-time-between-failure (MTBF) estimates, which reflect the work done by a pump, instead of simply the total running hours.

Additionally, the relative load severity of different manufacturing processes can be inferred from the GMM developed for each individual process. The GMM for each process can be identified online, which means that the variations in operating profiles for different processes can be easily described. Finally, the developed statistical characterisation of typical pump loads can be used for developing simulations of potential future loads. These simulated loads could potentially be used in the development of model-based prognostics for vacuum pumps, in which the expected future operating load on a pump can be incorporated into the generation of RUL estimates.

5.5.2 Multi-Mode Exhaust Pressure Tracking

This section illustrates how the inferred pump mode signal can be used to track any of the other recorded pump signals which may exhibit multi-modal behaviour in response to changes in the pump mode. Figure 5.28 illustrates the EP signal, as seen previously in Figure 5.24, in which a build-up of deposits in the pump exhaust is gradually reducing the diameter of the pump exhaust pipe. Thus, when the pump is exposed to high gas loads, a significant rise in the pressure value recorded in the pump exhaust is observed.

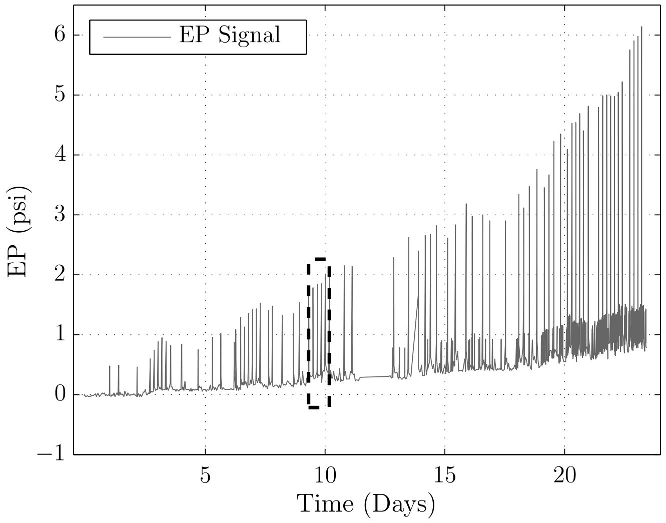


Figure 5.28: Exhaust pressure (EP) signal response to increasing deposit buildup

In Figure 5.28, a region of data covering approximately 15 hours is highlighted. Using the inferred pump mode at each sample time, all of those EP samples recorded when the pump was operating in pumpdown mode were selected. The distribution of those samples is illustrated in Figure 5.29. As illustrated in Figure 5.29, the distribution of the EP signal, during the pumpdown mode, exhibits a similar distribution to the TPU HF mode CT signal. Therefore, to track how the distribution of this signal changes in response to deposit buildup, the same approach to tracking the HF mode CT signal is employed, using a moving-window, where at each iteration the parameters of a 2 component GMM are identified. This is illustrated in Figure 5.29. Using the GMM parameter values identified at each iteration, the changes in the EP signal distribution can be tracked over time.

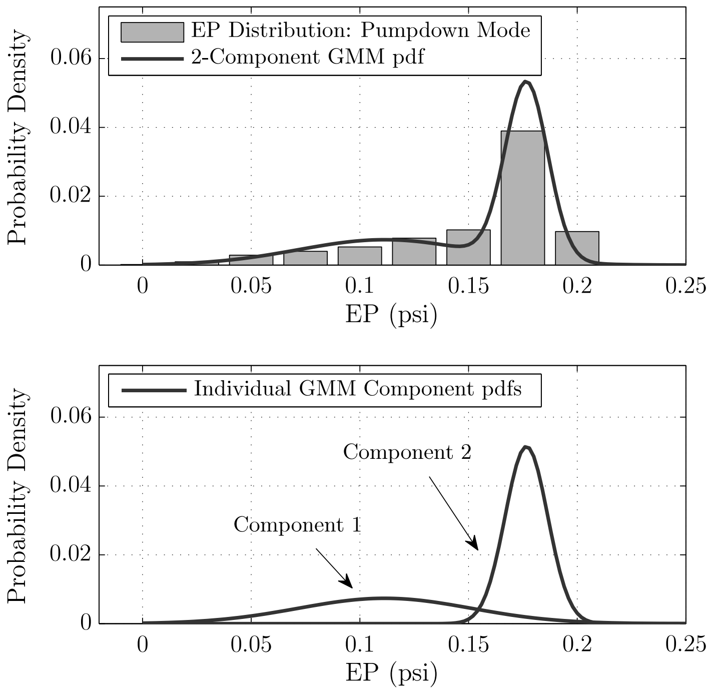


FIGURE 5.29: Exhaust pressure signal distribution (pumpdown mode)

Figure 5.30 illustrates the tracking of the EP signal in two of the pump operating modes. The pumpdown mode tracking signal is simply the the mean value of Component 2 in Figure 5.29, identified at each iteration of the moving window. The idling mode tracking signal is computed by simply calculating the moving-average of the EP values recorded during periods of idling mode operation, as the EP signal follows a Gaussian distribution in this pump mode.

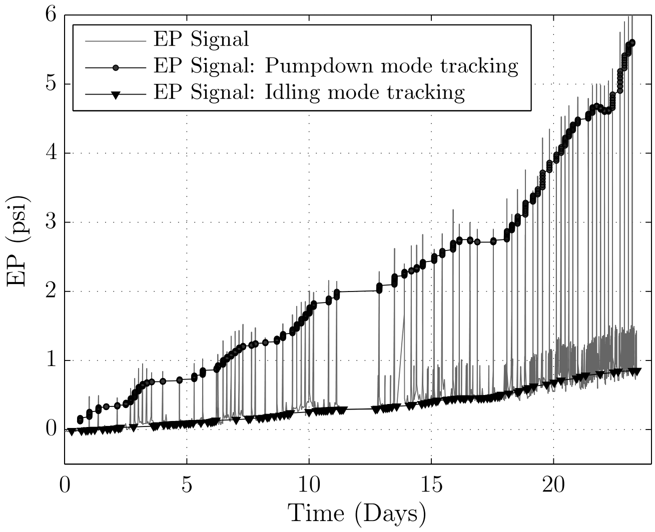


FIGURE 5.30: Exhaust pressure signal tracking in different pump operating modes

The ability to identify the pump mode is critical to permit the accurate tracking of the multi-modal signals which are generated by the changing gas loads imparted on the pump. The approach described above, using GMMs and Bayesian inference to identify the pump mode at each sample time, provides a robust approach to inferring such information. Furthermore, by modelling the distribution of the BP signal using GMMs, it is possible to identify the operating characteristics generated by pumps running on different manufacturing processes in an on-line fashion, where the number, and frequency, of different pump modes can vary significantly.

5.6 Conclusions

This chapter has presented the development of a condition monitoring algorithm for TPU systems. The primary challenge addressed is the multi-modal, non-Gaussian distribution signals, which are generated by TPU systems. These signals are generated as a result of the TPU switching between operating modes. To address this challenge, a GMM-based solution was developed. The main benefit of using a GMM-based approach to addressing this challenge, is the ability to parameterise the underlying distribution of the TPU signals. Then, by tracking the changes in the parameter values of the GMM used to model the signal distributions, robust and accurate condition monitoring is possible. This ability was demonstrated on historical TPU failure examples. The GMM-based multi-mode signal tracking algorithm, developed for tracking the buildup of deposits within a TPU combustion chamber, also generates a suitable signal, or fault indicator, for use in developing prognostic capabilities for TPU systems. The development of these prognostic capabilities is presented in the following chapter.

In addition to addressing the challenge of tracking deposit buildup in the TPU combustion chamber, the GMM representation of TPU signals is also demonstrated to have other applications. Section 5.4.4 presented the development of a novelty detection algorithm for TPU systems. By modelling the fault-free distribution of the TPU signals, potential fault conditions can be identified. The novelty detection algorithm, demonstrated on a fault in the quench-unit of a TPU, has the capability to detect potential fault conditions as soon as they occur, thus avoiding the potential propagation of a fault, and minimising potential equipment downtime.

Finally, in Section 5.5, another application of GMMs for dry vacuum pump condition monitoring was presented. By modelling the distribution of the booster power signal using a GMM, simple Bayesian inference can then be applied to identify the operating mode of a pump, at any instant in time. In addition, by modelling the underlying distribution of the booster power signal, using data collected on-line, models for dry vacuum pumps operating on different manufacturing processes can be identified online. In this way, the multi-modal signal tracking approach is adaptive to different manufacturing processes, which generate different load profiles on dry vacuum pumps. The benefit in identifying the dry vacuum pump mode at any instant in time is also demonstrated using an historical pump failure example. The response generated by pump signals, which can be used to identify and track the development of a fault condition, is often dependant on the load imparted on the pump at any instant in time. By having knowledge of the pump mode at each sample time, the development of fault conditions in each mode can be investigated. The benefit of this capability is demonstrated in Section 5.5.2, where tracking the exhaust pressure signal, under high load conditions, provides a means to accurately identify and track the development of an exhaust pressure blockage.

Chapter 6

TPU Prognostics

6.1 Introduction

Prognostics has been identified as the key enabler of a truly condition-based maintenance approach, providing maintenance personnel with visibility on upcoming maintenance issues so that corrective action can be taken to avoid equipment failure. However, the key difficulty encountered in the development of prognostic algorithms is the inherent level of uncertainty associated with the generation of long-term predictions of equipment health. In this chapter, the development of a model-based prognostic algorithm for predicting the remaining-useful-life (RUL) of TPU systems is presented. The prognostic technique investigated is particle filtering, which has emerged in recent years as the defacto state-of-the-art technique for prognostics, with capabilities for representing and managing the inherent uncertainty involved in predicting the future behaviour of degrading equipment.

In reviewing published literature on the application of particle filtering to prognostics, a notable observation is the lack of real-world application examples, using data collected from equipment operating in an industrial setting. Indeed, the requirement for applying and testing newly developed prognostic algorithms on real world data, as opposed to simulated data, or data collected on a test rig, has been identified previously [101]. In this study, particle filtering for prognostics is investigated using real-world data collected from a large semiconductor manufacturing facility. The benefit of using real-world data, in developing prognostic algorithms, is that such data captures and illustrates how different sources of uncertainty influence how a fault evolves over time. Such issues can often be difficult to anticipate, capture, or represent, when using simulated or test-rig data. This chapter focuses on developing and demonstrating prognostic solutions for

TPU systems. The proposed solution addresses the uncertainty challenges presented by the semiconductor manufacturing environment.

The layout of this chapter is as follows. Section 6.2 introduces prognostics for TPU systems and discusses the potential benefits, and the specific challenges, in implementing prognostics for TPU systems. Section 6.3 then introduces particle filtering for TPU prognostics. A review of the implementation details involved in applying particle filtering to TPU prognostics is presented. Section 6.4 then illustrates the application of particle filtering to TPU prognostics, and discusses how the various sources of uncertainty within the semiconductor manufacturing environment affect the ability to accurately predict TPU RUL. Finally, in Section 6.5, a multiple model particle filtering approach is introduced to address the uncertainty challenges presented by the semiconductor manufacturing environment. The capabilities and performance of the multiple model particle filtering approach are demonstrated on historical TPU failure examples.

6.2 Prognostics for TPU Systems

6.2.1 Benefits of Prognostics for TPU Systems

This section discusses how prognostic technologies have the potential to deliver major improvements in how TPU systems are maintained, and lead to a significant overhaul of current maintenance approaches and practices for TPU systems. The potential benefits of prognostic solutions include increasing equipment uptime and availability, a reduction in the frequency of scheduled downtime, reducing instances of in-service failure, and reducing overall operating and maintenance costs. For example, as discussed previously in Section 5.3.2.5, some manufacturers operate two TPU systems in parallel on certain processes, to provide immediate redundancy in the event one of the TPU systems failing. Utilising prognostic technologies, it might be possible to reduce the risk of in-service failure sufficiently, that it no longer makes economic sense to operate two systems in parallel, which effectively doubles the capital, operational, and maintenance costs of operating a thermal abatement system.

Prognostics may also lead to a significant reduction in ongoing maintenance costs for TPU systems. Maintenance personnel have traditionally utilised time-based intervals or derived statistics such as mean-time-between-failures (MBTF) to schedule maintenance activities. Such approaches often choose conservative estimates for MTBF to try and avoid in-service failures. However, this approach often results in components being replaced before they have reached the end of their serviceable life, or occasionally, failing in-service. The cost of TPU system components, such as the ceramic liner within a TPU combustion chamber, are not insignificant, and the premature replacement of these parts results in potentially avoidable costs. The cumulative costs of premature component replacement within a manufacturing facility are all the more significant when you consider that a large semiconductor plant may have in excess of 50 TPU systems in operation over a 20 year lifespan of a plant. Accurate prognostics would improve the lifetime utilisation of key components, so that the in-service life of components, which are swapped out as part of necessary maintenance, is maximised. Perhaps the most significant benefit which prognostic technologies for TPU systems might provide is in the overall management and planning of maintenance activities within both the sub-fab and fab environments. In the absence of any prior warning of a TPU system failing, maintenance actions will always be reactionary in nature. In such situations, the necessary maintenance personnel, spare parts, tools and supplies must be organised, often resulting in considerable delays before remedial action can be undertaken. In a large semiconductor manufacturing facility, these delays often results in downtime for all of those processing tools connected to the failed TPU, which is all the more significant when you consider that a TPU system can be connected to up to four individual wafer processing chambers.

The cost of process tool downtime usually far outweighs the cost of rectifying maintenance issues. Unexpected tool downtime can have a major impact on production scheduling, reducing the overall wafer throughput. In an increasingly lean and competitive business environment [96], manufacturers are striving to maximise tool availability and minimise instances of in-service equipment failure [102]. Prognostics represents a key enabling technology in achieving these goals.

Prognostic algorithms are designed to provide maintenance staff with accurate and reliable estimates of the RUL of a TPU suffering from a buildup of deposits. This will provide maintenance personnel with sufficient warning of upcoming maintenance events, so that necessary remedial maintenance work can be planned in advance, with the necessary resources and personnel allocated as necessary. Additionally, with the availability of sufficient lead times before failure, the ability to perform *opportunistic maintenance* becomes possible. This means for example, that if a TPU needs to be taken down for maintenance, the upstream process tool maintenance staff are provided with an opportunity to perform any necessary maintenance, or any other work that might be necessary on the processing tool, at the same time. The objective is to increase the availability of the entire manufacturing process.

Whilst the potential benefits of prognostics help explain the growth in related research in recent years [16], there are many reasons that such technologies have not yet become commonplace. The primary difficulty associated with the development of prognostics is the large level of uncertainty, associated with the generation of long-term predictions of equipment health. Indeed, prognostics has been described as the Achilles' heel of a true prognostic and health management (PHM) system [103]. How this uncertainty is represented and managed is is key to the success of any future prognostic technologies.

6.2.2 Challenges in Developing Prognostics for TPU Systems

The primary difficulty encountered in the development of prognostic algorithms is the significant uncertainty associated with the generation of long-term predictions, describing the evolution of a signal of interest, or fault indicator measurement, into the future. This uncertainty arises from a range of different sources. One of the primary sources of uncertainty arises in estimating the current level of system degradation. Any errors at this stage will be propagated into the future, with the margin of error increasing as the prediction horizon increases. To address and reduce this uncertainty it has become common to represent the current level of degradation as a random variable, with an associated probability distributions. This allows for uncertainty regarding the current level of degradation to be represented and defined. It also allows for the uncertainty regarding the current degradation level to be predicted into the future, facilitating the estimation of RUL of the system in terms of a probability distribution, from which further attributes such as confidence intervals on the prediction can be computed.

In addition to uncertainty associated with monitoring or estimating the current degradation level of a system, there are many other sources of uncertainty in generating long-term predictions of equipment health. One of the most significant sources is modelling error. The generation of long-term predictions often requires a model describing the future evolution of a fault indicator. However, the dynamics of fault processes are often non-linear, with uncertainties that are often non-Gaussian [29], which make the challenge of modelling such processes more difficult. This introduces further uncertainty regarding the accuracy of predictions generated. Another major source of uncertainty concerns the future load profile on the system. This issue is of particular importance since the RUL of a system is likely to be a major function of the sequence and nature of the future loads imparted on the system.

The issue of uncertainty, in the generation of long-term predictions, leads to two key concepts which a prognostic algorithm must address, *uncertainty representation* and *uncertainty management*. Uncertainty representation implies the ability to model various forms of uncertainty stemming from a variety of sources, whereas uncertainty management concerns the methodologies and tools needed to continuously “shrink” the uncertainty bounds as a fault evolves [3]. One of the most successful and promising approaches to prognostics, to date, is particle filtering, which provides a framework for addressing the issues of uncertainty management and representation. Indeed, particle filtering has been described as the de facto *state of the art* in failure prognosis [31] The issues of uncertainty management and representation are key in the development of prognostic technologies within the semiconductor manufacturing environment in which there are many factors, which can influence future equipment behaviour, which cannot be predicted *a priori*. For this reason, particle filtering was chosen as a technique to investigate in the development of a prognostic solutions for TPU systems.

6.3 Dynamic Model-Based Prognostics

Prognostics is understood to be the generation of long-term predictions, describing the evolution of a signal of interest, or fault indicator, for the purpose of estimating the RUL of a failing system [27]. Inherent in the generation of such long-term predictions is a significant level of uncertainty associated with the propagation of the current degradation level into the future, in the absence of any future measurements. This uncertainty is propagated at each future horizon, resulting in increasing levels of uncertainty as the prediction horizon increases. Thus, one of the most important attributes of any prognostic technique is the ability to manage, and reduce, this uncertainty as well as possible.

In addition to the uncertainty associated with generating long term predictions, another major source of uncertainty arises in estimating the current level of degradation. In some cases, the current level of degradation may not be directly quantifiable from system measurements and must, instead, be inferred from system measurement data. This can lead to errors in inferring the current degradation state. Additionally, the original measurement data may also be affected by noise and disturbances. A sensible approach to addressing this uncertainty is to represent the current level of degradation as a random variable, the *degradation state*, with an associated probability distribution which must be inferred from the sequence of measurement data received from the system under observation. This suggests the use of a dynamic model-based approach to inferring the current degradation state, using recursive Bayesian estimation techniques.

Dynamic model-based approaches to prognostics employ a state-transition model which describes the evolution of the degradation state of a system. The predictions from the state-transition model are combined with the sequence of measurement data received from the system under observation to generate *a posteriori* estimates of the degradation state PDF. As new measurement data arrives sequentially, the *posterior* degradation state PDF is recursively updated. Figure 6.1 illustrates the key concepts of a dynamic model-based approach to prognostics, based upon the use of recursive Bayesian estimation techniques. There are two distinct stages in the process; 1) *state estimation* and 2) *long-term prediction*.

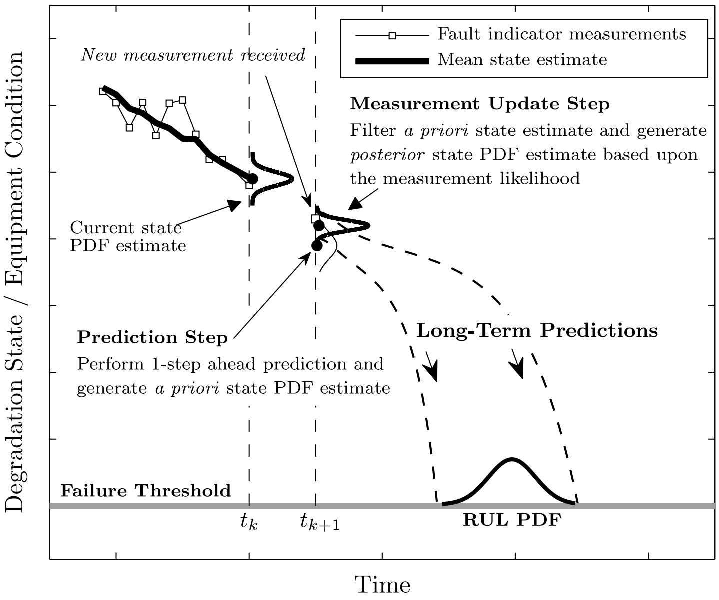

Fault indicator measurements
Mean state estimate
New measurement received
Measurement Update Step
Filter a priori state estimate and generate posterior state PDF estimate based upon the measurement likelihood
Current state PDF estimate
Prediction Step
Perform 1-step ahead prediction and generate a priori state PDF estimate
Long-Term Predictions
Failure Threshold
t_k
t_{k+1}
RUL PDF
Degradation State / Equipment Condition
Time

Figure 6.1: Principle of dynamic model-based prognostics

The first stage, state estimation, is carried out by performing two sequential steps; *prediction* and *update* [71]. In the prediction step, the current state PDF estimate at time t_k, and the state-transition model (which describes the evolution of the degradation process), are used to generate an *a priori* state PDF estimate at time t_{k+1}. The update step then combines the latest measurement data received at time t_{k+1}, the *a priori* state PDF estimate, the *measurement likelihood* function, and using Bayes' law, generates the *a posteriori* degradation state PDF estimate at time t_{k+1} [27]. This process of recursively estimating the degradation state PDF, using a prediction step followed by an update step, is commonly known as a *filtering* problem.

Once the latest *posterior* state PDF estimate has been generated at time t_{k+1}, the second stage of the prognostic process can be carried out; the generation of *long-term predictions* describing the future evolution of the degradation state. To generate these long-term predictions, the current *posterior* state PDF estimate is propagated into the future in a recursive manner, using the state-transition model which describes the evolution of the degradation state. The degradation state is propagated until its value exceeds a defined threshold value. Using these predictions, a RUL PDF is then generated which describes the probability of system failure at different times in the future. Corrective maintenance activities must then be performed, before the just-in-time-point (JITP) (see Section 2.5.1 for details), to avoid a potential in-service failure of the system.

The primary benefit of the dynamic model-based approach to prognostics is the ability to quantify the level of uncertainty in the estimate of the current degradation state, which is represented by the *posterior* state PDF. This uncertainty can then be incorporated into the prediction of the RUL of the system. As the degradation process evolves, it is reasonable to assume that new measurement data will continue to be received, for some period of time, before system failure. This sequence of new measurement data can be used to continuously refine the RUL estimate, via recursive updating of the current system state and the generation of new long-term predictions.

Within the domain of recursive Bayesian estimation techniques, the Kalman filter is by far the most commonly employed technique, and provides the optimal solution (in the sense of estimating the *posterior* state PDF), provided that the underlying process can be modelled using a linear state-transition model, and that all the associated noise processes are Gaussian. However, the dynamics of degradation processes are typically non-linear, and/or the associated noise processes may be non-Gaussian [29]. This often results in a *non-linear filtering* problem for which the Kalman filter is no longer optimal. In such circumstances, approximations to the optimal solution can be made using techniques such as the Extended Kalman Filter (EFK).

In recent times, sequential Monte Carlo methods, more commonly known as particle filtering, have grown in popularity, for their flexibility and ease of design in tackling *non-linear filtering* problems [17]. Indeed, particle filtering has emerged as the de facto *state of the art* technique in failure prognostics [31], and has been demonstrated to outperform EKF based approaches in prognostic applications [27, 30]. The basic principle of particle filtering is to approximate the state PDF with a set of *particles*, representing state values, and an associated set of particle weights, which represent the discrete probability masses of the individual particles. The particles can be generated and updated recursively using a non-linear state-transition model which describes the evolution of the process under observation. Section 6.3.1 introduces the application of particle filtering for model-based prognostics of TPU systems.

6.3.1 Particle Filtering for Prognostics

Particle filtering, also known as sequential Monte Carlo (SMC) methods [72], is a technique for implementing a recursive Bayesian filter by Monte Carlo simulations. The basic principle of particle filtering is to represent the *posterior* state PDF by a set of random samples or "particles", each with an associated weight, and to compute estimates based on these samples and weights. In this study, a type of particle filter known as the sampling importance resampling (SIR) particle filter was used. The SIR particle filter approximates the *posterior* state PDF $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$ by a set S of N_s weighted particles,

$$S = \{\mathbf{x}_{0:k}^{(i)}, w_k^{(i)}\}_{i=1}^{N_s} \tag{6.1}$$

where $\{\mathbf{x}_{0:k}^{(i)}, i = 1, ..., N_s\}$ is a set of particles representing state values, with an associated set of importance weights $\{w_k^{(i)}, i = 1, ..., N_s\}$ which are approximations to the relative posterior probabilities of the particles, and $\mathbf{x}_{0:k} = \{\mathbf{x}_j, j = 0, ..., k\}$ is the set of all states up to time k.

The weight values are also normalised such that,

$$\sum_i w_k^{(i)} = 1 \tag{6.2}$$

The *posterior* state PDF can then be approximated as

$$p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) \approx \sum_{i=1}^{N_s} w_k^{(i)}\, \delta(\mathbf{x}_{0:k} - \mathbf{x}_{0:k}^{(i)}) \tag{6.3}$$

This results in a discrete weighted approximation to the true *posterior* state distribution $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$. As the number of samples/particles $N_s \to \infty$, this approximation will approach the true *posterior* distribution [71]. The particles can be generated and updated recursively using a state-transition model, describing the evolution of the process under observation

The process of applying the principles of particle filtering to prognostic applications is known as the *particle filtering framework* for prognostics [27]. This framework describes the process of generating RUL estimates from a sequence of measurements received from a process under observation. A review of the background, theory and application of particle filters for prognostics was presented previously in Section 3.4, which should be referred to as necessary. In the following sections, a brief review of particle filtering for TPU prognostics is presented. Firstly, in Section 6.3.2, a description of the model developed to describe the evolution of the TPU degradation state is presented. This is then followed in Section 6.3.3 by a description of the implementation steps involved in the application of the particle filtering framework for TPU prognostics.

6.3.2 Modelling TPU Degradation

Key to the implementation of a dynamic model-based approach to prognostics is a model which describes the evolution of the degradation process. Ideally such a model would be based upon the physics-of-failure and derived from first principles. In recent years a number of applications of particle filtering for prognostics have focused on applications such as predicting fatigue crack growth in structural elements, which have incorporated physics-of-failure models [29, 33]. However, in most real-world situations such models are unavailable, and the technical difficulty and cost of developing such high-fidelity models of complex physical processes is often prohibitive. In such scenarios, models can instead be derived, which describe the observed behaviour of historical failure examples.

In this study, the objective is to develop a prognostic algorithm to estimate the RUL of TPU systems which suffer from deposit buildup. The ability to infer the current level of deposit buildup is provided by the fault indicator measurements, which are derived from the original combustor temperature signal (see Section 5.4.3.4 for details). Figure 6.2 illustrates the CT signal recorded from a TPU which suffered an in-service failure due to deposit buildup within the combustion chamber. Overlayed on the CT signal, in Figure 6.2, are the fault indicator measurements, generated by the multi-modal CT tracking algorithm. The objective, then, is to develop a state-transition model of the TPU degradation process, as described by the set of fault indicator measurements.

Figure 6.3 illustrates the typical CT signal profile of a TPU failure example. The evolution of the TPU degradation state is characterised by two distinct stages/periods of decay. Initially, when the fault is first detected at time t_D, the degradation state decays at a slow rate. After a period of some days, the rate of degradation gradually

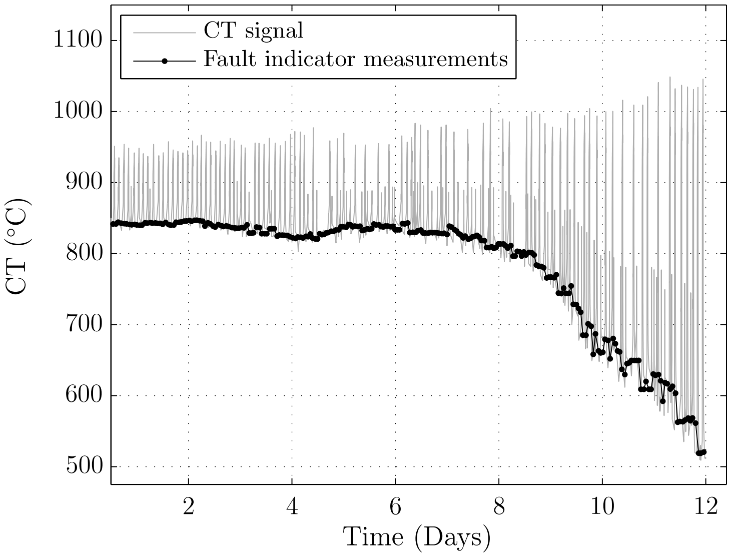


FIGURE 6.2: CT signal fault indicator measurements

accelerates until either the TPU suffers an in-service failure, or undergoes necessary maintenance.

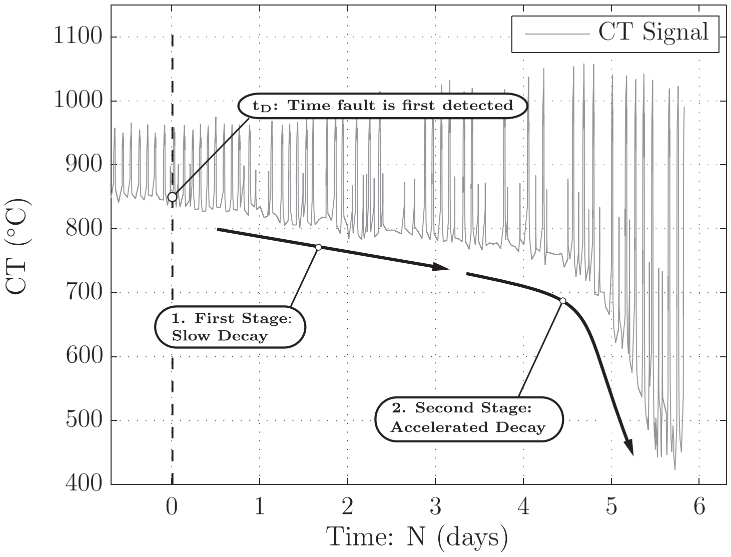


FIGURE 6.3: Typical CT signal profile in response to a gradual build-up of deposits

To apply model-based prognostics to predicting the RUL of a TPU, it is necessary to develop a model which describes the evolution of the TPU degradation state. In this study, the structure of the state-transition model used to describe the evolution of the TPU degradation state was adapted from a model developed by Saha & Goebel [104]. The model developed by Saha & Goebel was used to describe the evolution of lithium-ion battery discharge cycles, with a view to predicting the RUL of lithium-ion batteries, before the end-of discharge (EOD). The profile of these discharge cycles was observed to be very similar to the profile of the fault indicator measurements from a TPU, as seen in Figure 6.2. Additionally, the parameterisation of the model developed by Saha allowed it to be easily adapted, and tuned, to model the evolution of the TPU degradation process.

By adapting the model originally described by Saha, the evolution in time of the TPU degradation state is described by the following state-transition model

$$x_k = x_{k-1} \underbrace{- \alpha_1 \exp\left[\frac{-\alpha_2/t_k}{t_k^2}\right]}_{\text{Component 1}} \underbrace{- \alpha_3 \exp[\alpha_4\, t_k]}_{\text{Component 2}} \underbrace{- \alpha_5 t_k}_{\text{Component 3}} + \omega_k \tag{6.4}$$

where x_k is the TPU degradation state at time t_k, ω_k is the process noise PDF, and $\alpha_1, ..., \alpha_5$ are model parameters which can be tuned to adapt the behaviour of the model. The state-transition model in (6.4) is composed of three components which model the different stages of the TPU degradation process. Figure 6.4 illustrates the contribution of each of the model components in modelling the TPU degradation process.

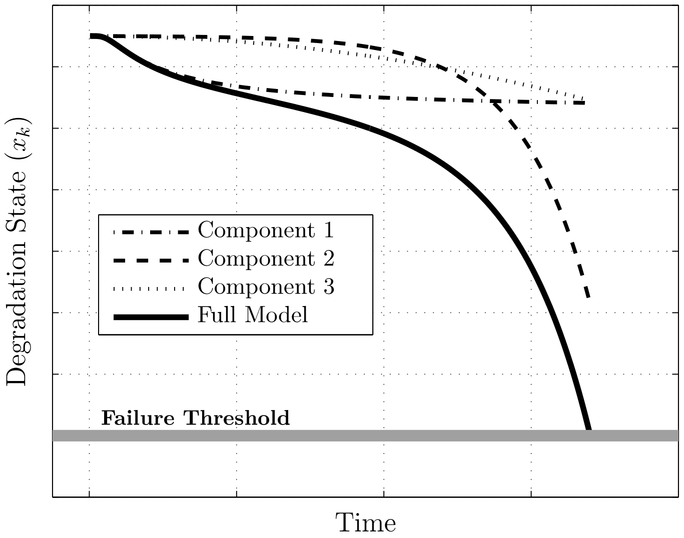


FIGURE 6.4: Illustration of the contribution of each of TPU degradation process model components

Component 1 models the initial stage of TPU degradation, where the rate of degradation is characterised by a slow decay. As time evolves, the t_k^2 term in component 1 grows at a faster rate than the $-\alpha_2/t_k$ term, which results in the contribution of this component in modelling the TPU degradation process, gradually decaying to zero. Component 2 then models the dynamics of TPU degradation during the second stage of decay which is characterised by an accelerated rate of TPU degradation. The final model component, component 3, provides further flexibility to tune the rate of degradation.

Figure 6.5 illustrate an example of the TPU degradation process model tuned to fit an historical TPU failure example. As seen in Figure 6.5, the model dynamics match the observed behaviour quite accurately. Using the structure of the state-transition model described in (6.4), it was possible to tune the model parameters to fit all of the available historical TPU failure examples.

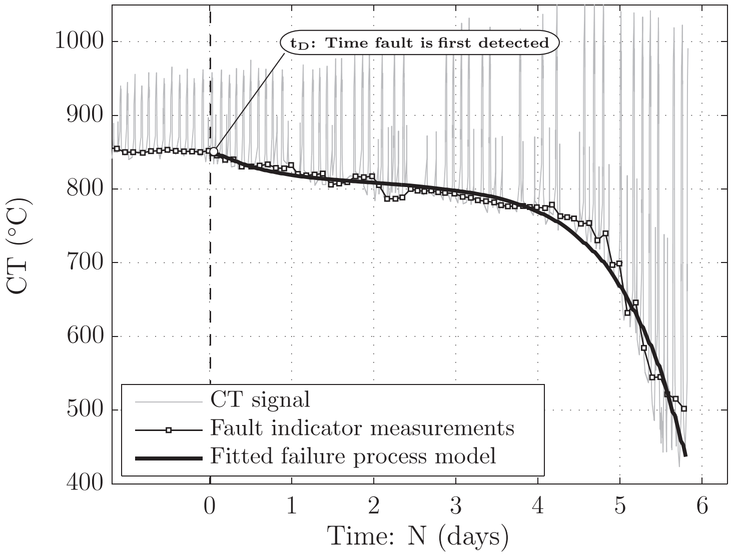
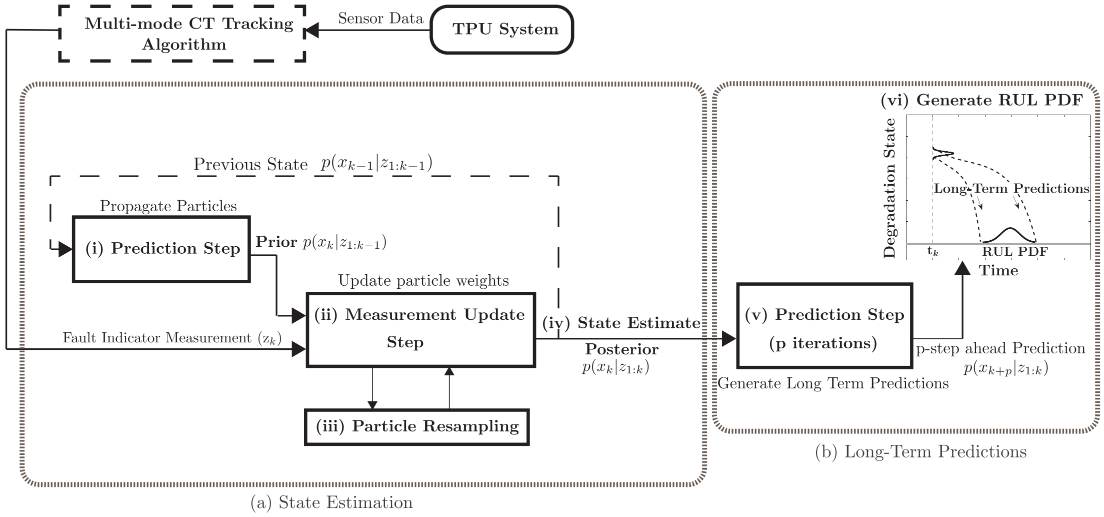


FIGURE 6.5: TPU degradation process model tuned to fit historical TPU failure example

6.3.3 The Particle Filtering Algorithm for TPU Prognostics

This section describes the implementation details in applying particle filtering to prognostics. The algorithm describes the entire process of using sequential fault indicator measurements to generate a RUL PDF of a system under observation. The algorithm has previously been described as the *particle filtering framework* for prognostics [73]. Figure 6.6 presents a flow chart illustrating the particle filtering algorithm for prognostics. There are two key stages within the particle filtering algorithm for prognostics which are highlighted in Figure 6.6, (a) *state estimation* and (b) *long-term predictions*.


Multi-mode CT Tracking Algorithm
Sensor Data
TPU System
(vi) Generate RUL PDF
Degradation State
Long-Term Predictions
t_k
RUL PDF
Time
Previous State $p(x_{k-1}|z_{1:k-1})$
Propagate Particles
(i) Prediction Step
Prior $p(x_k|z_{1:k-1})$
Update particle weights
(ii) Measurement Update Step
Fault Indicator Measurement (z_k)
(iv) State Estimate
Posterior $p(x_k|z_{1:k})$
(iii) Particle Resampling
(v) Prediction Step (p iterations)
p-step ahead Prediction $p(x_{k+p}|z_{1:k})$
Generate Long Term Predictions

(b) Long-Term Predictions

(a) State Estimation

FIGURE 6.6: The particle filtering algorithm for TPU prognostics

The purpose of the state estimation stage is to generate a *posterior* estimate of the degradation state from the sequence of fault indicator measurements received. The second stage, long-term predictions, use the current posterior state estimate as a starting condition and recursively iterates the current degradation state into the future, using the degradation process model. Using these predictions, an estimate of the RUL PDF is generated. In the current application, the TPU degradation process is modelled by the following state transition model, detailed in Section 6.3.2, which describes the evolution of the TPU degradation process once a fault is first detected, as

$$x_k = x_{k-1} - \alpha_1 \text{exp}\left[\frac{-\alpha_2/t_k}{t_k^2}\right] - \alpha_3 \exp\left[\alpha_4\, t_k\right] + \alpha_5 t_k + \omega_k \tag{6.5}$$

$$z_k = x_k + \upsilon_k \tag{6.6}$$

where x_k represents the TPU degradation state, z_k is the fault indicator measurement received from the multi-mode CT tracking algorithm, and ω_k and υ_k are independent zero-mean Gaussian noise terms, representing the process noise and the measurement noise respectively. The α values are model parameters which can be tuned to fit the observed behaviour. Equation (6.5) describes the state-transition function, and equation (6.6) describes the measurement likelihood function, which relates the fault indicator measurements to the system state estimate. Sections 6.3.3.1 and 6.3.3.2 briefly describe the specific functionality of each of the blocks in the flowchart in Figure 6.6 for each of the key stages in the particle filtering algorithm for prognostics, (a) *state estimation* and (b) *long-term predictions.*

6.3.3.1 (a) State Estimation

Once a fault is first detected in the CT signal, at time t_0, an initial estimate of the TPU degradation state is generated and represented as a set of particles and weights, S_0

$$S_0 = \{x_0^{(i)}, w_0^{(i)}\}_{i=1}^{N_s} \tag{6.7}$$

where $x_0^{(i)}$ and $w_0^{(i)}$ represent the individual particle values and weights, at time t_0, respectively, with S_0 representing an approximation to the true TPU degradation state x_0 at time t_0. The value of N_s determines the number of particles in the set S, which must be chosen *a priori.* Once the fault has been detected and an initial set of particles and weights has been generated, the *posterior* TPU degradation state can be recursively estimated from the sequence of fault indicator measurements by applying the following steps. The interaction of the different steps in the state estimation process is illustrated in Figure 6.6.

(i) Prediction Step

At each iteration the set of particles S_{k-1} are propagated according to the degradation process model to generate a new set of particles S_k, which represent the *a priori* TPU degradation state estimate

$$x_k = x_{k-1} - \alpha_1 \exp\left[\frac{-\alpha_2/t_k}{t_k^2}\right] - \alpha_3 \exp\left[\alpha_4\, t_k\right] + \alpha_5 t_k + \omega_k \tag{6.8}$$

(ii) Measurement Update Step

Once a new set of particles has been generated, the particle weights are updated by computing the likelihood of observing the latest fault indicator measurement value z_k, with respect to each predicted particle value. Using the SIR particle filter, the *importance density* is set equal to the *a priori* state PDF $p(x_k|x_{k-1})$. This leads to the simplification of weight update formula such that the non-normalised weight update formula is given by (see section 3.4.3.1 for details)

$$w_k^{(i)} = w_{k-1}^{(i)} p(z_k|x_k^{(i)}) \tag{6.9}$$

In this way, the weights for the newly generated particles $x_k^{(i)}$ are evaluated from the measurement likelihood function for new observations $p(z_k|x_k^{(i)})$, which is defined by equation (6.6). Since the only probabilistic component of the measurement likelihood function is a zero-mean Gaussian term, the measurement likelihood function is computed using a Gaussian kernel function, so that

$$p(z_k|x_k^{(i)}) = p(\upsilon_k) = \frac{1}{\sqrt{2\pi\sigma_\upsilon^2}} \exp\left\{\frac{-(z_k - x_k^{(i)})^2}{2\sigma_\upsilon^2}\right\} \tag{6.10}$$

Once the particle weights have been updated it is necessary to normalise the particle weights to ensure they sum to 1

$$w_k^{(i)} = w_k^{(i)} / \sum_{i=1}^{N_s} w_k^{(i)} \tag{6.11}$$

(iii) Particle Resampling

The objective of the resampling operation is to eliminate those particles with small weights and focus on those particles with larger weights. The level of particle degeneracy is determined by computing the effective particle sample size

$$\hat{N}_{eff} = \frac{N_s}{\sum_{i=1}^{N_s} (w_k^{(i)})^2} \tag{6.12}$$

A resampling operation is then performed whenever $\hat{N}_{eff} < \hat{N}_{thres}$. The resampling operation generates of a new set of particles $\{x_k^{(i*)}\}_{i=1}^{N_s}$ by sampling (with replacement) N_s times from the current discrete approximation of $p(x_k|z_{1:k})$, such that $Pr(x_k^{(i*)} = x_k^{(i)}) = w_k^{(i)}$. The new particle population thus represents a set of independent and identically distributed set of samples from the current discrete approximation of the degradation state PDF, and therefore the particle weights are reset so that $w_k^{(i)} = 1/N_s$.

(iv) State Estimate

The output at each iteration is a discrete weighted approximation to the true *posterior* degradation state $p(x_{0:k}|z_{1:k})$, which is given by

$$p(x_k|z_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)} \delta(x_{0:k} - x_{0:k}^{(i)}) \tag{6.13}$$

where δ is the dirac-delta function.

This process is then repeated at iteration to recursively estimate the *posterior* TPU degradation state.

6.3.3.2 (b) Long-Term Predictions

The second stage within the particle filtering framework for prognostics involves predicting the RUL of the system under observation. Firstly, long-term predictions of the TPU degradation state are generated, from which a RUL PDF for the system is then computed.

(v) Prediction Step (p Iterations)

To generate long-term predictions of the TPU degradation state, the set of particles and weights which define the current *posterior* TPU degradation state estimate, at the time the long-term predictions are generated t_k, are used as the initial conditions. Each particle is individually progagated into the future, by recursively applying the state-transition model, which is given by

$$x_k = x_{k-1} - \alpha_1 \text{exp}\left[\frac{-\alpha_2/t_k}{t_k^2}\right] - \alpha_3 \exp\left[\alpha_4\, t_k\right] + \alpha_5 t_k + \omega_k \tag{6.14}$$

Each particle is propagated until the value of each particle enters the defined hazard zone. The hazard zone defines a range of critical values of the degradation state at which the system under observation is deemed to have reached the end of its serviceable life, and there is a significant risk of equipment failure in continuing to operate the equipment beyond the range of values defined by the hazard zone. Typically the lower and upper hazard zone bounds (H_{lb} and H_{ub} respectively) are determined statistically from historical failure data, or inferred using expert knowledge gathered from domain experts, such as maintenance personnel.

(vi) Generate RUL PDF

Once the projected path for each particle has been generated, $\hat{x}^{(i)}_{k+p}$, and the time at which each particle enters the hazard zone has been identified, this information is then combined with the weight of each particle, $w^{(i)}_{k}$, at the time the predictions are generated, t_k, to generate a RUL PDF for the system. The RUL PDF can be computed by applying the law of total probabilities [27], whereby

$$p_{ttf}(k+p) = \sum_{i=1}^{N_s} Pr(Failure|X = \hat{x}^{(i)}_{k+p}, H_{lb}, H_{ub})\ .w^{(i)}_{k+p} \tag{6.15}$$

where $p_{ttf}(k+p)$ is the probability of equipment failure at time t_{k+p} This equation essentially states that the probability of time-to-failure, t_{ttf}, occurring at time t_{k+p} is equal to the sum of the weights of those projected particles which enter the hazard zone at time t_{k+p}. The overall system RUL PDF is then defined as sum of the individual failure probabilities at each future time instant.

6.4 Particle Filtering for TPU Prognostics

6.4.1 Application Example - Single Model Approach

In applying the particle filtering algorithm for prognostics to the prediction of TPU RUL, there are a number of parameters which must first be defined. The first parameter to define are the bounds of the hazard zone. In each of the examples which will be presented, the hazard zone for the TPU degradation state was defined as $H_{lb} = 580^\circ$ C and $H_{ub} = 600^\circ$ C, where H_{lb} and H_{ub} define the lower and upper bounds of the hazard zone. Once the value of the TPU degradation state decays to within this range of values, the TPU system is deemed to be beyond its useable life.

To illustrate the application of the particle filtering algorithm to TPU prognostics, two historical failure examples, which exhibited similar degradation profiles, in terms of rate of decay and time-to-failure, were selected. The TPU degradation process model parameters were tuned to fit one of these failure examples. The identified model was then used as the state-transition model in applying the particle filtering framework to the second TPU failure example. The two specific failure examples where chosen so that the model derived from the first failure example would be sufficiently descriptive of the observed behaviour of the second example, which is used to provide an illustrative example of the application of the particle filtering framework to TPU prognostics in this section.

Figure 6.7 illustrates the application of the particle filtering algorithm to a TPU failure example. A total of 50 particles were used to approximate the TPU degradation state in all of the application examples which will be shown. The number of particles represents a trade off between computational cost and performance and using 50 particle was identified as sufficient to approximate a RUL PDF. Figure 6.7 illustrates the output generated by the two stages within the particle filtering algorithm, (1) *state estimation* and (2) *long-term predictions.*

(1) State Estimation

At time t_D a fault condition is first detected. This initiates the particle filtering framework. The particle filtering algorithm iterates at 15 minute intervals as new fault indicator measurements are generated by the multi-mode CT tracking algorithm. The latest fault indicator measurement, z_k, is combined with the state-transition model predictions, to generate the *posterior* TPU degradation state estimate.

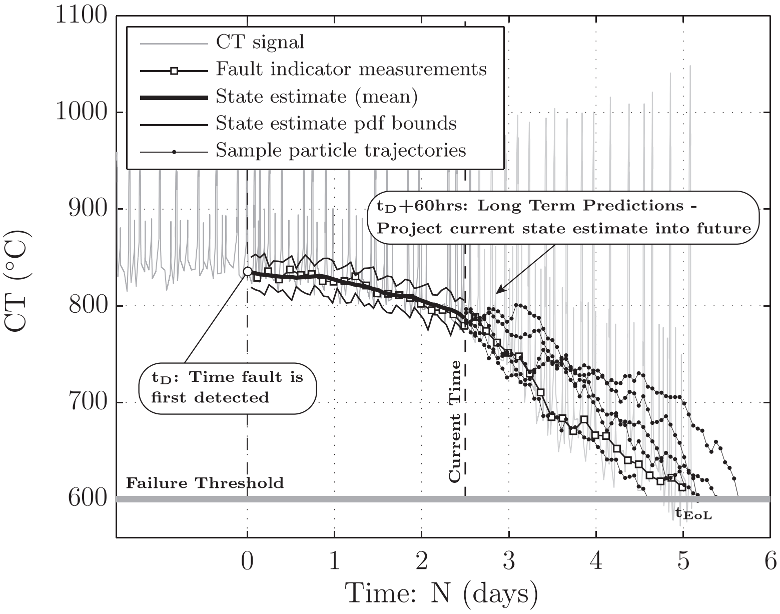


FIGURE 6.7: Particle filtering for TPU degradation state estimation & generation of long-term predictions

Figure 6.7 illustrates tracking of the TPU degradation state between times t_D and $t_D + 60$ hours. The mean state estimate is computed from the set of particles and weights at each iteration, and the state estimate bounds, which represent the level of uncertainty regarding the current state estimate, are defined by the highest and lowest valued particles generated at each iteration.

(2) Long-Term Predictions

At time $t_D + 60$ hours in Figure 6.7, the latest set of particles and weights S_k, which define the *posterior* degradation state PDF estimate, are generated. To generate long-term predictions each particle, $x_k^{(i)}$, in S_k (using its value as the initial condition), is projected recursively into the future, using the state-transition model, until the predicted value of each particle enters the hazard zone. Figure 6.7 shows the projected paths taken by 5 randomly selected particles in S_k, at time $t_D + 60$ hours. Once all of the particles have been projected into the future the final task is to generate the RUL PDF. The RUL PDF is generated according to equation (6.15). In the generation of the long-term predictions, the weights of each projected particle remains unchanged from their values at the time the predictions are generated.

Figure 6.8 illustrates the RUL PDF generated from long-term predictions made at two different times. The first RUL PDF was generated from long-term predictions made at time $t_D + 1.5$ days, and the second RUL PDF was generated from long-term predictions made at time $t_D + 4$ days. An important observation to note is how the RUL PDF becomes more precise as the prediction horizon decreases. It is also noted that the true TPU end-of-life, t_{EoL}, occurs within the bounds of the RUL PDF in both instances.

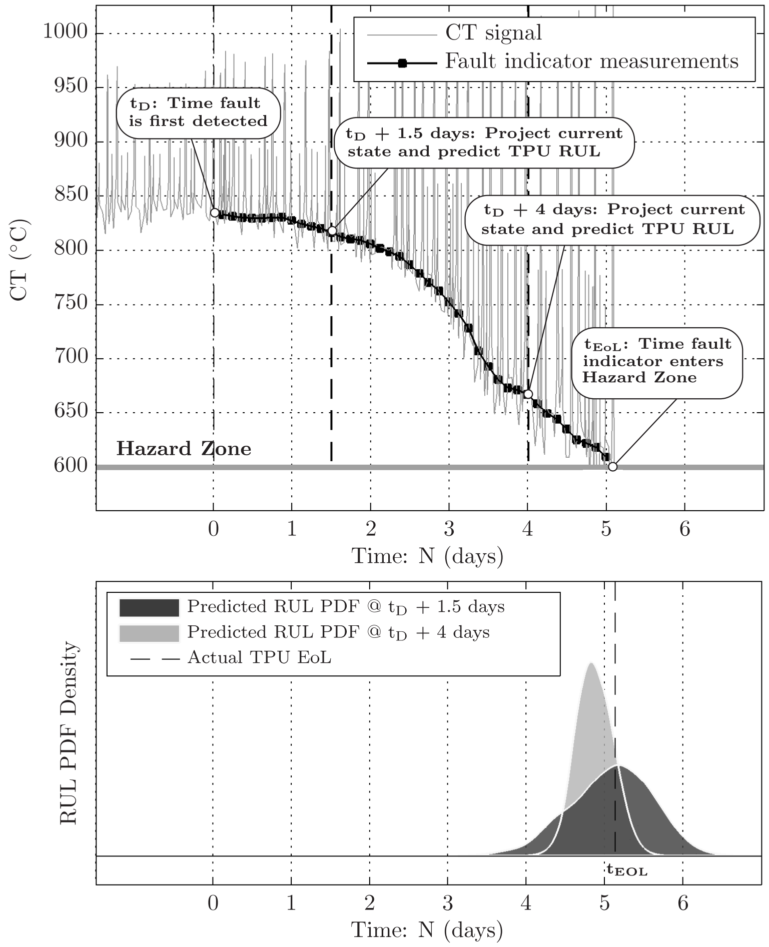


Figure 6.8: TPU RUL estimation using particle filtering

Once the RUL PDF has been generated, the issue of how to incorporate the RUL PDF within an automated maintenance decision support system must be considered. The key parameter for incorporation within an automated maintenance decision support system is the lead-time-interval (LTI). The LTI defines the remaining time before maintenance actions must be undertaken, to avoid operating the equipment beyond the maximum allowable probability of failure (PoF), as identified by the just-in-time-point (JITP) (See Section 2.5.1 for further details). Figure 6.9 illustrates the evolution of the LTI value for the example shown in Figure 6.8, between when the fault is first detected at time t_D, and the true TPU end-of-life (EoL), t_{EoL}. In the example shown a maximum allowable

PoF value of 5% was selected for defining the JITP.

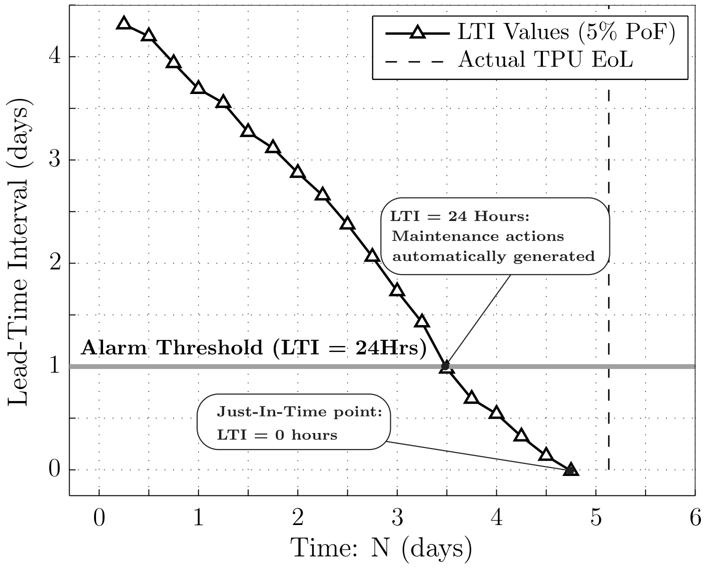


FIGURE 6.9: Evolution of lead-time interval estimates as TPU fault progresses

As time evolves and the TPU EoL approaches, the LTI can be used to automatically generate maintenance actions. For example, an alarm threshold of 24 hours is illustrated in Figure 6.9 at which point maintenance actions are generated. In this way, maintenance personnel are provided with a sufficient window within which corrective action can be taken, and equipment failure avoided and, additionally, the serviceable life of the equipment is maximised.

6.4.2 Uncertainty in TPU Prognostics

The primary challenge in the development of prognostic algorithms is addressing the level of uncertainty inherent in the generation of long-term predictions which describe the evolution of a fault indicator. Much of this uncertainty arises from the stochastic nature of the process, but additionally, uncertainty about how a system might be operated in the future, or the types of loads it might be operated under, significantly contributes to the overall level of prognostic uncertainty. In this section, the various sources of uncertainty which effect to ability to accurately predict TPU RUL are presented and existing and potential solutions to addressing these uncertainty challenges are discussed.

From a TPU prognostics perspective the issue of prognostic uncertainty is illustrated by the set of historical TPU failure examples. Figure 6.10 illustrates the fault indicator measurements derived from each of the historical TPU failure examples available. The immediate observation is the large variability in the time-to-failure t_{ttf} value for each failure example. The t_{ttf} represents the time between when a fault is first detected at time t_D and when a TPU reaches the end of its serviceable life t_{EoL}.

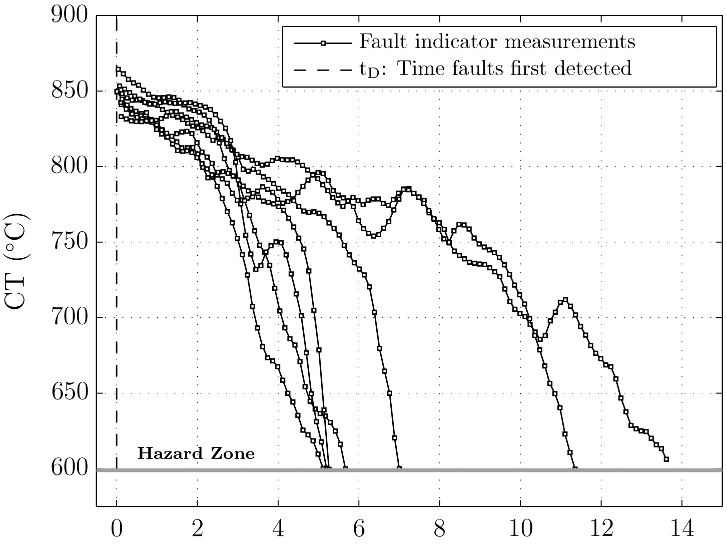


FIGURE 6.10: Fault indicator measurements extracted from 7 available historical TPU failure examples

The variability in the observed t_{ttf} values is largely due to differences in the process load imparted on each individual TPU during its serviceable life. The process load imparted on a TPU over any time period will be a function of the types, volumes, and mixtures of different process gases abated by the TPU. The process load is essentially a function of the number of wafers processed and the specific type of manufacturing process each of those wafers undergoes, since different manufacturing processes utilise different process chemistries. Additionally, the wafer throughput rate, and the cumulative effect of different sequences of high rate and low rate wafer throughput are likely to influence the rate at which deposit build-up occurs within the TPU combustion chamber.

Addressing these sources of uncertainty requires the development of an understanding of the relationship between the different types of process chemistries abated and the rate of deposit build-up. Additionally, knowledge of historical processing information and wafer throughput rates would be necessary to relate the effects of different wafer throughput rates to TPU RUL. These issues are common in the development of any prognostic algorithm that does not assume a fixed constant load in the future. However, in developing prognostic technologies for TPU systems, addressing these issues is made more challenging due to the industry in which TPU systems are operated; namely semiconductor manufacturing.

Semiconductor manufacturing is generally at the cutting edge of modern technology, requiring multi-billion dollar investments in facilities, equipment, and R&D. To protect these investments, and their intellectual property, semiconductor manufacturers maintain strict controls over any process details, data, or information, that might be considered proprietary or of a commercially sensitive nature. Considering that TPU systems (and many other types of support equipment used in semiconductor manufacturing) are often operated and maintained on-site by OEMs (who in many cases also develop and supply the condition monitoring and prognostic software), this presents additional challenges and further sources of uncertainty in the development of prognostic technologies for TPU systems. These same issues are also relevant in the development of prognostic technologies for vacuum pump systems used in semiconductor manufacturing.

In general, semiconductor manufacturers do not make available any information, either historically or in real-time, regarding the rate of upstream wafer processing. Even more tightly controlled is sufficiently specific information on the types, and volumes, of different process gases utilised, which might be abated by any individual TPU system. This lack of information sharing makes it more difficult to address and manage the different sources of uncertainty which arise from the variability in the input load an a TPU.

In the application of the particle filtering framework for prognostics there are two key, interrelated, sources of uncertainty to consider, model uncertainty and future load profile uncertainty. In the example application of the particle filtering framework for TPU prognostics, shown in Figure 6.8, the state-transition model used to describe the evolution of the TPU degradation process was sufficiently representative of the observed behaviour that accurate TPU prognostics was possible. However, as illustrated by the set of failure examples shown in Figure 6.10, it is difficult to capture the variability in the observed behaviour in a single model.

In Sections 6.4.2.1 and 6.4.2.2, the issues of model uncertainty and future load uncertainty in TPU prognostics are discussed with reference to relevant publications addressing these issues. Finally, in Section 6.4.2.3, a proposed solution to addressing the issues of model uncertainty and future load profile uncertainty in TPU prognostics is introduced, which attempts to address the uncertainty challenges presented by the semiconductor manufacturing environment.

6.4.2.1 Modelling Uncertainty

One of the primary sources of uncertainty, in the application of the particle filtering framework for TPU prognostics, is errors in the model used to describe the evolution of the degradation state. Ideally, such a model would be based upon the physics-of-failure but, in general, the development of such high-fidelity models is often difficult and costly. Instead, models are often developed which simply describe the observed behaviour of historical failure examples. In the case of TPU degradation, the model employed simply describes the evolution of the TPU degradation state, as time evolves. Ideally this model would incorporate information describing how factors such as the volumes, flow rates, and types of different process gases used, influence the rate of deposit buildup within a TPU combustion chamber. However, developing such models is difficult considering the strict controls maintained over such information. Additionally, the transferability of such models between different manufacturers and different processes might be difficult considering the variability in process chemistries between different manufacturers.

Within the particle filtering framework for prognostics the issue of model uncertainty has been considered by a number of authors [27, 32, 101, 105]. The most common approach identified in the literature, to addressing model uncertainty, is to consider the model parameters as time varying and to incorporate them as part of the state vector. In this way the particle filtering algorithm performs both model identification and state estimation [32]. Consider the model used to describe the evolution of the TPU degradation process

$$x_k = x_{k-1} - \alpha_1 \mathrm{exp}\left[\frac{-\alpha_2/t_k}{t_k^2}\right] - \alpha_3 \exp\left[\alpha_4\, t_k\right] + \alpha_5 t_k + \omega_k \tag{6.16}$$

The most common technique for incorporating the model parameters as part of the state vector is to consider the state transition model which describes the evolution of each of the model parameters as a Gaussian random walk, so that

$$\alpha_{i,k} = \alpha_{i,k-1} + \phi_{i,k-1} \tag{6.17}$$

where $\phi_{i,k-1}$ is a sample drawn from a zero-mean Gaussian distribution with variance σ_{ϕ}^2. Provided a suitable initial value for $\alpha_{i,0}$ and σ_{ϕ}^2 is chosen, the principle of the model identification process is that the α_i estimates will converge on the actual parameter value by the *law of large numbers* [105].

A second technique for updating model parameters presented by Orchard [27] utilises an *outer correction loop* in which short-term prediction errors are used to modify the variance of ϕ_i in (6.17) with respect to the margin of the prediction error. The principle of the *outer correction loop* is to allow the value of the unknown model parameters to be updated rapidly if the prediction error becomes large which might indicate that the model no longer describes the dynamics of the observed behaviour. In this way the model parameter values can respond rapidly to a significant change in the operating conditions of the system (by increasing the variance of ϕ_i), and stabilise the model parameter values when the model behaviour matches the predicted behaviour more closely (by reducing the variance of ϕ_i).

Incorporating the model parameters as part of the state vector within the particle filtering framework means that both the *posterior* degradation state, and the model parameters, are updated recursively at each iteration. However, once the particle filtering framework switches from the state estimation to the generation of long-term predictions, the model parameters generally remain fixed as the degradation state is iteratively projected into the future. This raises the question of how changes in the model parameter values effect the projected evolution of the degradation state.

The incorporation of the model parameters as part of the state vector was investigated for TPU prognostics. Two main issues arose with this approach to addressing model uncertainty. The first issue was the difficulty in selecting initial values for $\alpha_{i,0}$ and $\sigma_{i,0}^2$, for each model parameter. The second issue which arose related to the generation of long-term predictions. At each iteration, when the model parameters were updated, the new set of parameters remain fixed for the purpose of generating long-term predictions. This often resulted in significant differences in the RUL projections generated at successive iterations. Furthermore, given the uncertainty in the future load profile, the accuracy of the long-term predictions generated by the updated model parameters at each iteration often failed to match the observed behaviour, as the prediction horizon increased.

6.4.2.2 Future Load Uncertainty

Another major source of uncertainty in TPU prognostics arises from the variability in future load profiles on TPU systems. As discussed previously, the future load profile of a TPU system depends upon the volume and type of process gases that a TPU system will abate over a future horizon. This is essentially a function of the number of wafers processed and the type of manufacturing processes those wafers undergo. Both of these factors can vary significantly in any upstream processing chamber which might be connected to an individual TPU system.

The variability in future load profiles on TPU systems arises for a variety of reasons. One of the primary reasons is that wafer processing tools are regularly taken down for preventative, or corrective, maintenance, as necessary. Additionally, the wafer throughput rate can be affected by a range of short and longer term influences. In the short-term, production is organised in real-time to match customer demand for different products, and to respond to the availability, or otherwise, of different equipment. In the longer term, conditions such as the worldwide economic climate can significantly influence demand and, hence, the rate and volume of wafers processed. The issue of future load profile uncertainty is also further complicated by the fact that a TPU can be connected to up to four individual processing chambers, each of which operate independently.

Ideally, in the development of TPU prognostics, both historical and real-time data on wafer processing rates and processing tool status would be available to incorporate in estimating TPU RUL. The historical data would permit an analysis of the relationship between wafer processing rates and deposit buildup on individual TPUs. Any identified relationship could then be incorporated in modelling the evolution of the degradation state so that, ideally, a state transition model of the form described in (6.18) could be derived

$$x_k = f_k(x_{k-1}, u_{k-1}, \omega_k) \tag{6.18}$$

where x_k is the current degradation state, ω_k is the process noise, and u_{k-1} represents the input/load on the TPU. With access to historical wafer processing information, including processing tool status information, statistical load profiles for various possible future wafer load scenarios could be generated from historical analysis of wafer processing rates, and the model could then be simulated using the different potential future load-profiles, improving the overall prognostic performance. Currently, both real-time and historical wafer processing information is generally not made available as such information is often considered to be commercially sensitive. However, there is presently a move toward greater data sharing between processing tools and support equipment. This is driven by a move to implement more energy efficient practices, whereby processing tools and support equipment switch into semi-standby mode during periods when no wafers are being processed. The objective is to reduce the wasteful consumption of utilities, fuel and gases during these periods.

To automate the process of having support equipment reduce consumption during periods of no processing requires a greater level of communication between processing tools and support equipment. In the future this may provide a method for accessing real-time and historical processing data, which could be incorporated into the development of TPU prognostics. The issue of reducing future load uncertainty by incorporating historical load profiles into RUL predictions is considered an area in which to focus ongoing research activities into TPU prognostics. Such research, based upon wafer load profile characterisation, could also be incorporated into the development of prognostic algorithms for vacuum pumps. Additionally, with the development of pump mode tracking algorithms, as described previously in Section 5.5.2, this may provide another method for developing statistical characterisations of typical TPU load profiles online, which could potentially be incorporated into the development of TPU prognostics.

6.4.2.3 Addressing Uncertainty in TPU Prognostics

Sections 6.4.2.1 and 6.4.2.2 have illustrated how issues of uncertainty, arising from modelling errors and future load profile uncertainty, are compounded by the semiconductor manufacturing environment in which TPU systems are operated. This environment, in which most process details are considered proprietary or commercially sensitive, present additional challenges from an uncertainty management and representation perspective. The challenge becomes how to capture and propagate this uncertainty in the generation of long-term predictions of TPU RUL, in the absence of many process details which would normally be available in the development of prognostic solutions.

Without improved process models, or information on future load profiles, the primary source of information on how the variability in TPU operating characteristics influences the evolution of the TPU degradation state is captured in the set of available historical

TPU failure examples, illustrated in Figure 6.10. The set of TPU failure examples represents samples from the set of all possible failure paths, considering the variability in the process gas load and operating profiles. By tuning the TPU degradation process model to each of these failure examples, this generates a set of 7 degradation process models. These seven models which describe the path taken by the TPU degradation process represent samples from the set of all possible paths which the TPU degradation state might follow.

In Section 6.5, a solution to addressing the issues of uncertainty representation and management in TPU prognostics is presented. The solution is based upon the use of a multiple model particle filtering approach. The basic principle of the multiple model approach is to maintain a set of candidate models, where each candidate model describes how the TPU degradation state might evolve. As a real TPU fault evolves and new fault indicator measurement data is received, the basic idea is to use the sequence of new measurement data received to recursively estimate the plausibility that each model is representative of the observed behaviour. Each model is then weighted accordingly, and the overall RUL PDF is a weighted sum of the individual RUL PDFs generated by each model.

6.5 Multiple Model Particle Filtering for Prognostics

It was shown previously in Section 6.4.2 how various sources of uncertainty effect to ability to accurately predict TPU RUL. One option to address future load profile uncertainty is to increase the magnitude of the process noise term, used in the generation of long-term predictions. In this way, the predictions can account for the uncertainty in future load profile. However, increasing the process noise settings also increases the prognostic uncertainty, which is of particular concern when the uncertainty is propagated over a long prediction horizon [101]. In this section, a potential solution to addressing the uncertainty challenges in TPU prognostics is presented.

To address the issue of model uncertainty, Tang et. al. [106] introduced the idea of the multiple model approach for model-based prognostics. The original premise for the multiple model approach is the situation in which multiple failure modes may exist for a system which might compete for dominance as a failure evolves. The conceptual situation presented by Tang, in which such a scenario might occur, is the prediction of metal corrosion in rolling element bearings [106]. The evolution of a metal corrosion process can potentially take the form of uniform corrosion, pitting, crevice or galvanic corrosion. During the initial stages of corrosion, it can be difficult to identify which type of corrosion which is occurring from the measurement data taken from the process. The principle of the multiple modelling approach is to utilise a set of candidate models, in which each model describes how the corrosion process might evolve for each of the potential failure modes. As time evolves and new measurement data is received, the idea is to use the sequence of new measurement data received to recursively update the plausibility that each model is representative of the observed behaviour. In addition to the application of multiple modelling approach for situations in which multiple failure modes exist, Tang also makes the observation that failure data collected in the field is usually insufficient to faithfully construct a model online, but is often sufficient to be used to select candidate approximating models. Furthermore, the output of the multiple models can be fused to improve the overall prediction accuracy [106].

In this study, the set of candidate approximating models are derived from the set of available historical failure examples. The idea is to generate a set of models in which each model represents a description of how the TPU degradation state might evolve, given given the uncertainty in the types of gases abated and the load profile imparted on the TPU. The predictions generated by each of the models can then be fused to improve the overall prediction accuracy. Furthermore, by using a set of candidate models which represent approximations to the potential behaviour of the system, the process noise term for each model can be made much smaller, thus improving the prediction accuracy over a longer prediction horizon. Section 6.5.1 introduces the principle of the multiple model particle filtering approach, and illustrates the application of the technique to addressing uncertainty in TPU prognostics.

6.5.1 Multiple Model Particle Filtering Algorithm

The principle of the multiple model particle filtering approach for prognostics is to maintain a bank of particle filters which each operate in parallel. The state transition model in each of the filters describes a possible failure mode of the system, or in the current application, represents a *candidate model* describing how the TPU degradation state might evolve, given the uncertainty in the future load profile.

Figure 6.11 presents a flow diagram illustrating the different stages in the multiple model particle filtering algorithm. Figure 6.11 presents the conceptual case of two candidate models for ease of visualisation, though the algorithm can easily be scaled to the required number of candidate models without any change in functionality.

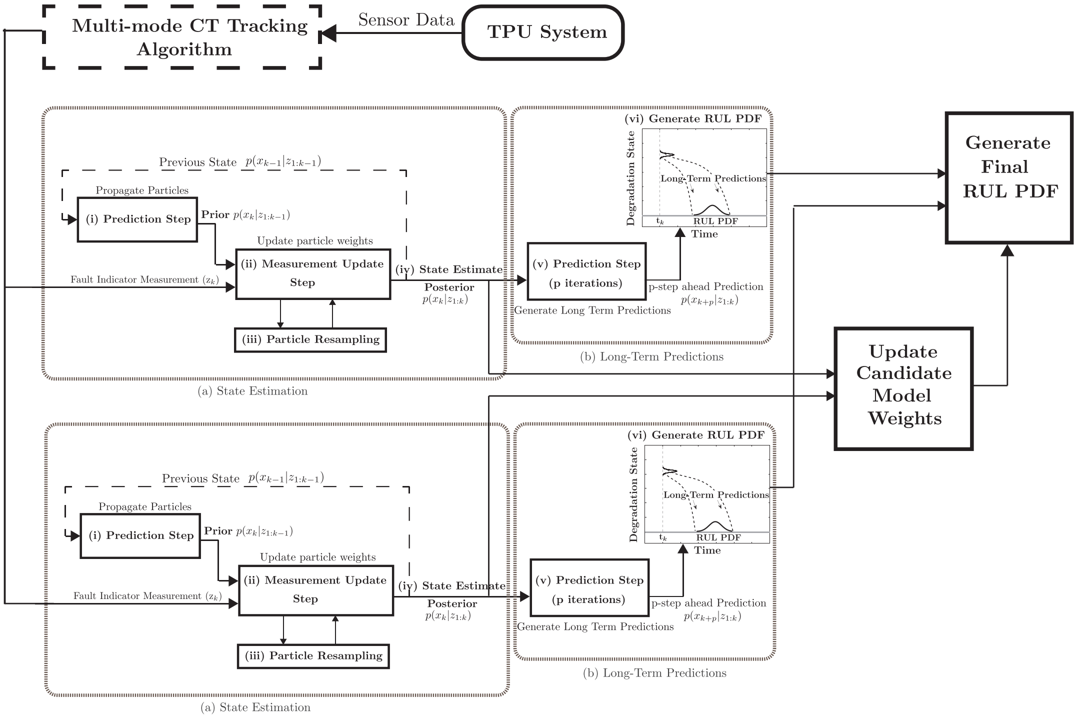

Multi-mode CT Tracking Algorithm
Sensor Data
TPU System
Previous State $p(x_{k-1}|z_{1:k-1})$
Propagate Particles
(i) Prediction Step
Prior $p(x_k|z_{1:k-1})$
Update particle weights
(ii) Measurement Update Step
Fault Indicator Measurement (z_k)
(iii) Particle Resampling
(iv) State Estimate
Posterior $p(x_k|z_{1:k})$
(a) State Estimation
(v) Prediction Step (p iterations)
p-step ahead Prediction $p(x_{k+p}|z_{1:k})$
Generate Long Term Predictions
(vi) Generate RUL PDF
Degradation State
Long-Term Predictions
t_k
RUL PDF
Time
(b) Long-Term Predictions
Generate Final RUL PDF
Update Candidate Model Weights

Figure 6.11: The multiple model particle filtering algorithm for TPU prognostics

In the multiple model particle filtering algorithm, each of the filters operates independently, using the sequence of fault indicator measurements to estimate the posterior TPU degradation state at each iteration. The TPU degradation state, which is defined in terms of a set of particles and weights, is then propagated into the future to generate long-term predictions of system health, from which a RUL PDF is then generated. Initially, when a fault is first detected, each of the particle filters is initialised and begins the process of estimating the posterior TPU degradation state recursively, as each new fault indicator measurement is generated. This process, of estimating the current degradation state and projecting into the future, is the same for each candidate model and is implemented exactly as the particle filtering algorithm, described in Section 6.3.3, is implemented.

Beyond implementing the particle filtering algorithm for each candidate model, there are two new stages involved in implementing the multiple model particle filtering algorithm, updating the candidate model weights and generating the final RUL PDF, as illustrated in Figure 6.11. The implementation details of these two stages are described below.

6.5.1.1 Updating Candidate Model Weights

Each filter within the multiple model particle filtering framework uses a candidate state-transition model $m^{(j)}$ from the set of candidate models $M = \{m^{(1)}, ..., m^{(N_M)}\}$, where N_M is the number of candidate models. Initially, when a fault is first detected at time t_0, the weighting, or plausibility, of each candidate model $m^{(j)}$ describing the (as yet unobserved) behaviour is considered equal, so that

$$\mu_0^{(j)} = \frac{1}{N_M} \tag{6.19}$$

where $\mu_0^{(j)}$ represents the *a priori* plausibility that candidate model $m^{(j)}$ describes the as yet unobserved behaviour at time t_0. To satisfy the requirement that the $\mu^{(j)}$ values represent probabilities, they are subject to a number of constraints, such that

$$\sum_{j=1}^{N_M} \mu^{(j)} = 1 \quad \text{and} \quad 0 \leq \mu^{(j)} \leq 1 \tag{6.20}$$

As the fault evolves and new measurement data is received, the basic idea of the multiple modelling approach is to use the sequence of new measurement data received to update the relative plausibility that each of the candidate models is descriptive of the observed behaviour. At time t_k, the plausibility of each candidate model $m^{(j)}$ in describing the observed behaviour is given by

$$\mu_k^{(j)} = p(m^{(j)}|z_{1:k}) \tag{6.21}$$

which describes the probability that candidate model $m^{(j)}$ is in force, conditioned of the sequence of fault indicator measurements $z_{1:k}$ received up to time t_k. As a fault evolves, the plausibility of each candidate models in describing the evolution of the TPU degradation state can be computed recursively using Bayes' rule [101], whereby

$$\mu_k^{(j)} = \frac{p(z_k \,|\, m^{(j)}, z_{1:k-1})\, \mu_{k-1}^{(j)}}{\sum_{j=1}^{N_M} p(z_k \,|\, m^{(j)}, z_{1:k-1})\, \mu_{k-1}^{(j)}} \tag{6.22}$$

where $p(z_k \,|\, m^{(j)}, z_{1:k-1})$ is the model likelihood function which describes the likelihood of observing the latest fault indicator measurement z_k, conditioned on model $m^{(j)}$ and the history of fault indicator measurements $z_{1:k-1}$ received up to time t_{k-1}. In a situation where Kalman filters are being utilised, the function $p(z_k \,|\, m^{(j)}, z_{1:k-1})$ can be computed analytically. However, in the situation where particle filters are utilised, the function $p(z_k \,|\, m^{(j)}, z_{1:k-1})$ must be computed numerically from the particles [101]. One method to approximate $p(z_k \,|\, m^{(j)}, z_{1:k-1})$ is to recursively update the model likelihood function as follows

$$p(z_k \,|\, m^{(j)}, z_{1:k-1}) = p(z_k | x_k^{(m^{(j)})}) \prod_{i=1}^{k-1} p(z_i | x_i^{(m^{(j)})}) \tag{6.23}$$

where $p(z_i | x_i^{(m^{(j)})})$ is the measurement likelihood function, which describes the probability of observing the measurement z_i at time t_i, conditioned on $x_i^{(m^{(j)})}$,which represents the *posterior* state PDF estimate generated by the filter using candidate model $m^{(j)}$. Since particle filters are used to estimate the posterior state PDF at each iteration, the function $p(z_i | x_i^{(m^{(j)})})$ cannot be computed analytically and instead must must be computed numerically from the particles. To approximate this function, the mean posterior state estimate $\bar{x}_k^{(m^{(j)})}$ is computed from the set of particles and weights which define the *posterior* TPU degradation state estimate at each iteration. The measurement likelihood function for this purpose is defined in the same fashion as the particle measurement update function, so that

$$z_i = \bar{x}_i + \psi_i \quad \leftrightarrow \quad p(z_i | x_i^{(m^{(j)})}) \tag{6.24}$$

where ψ_i is a zero-mean Gaussian measurement noise process with variance σ_ψ^2. The value of the variance term σ_ψ^2 can be tuned to influence how the degree of error, between the predicted state PDF and the latest measurement, is penalised in computing the model likelihood function as the fault evolves.

6.5.1.2 Update Final RUL PDF

The Bayesian model weighting approach, described above, provides a method for inferring the plausibility that each candidate model is representative of the observed behaviour. As more fault indicator measurements become available, this information can be used to update the plausibility that each model is representative of the observed behaviour. In this way, those models which better describe the observed behaviour gain greater weighting. This weighting is then incorporated into the generation of RUL PDFs for each model. The final RUL PDF is then defined as a weighted combination of each of the individual model RUL PDFs

$$p_{ttf}(k+p) = \sum_{j=1}^{N_M} \mu_k^{(j)} \sum_{i=1}^{N_s} Pr\big(Failure|X = \hat{x}_{k+p}^{(i)}, H_{lb}, H_{ub}, m^{(j)}\big)\, w_k^{(i)} \tag{6.25}$$

where $p_{ttf}(k+p)$ is the probability of time-to-failure occurring at at time t_{k+p} where t_k is the current time. H_{lb} and H_{ub} represent the bounds on the hazard zone, $\hat{x}_{k+p}^{(i)}$ and $w_k^{(i)}$ represent the predicted particle values and their initial weights respectively, which are generated by the filter using candidate model $m^{(j)}$. The $\mu_k^{(j)}$ values represent the individual candidate model weights at the time the predictions are generated, t_k, which are recursively update as new measurement data becomes available.

6.5.2 Candidate Model Generation

Before applying the multiple model particle filtering framework, it is first necessary to generate a set of candidate models from the set of historical TPU failure examples available. To derive a set of candidate models, the TPU degradation process model was first tuned to fit each of the available TPU failure cases. Using the set of model parameter values identified for each failure cases, the maximum and minimum value for each α_i parameter was identified. To generate a potential candidate model, a value for each α_i parameter was chosen by sampling uniformly from the identified range of values for each α_i parameter. This process of generating a potential candidate model was repeated several thousand times to generate a large set of potential candidate models.

The second step then initialises each potential candidate model with the same initial state value x_0. An initial state value of 850° was selected as this was the average degradation state value at which a fault was first detected in each of the historical TPU failure examples. The initial state value was propagated into the future in a recursive manner, using each of the potential candidate models. At this stage, it was noted that the combination of certain α_i parameter values resulted in the generation of many improbable degradation paths, from an intuitive perspective, being generated.

The next step was to select a set of candidate approximating models to be used within the multiple model particle filtering framework for TPU prognostics. Figure 6.12 again illustrates the set of historical TPU failure examples. As can be seen, the historical failures have occurred between 5 and 14 days after the time the fault was first detected, t_D. In selecting a set of candidate approximating models, all those potential candidate models which crossed the hazard zone anytime between 4 and 15 days were first selected. This was deemed to be sufficient, at this stage, to capture the potential behaviour of future failure cases. In reality, as more failure data becomes available, the set of candidate models can be continuously refined to reflect future observed behaviour.

A final set of 500 candidate approximating models were chosen by sampling uniformly within the range of candidate models which crossed the failure threshold within 4 to 15 days. Figure 6.13 illustrates the evolution of the TPU degradation state, without incorporating any process noise, as described by a sample of 50 of the selected candidate approximating models. As can be seen, the set of of candidate models represent approximations to potential degradation paths taken by future failure examples. Whilst the set of available historical failure examples represent samples from the range of possible paths taken by the TPU degradation state, the set of candidate approximating models

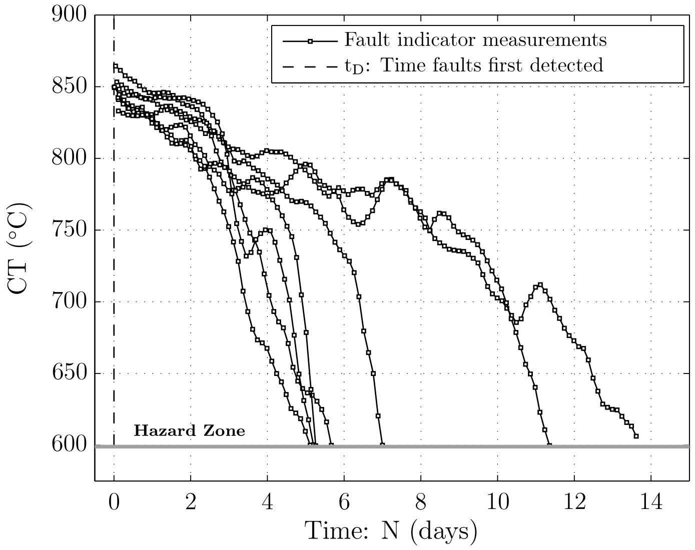


FIGURE 6.12: Fault indicator measurements extracted from 7 available historical TPU failure examples

are designed to describe the full spectrum of potential degradation paths, as can be best inferred from the set of historical failure examples.

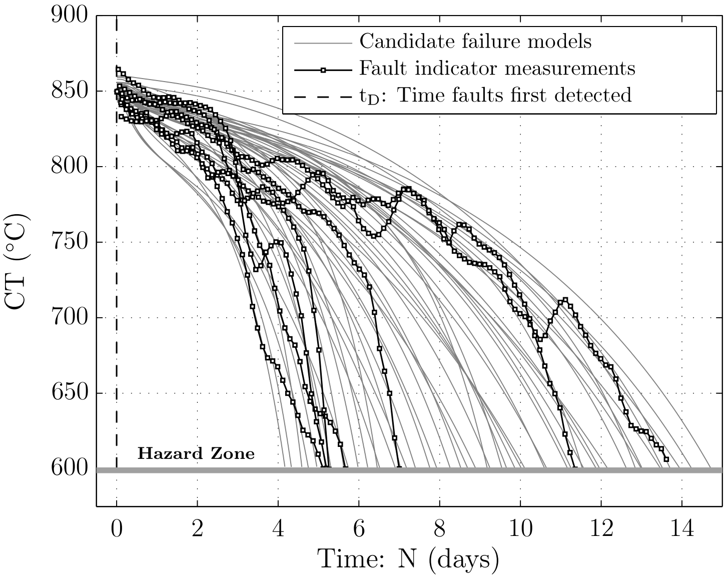


FIGURE 6.13: Illustration of sample candidate models generated from set of historical TPU failure examples

Using the approach described above, a total of 500 candidate approximating models were initially generated, for the purpose investigating how many candidate models should be used within the multiple model particle filtering framework for the current application. Selecting the number of candidate models to use essentially represents a trade-off between prognostic performance and computational cost. As the number of candidate models is increased, this means that the value of the process noise term used in each model for the generation of long-term predictions can be reduced, thus improving the precision of long-term predictions. Furthermore, a greater number of models allows for a greater coverage of potential failure paths which might occur. However, each extra candidate model increases the computational cost of the approach, associated with performing both state estimation and the generation of long-term predictions for each candidate model.

In the current application, the multiple model particle filtering approach was tested across the set of historical failure examples, using combinations of different numbers of candidate models and process noise values. The trade-off between performance and computational cost was investigated for various combinations. The objective was to identify a satisfactory combination of both the number of candidate models, and process noise values, which generate both accurate and precise RUL estimates, and ensure a smooth evolution of the RUL PDF as the fault evolves. In each of the subsequent examples shown, a subset of 150 candidate models was chosen, which was identified as sufficient to provide good coverage of the state-space, whilst maintaining a small process noise value in each of the candidate models.

6.5.3 Application Example: Multiple Model Particle Filtering

Figure 6.14 shows an example of the multiple model particle filtering framework applied to predicting the RUL of a TPU failure example. Figure 6.14 illustrates how the the RUL PDF evolves in response to the observed behaviour of the fault signal. In this example, the fault signal evolves at a relatively slow rate of decay, which implies a reduced likelihood of equipment failure over a shorter prediction horizon. This process is reflected in the evolution of the RUL PDF in Figure 6.14, whereby the RUL PDF density remains concentrated over a longer prediction horizon as the fault evolves, gradually becoming more accurate and precise as the equipment approaches the end of its serviceable life, t_{EoL}.

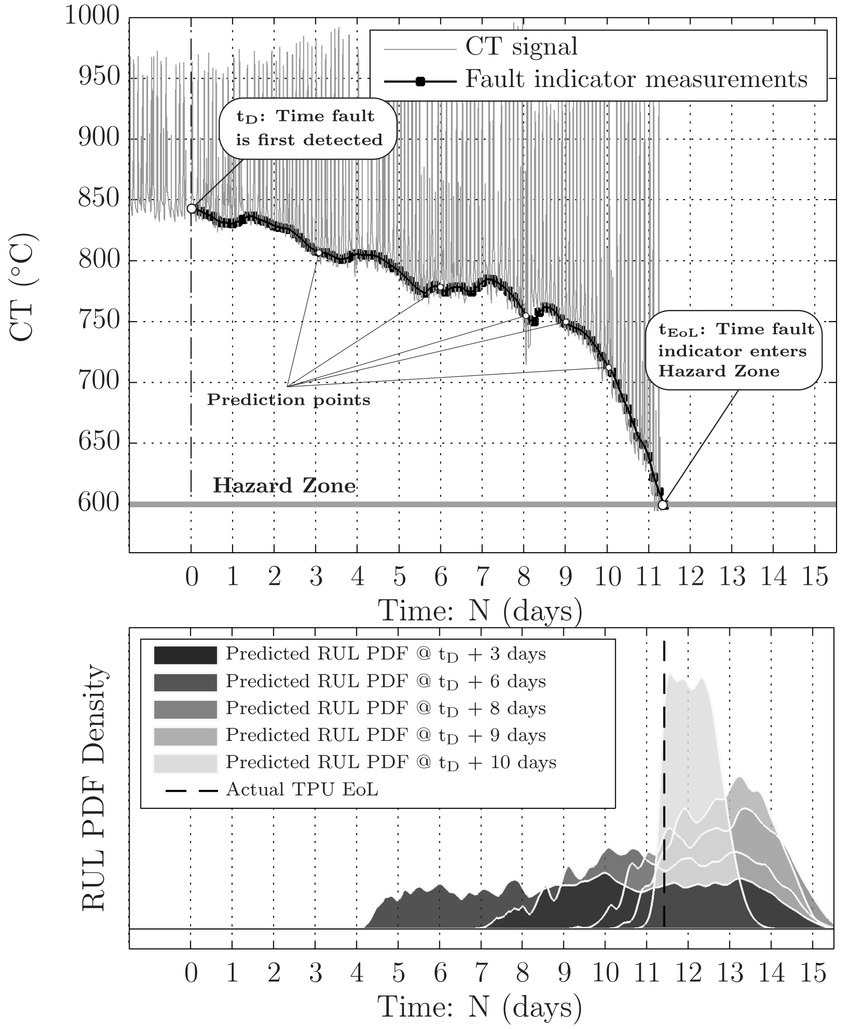


FIGURE 6.14: Illustration of evolving RUL PDF estimate, generated by multiple model particle filtering approach for TPU prognostics, in response to observed behaviour of TPU fault

The key output, generated from the evolving RUL PDF, which can used to automatically trigger maintenance actions is the lead-time interval (LTI), which provides maintenance staff with an estimate of the remaining time before corrective maintenance action must be taken to avoid operating equipment beyond the maximum allowable PoF. Figure 6.15 illustrates the evolution of the LTI value as computed from the evolving RUL PDF, generated by the multiple model particle filtering framework, in Figure 6.14.

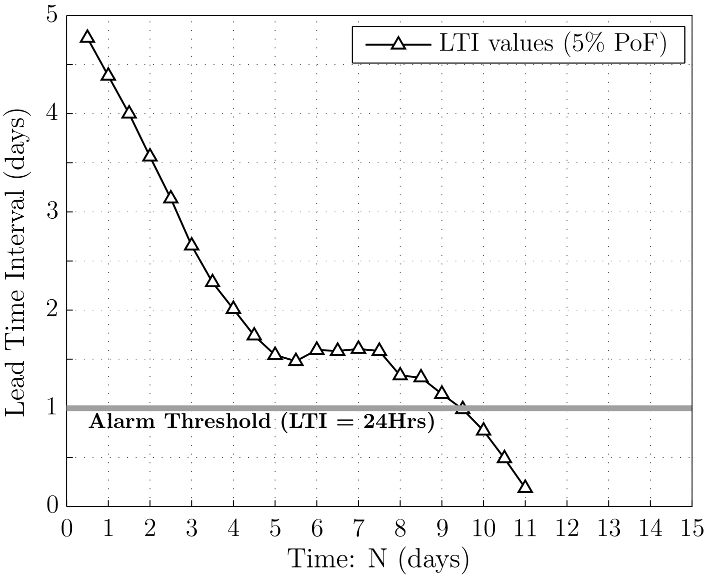


FIGURE 6.15: Evolution of LTI estimate as TPU fault evolves

Figure 6.15 illustrates how the LTI value estimate is updated as the fault evolves, providing maintenance staff with significant visibility on an upcoming maintenance issue. As the fault approaches, the passing of the estimated LTI values through a specified threshold can be used to automatically generation maintenance actions which must be performed to avoid risk of equipment failure. The alarm level can be chosen so to provide sufficient warning so that maintenance actions can be planned at a suitable time, to minimise overall manufacturing downtime. In applying the multiple model particle filtering framework, the evolution of the RUL PDF and the LTI estimate is essentially a function of the weights of the individual candidate model weights changing, as those filters using candidate models which better describe the observed fault behaviour have their weights increased, and vice versa. To demonstrate this behaviour, Figure 6.16 illustrates the evolution in time of each of the candidate model weights from the example in Figure 6.14.

Once a fault is first detected at time t_D, the candidate model weights $\mu^{(j)}$ are initialised

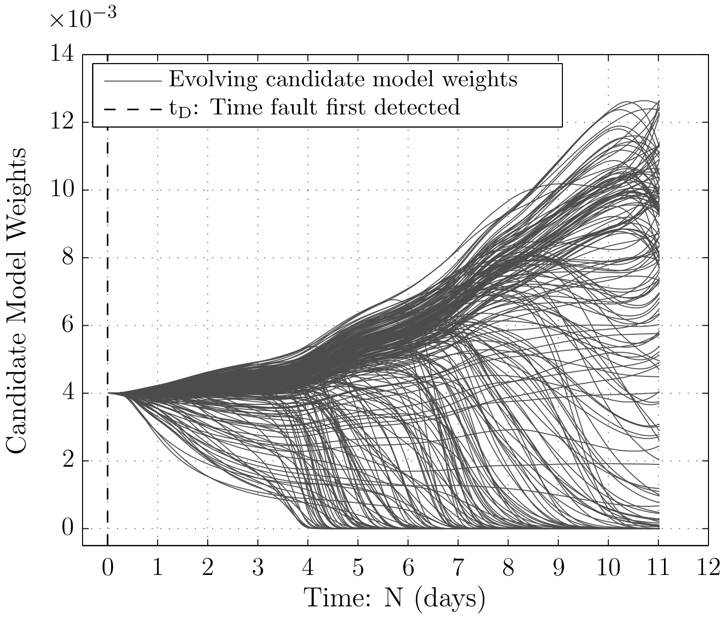


Figure 6.16: Evolving candidate model weights in response to observed behaviour of TPU fault example

with equal weighting. As the fault evolves, the plausibility of each candidate model describing the observed behaviour is updated recursively, and the candidate model weights change to reflect this. As can be seen in Figure 6.16, immediately after a fault is first detected, the model weights begin to respond as more fault indicator measurements are received. As the fault continues to evolve, an increasing number of the model weights decay to zero, indicating that the dynamics of these models do not reflect the dynamics of the observed behaviour.

6.5.4 Adaptive Candidate Model Weighting

As a TPU fault evolves, the model likelihood function is computed over the set of fault indicator measurements to update the plausibility that each candidate is descriptive of the observed behaviour. During testing of the multiple model particle filtering framework, across the set of historical TPU failure examples, it was noted that the evolution of the candidate model weights was heavily influenced by the value of the model likelihood function computed during the initial stages of TPU degradation. In computing the model likelihood function, equal weighting is given to each of the fault indicator measurements. However, this presents some issues in updating the candidate model weights.

During the initial stages of TPU degradation, the dynamics of each of the candidate models is similar, which reflects the typical dynamics of a TPU failure example, which is characterised by an initial period of slow decay of the TPU degradation state. This often results in larger measurement likelihood values for each candidate model, in response to the fault indicator measurements generated during the early stages of a fault. As a fault evolves, the model likelihood function value for each candidate model is heavily weighted by the measurement likelihood values, computed during the initial stages of a TPU fault. As a result, the evolution of the candidate model weights is often overly influenced by how well each candidate model describes the observed behaviour during the early stages of fault evolution. This can then result in some candidate model weights taking longer to decay to zero, as their dynamics deviate from the observed behaviour.

To address this issue, it is proposed to compute the model likelihood function with greater weighting applied to more recent fault indicator measurements, as opposed to applying equal weighting across all fault indicator measurements. In this way, the plausibility of each candidate model in describing the observed dynamics will be computed with greater weighting applied to how well each model describes the more recent behaviour, as opposed to applying equal weighting across the full history of fault indicator measurements. To achieve this, it is proposed to use exponential weighting in computing the model likelihood function. By incorporating exponential weighting, the model likelihood function can be computed as

$$p(z_{1:k}|m^{(j)}) = \prod_{i=1}^{k} \lambda^{i-k}\, p(z_i|x_i^{(m^{(j)})}) \tag{6.26}$$

where $\lambda \in [0, 1]$ can be tuned to govern the rate of exponential decay applied to the set of historical measurement likelihood values. In Section 6.5.4.1, an example of the multiple model particle filtering framework, using an exponential weighted likelihood function, is illustrated. To investigate if the suggested approach results in an improvement in prognostic performance, the results generated using both approaches to updating the model weights, were compared using available prognostic metrics. The results are presented in the following section.

6.5.4.1 Application Example: Exponential Model Weighting

Figure 6.17 demonstrates the application of the multiple model particle filtering framework to another historical TPU failure example. In this example, exponential weighting was incorporated within the model likelihood function. In Figure 6.17, the RUL PDF evolves in response to observed behaviour of the system, with the predictions becoming more accurate and precise as the equipment approaches the end of its serviceable life.

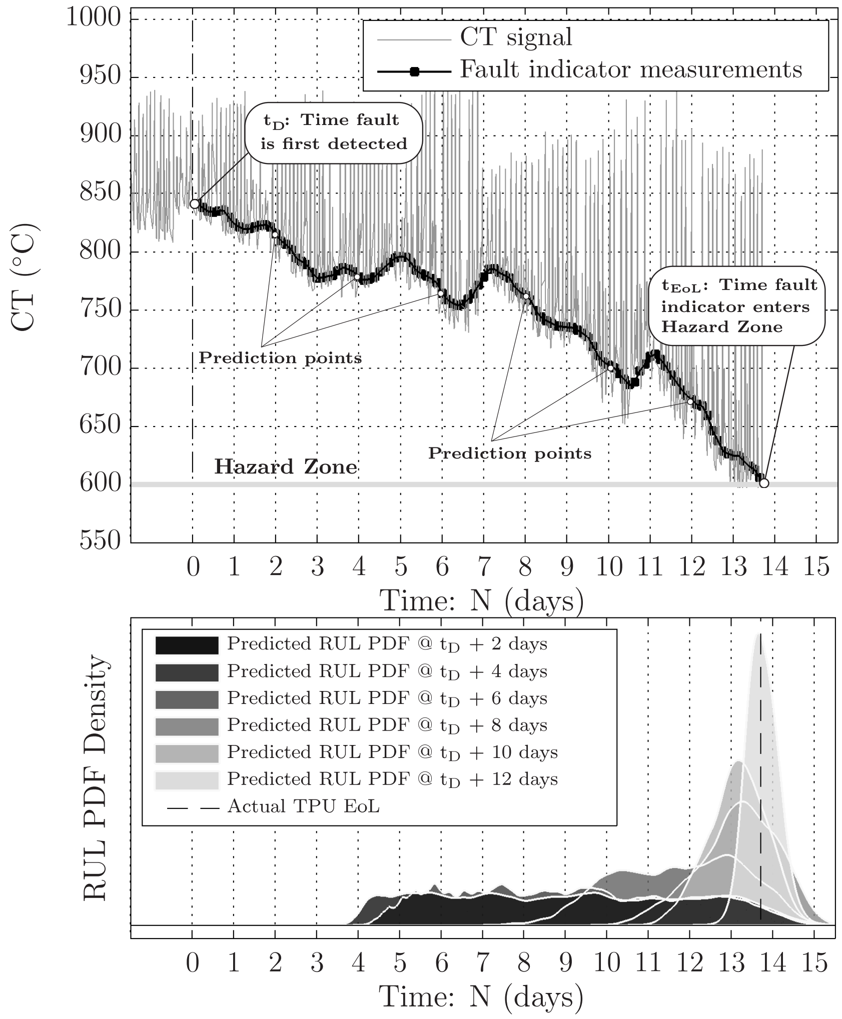


Figure 6.17: Illustration of evolving RUL PDF estimate, generated by multiple model particle filtering approach for TPU prognostics using exponential weighting within model likelihood function, in response to observed behaviour of TPU fault

The overall objective of the exponential weighting approach is to identify those models which better reflect the observed dynamics in the more recent past, and to discount those models which may have described the failure dynamics during the early stages of the fault, but diverge from the observed behaviour as the fault continues to evolve. To investigate if the exponential weighting scheme improves the performance of the RUL predictions generated, the results, using both uniform and exponential weighting within the model likelihood function were investigated, using available prognostic metrics.

Figure 6.18 compares the evolution of the candidate model weights using both the uniform weighting and the exponentially weighted methods. The evolution of the weights is presented in histogram format to compare the general trends in the evolution of the model weights. As can be seen, at time $t_D + 1$ day, the distribution of the candidate model weights are similar using both approaches, with each candidate model having the same approximate weight value.

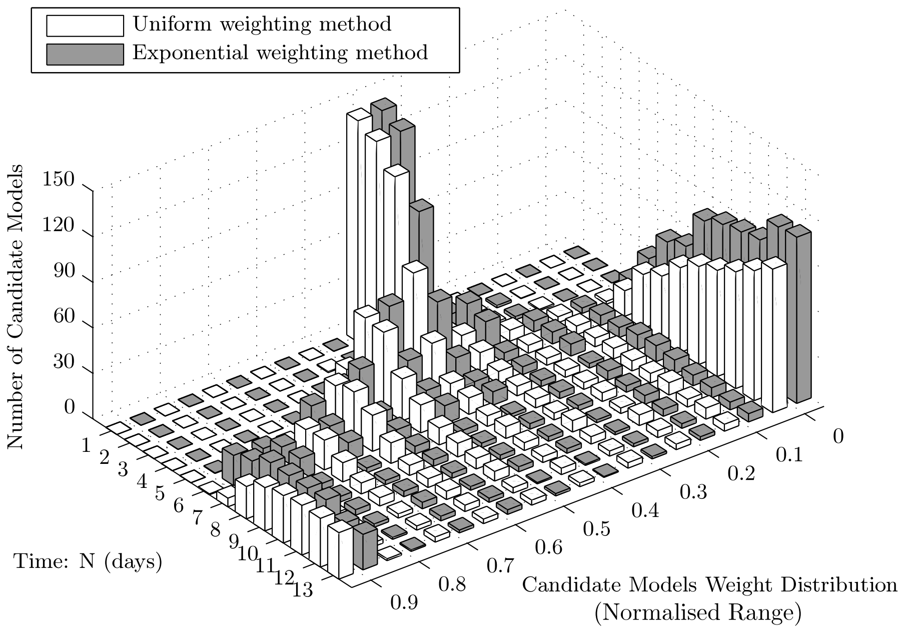


Figure 6.18: Comparison of candidate model weight distribution, using both uniform weighting and exponential weighting methods, as TPU fault continues to evolve

At the fault evolves, it can be seen that in both cases the weight assigned to the candidate models begin to diverge, as those models which describe the observed dynamics, result in a weight increase, and likewise, those models which do not describe the observed dynamics result in a weight decrease. In comparing the performance of both approaches, the main feature to note in Figure 6.18 is the rate at which the candidate model weights decay to zero as the fault evolves. It can clearly be seen in Figure 6.18 that using the exponentially weighted approach, the rate at which candidate model weights decay to zero is significantly faster than is the case using uniform weighting. As a greater number of candidate model weights decay to zero, a more precise RUL PDF should be generated since the RUL PDF becomes a weighted sum of a smaller number of individual RUL PDFs.

6.5.4.2 Performance Evaluation

To determine if the exponential weighting scheme improves the prediction performance, the predictions generated by the two approaches were compared using available prognostic performance metrics. The development of prognostic performance metrics has been an active research area in recent years. Much of the work on performance metrics has been developed and reviewed in a series of publications by Saxena *et al.* [51]. These techniques were developed to address the perceived shortcomings in this area, particulary the unsuitability of standard forecasting metrics for prognostic problems.

Much of the recent research on prognostic performance metrics has focused on the development of techniques for comparing the performance of different prognostic algorithms, when applied to the the same prognostic problem. A brief review of prognostic metrics was presented previously in Section 2.5.3. Techniques developed include metrics such as the prognostic horizon (PH), α-λ performance, and relative accuracy (RA). PH evaluates how far in advance of system failure does an algorithm generate predictions within the desired accuracy range, around the true EoL, whilst α-λ performance evaluates if an algorithms stays within desired performance levels, relative to the actual RUL at a given time [52].

In the current study, techniques such as prognostic horizon and α-λ performance are not relevant, as the objective is not to compare the performance of different techniques, but to compare the performance of adaptations to the same technique. Two metric were evaluated to compare the performance of the uniform and exponential weighting schemes, *accuracy* and *precision.* Accuracy provides a measure of how close a point estimate of RUL is to the actual RUL, whilst precision quantifies the narrowness of the time interval which the expected RUL covers, which is essentially a measure of the spread, or width, of the the RUL PDF [3].

To compare prediction accuracy, the recently developed relative accuracy (RA) metric was used [50]. The RA metric quantifies the accuracy of a prognostic algorithm at a given time, where the output is normalised by the actual RUL, so that predictions made closer to the true failure time, t_{EoL}, should demonstrate greater accuracy. The range of values for the RA metric is $[0, 1]$, where 1 implies a perfect score. The RA metric is defined as

$$\mathrm{RA}(t_p) = 1 - \frac{|r_*(t_p) - \langle r(t_p) \rangle|}{r_*(t_p)} \tag{6.27}$$

where t_p is the time index at which the RUL predictions are generated, $r_*(t_p)$ is the actual RUL at time index t_p, and $\langle r_*(t_p)\rangle$ is a point estimate of the RUL, generated from predictions made at time index t_p. In the situation where the RUL PDF is described by a parametric distribution, the point estimate of the RUL is defined by the distribution mean. In the case of the multiple model particle filtering approach, the RUL PDF cannot be described analytically, and so the RUL point estimate of the RUL is inferred from the cumulative distribution function representation of the RUL PDF, which is defined by the set of predicted particles and weights, according to equation (6.25).

The precision metric quantifies the narrowness of the time interval over which the expected RUL falls. In the current application, the time interval over which the expected RUL falls is inferred from the cumulative distribution function representation of the RUL PDF. The upper and lower time indices, which approximate the 95% confidence limits on the RUL PDF are used to define the time interval over which the expected RUL falls. The precision metric is defined as

$$\text{Precision } (t_p) = \exp(\frac{-R_i}{R_0}) \tag{6.28}$$

where R_i is the time interval over which the expected RUL falls, and R_0 is a normalising factor, whose value is dependant on the application. Precision is defined over the interval $[0, 1]$ with a value of 1 denoting a higher precision.

To investigate if the exponential weighting scheme within the model likelihood function improves the accuracy and precision of the RUL estimates, the multiple model particle filtering framework was applied to the failure example illustrated in Figure 6.17. The algorithm was tested separately using both the uniform and exponential weighting schemes in computing the model likelihood function. Figure 6.19 illustrates the values of the RA and precision metrics computed from the predictions generated by the multiple model particle filtering approach using both model likelihood weighting schemes. As illustrated in Figure 6.19, the incorporation of exponential weighing scheme within the model likelihood function improves both the prediction accuracy, as defined by the RA metric, and the precision of the RUL predictions. This is due to the faster rate at which candidate model weights decay to zero, in response to their dynamics deviating from the observed fault behaviour.

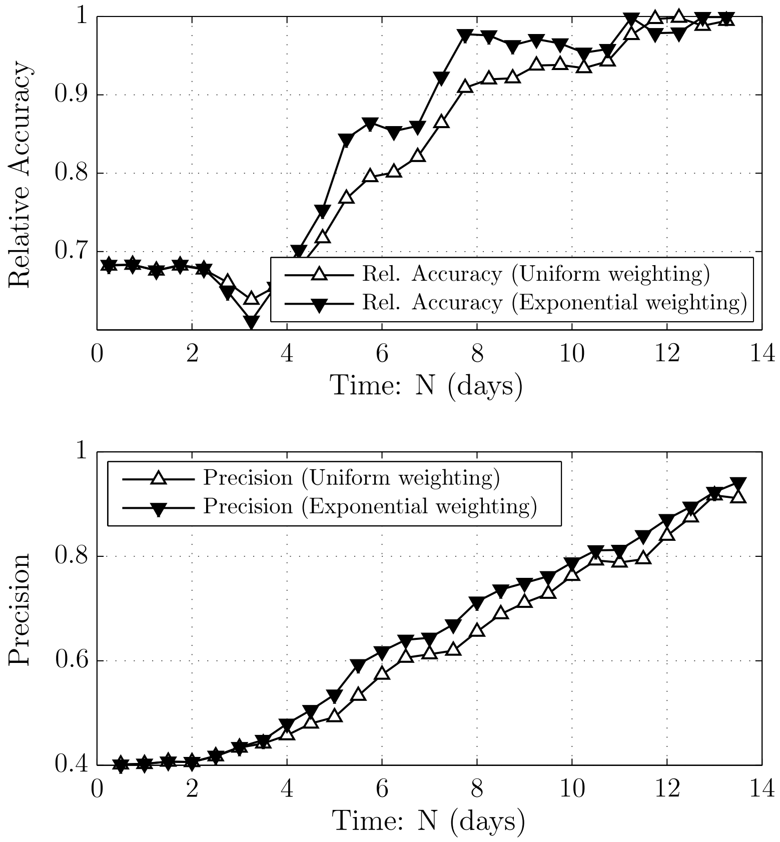


Figure 6.19: Comparison of prognostic performance as measured by relative accuracy and precision metrics, using both uniform weighting and exponential weighting methods, as TPU fault continues to evolve

6.6 Conclusions

The primary challenge in developing prognostic capabilities is the inherent uncertainty in predicting the future behaviour of degrading equipment, which will be subject to stochastic processes which have not yet occurred. Addressing this uncertainty is challenging, especially within the semiconductor manufacturing environment. Within this environment, many process details, such as the gas types abated, the specific flow rates, and expected future load profiles, are unknown. To address these challenges this chapter has detailed the development of a multiple model particle filtering approach for TPU prognostics. To the best of the authors knowledge this is the first time in which a multiple model approach has been applied to a real-world prognostic problem, using data collected from a large semiconductor manufacturing facility.

Section 6.1 and 6.2 introduced the potential maintenance and operational benefits, and the challenges, in developing prognostic capabilities for TPU systems. Section 6.3 then detailed the basic principles and implementation steps involved in applying particle filtering for TPU prognostics. The application of single model particle filtering to an historical TPU failure example was then illustrated in Section 6.4. Section 6.4 then discussed the challenges in developing prognostic capabilities for TPU system operating within the semiconductor manufacturing environment. These challenges motivated the development of the multiple model particle filtering approach which is subsequently presented in Section 6.5. Within Section 6.5, the details, and implementation steps, involved in developing a multiple model particle filtering approach were presented, which also included a description of how the set of candidate failure models were derived from the set of historical TPU failure examples. The performance of the multiple model approach was then demonstrated on another historical TPU failure example. The dynamics of the evolving candidate model weights, which respond to the observed behaviour of a TPU fault, were also demonstrated.

Finally, in Section 6.5.4, a modification to the model weighting scheme was introduced, which incorporates exponential weighting within the model likelihood function, used for computing the weight of each candidate model. In this way the individual candidate model weights are more heavily influenced by how well each candidate model describes the more recent fault behaviour, as opposed to being weighted on the basis of how well each candidate model describes the full fault history, up to the point at which RUL predictions are generated. The performance, and the improvement in prognostic accuracy and precision of the exponential model weighting scheme, are then demonstrated on another historical TPU failure example.

While the multiple model approach, presented in this chapter, was developed and tested on TPU systems used in semiconductor manufacturing, the presented approach could also be applied in many other prognostic application domains, in which there may be little understanding, or ability, to model the underlying degradation process, but for which a set of historical failure examples are available. The set of historical failures examples can then be used to generate a set of candidate approximating models as demonstrated in this chapter. As more historical failure examples become available, and the understanding of the underlying process improves, the set of candidate models can be continuously refined to reflect this better understanding. In the next chapter, the potential application of the multiple model particle filtering to another application domain, wind turbine prognostics, is demonstrated.

Chapter 7

Wind Turbine Condition Monitoring

7.1 Introduction

Over the past decade, the deployed wind power generating capacity worldwide has increased rapidly. By the end of 2010, wind generating capacity reached approximately 196,630 MW [107]. In addition, the size and generating capacity of individual wind turbines also continues to increase, with increasing numbers of wind turbines with >5MW generating capacity becoming standard in offshore wind farms. For the economic exploitation of wind energy, high reliability of wind turbines and their components is necessary. Wind turbines typically operate in harsh environments and are continuously exposed to changing, and sometimes extreme, environmental conditions. Over many years of operation, the constantly changing loads imparted by changing wind speeds and directions generate significant loads on wind turbine blades, which is transferred to the transmission system within a turbine nacelle. This can lead to the failure of bearings and other components within the gearbox, main bearing, and generator. As a result, the maintenance of wind turbines and the avoidance of component failures is of critical importance to wind farm operators.

Due to the better wind conditions available and the lack of shore based sites, wind turbines, of increasing size and generating capacity, are more commonly being located offshore. As a result, avoiding turbine downtime and minimising corrective maintenance due to component failure is even more important, as accessing failed turbines is far more costly. For offshore wind farms, studies have suggested that maintenance costs are about

20 to 25% of the total income generated, and that a considerable percentage of these costs are due to unexpected equipment failure, which require corrective maintenance [11]. In an effort to reduce the maintenance costs for wind turbines, wind farm operators are increasingly embracing condition-based maintenance philosophies in an effort to reduce maintenance costs, increase turbine reliability, increase the lifetime of turbine components, and reduce turbine downtime and associated loss of revenue. Various approaches to monitoring the condition of turbines have been considered. Most modern turbines incorporate onboard supervisory control and data acquisition (SCADA) systems for performance monitoring and supervision. As these system are already installed as standard, wind farm operators are interested in better exploiting this data for condition monitoring, fault diagnostics, and fault prognostics.

In this chapter, SCADA system data is used for a short feasibility study into the development of a condition monitoring solution for the main bearing, on a large utility scale wind turbine. The approach taken is to develop a model which describes the fault-free behaviour of the main bearing temperature signal. By modelling the fault-free behaviour of the main bearing temperature, future behaviour can then be estimated and, by analysing the difference between the estimated and actual behaviour observed, potential fault conditions can be identified.

The layout of this chapter is as follows. Section 7.2 briefly reviews the design of modern wind turbines and reviews the different approaches which have been investigated for condition monitoring of wind turbines. A description of the problem of main bearing failures, which is investigated in this chapter, is also presented. Section 7.4 describes the process of modelling the fault-free behaviour of the main bearing temperature on a wind turbine with greater that 2MW generating capacity. Due to data sensitivity concerns, the specific type of turbine investigated is not disclosed. In Section 7.5, the model developed in Section 7.4 is applied to two historical main bearing failure examples, to demonstrate the performance of the approach and the ability to detect the presence of incipient fault conditions. In addition, an approach to tracking the evolution of the behaviour of the fault signal within different turbine operating modes is demonstrated, which improves the ability to identify and track the development of a fault condition. Finally, in Section 7.6, the development of a prognostic solution for predicting the remaining useful life (RUL) of failing main bearings is demonstrated. The multiple model particle filtering framework for prognostics, developed in the Chapter 6, is used for this task, to demonstrate its capabilities in addressing the uncertainty in predicting the RUL of turbine components subject to uncertain future load profiles. Due to a lack of historical failure examples, the development of prognostic capabilities for the main bearing is presented simply as a proof of concept. However, with a greater number of historical failure examples, the approach presented has the potential to be improved and to be utilised in developing prognostic capabilities for other wind turbine components.

7.2 Wind Turbines

7.2.1 Wind Turbine Design

In this study, data collected from a utility scale wind turbines, with a generating capacity in excess of 2MW, is used to develop a condition monitoring solution for the main bearing. The rotor blades on the type of wind turbine investigated are of a three blade cantilever construction, mounted upwind of the supporting tower. The blades are manufactured using composite material, and are manufactured as a single piece. Figure 7.1 illustrates the typical layout and design of the type of wind turbine investigated [9].

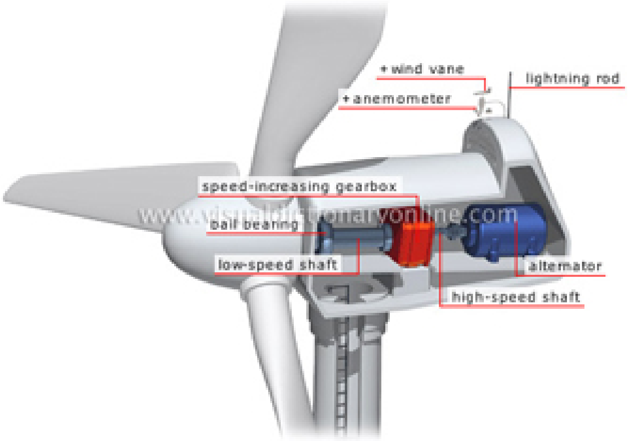


FIGURE 7.1: Wind turbine design and component layout [9]

The rotor blades are mounted on the main shaft, referred to as the low-speed shaft. The main shaft is supported at the front of the nacelle by a single main bearing, labelled ball bearing in Figure 7.1. The torque generated by the rotating blades is transferred to the speed increasing gearbox by the main shaft. The gearbox is a 3-stage planetary helical design, which increases the rotational speed of the main shaft. Splashed lubrication is used to lubricate the gearbox and an onboard oil cooling system is used to cool the gearbox. A hydraulic disk brake is located on the high-speed shaft, which connects the gearbox to the generator, to reduce the rotational speed of the main shaft under high-wind conditions. An asynchronous generator is located at the rear of the nacelle, with a rated power of greater than 2 MW.

7.2.2 Wind Turbine Condition Monitoring - A Review

In comparison to dry vacuum pumps and thermal abatement devices, investigated in Chapters 4 - 6, condition monitoring for wind turbines is an area of widespread research activity. Researchers and practitioners have been investigating condition monitoring solutions for many of the different components on wind turbines using a wide variety of different approaches. A number of publications are available which provide a review of the different techniques and approaches which have been investigated [108–110].

Wind turbines contain multiple rotating components, including the main shaft, the multi-stage gearbox, and the generator. As a result, vibration monitoring has been investigated for application to wind turbine condition monitoring and fault diagnostics [111, 112]. Within the domain of vibration monitoring for wind turbines, wavelet methods have seen widespread application due to their inherent ability to provide time-frequency resolution of the monitored vibration signal. This is important due to the variable-speed operation of wind turbines [113, 114]. Acoustic emissions (AE) analysis has also been proposed for wind turbine condition monitoring. The relatively low-speed operations for wind turbines can place limitations on vibration monitoring. AE based approaches are designed to detect the stress waves generated by the rubbing of rotating components. An example of an application of AE analysis for gearbox components can be found in [115].

To address cracking and delamination in large wind turbine blades, which are usually constructed of composite materials, a range of approaches have been considered. Strain gauges, located at different stress points, have been used to monitor peak strains, with the objective of identifying locations where structural damage may have occurred. More recently, fibre optic strain gauges have become more widespread due to greater capabilities and accuracy over electrical strain gauges [110, 116].

Wind turbine power curve analysis is a common method for providing a universal measure of wind turbine performance, and as an indicator of overall wind turbine health [117]. The wind power curve, for a specific wind turbine device, relates the turbine power output for a given wind speed and air density. Given the current wind conditions and air density, differences between the expected power output, as estimated by the power curve, and the actual power output are identified. The difference, which is often called the *power residual*, can be used to indicate potential operational issues, such as the overall blade condition [118], and gearbox faults [117].

The automatic monitoring and collection of operational data, in the form of sensor measurements and operational status information, is now common in many industries. In most modern wind turbines, supervisory control and data acquisition (SCADA) systems are now common. These systems monitor and sample a wide range of different sensor variables, including bearing temperatures, pressures, meteorological conditions, and power and electrical measurements. These measurements are then stored in a database, usually in the form of 5-minute or 10-minute aggregate data, which records the mean, min, max, and standard deviation of each sensor variable over the relevant interval. The data can the be analysed online or offline, to monitor the health and performance of each individual wind turbine. Some more recent SCADA systems also monitor overall vibration levels within critical components [119].

Since SCADA systems are already installed as standard on most wind turbines and, in many cases, years of historical data recorded by the SCADA system are available, wind farm operators and original equipment manufacturers are increasingly interested in better exploiting this data for real-time condition monitoring and fault diagnostics. One of the primary drivers for using SCADA data for condition monitoring is that the data collection and sensor networks are already in place, which makes such approaches significantly cheaper. In contrast, comprehensive vibration monitoring and AE approaches require high-frequency data which might be sampled at up to 20 kHz. The costs of the required sensor network and the data collection, storage, and processing capabilities required to analyse and extract relevant features from the high-bandwidth raw vibration signals are significant. Furthermore, due to the high-bandwidth data required for advanced vibration monitoring, all significant signal processing and feature extraction must be performed locally at each turbine, with the extracted features and measurements then transmitted to a central location for further analysis and decision making [111]. As a result, the cost per turbine of installing such a system is often significant.

SCADA data is often used for power curve monitoring and to estimate power curves online for individual turbines within a wind farm. Another application of SCADA data is to develop models which describe the fault-free behaviour of different turbine components. Differences between the estimated behaviour and observed behaviour can then be used to identify the presence of potential fault conditions. A number of authors have considered such approaches for gearbox and generator components [119, 120], which will be discussed further in Section 7.2.3.

7.2.3 Problem Description

In this chapter, data collected from a SCADA system, onboard large utility scale turbines, is used to develop a condition monitoring solution for the main bearing. The main bearing is responsible for supporting the rotor shaft, on which the turbine blades are mounted, and for transmitting torque to the turbine gearbox. Due to constantly changing wind conditions, the bearing loads can vary considerably and, during large wind gusts, very significant loads can be exerted on the main bearing by the rotor and blades. Continuous changes in operating loads and environmental conditions can eventually lead to damage to the main bearing, in the form of fretting on the inner and outer races and spalls and cracks in the rolling elements. In the specific type of turbine investigated, spherical roller bearings are employed in the main bearing which support the small angular elevation of the main shaft.

The objective of this study is to investigate how existing data collection and monitoring capabilities, in the form of onboard SCADA systems, can be exploited to provide greater insight into turbine health and provide maintenance personnel with visibility on developing maintenance issues. The focus of this work is to develop a data-driven model which describes the fault-free behaviour of the main bearing temperature signal. Then, using the developed model, the difference between the estimated main bearing temperature and the observed main bearing temperature, which is called the *residual*, is used to detect the presence of a potential incipient fault condition. Such approaches to modelling fault-free behaviour of wind turbine components, using data-driven models, have been considered previously. Zaher *et. al* [119] used artificial neural networks (ANNs) to model the fault-free behaviour of the gearbox bearing temperature, gearbox cooling oil temperature, and the generator temperature. Excessive values of the residual signal between the estimated and observed measurements are used to identify a fault condition. Garcia *et. al* [120] also use ANNs to model fault-free behaviour of several gearbox components. Guo *et al.* [121] employed a nonlinear state estimate technique (NSET) to model the fault-free behaviour of a turbine generator. A moving average filter was then used to detect statistically significant changes in the mean and variance of the residual signal, which might indicate the presence of a fault condition.

In this study, a similar approach to those described in the previous paragraph is investigated, whereby SCADA data is used to detect developing faults within the main bearing, by first modelling the fault-free behaviour of the main bearing temperature signal. Within the literature, SCADA data has not previously been investigated for condition monitoring of the main bearing on wind turbines. Another feature of this current study, not previously investigated for wind turbine condition monitoring, is how, under fault conditions, the evolution of the residual signal is also a function of the varying operating loads on the turbine. The publications described in the previous paragraph do not consider the effects of varying operating loads on the magnitude of the residual signal. In this study, by monitoring the evolution of the residual signal, under different operating loads, it will be demonstrated how significantly improved error tracking can be achieved. Indeed, it will also be demonstrated how, by tracking the evolution of the residual signal in different operating modes, the ability to develop real predictive prognostic capabilities for wind turbines is improved.

Real predictive prognostics involves predicting the evolution of a signal of interest, or fault indicator, until its value exceeds a predefined threshold. In many situations, identifying or extracting an appropriate signal for predictive prognostics is difficult. To date, most wind turbine condition monitoring research has focused almost exclusively on condition monitoring and fault diagnostics. In contrast, the development of real predictive prognostic capabilities for wind turbines is still in its relative infancy. This is due in large part to the difficulties associated with predicting the remaining useful life of turbine components, particularly when the future operating load is so uncertain. However, prognostic capabilities are essential to enable the benefits of a truly condition based maintenance philosophy for wind turbines.

The provision of a lead-time, between the detection of an incipient fault condition and actual equipment failure, would provides major benefits from a wind turbine operations and maintenance (O&M) perspective. The size, weight, and costs of many wind turbine components mean that most wind-farm operators do not maintain a large inventory of spares on-site and, in general, new components must often be ordered from suppliers once a fault condition is first detected. Furthermore, the actual replacement of specific components can often require the presence of domain experts and specialised equipment, including cranes capable of handling the relevant loads and reaching the required heights. Another consideration is the issue of weather-windows. The ability to operate large cranes is often dependant upon the prevailing weather conditions and particulary the wind speed. To ensure the safety of maintenance personnel and equipment, the weather must be satisfactory to enable corrective maintenance actions to be performed. Thus, during inclement weather conditions, maintenance actions must often be delayed until the weather improves, increasing both the maintenance costs and the non-revenue generating turbine downtime.

The provision of accurate RUL estimates can potentially be of major benefit for wind farm operators. By combining RUL estimates with weather forecasts, the potential for identifying suitable periods in which maintenance can be performed is improved. In this way, equipment failure, including potential damage propagation to secondary components, can be avoided and downtime and loss of revenue associated with performing corrective maintenance can be minimised. However, similar to the development of prognostic solutions for TPU systems in the previous chapter, and perhaps more challenging, is the uncertainty regarding the future load profile under which a turbine will operate. The future load profile will depend upon the wind conditions, which is a very stochastic process. Thus, as in the development of any prognostic algorithm, how this uncertainty is managed and represented will be key to exploiting the potential benefits of prognostic capabilities for wind turbines.

In this study, real predictive prognostics for the main bearing are also investigated. Due to the lack of historical failure examples, the investigation into main bearing prognostics is intended to simply serve as a proof of concept. In the investigation into main bearing prognostics, to be presented in Section 7.6.2, many assumptions regarding the future behaviour and characteristics of main bearing failure examples are made, which may be incorrect. However, as more historical failure examples become available, the investigated approach can be refined and improved as the level of understanding of the underlying process increases. In addition, with a greater number of historical failure examples, some of the assumed values regarding how the level of the error signal relate to failure conditions will also be able to be improved. To investigate prognostics for the main bearing, the multiple model particle filtering framework, developed in the Chapter 6, is utilised. The capabilities of the multiple model particle filtering approach are particulary suited for wind turbine prognostics, where the uncertainty over future load profiles, dictated by the wind conditions, requires that any prognostic approach is capable of representing and managing this associated uncertainty.

7.3 Wind Turbine Condition Monitoring Algorithm

This section introduces a proposed algorithm for main bearing condition monitoring and remaining useful life (RUL) prediction. A model-based approach is proposed. Figure 7.2 presents a flow chart illustrating the different stages in the proposed condition monitoring and prognostic algorithm.

The basic principle of the model-based approach is to develop a model for each turbine which describes the fault-free behaviour of the main bearing temperature. The estimated main bearing temperature generated by the fault-free model is then compared with the actual main bearing temperature at each iteration of the algorithm. The difference between the estimated and actual main bearing temperature, known as the residual, is then evaluated. Assuming the turbine remains fault-free, the residual signal should generate a Gaussian distributed signal with a mean of zero and a small variance. Once a fault develops, the residual signal may change and no longer be zero-mean. Analysis of the residual signal is performed by the residual processing and decision logic stages. Assuming no fault condition is detected, the algorithm continues to iterate at 10-minute intervals as data is recorded by the SCADA system. Assuming a fault condition is detected, this generates an alarm and initialises the prognostic stage, which estimates the RUL of the main bearing. The algorithm then continues to iterate and the RUL predictions are recursively updated using the latest filtered value of the residual signal.

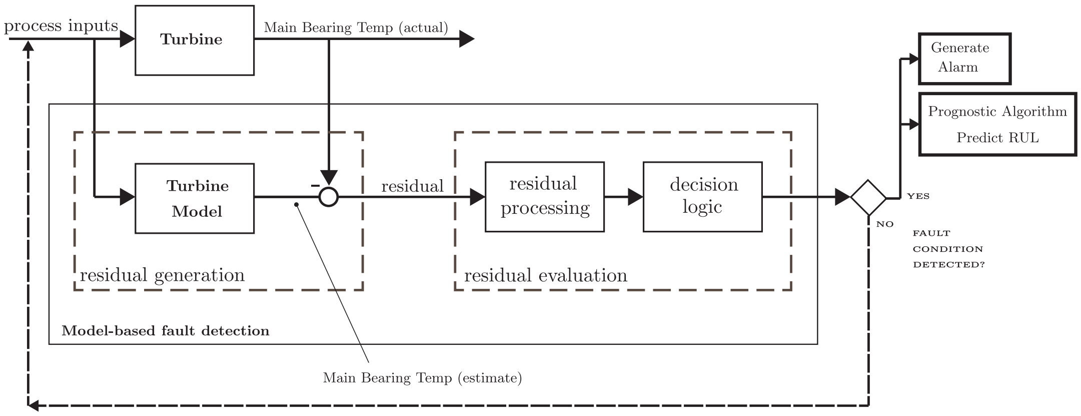

process inputs
Turbine
Main Bearing Temp (actual)
Generate Alarm
Prognostic Algorithm Predict RUL
Turbine Model
−
residual
residual processing
decision logic
YES
NO
FAULT CONDITION DETECTED?
residual generation
residual evaluation
Model-based fault detection
Main Bearing Temp (estimate)

Figure 7.2: Main bearing model-based condition monitoring algorithm

The remainder of this chapter describes in detail the different stages illustrated in Figure 7.2. Section 7.4 describes the process of modelling the fault-free behaviour of the main bearing temperature. Section 7.5 then describes how the developed models can be used to detect fault conditions in the main bearing by analysing the residual signal. Section 7.6 then describes the decision logic stage which is used to determine whether a fault condition is present and also describes the proposed RUL prediction approach and the results generated.

7.4 Modelling Main Bearing Temperature

In this section, the detailed steps in developing a model to describe the fault-free behaviour of the main bearing temperature are presented. Firstly, in Section 7.4.1, a description of the data set available for this study is presented. Section 7.4.2 then introduces the proposed modelling approach, including the inputs and model structure to be used to model the behaviour of the fault-free main bearing temperature. The motivation for using sparse Bayesian learning for regression to model the behaviour of the main bearing temperature is also discussed. Section 7.4.3 then discusses data detrending, to address the variability in sensor values caused by changes in the ambient temperature, as the seasons change. Finally, in Section 7.4.4, the proposed model is trained and tested on historical turbine data to demonstrate the performance of the trained model on fault-free turbine data.

7.4.1 Data Set Description

For this study, data from a large wind-farm was made available. For each turbine, the complete history of sensor information and turbine status information, for a period of 11-months, was available. The onboard SCADA system for each turbine records 10-minute averages of each monitored sensor variable. In addition, the maximum, minimum, and standard deviation of each of these sensor values, over each 10-minute period, is also recorded by the SCADA system. In addition to the values of the onboard sensors, status information, such as generator start and stop times, are also recorded by the SCADA system.

7.4.2 Modelling Approach

To develop a condition monitoring solution for main bearings, using available SCADA data, it is proposed to develop a model which describes the behaviour of the main bearing temperature under normal, fault-free, operation. Once the model is developed, it can then be used to estimate future main bearing temperature values, using other turbine variables as inputs. Assuming the main bearing remains fault free, the residual signal, representing the difference between the estimated main bearing temperature and the actual main bearing temperature, should be a zero-mean random variable with, ideally, a small variance. Once a fault develops in the main bearing, in the form of spalling, fretting, or cracks, the increased friction should result in an increase in the main bearing temperature, beyond the estimated value. As a result, the residual signal, and its statistical characteristics, should begin to change, reflecting the development of an incipient fault condition within the main bearing.

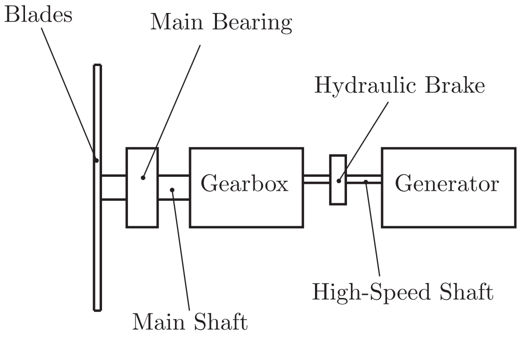


Figure 7.3: Turbine components layout

Modelling the behaviour of the fault-free main bearing temperature required selection of appropriate input variables, which can be used to estimate the fault-free main bearing temperature under varying load conditions. A range of different variables were investigated to identify their usefulness in estimating the main bearing temperature. The final set of input variables, selected for inclusion in the feature vector for estimating the main bearing temperature are described below. Figure 7.3 illustrates the location of the different components whose sensor values are described below.

Main Shaft RPM The heat generated in the main bearing will be a function of the load on the bearing. The main shaft RPM describes the load exerted on the main bearing under varying wind conditions.

Hydraulic Brake Temperature The turbine brake is located on the high-speed shaft, which connects the gearbox to the generator, and analysis has demonstrated that, under fault-free conditions, the brake temperature is closely correlated with the main bearing temperature.

Hydraulic Brake Pressure The average hydraulic brake pressure over a ten-minute interval provides a measure of the brake friction applied to the high-speed shaft, which in turn generates friction within the main bearing, resulting in a response in the main bearing temperature

Blade Pitch Position All modern turbines employ pitch control to pitch the blades under high-wind conditions. While the main shaft RPM may remain constant, the load imparted on the main bearing will vary with the blade pitch position.

The model used to describe the behaviour of the fault-free main bearing temperature signal is of the form

$$\hat{r}(k) = f(r(k-1), u_1(k), u_1(k-1), u_2(k), u_3(k), u_4(k)) \tag{7.1}$$

where $\hat{r}(k)$ is the estimated main bearing temperature at time k, $r(k-1)$ is the actual main bearing temperature at time $k-1$, and $u_i(k)$, $i = 1, ..., 4$, is the value of input i, at time k. The input variables, u_i, represent the following turbine variables, which are described above

- u_1: Main Shaft RPM
- u_2: Hydraulic Brake Temperature
- u_3: Hydraulic Brake Pressure
- u_4: Blade Pitch Position

To model the relationship between the main bearing temperature and the input variables, described by Equation (7.1), sparse Bayesian learning for regression [74, 75] was used. A review of the background and theory of sparse Bayesian learning for regression is presented in Section 3.5. In previous similar studies [119, 120], ANNs have been widely employed for modelling the fault-free behaviour of turbine components. In this study, sparse Bayesian learning for regression was considered for a number of reasons.

Preliminary modelling studies demonstrated that it is not possible to develop a single model which could be used to describe fault-free behaviour of every turbine in a wind farm. Due to manufacturing variations, maintenance history, historical load conditions, and variations on the current condition of different turbine components, the need to model the fault-free behaviour of individual turbines was identified. Thus, for each turbine within a wind farm, an individual model would be required to describe the fault-free behaviour of the main bearing, using historical SCADA data for each turbine. Thus, in training and validating models for each turbine, where wind farms often contain in excess of 100 wind turbines, training time is a major consideration. Sparse Bayesian learning, and in particular the fast marginal likelihood maximisation approach developed by Tipping *et. al* [76], provide extremely fast training times for model development.

In addition, sparse Bayesian learning models have excellent generalisation capabilities on unseen data, owing to their sparse representation. Furthermore, the outputs generated by sparse Bayesian learning models are probabilistic, providing variances estimates on the generated predictions. This capability is potentially of major benefit from a fault detection and diagnosis perspective, in that confidence estimates provide a method to infer how confident the model is in the generated predictions. Although this capability was not actually exploited in this study, future work will likely investigate how probabilistic outputs might be exploited to improve condition monitoring capabilties.

7.4.3 Data Detrending for Ambient Temperature

In modelling wind turbine behaviour, a major consideration is the effect of ambient temperature. Turbine sensor variables, and particularly temperature sensor variables, are a function of both the current operating conditions and the ambient temperature. This presents some issues when trying to model fault-free turbine behaviour. For a model to be sufficiently descriptive of turbine behaviour, across all seasons and weather conditions, significant volumes of historical data would be required to capture the turbine responses under varying conditions. Alternatively, some approach to detrending the data, to remove the ambient temperature relationship, must be considered. In this study, an approach to detrending turbine temperature variables, suggested by Wiggelinkhuizen *et. al* [122, 123], was employed.

To detrend the turbine temperature values, and remove the ambient temperature contribution, each relevant turbine signal was linearly corrected for ambient temperature, using data collected when the turbine was operating under a small rotational speed. Figure 7.4 shows a scatter plot of ambient temperature and main bearing temperature, over an 11-month period from January to November. The samples shown were recorded when the turbine was operating under a small rotational speed, between 0.1 and 1 RPM. Figure 7.4 illustrates the linear relationship between main bearing temperature and ambient temperature under low rotational speed. Note, due to data sensitivity concerns, the values of all turbine sensor variables have been mapped to an arbitrary interval, though meteorological variables have been left on their original scale.

Using the observed relationship, the main bearing temperature was then linearly corrected for ambient temperature. Figure 7.5 illustrates the main bearing temperature signal from a randomly selected, fault-free, turbine. Figure 7.5 (a) illustrates the original main bearing temperature signal and Figure 7.5 (b) shows the main bearing temperature detrended, with the contribution from the ambient temperature removed. In this way, the effects of ambient temperature can be removed and the process of modelling the main bearing temperature can proceed without concern for the ambient temperature contribution. All other turbine variables employed, whose response is influenced by the ambient temperature, were also normalised in the same fashion as described above.

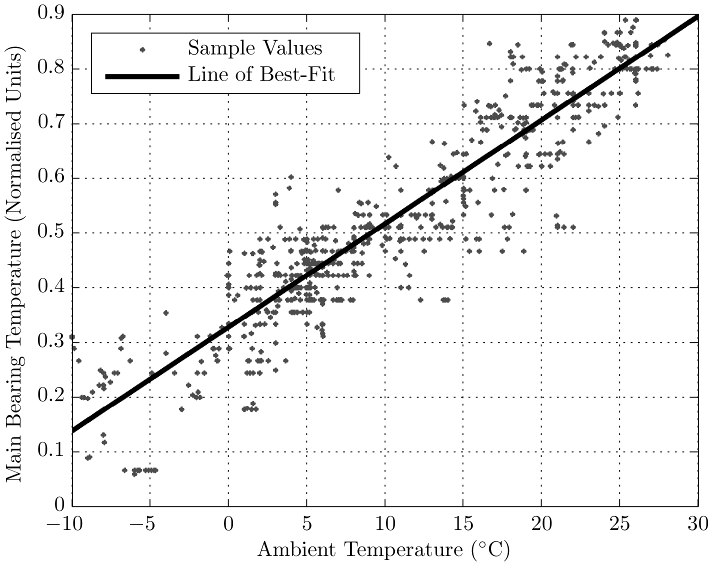


FIGURE 7.4: Relationship between main bearing temperature and ambient temperature, under low-load conditions

7.4.4 Model Training and Validation

The dataset available for this study comprised 11 months of historical data. Thus, at a 10-minute sampling interval, approximately 48,000 samples were available for each turbine. To validate the proposed modelling approach, using the structure of the model described in Equation (7.1), three fault-free turbines were randomly selected from the available data set. For each fault-free turbine, approximately 12,000 samples, representing approximately 3 months of data, were selected for *model training.* A feature of spare Bayesian learning for regression is that no data is required for validation, only training and test data is required. Thus, the remaining samples for each turbine were then used for *model testing.* In developing the model describing fault-free behaviour for each turbine, the data selected for training was taken from a different three month period, to validate that the data normalisation approach, described in Section 7.4.3, accurately accounts for the effects of ambient temperature variations.

Once each of the models, describing normal fault-free main bearing temperature behaviour, were trained, they were each tested on the remaining previously unseen samples for each turbine. Figure 7.6 shows the performance of the first fault-free turbine model over a 20-day period of previously unseen data. Figure 7.6 (a) shows the model estimate and the actual main bearing temperature, and Figure 7.6 (b) shows the residual term, which is the difference between the estimated and actual main bearing temperature.

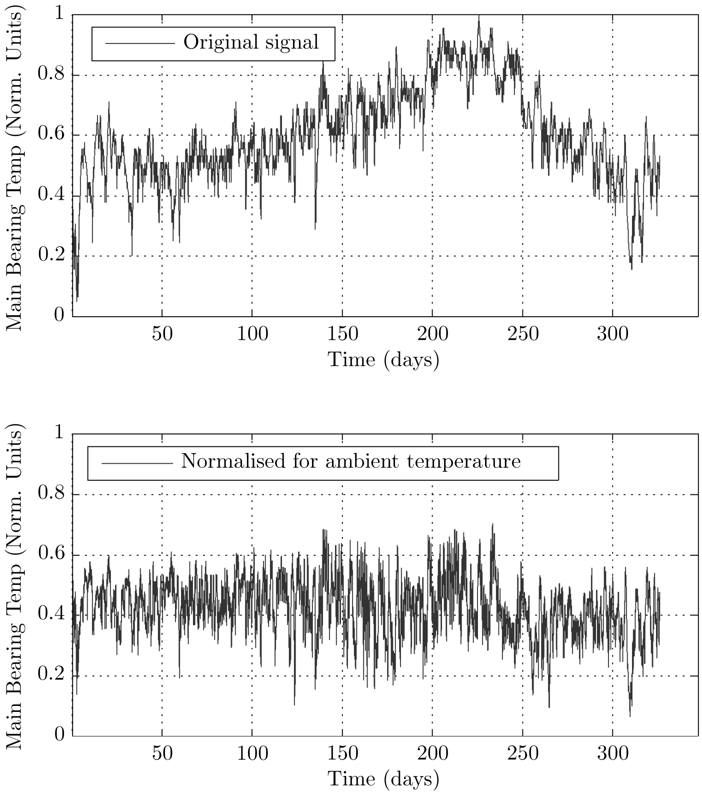


Figure 7.5: Main bearing temperature: original signal ((a) upper plot)and normalised for ambient temperature signal ((b) lower plot)

The model error term is sufficiently small that it is difficult to distinguish the model estimate from the actual main bearing temperature in Figure 7.6 (a).

To confirm that the error signal is zero-mean and Gaussian distributed, the distribution of the error signal for the second turbine used for model validation, for the complete 8-months of test data, is plotted as a histogram in Figure 7.7. As can be seen in Figure 7.7, the error-term is zero-mean and Gaussian distributed.

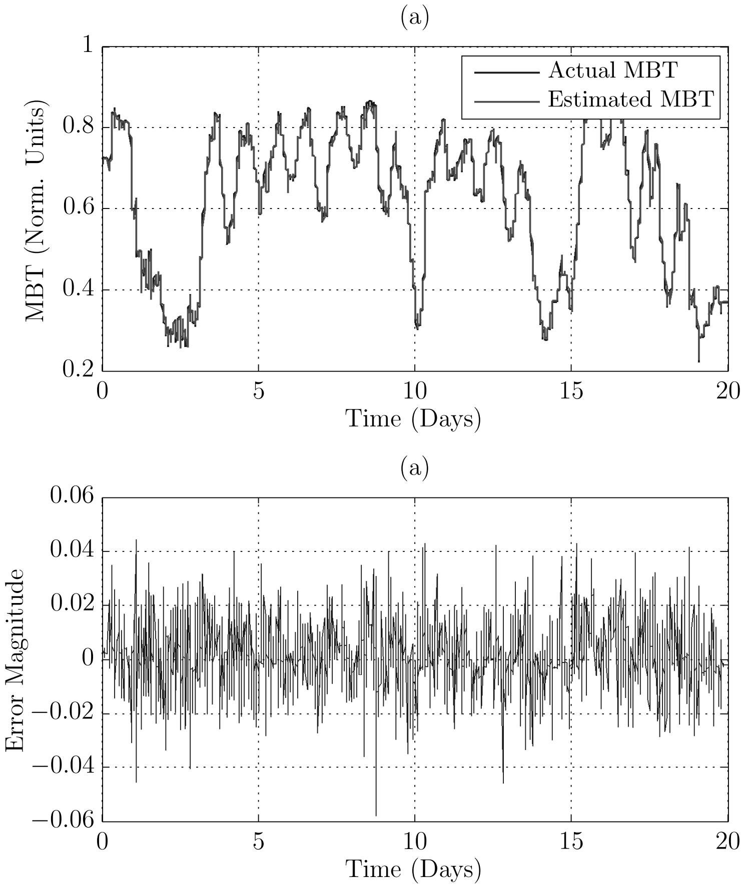


FIGURE 7.6: Main bearing temperature estimation (a) and generated residual signal (b) indicating error magnitude at each sample time (fault-free turbine)

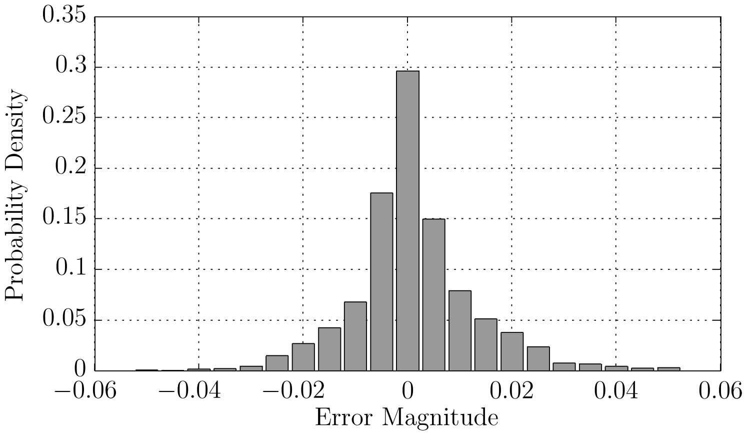


FIGURE 7.7: Distribution of residual signal between estimated and actual main bearing temperature (fault-free turbine)

7.5 Main Bearing Fault Detection

In Section 7.4.4, the ability to accurately estimate the behaviour of the main bearing temperature, under fault-free operation, was demonstrated. In the presence of a main bearing fault, in the form of corrosion, pitting, or fretting, it might be expected that the additional friction may generate excessive heat within the main bearing, beyond what might normally be expected within a fault-free bearing. The next stage is to investigate if the main bearing temperature model, designed in the previous section, is capable of identifying, and tracking, the development of a fault in the main bearing.

In the available data set, two turbines suffered issues with the main bearing. In the first case, which will be referred to as Turbine A, an issue with the main bearing was identified during a routine visual inspection and, some time later, the turbine was removed from service and the main bearing replaced. In the second case, referred to as Turbine B, the main bearing temperature exceeded the acceptable operating limits and the turbine control system automatically shut the turbine down during service. This turbine remained out of service for a number of weeks while the main bearing was replaced.

Figure 7.8 (a) illustrates the main bearing temperature in the final 225 days of operation of Turbine A, which was removed from service following a visual inspection of the main bearing. Figure 7.8 (b) shows the residual term generated between the estimated and actual main bearing temperature. Visual analysis appears to show some changes in the characteristics of the residual signal after approximately 150 days. Before Turbine A was removed from service, the mean of the residual signal is clearly above 0. Using the residual signal in Figure 7.8 (b), in its raw form, the ability to make informed maintenance decisions, regarding when to perform maintenance, is clearly difficult. Guo *et al.* [121] suggest using a moving average (MA) filter to detect statistically significant changes in the mean and variance of the residual signal. Figure 7.9 (b) illustrates a moving-average filter applied to the residual signal from Figure 7.8 (b). A two-day window, comprising 288 samples, was used to generate the signal shown.

As Figure 7.9 (b) clearly illustrates, the statistical characteristics of the residual signal show a significant, and relatively sustained, change after approximately 175 days. Over the following 50 days, prior to removal from service, the peak magnitude of the filtered residual signal continues to increase. Another noticeable feature is the large fluctuations on the value of the filtered residual signal over this period, with no clear monotonic increase in the value of the filtered residual signal. However, the results in Figure 7.9 clearly illustrate the ability of the modelling approach in detecting the presence and

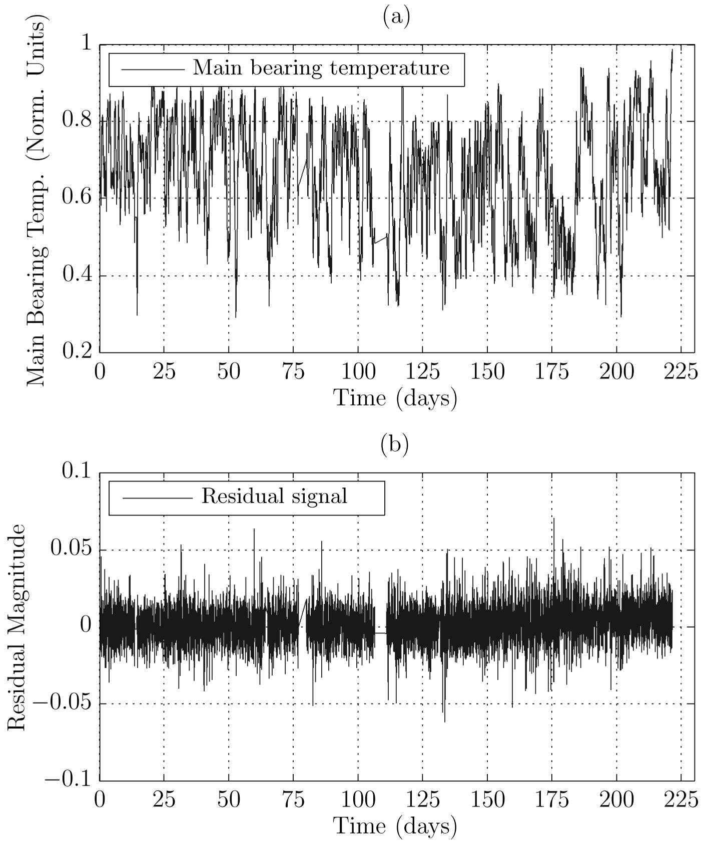


Figure 7.8: Main bearing temperature and residual signal (Turbine A)

increasing magnitude of an incipient fault condition within the main bearing, potentially negating the requirement for regular visual inspections to detect such conditions.

Figure 7.10 (a) illustrates the main bearing temperature in the final 200 days of operation of Turbine B, which failed in-service due to an excessively high main bearing temperature value. Figure 7.10 (b) shows the residual term generated between the estimated and actual main bearing temperature. As clearly illustrated in Figure 7.10 (b), a major increase in the value of the residual term occurred some 5 days prior to failure. To detect significant changes in the statistical characteristics of the residual signal for Turbine B, a moving average filter was applied. Figure 7.11 (b) shows the evolution of the filtered residual signal using 2-day and 5-day moving averages. The filtered residual signals clearly demonstrate the capability to identify the presence of a fault condition which, if presented to maintenance personnel, would result in the turbine being removed from service and avoid damage potentially propagating to other turbine components.

An identifiable shortcoming of applying a simple MA filter to the residual signal is

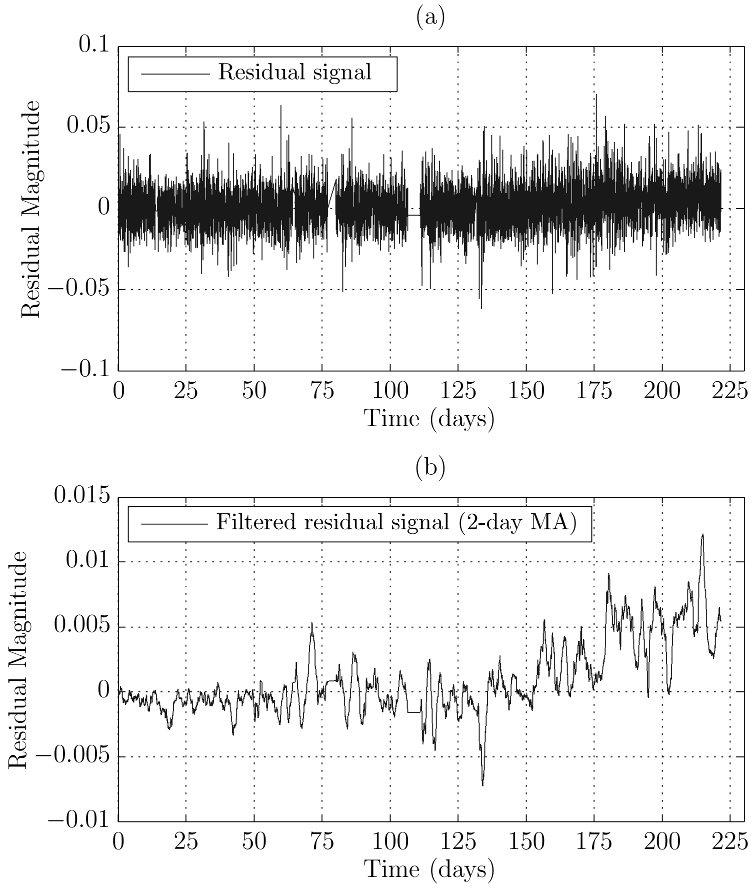


Figure 7.9: Main bearing temperature residual (a) and filtered residual signal (b) (Turbine A)

that no consideration is given to how the different operating modes of the turbine may influence the magnitude of the residual signal. Furthermore, the filtered residual signals exhibit large fluctuations, making it difficult to determine, confidently, the presence of a fault condition. To address the fluctuations in the filtered residual signal, one option is to increase the length of the moving-window, effectively increasing the time constant of the MA filter. However, this reduces the ability to detect sudden significant changes in the characteristics of the filtered residual signal. In the following section, analysis of the residual signal, based upon the operating mode of the turbine, is investigated as a method for improving the error tracking accuracy, and for early detection of a fault condition.

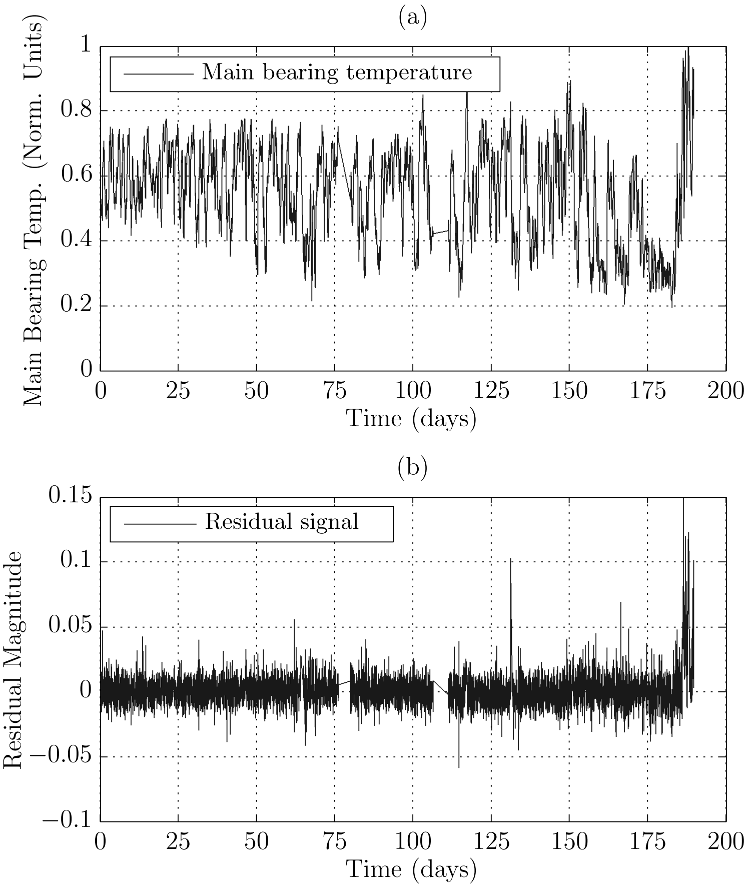


FIGURE 7.10: Main bearing temperature and residual signal (Turbine B)

7.5.1 Fault Tracking by Turbine Operating Mode

Section 7.5 demonstrated how, by modelling the fault-free behaviour of the main bearing temperature, it is possible to detect the presence of incipient fault conditions and to track the increasing magnitude of the identified fault. However, using a simple moving-average filter to track the evolution of the residual signal takes no account of how the magnitude of the error signal may relate to specific operating modes, or operating regions, of the turbine. If the residual error magnitude is related to operating condition, then the evolution of moving-average signal will also depend upon the varying operating conditions, instead of solely being a function of the current level of component degradation.

In this section, a brief investigation into operating modes of the main bearing is undertaken. The objective is to identify if improved error tracking can be achieved by tracking the error signal while the turbine is operating within a specific region. Wind turbines do not have specific discrete operating modes, such as the high-fire and low-fire

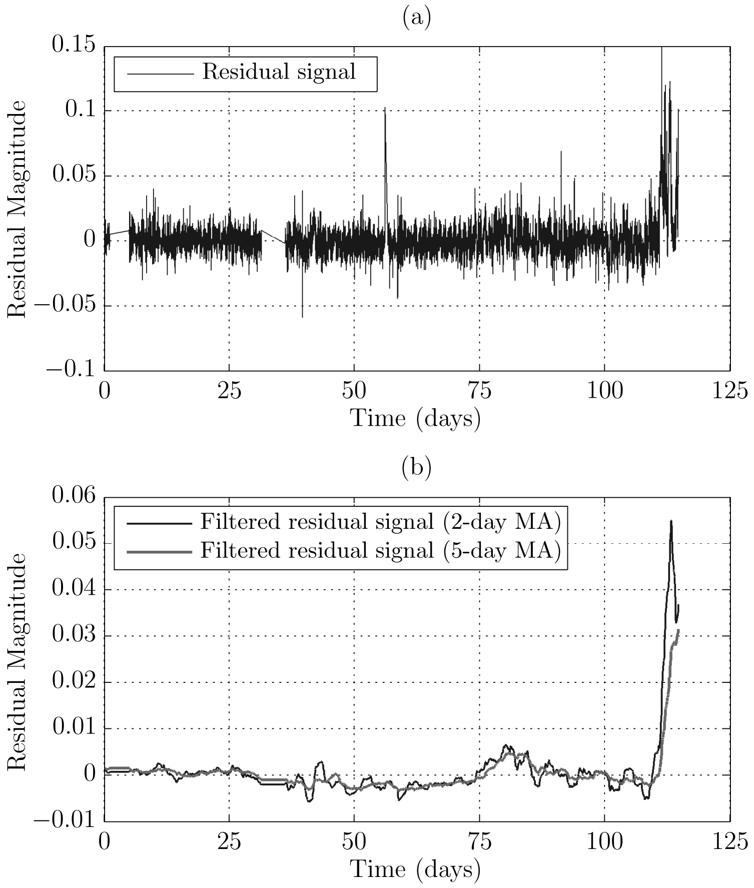


Figure 7.11: Main bearing temperature residual and filtered residual signal (Turbine B)

operating modes of a thermal processing unit in Chapter 5. Instead, the operating mode of a wind turbine can be loosely defined by the response to different wind conditions. In an attempt to identify turbine operating modes, a number of available signals were investigated. The obvious signal for classifying the current turbine “operating mode” is the power output. The power output is a function of the generator RPM, which is a direct function of the main shaft RPM. In addition, the generator and gearbox bearings, and the gearbox oil temperature, provide a means to identify the current turbine operating modes. Under high wind speeds, the gearbox and generator bearing temperatures respond and increase significantly in value, compared to no-wind, low load, operations.

However, using these variables to identify specific turbine operating modes, which could potentially be used to track the development of the main bearing temperature error signal within each of the identified modes, is difficult. This is due to the different thermal conductivity of wind turbine components. The size of the main bearing is many orders of magnitude greater than the bearings within either the gearbox or the generator. Due to the greater mass of the main bearing, the thermal conductivity of the main bearing is significantly smaller than many of the other available signals, which means the main bearing temperature is much slower to respond to changes in wind load, as opposed to the other variables.

In identifying turbine operating modes, which can be used to track the magnitude of the error signal in the different operating modes, the task is to identify specific, frequently occurring, conditions at which times the magnitude of the error signal can be evaluated. The objective in trying to identify turbine operating modes is to improve the ability to track the evolution of main bearing degradation. The primary variable which directly describes the load (i.e. friction generating force) on the main bearing is the main shaft RPM. To investigate the relationship between the main shaft RPM and the main bearing temperature, the joint distribution between the main bearing temperature and main shaft RPM was investigated. Figure 7.12 illustrates this distribution, over a 6-month fault-free period, from a randomly selected turbine. The joint distribution of the main shaft RPM and main bearing temperature, in Figure 7.12, was identified using a kernel density estimator (KDE) [124], a non-parametric method for estimating probabiity density functions. A KDE was used as the distribution in Figure 7.12 can clearly not be approximated by any parametric distribution function.

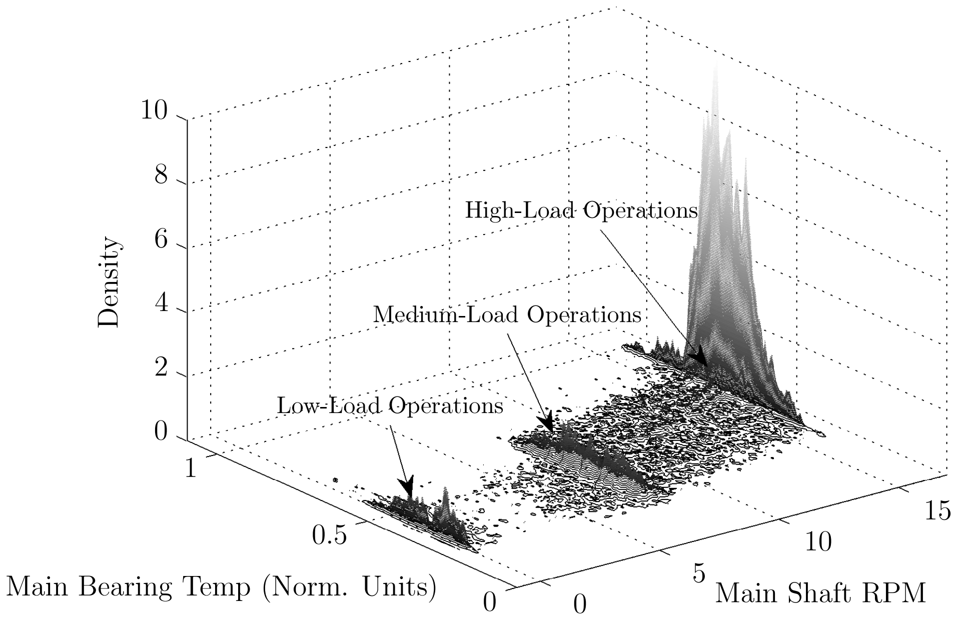


FIGURE 7.12: Joint distribution of main shaft RPM and the main bearing temperature

Within Figure 7.12, there are clearly three distinct "operating regions" identifiable in the data, labelled Low-Load, Medium-Load, and High-Load. The identified operating regions are primarily a function of the main-shaft RPM and, within each of the three identified operating regions, the main bearing temperature is distributed across a range of values. As a result, it is difficult to further sub-classify each of the three operating regions using the main bearing temperature. Using the operating regions highlighted within Figure 7.12, turbine operating modes based solely upon the main shaft RPM were selected. Table 7.1 shows the main shaft RPM range over each of the three identified operating regions.

Operating Region	Low-Load	Medium-Load	High-Load
RPM Range	0.1 - 2	7 - 9	15 - 16

Table 7.1: Main bearing operating modes: RPM Range

By identifying the turbine operating region at each iteration, using the main shaft RPM, the evolution of the residual signal in each of the three operating was investigated. By far the greatest error tracking performance, in terms of a relatively smooth monotonic increase in the value residual signal within that operating regions, was achieved whilst the turbine was operating in the Low-Load operations region. A possible reason for this behaviour may be that, during low RPM operations, the load on the main bearing, and the friction generated, may be more consistent than during higher load operations. During low-wind conditions, the wind speed variance is generally lower, and thus the load on the main bearing is possibly more consistent. During periods of higher-wind conditions, the wind speed variance is generally higher, resulting in varying loads on the main shaft RPM and thus greater variance in the value of the residual signal.

For the available main bearing failure example, Turbine B, Figure 7.13 illustrates the evolution of the filtered residual signal, generated by filtering only those samples recorded when the main shaft RPM was within the range of Low-Load operations, as defined in Table 7.1. The times at which the turbine was under Low-Load operations are also indicated in Figure 7.13.

Another feature, not utilised in previous similar work [121], is the use of an exponentially weighted moving-average (EWMA) filter to track changes in the statistical characteristics of the residual signal. The principle of an EWMA filter is similar to that of a simple moving average filter except that equal weighting is not applied to each sample within the window. Instead, the weighting for older data points decreases exponentially, giving more importance to recent observations, whilst not discarding older observations entirely. Equation (7.2) describes the operation of an EWMA filter

$$s(t) = \alpha x(t) + (1 - \alpha)s(t - 1) \tag{7.2}$$

where $s(t)$ is the output of the filter and $x(t)$ is the original signal value at time t. The filter constant α is the *smoothing factor* which controls the degree of smoothing applied to the input signal. The lower the value of α the greater the level of smoothing applied to the data. The smoothing constant α is analogous to the cut-off frequency in a low-pass filter.

To demonstrate the improved error tracking performance, Figure 7.13 also illustrates the same EWMA filter applied across all values of the residual signal, regardless of the turbine operating mode. As clearly illustrated, tracking of the residual signal during Low-Load operations significantly improves the error tracking performance, and identifies the presence of an incipient fault condition long before a fault is detected by filtering the full residual signal.

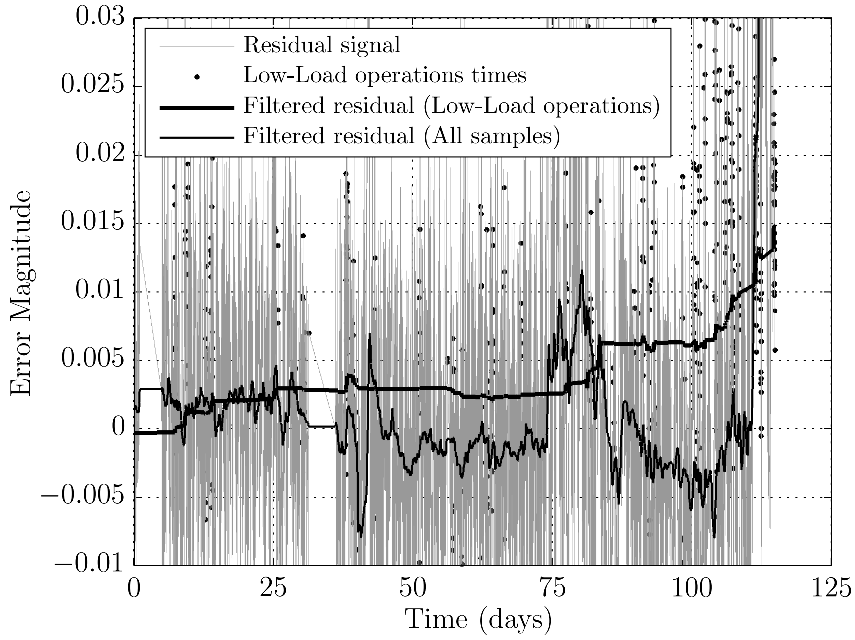


FIGURE 7.13: EWMA filtered residual signal by turbine operating mode

7.6 Fault Prognostics for Wind Turbines

In this section, the potential for the application of the multiple model particle filtering framework, developed previously in Chapter 6 for TPU prognostics, is investigated for wind turbine prognostics. Due to having only a single historical failure example, the work presented in this section is presented simply as a proof of concept and a number of assumptions are made. However, the general concept and approach, and the generated results, clearly illustrate the potential for the development of prognostic capabilities for wind turbines. In addition, the multiple model particle filtering framework is demonstrated as a potential enabling technology for wind turbine prognostics, with the necessary capabilities for managing and representing the associated uncertainty involved in predicting the future behaviour of systems subject to both stochastic degradation processes and uncertain future load profiles.

7.6.1 Incipient Fault Detection

The first task in developing prognostic capabilities for the main bearing is identifying, with confidence, the presence of an incipient fault condition. In Section 7.5.1, an EWMA filter is used to measure changes in the statistical characteristics of the residual signal, during Low-Load operations. Deviations in the value of this signal can be used to identify the occurrence of a incipient fault condition and track the evolution of the fault condition over time.

To identify an appropriate threshold of the EWMA filtered residual signal, which confirms the presence of an incipient fault condition, the three fault-free turbines, used for model development and testing in Section 7.4.4, were selected. Using the model developed for each turbine, each model was tested on the previously unseen test data. The residual signal generated was then filtered, using the same EWMA filter as used to generate the filtered residual signals in Figure 7.13, using the values of the residual signal generated when the turbine was operating in Low-Load operations. The distribution of the three EWMA filtered residual signals was then analysed. Figure 7.14 illustrates the distribution of the three filtered residual signals, during Low-Load operations, for each of the three fault-free cases.

The distribution of the EWMA filtered fault-free residual signal, during Low-Load operations, can be approximated by a Gaussian distribution, as illustrated in Figure 7.14. The distribution is approximately zero-mean, as might be expected during fault-free

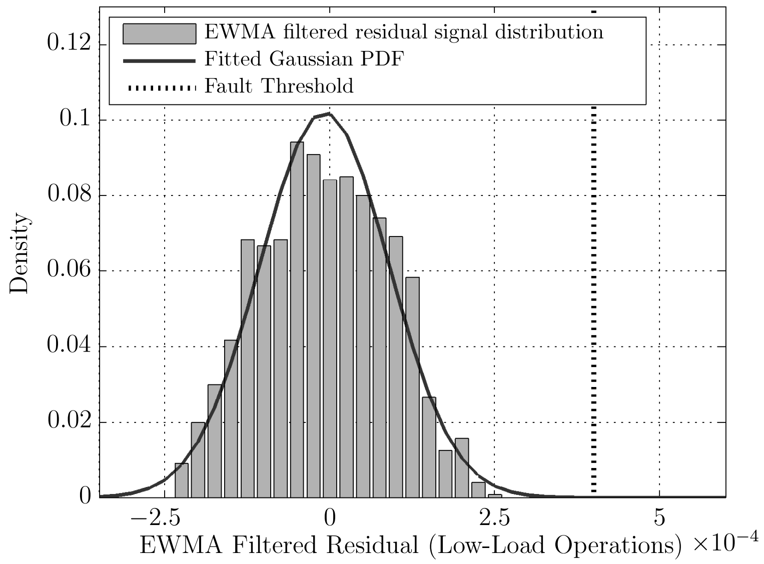


Figure 7.14: Distribution of EWMA filtered residual signal during Low-Load operations for 3 fault-free turbines

operation, with a standard deviation (σ) of approximately 0.0097. For a Gaussian distribution, the 99% confidence limits are defined by approximately $\pm$ 3σ. Therefore, to provide a sufficient separation between the expected limits of “normal” fault-free operation, a value of 0.004 ($> 4\sigma$) was chosen to define the threshold at which a fault condition is confirmed. The location of the fault threshold is illustrated in Figure 7.14.

Having selected an appropriate fault threshold for the EWMA filtered residual signal, during Low-Load operations, Figure 7.15 illustrates the point at which the fault condition is first identified in the available historical main bearing failure example, i.e. Turbine B. Using the selected fault threshold value, a fault is first detected approximately 32 days prior to failure. Predicting the evolution of the filtered residual signal, which henceforth is described as the *fault indicator*, defines the realm of prognostics. To address this problem, the multiple model particle filtering framework, previously utilised in Chapter 6, is used to demonstrate the potential application of prognostics for a wind turbine main bearing. The details of this work are presented in the following section.

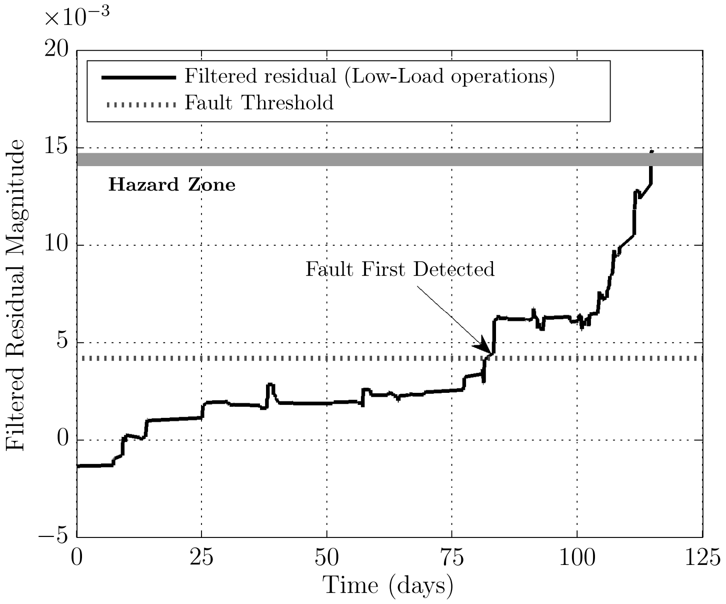


FIGURE 7.15: Evolution of EWMA filtered residual signal during Low-Load operations for faulty main bearing (Turbine B)

7.6.2 Multiple Model Particle Filtering for Wind Turbine Prognostics

Particle filtering has emerged in recent years as the de facto state of the art technique for real predictive prognostics. Particle filtering employs recursive Bayesian estimation techniques to infer the current level of component degradation/damage, by combining model estimates and online measurements to estimate the current level of degradation as a random variable, the degradation state. The distribution of the current degradation state is defined by a set of particles, representing state values, and an associated set of particles weights, representing the discrete probability masses of the individual particles. A more comprehensive review of recursive Bayesian estimation techniques and the particle filtering framework for prognostics was presented previously in Section 6.3 and Section 6.3.3, respectively, which should be referred to as necessary. In the following paragraph, some of the general, and more important, principles of particle filtering and, in particular, multiple model particle filtering for prognostics, are briefly reviewed.

The foundation of particle filtering for prognostics is the use of a state-transition model, which describes the evolution of the degradation process under observation. In the current application, the degradation process refers to the the increasing magnitude of the fault within the main bearing, as described by evolution of the EWMA filtered residual signal shown in Figure 7.15.

The application of particle filtering for prognostics involves two distinct stages 1.) *state estimation* and 2.) *long-term predictions*. In the first stage, predictions generated by the state-transition model are combined with fault indicator measurements (i.e. EWMA filtered residual signal values), to generate a *posterior* estimate of the current degradation state. This process is repeatedly recursively as new fault indicator measurements are generated. Once the current degradation state is estimated, the second stage can be carried out; long-term predictions. Using the state transition model, the set of particles defining the current degradation state estimate can be propagated into the future, until the value of the degradation state exceeds a predefined threshold. The predefined threshold is defined by the hazard zone specified for the current application. Figure 7.15 illustrates the hazard zone chosen for the current application.

The performance of a particle filtering approach relies upon the ability to accurately model the degradation process. However, without a physics-of-failure model, developing an accurate model with sufficient fidelity to describe the likely behaviour of all future failure examples is difficult. In addition, uncertainty regarding the future load profile, which in the case of wind turbines depends upon future weather conditions, introduces a significant level of uncertainty regarding the future behaviour of the degradation process. To address this challenge, a multiple model particle filtering approach is considered. By generating a large set of candidate models, designed to approximate the possible behaviour of future failure examples, the predictions of each of the models can be combined and, as the fault evolves, the plausibility that each model is descriptive of the observed behaviour can be computed. The mathematics involved in updating the candidate model weights and generating RUL estimates which are a weighted combination of RUL estimates generated by each model, were presented previously in Section 6.5.1, and should be referred to as necessary. In Section 7.6.2.1, the process of developing a set of candidate models to describe the potential future behaviour of main bearing faults is presented. A demonstration of the multiple model approach applied to the main bearing failure example available is then presented in Section 7.6.2.2.

7.6.2.1 Modelling Turbine Main Bearing Degradation

In applying the particle filtering framework for main bearing prognostics, the first task is to identify a suitable model to describe the evolution of the main bearing degradation process, as described by the fault indicator signal. With only a single failure example available, some significant assumptions regarding the degradation behaviour of future main bearing failures, must be made. To model the evolution of the main bearing fault indicator, the model used previously to describe the evolution of the TPU degradation process in Section 6.3.2 was adapted for this task. The form of the model used to describe the evolution of the main bearing fault indicator is given by

$$x_k = x_{k-1} + \alpha_1 \mathrm{exp} \left[\frac{-\alpha_2/t_k}{t_k^2} \right] + \alpha_3 \exp \left[\alpha_4 \, t_k \right] + \omega_k \tag{7.3}$$

where x_k represents the degradation state at time t_k, the α_i values represent model parameters which can be tuned to fit the model to describe specific behaviour, and ω_k is a zero-mean Gaussian distribution representing the process noise term. The structure of the model described by Equation (7.3) provides great flexibility in tuning the model to describe observed behaviour.

With only a single historical failure example available, generating a set of candidate models, which are designed to describe the potential behaviour of future examples, is difficult. To address this task, the model parameters in Equation (7.3) were first tuned to fit the available historical example. Using the identified value of each α_i parameter, a distribution of values for each α_i parameter was generated, using the identified α_i parameter value as the mean of the distribution. By setting a range of values for each α_i parameter and sampling randomly from each distribution, a large set of candidate models was generated to describe the behaviour of future failure examples. By appropriate tuning of the distribution from which each α_i value is sampled, a set of candidate models which were deemed sufficient to describe future failure examples, given the lack of current understanding, were generated. Figure 7.16 illustrates the set of 100 candidate models generated to describe the potential future behaviour of main bearing failures, with the available historical failure example overlayed.

7.6.2.2 Application Example

Once the set of candidate models have been generated, the multiple model particle filtering framework was applied for main bearing prognostics. Initially, when a fault is first detected, the model weight associated with each candidate model is the same. As the fault continues to evolve, the plausibility that each candidate model is descriptive of the observed behaviour of the fault indicator is recursively estimated. The weight of each candidate model is then updated to reflect how well each model describes the observed behaviour. At each iteration, long-term predictions, describing the evolution of the fault indicator are generated, with the RUL PDF generated by each candidate model multiplied by its respective weight.

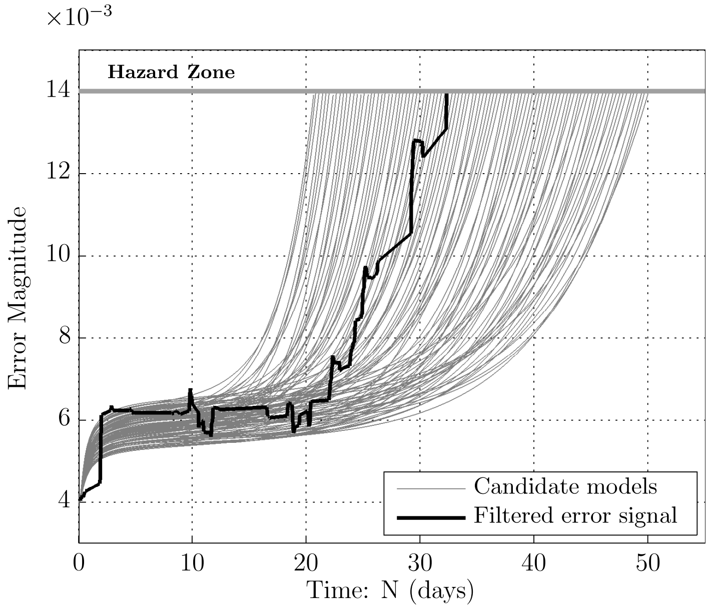


Figure 7.16: Candidate models for use in multiple model particle filtering framework which describe the potential evolution of future main bearing failure examples, as derived from the single available historical failure example (Turbine B)

Figure 7.17 illustrates the evolution of the RUL PDF for the single historical main bearing failure example, using the multiple model particle filtering framework for prognostics. As illustrated in Figure 7.17, as the fault continues to evolve the RUL PDF becomes more accurate and precise, as the behaviour of the degradation process evolves. By providing maintenance personnel with access to such information, informed maintenance decisions can be made and, by combining RUL predictions with weather forecasts, potential windows for performing corrective maintenance can be identified, thus minimising the risk of inclement weather conditions postponing maintenance activities and significantly increasing the associated costs. Furthermore, the overall turbine downtime can also be minimised, reducing the period for which an individual turbine is generating no revenue.

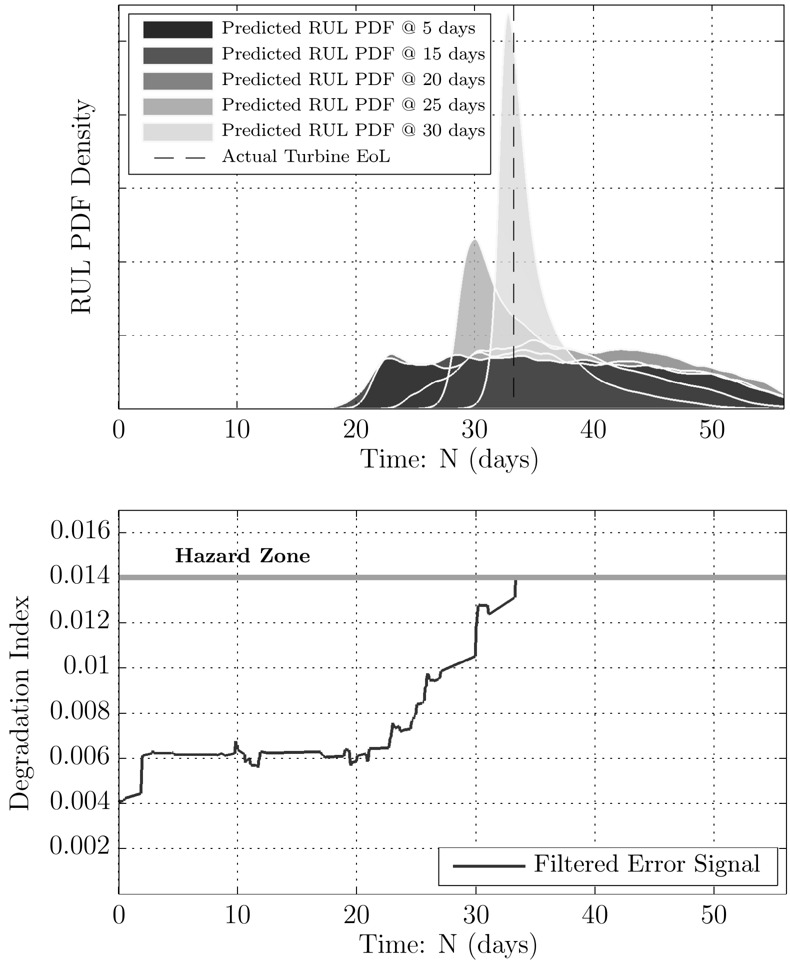


Figure 7.17: Evolution of RUL PDF using multiple model particle filtering framework for main bearing prognostics (Turbine B)

7.7 Conclusions

In this chapter, the development of a condition monitoring solution for the main bearing of a large utility scale wind turbine was developed. A major feature of the developed solution is that data already collected by onboard SCADA systems was used, which is of major interest to wind-farm operators as no additional sensors and monitoring equipment needs to be installed.

A sparse Bayesian learning scheme was utilised to model the fault-free behaviour of the main bearing temperature signal. A major benefit of this modelling scheme is the fast model training algorithm, which is extremely useful as it is necessary to develop a single model for each turbine in a wind farm. Using the developed models, it is possible for wind farm operators to identify wind turbines in which a potential fault within the main bearing is developing. Furthermore, by tracking the magnitude of the residual signal generated, it is possible to track and quantify the magnitude of the residual signal over time.

In addition, prognostic capabilities for the main bearing were also demonstrated as a proof of concept. The multiple model particle filtering framework for prognostics was used to demonstrate its inherent capabilities to manage and represent the uncertainty associated with predicting the future behaviour of degrading equipment subject to uncertain future load profiles.

Chapter 8

Conclusions and Future Directions

8.1 Conclusions

This thesis has investigated the development of algorithms for condition monitoring and prognostics of critical equipment within the domains of semiconductor manufacturing and wind turbine power generation. As discussed in Chapter 2, the development of such technologies is of major interest to manufacturers and operators of critical equipment, for the range of maintenance and operational benefits such technologies might provide. These include, reduced maintenance costs, reduced instances of equipment failure, a reduction in ongoing scheduled maintenance activities and costs, and improved equipment uptime and availability.

Key to enabling such benefits to be realised are robust algorithms capable of operating within real-world environments. In this thesis, all of the algorithms presented were developed exclusively using data collected from systems operating in real-world environments. Thus, the developed approaches have demonstrated applicability to the relevant domains for which they are developed and provide future practitioners with insight and guidance in the development of future condition monitoring and prognostic technologies.

In Chapter 4, the development of a condition monitoring solution for dry vacuum pumps was presented. The key conclusions of this chapter is the demonstrated benefits of greater information sharing between pumps suppliers and semiconductor manufacturers. As discussed in Chapter 4, such information sharing is not typical, as semiconductor manufacturers do not make available any information that might be considered proprietary or commercially sensitive. This chapter demonstrates how upstream process measurements, in the form of the foreline pressure signal, allows for the degradation of pump condition to be modelled and, subsequently, permits maintenance personnel to identify, track and predict pump RUL, using only the available pump signals. This then enables pump maintenance personnel to maximise the serviceable life of pumps and reduce instances of in-service pump failures.

In Chapters 5 and 6, the development of a comprehensive condition monitoring and prognostic solution for thermal abatement devices is presented. A number of general issues and conclusions can be identified. In developing condition monitoring and, in particular, prognostic capabilities, the key consideration is how such systems will enable maintenance personnel to make informed maintenance decisions regarding when to perform maintenance. The key output which enables maintenance personnel to make such decisions is an accurate RUL PDF. This information allows maintenance personnel to make risk-based assessments regarding when to perform necessary corrective maintenance. The multiple model particle filtering algorithm presented in Chapter 6, delivers this capability directly to maintenance personnel, allowing for a range of maintenance benefits to be realised.

In developing the multiple model particle filtering solution, a number of key challenges, which are generic across many application domains, were addressed. In particular, uncertainty about future equipment operating loads and utilisation rates adds a greater level of uncertainty to an already difficult task - predicting the evolution of a stochastic degradation process. To address these challenges, a multiple model particle filtering approach was developed, which exploits and expands upon the existing particle filtering framework for prognostics. In particular, the challenges presented in addressing model uncertainty and future load uncertainty are addressed by the multiple model approach, thus accurately representing the uncertainty in RUL predictions to maintenance personnel.

Finally, a condition monitoring solution for the main bearing on utility scale wind turbines was presented in Chapter 7. The key feature of this chapter is that the approach developed exploits data collected by SCADA systems, which are installed as standard on most modern wind turbines. The benefit of using such data in developing condition monitoring solutions is that no additional hardware, in terms of sensors, data collection, storage, and processing capabilities are required, thus enabling wind farm operators to better exploit already installed data collection and monitoring systems.

This thesis has focused on the development of condition monitoring and prognostic algorithms for dry vacuum pumps, thermal abatement devices, and wind turbines. While the developed algorithms are specific to the individual problems addressed, a number of more general issues can be concluded. The first issue, which is common across all of the investigated application domains, is the importance of investigating equipment response to degradation under different operating modes. The response in equipment sensor data usually varies in response to the operating load. For example, in the case of dry vacuum pumps, the operating load is a function of the gas load from the upstream processing chamber. In developing condition monitoring and prognostic solutions, one of the primary challenges is identifying, or inferring, an appropriate signal of interest, or fault indicator, for quantifying the current level of equipment degradation and for use in predicting equipment RUL. To track the evolution of an identified fault indicator over time requires that the operating conditions at which times the fault indicator is evaluated are consistent. In this way, the only issue which can be responsible for changes in the value of the fault indicator is equipment degradation, thus enabling the evolution of the degradation process to be tracked.

The importance of identifying equipment operating modes, to enable the evolution of a fault indicator to be tracked, is demonstrated in each application presented in this thesis. In chapter 4, the response of the booster temperature signal to changes in upstream gas loads required that a suitable approach to tracking this signal was developed. In Chapter 5, by tracking the changes in the distribution of the CT signal, during low-fire operating mode, it is possible to infer the level of deposit buildup within the combustion chamber and, more importantly, identify an appropriate fault indicator which can then be predicted to estimate the RUL of a TPU. Section 5.5 also illustrates how, by inferring dry vacuum pump operating modes using the booster power signal, in its possible to track the evolution of deposit buildup within the pump exhaust, in each identified operating mode independently, greatly improving signal tracking capabilities. Finally, in Chapter 7, it is illustrated how tracking the evolution of the residual error signal during low-rpm operations, the generated fault indicator signal identifies the fault at an earlier stage, and generates a suitable signal for prognostics. Thus, it is possible to conclude the importance of investigating the evolution of equipment degradation under different operating conditions, to improve overall degradation tracking performance. Equipment operating modes can be discrete and recorded in sensor data (e.g. TPU low-fire mode), or must be inferred from system data (e.g. pump mode inference from booster power signal). Alternatively, operating modes can be continuous (e.g. turbine response to varying wind conditions). Regardless, the importance of investigating and identifying suitable operating conditions at which times to track fault evolution are clearly important.

The development of prognostic technologies is a difficult task, and goes some way to explaining why such technology has not yet seen widespread application. The primary challenges facing practitioners is the inherent uncertainty in predicting the evolution of a stochastic process. In developing prognostic solutions, the applicability of the different algorithms across different application domains must be considered. As described by Vachtsevanos [3], and illustrated visually by Figure 2.6, the cost of increased prognostic capabilities generally comes at the expense of reduced applicability. This is often due to the requirement that high-fidelity models of degrading processes are required to develop advanced capabilities.

The primary contribution of the thesis is the development and demonstration of a multiple model particle filtering algorithm for prognostics. The developed algorithm has a number of desirable properties, which enable it to be adapted and applied across different application domains. This is demonstrated in the thesis, whereby the approach developed in Chapter 6 is then applied for prognostics of the main bearing on a large utility scale wind turbine in Chapter 7. By using the available historical failure examples, and generating a set of candidate models designed to approximate the future behaviour of historical failure examples, it is possible to exploit the potential of particle filtering for prognostics in many other application domains, with the ability to address the uncertainty challenges presented within each potential application domain.

8.2 Future Work

The potential for taking the work developed in this thesis forward is clearly demonstrated by the successful commercialisation and licensing of the developed technologies. The industrial partner to whom the technologies were licensed have also expressed their desire to take this work forward and are actively involved in investigating avenues for future research, in partnership with academic institutions. In particular, with the proliferation of ethernet-based network communications, the capabilities of the data collection systems used for vacuum pumps and abatement systems will enable the work presented in this thesis to be taken forward and improved. The improved signal resolution and sampling rates achievable using ethernet communications will be a key enabler.

Key areas for future focus will include the application of particle filtering for other applications within the semiconductor manufacturing space. The prediction of pump RUL for fluorine induced degradation and exhaust blockages are two immediate areas in which the multiple model particle filtering algorithm may be focused. In addition, the potential for the incorporation of *a priori* information into the generation of the set of candidate models is an area for focus. As greater knowledge and understanding of a process becomes available, it may be possible to incorporate this information into the set of candidate models, so that the initial RUL PDF might not be uniformly distributed, and instead be more densely distributed about a specific region, with the ability to adapt in response to the evolution of the fault indicator signal.

From a wind turbine perspective, the presented work has demonstrated the potential for the development of prognostic capabilities for wind turbines. As the number of historical failure examples grows, the demonstrated capabilities can continue to be improved. As the size of wind turbines continue to expand, and with wind farms increasingly being located offshore, the potential benefits of prognostic capabilities will also continue to grow.

In addition, the developed approach to modelling the fault-free behaviour of the main bearing can also be replicated for other turbine components, such as the gearbox and generator, enabling wind farm operators to better exploit already available information. Furthermore, the spare Bayesian learning scheme for regression has clear potential in the wind turbine domain, as it has been identified that it is necessary to model the fault-free behaviour of each individual turbine. The fast marginal likelihood maximisation scheme developed by Tipping [76] enables the fast training of such models, which is of particular use across a wind farm with potentially hundreds of turbines.

References

[1] M. S. Lebold, K. M. Reichard, D. Ferullo, and D. Boylan. Open systems architecture for condition-based maintenance: Overview and training manual. Technical report, Penn State University/Applied Research Laboratory, 2003.

[2] A. Hess, G. Calvello, P. Frith, S.J. Engel, and D. Hoitsma. Challenges, issues, and lessons learned chasing the "big p". real predictive prognostics. part 2. In *Aerospace Conference, 2006 IEEE*, pages 1–19, 2006.

[3] G. Vachtsevanos, F. L. Lewis, M. Roemer, A. Hess, and B. Wu. *Intelligent Fault Diagnosis and Prognosis for Engineering Systems.* Wiley, 2006.

[4] S. Lynn. *Virtual Metrology for Plasma Etch Processes.* PhD thesis, National University of Ireland, Maynooth, 2011.

[5] Edwards. *Instruction Manual - TCS, TPU and Kronis Systems.* Edwards Ltd, 2009.

[6] BOC Edwards. Integrated solution for the semiconductor industry. Technical report, BOC Edwards, 2006.

[7] P. Holland, M. Percy, and J. Boegner. Designing safe , low-cost vacuum and exhaust management systems for semiconductor processes. *Semiconductor Manufacturing China*, N/A:N/A, 2007.

[8] P. Fancourt, S. Ishaq, and M. Czerniak. Modified pump, trap system cuts down pecvd maintenance. *Solid State Technology*, 43:149–156, 2000.

[9] Merriam-Webster Visual Dictionary Online. Wind turbine nacelle cross section. `http://visual.merriam-webster.com/energy/wind-energy/wind-turbines-electricity-production/nacelle-cross-section.php`. Last Accessed: March, 2012.

[10] R. Ochsner. Factory integration. *Future Fab International*, 40:73–78, Jan 2012.

[11] D. McMillan and G. W. Ault. Quantification of condition monitoring benefit for offshore wind turbines. *Wind Engineering*, 31:267–285, 2007.

[12] A. K. S. Jardine, D. Lin, and D. Banjevic. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, 20(7):1483–1510, 2006.

[13] S. J. Engel, B. J. Gilmartin, K. Bongort, and A. Hess. Prognostics, the real issues involved with predicting life remaining. In *Aerospace Conference Proceedings, 2000 IEEE*, volume 6, pages 457–469, 2000.

[14] A. Hess, G. Calvello, and P. Frith. Challenges, issues, and lessons learned chasing the "big p". real predictive prognostics. part 1. In *Aerospace Conference, 2005 IEEE*, pages 3610–3619, March 2005.

[15] A. Neely. Exploring the financial consequences of the servitization of manufacturing. *Operations Management Research*, 1(2):103–118, 2009.

[16] J. Z. Sikorska, M. Hodkiewicz, and L. Ma. Prognostic modelling options for remaining useful life estimation by industry. *Mechanical Systems and Signal Processing*, 25(5):1803–1836, 2011.

[17] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri. A review of process fault detection and diagnosis: Part i: Quantitative model-based methods. *Computers & Chemical Engineering*, 27(3):293–311, 2003.

[18] V. Venkatasubramanian, R. Rengaswamy, and S. N. Kavuri. A review of process fault detection and diagnosis: Part ii: Qualitative models and search strategies. *Computers & Chemical Engineering*, 27(3):313–326, 2003.

[19] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, and K. Yin. A review of process fault detection and diagnosis: Part iii: Process history based methods. *Computers & Chemical Engineering*, 27(3):327–346, 2003.

[20] R. Isermann and P. Balle. Trends in the application of model-based fault detection and diagnosis of technical processes. *Control Engineering Practice*, 5(5):709–719, 1997.

[21] R. J. Patton and J. Chen. Review of parity space approaches to fault diagnosis for aerospace systems. *Journal of Guidance, Control, and Dynamics*, 17(2):278–285, March 1994.

[22] N. B. Gallagher, B. M. Wise, S. W. Butler, D. D. White, and G. G. Barna. Development and benchmarking of multivariate statistical process control tools

for a semiconductor etch process: Improving robustness through model updating. In *Process: Impact of Measurement Selection and Data Treatment on Sensitivity, Safeprocess 97*, pages 26–27, 1997.

[23] J. F. MacGregor. Multivariate statistical approaches to fault detection and identification. In *IFAC SAFEPROCESS*, pages 579–584, 2003.

[24] J. S. Qin. Data-driven fault detection and diagnosis for complex industrial processes. *Proceeding of the 7th IFAC Symposium on Fault Detection Supervision and Safety of Technical Processes (SAFEPROCESS)*, pages 1115–1125, 2009.

[25] M. Markou and S. Singh. Novelty detection: a review part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003.

[26] M. Markou and S. Singh. Novelty detection: a review part 2:: neural network based approaches. *Signal Processing*, 83(12):2499–2521, 2003.

[27] M. E. Orchard. *A Particle Filtering based Framework for On-Line Fault Diagnosis and Fault Prognosis.* PhD thesis, Georgia Institue of Technology, 2007.

[28] A. Ray and S. Tangirala. Stochastic modeling of fatigue crack dynamics for on-line failure prognostics. *Control Systems Technology, IEEE Transactions on*, 4(4):443–451, July 1996.

[29] F. Cadini, E. Zio, and D. Avram. Monte carlo-based filtering for fatigue crack growth estimation. *Probabilistic Engineering Mechanics*, 24(3):367–373, 2009.

[30] B. Saha, K. Goebel, and J. Christophersen. Comparison of prognostic algorithms for estimating remaining useful life of batteries. *Transactions of the Institute of Measurement and Control*, 31:293–308, 2009.

[31] D. Edwards, M. Orchard, L. Tiang, K. Goebel, and G. Vachtsevanos. Impact of input uncertainty on failure prognostic algorithms: Extending the remaining useful life of nonlinear systems. In *Annual Conference of the Prognostics and Health Management Society*, 2010.

[32] B. Saha and K. Goebel. Modeling li-ion battery capacity depletion in a paticle fitering framework. In *Annual Conference of Prognostics and Health Management Society*, 2009.

[33] M. E. Orchard and G. J. Vachtsevanos. A particle-filtering approach for on-line fault diagnosis and failure prognosis. *Transactions of the Institute of Measurement and Control*, 31:221–246, 2009.

[34] E. Zio and G. Peloni. Particle filtering prognostic estimation of the remaining useful life of nonlinear components. *Reliability Engineering & System Safety*, 96(3):403 – 409, 2011.

[35] M. J. Daigle and K. Goebel. A model-based prognostics approach applied to pneumatic valves. *Internaional Journal of Prognostics and Health Management*, 2, 2011.

[36] X. S. Si, W. Wang, C. H. Hu, and D. H. Zhou. Remaining useful life estimation: A review on the statistical data driven approaches. *European Journal of Operational Research*, 213(1):1–14, 2011.

[37] O. E. Dragomir, R. Gouriveau, F. Dragomir, E. Minca, and N. Zerhouni. Review of prognostic problem in condition-based maintenance. In *European Control Conference, ECC'09.*, pages 1585–1592, Budapest, Hungary, 2009.

[38] A. Heng, S. Zhang, A. C. C. Tan, and J. Mathew. Rotating machinery prognostics: State of the art, challenges and opportunities. *Mechanical Systems and Signal Processing*, 23(3):724–739, 2009.

[39] W. Wu, J. Hu, and J. Zhang. Prognostics of machine health condition using an improved arima-based prediction method. In *Industrial Electronics and Applications, 2007. ICIEA 2007. 2nd IEEE Conference on*, pages 1062–1067, May 2007.

[40] C. S. Byington, M. J. Roemer, and T. Galie. Prognostic enhancements to diagnostic systems for improved condition-based maintenance [military aircraft]. In *Aerospace Conference Proceedings, 2002. IEEE*, volume 6, pages 2815–2824, 2002.

[41] G. E. P. Box and G. Jenkins. *Time Series Analysis, Forecasting and Control.* Holden-Day, Incorporated, 1990.

[42] M. A. Herzog, T. Marwala, and P. S. Heyns. Machine and component residual life estimation through the application of neural networks. *Reliability Engineering and System Safety*, 94(2):479–489, 2009.

[43] G. Vachtsevanos and P. Wang. Fault prognosis using dynamic wavelet neural networks. In *AUTOTESTCON Proceedings, 2001. IEEE Systems Readiness Technology Conference*, pages 857–870, 2001.

[44] P. Wang and G. Vachtsevanos. Fault prognostics using dynamic wavelet neural networks. *Artif. Intell. Eng. Des. Anal. Manuf.*, 15:349–365, September 2001.

[45] Z. Tian. An artificial neural network method for remaining useful life prediction of equipment subject to condition monitoring. *Journal of Intelligent Manufacturing*, N/A:1–11, 2009.

[46] F. O. Heimes. Recurrent neural networks for remaining useful life estimation. In *Prognostics and Health Management. PHM 2008. International Conference on*, pages 1–6, Ocober. 2008.

[47] K. Goebel, B. Saha, and A. Saxena. A comparison of three data-driven techniques for prognostics. In *Proceedings of the 62nd Meeting of the Society For Machinery Failure Prevention Technology (MFPT)*, pages 119–131, 2008.

[48] X. Zhang, R. Xu, C. Kwan, S.Y. Liang, Q. Xie, and L. Haynes. An integrated approach to bearing fault diagnostics and prognostics. In *American Control Conference, 2005. Proceedings of the 2005*, pages 2750–2755, June 2005.

[49] W. Q. Wang, M. F. Golnaraghi, and F. Ismail. Prognosis of machine health condition using neuro-fuzzy systems. *Mechanical Systems and Signal Processing*, 18(4):813–831, 2004.

[50] A. Saxena, J. Celaya, B. Saha, S. Saha, and K. Goebel. On applying the prognostics performance metrics. In *Annual Conference of the Prognostics and Health Management Society*, 2009.

[51] A. Saxena, J. Celaya, B. Saha, S. Saha, and K. Goebel. Metrics for offline evaluation of prognostics performance. *International Journal of Prognostics and Health Managemant*, 1, 2010.

[52] A. Saxena, J. Celaya, B. Saha, S. Saha, and K. Goebel. Evaluating prognostics performance for algorithms incorporating uncertainty estimates. In *Aerospace Conference, 2010 IEEE*, pages 1–11, March 2010.

[53] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, 2003.

[54] W. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[55] P. V. Yee and S. S. Haykin. *Regularised radial basis function networks: Theory and Applications*. Wiley, 2001.

[56] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, January 1982.

[57] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the USA*, 79:2554–2558, 1982.

[58] R. J. Schalkoff. *Artificial neural networks.* McGraw-Hill, 1997.

[59] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning representations by back-propagating errors*, pages 696–699. MIT Press, Cambridge, MA, USA, 1988.

[60] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[61] I. D. Dinov. Expectation maximization and mixture modeling tutorial. Technical report, Statistics Online Computational Resource (University of California, Los Angeles), 2008. Retrieved from: http://escholarship.org/uc/item/1rb70972.

[62] M. A. T. Figueiredo and A. K. Jain. Unsupervised selection and estimation of finite mixture models. In *Pattern Recognition. Proceedings. 15th International Conference on*, volume 2, pages 87–90, 2000.

[63] G. McLachlan and D. Peel. *Finite Mixture Models.* John Wiley & Sons, 2000.

[64] M. A. T. Figueiredo, J. M. N. Leitao, and A. K. Jain. On fitting mixture models. *Energy Minimization Methods in Computer Vision and Pattern Recognition*, 1654:732–732, 1999.

[65] G. J. Mclachlan and D. Peel. Mixfit: An algorithm for the automatic fitting and testing of normal mixture models. In *Proceedings of the 14th International Conference on Pattern Recognition*, pages 553–557. IEEE Computer Society, 1998.

[66] M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, March 2002.

[67] G. Celeux, S. Chrtien, F. Forbes, and A. Mkhadri. A component-wise em algorithm for mixtures. *Journal of Computational and Graphical Statistics*, 10:697–712, 2001.

[68] J. A. DeCastro, L. Tang, K. A. Loparo, and K. Goebel. Exact nonlinear filtering and prediction in process model-based prognostics. In *Annual Conference of the Prognostics and Health Management Society*, 2009.

[69] M. Orchard, G. Kacprzynski, K. Goebel, B. Saha, and G. Vachtsevanos. Advances in uncertainty representation and management for particle filtering applied to prognostics. In *Prognostics and Health Management, PHM 2008. International Conference on*, pages 1–6, October 2008.

[70] A. J. Huag. A tutorial on bayesian estimation and tracking techniques applicable to nonlinear and non-gaussian processes. Mitre technical report, The MITRE Corporation, 2005.

[71] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *Signal Processing, IEEE Transactions on*, 50(2):174 –188, February 2002.

[72] A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

[73] M. Orchard, B. Wu, and G. Vachtsevanos. A particle filtering framework for failure prognosis. In *World Tribology Congress III*, 2005.

[74] M. E. Tipping. The relevance vector machine. *Advances in Neural Information Processing Systems*, 12:652–658, 1999.

[75] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[76] M. E. Tipping and A. Faul. Fast marginal likelihood maximisation for sparse bayesian models. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

[77] P. H. Singer. Vacuum: A high-tech commodity? *Semiconductor International*, 1999.

[78] M. Mooney and G. Shelley. Data collection and networking capabilities enable pump predictive diagnostics. *Solid State Technology*, 48:84–94, 2005.

[79] P. H. Singer. Vacuum pumps now run hotter and better. *Semiconductor International*, 1998.

[80] S. Konishi and K. Yamasawa. Diagnostic system to determine the in-service life of dry vacuum pumps [used in lpcvd semiconductor fabrication facility. *Science, Measurement and Technology, IEE Proceedings*, 146(6):270–276, November 1999.

[81] J. A. Twiddle, N. B. Jones, and S. K. Spurgeon. Fuzzy model-based condition monitoring of a dry vacuum pump via time and frequency analysis of the exhaust pressure signal. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 222:287–293, 2008.

[82] Nigel S. Harris. *Modern Vacuum Practice.* McGraw-Hill, 2005.

[83] BOC Edwards. *Instruction Manual: iH Dry Pumping System.* BOC Edwards, 2005.

[84] Edwards Vacuum. Fabworks ims. Online, 2011.

[85] M. Miskowicz. Send-on-delta concept: An event-based data reporting strategy. *Sensors*, 6(1):49–63, 2006.

[86] B. Ji, J. H. Yang, P. R. Badowski, and E. J. Karwacki. Optimization and analysis of nf3 in situ chamber cleaning plasmas. *Journal of Applied Physics*, 95(8):4452–4462, 2004.

[87] A. D. Johnson, W. R. Entley, and P. J. Maroulis. Reducing pfc gas emissions from cvd chamber cleaning. *Solid State Technology*, 43:103–4, 106, 110, 112, 114, 2000.

[88] W. Cheung, J. Lim, K. Chung, and S. Lee. Patent application: A precision diagnostic method for the failure protection and predictive maintenance of a vacuum pump, June 2006.

[89] B. L. Bowerman and R. T. O'Connell. *Forecasting and Time Series.* Duxbury Press, 1993.

[90] G. Kacprzynski, M. Gumina, M. Roemer, D. Caguiat, T. Galie, and J. McGroart. A prognostic modeling approach for predicting recurring maintenance for shipboard propulsion systems. In *ASME Turbo Expo*, 2001.

[91] U.S. Climate Change Technology Program. Semiconductor industry: Abatement technologies: Technology options for the near and long term. Technical report, U.S. Climate Change Technology Program, 2003.

[92] M. B. Chang and J. S. Chang. Abatement of pfcs from semiconductor manufacturing processes by nonthermal plasma technologies: A critical review. *Industrial and Engineering Chemistry Research*, 45:4101–4109, 2006.

[93] L. Beu. Reduction of perfluorocompound (pfc) emissions: 2005 state-of- the-technology report. Technical report, International SEMATECH Manufacturing Initiative, 2005.

[94] World Semiconductor Council. Joint statement of the 14th meeting of the world semiconductor council, May 2010.

[95] Gilliland, Cummins, and Ridgeway. S69 evaluation of an edwards/alzeta thermal processing unit designed to abate pfcs. Technical report, SEMATECH Technology Transfer 95113010B-ENG, 1995.

[96] C. M. Christensen, S. King, M. Verlinden, and W. Yang. The new economics of semiconductor manufacturing. *Spectrum, IEEE*, 45(5):24 –29, May 2008.

[97] Edwards Vacuum. Edwards advanced diagnostic services. `http://www.ediag.edwardsvacuum.com/ads.htm`. Last Accessed: March, 2011.

[98] J Yu and S J Qin. Multimode process monitoring with bayesian inference-based finite gaussian mixture models. *AIChE Journal*, 54:1811–1829, 2008.

[99] S. W. Choi, J. H. Park, and I. B. Lee. Process monitoring using a gaussian mixture model via principal component analysis and discriminant analysis. *Computers and Chemical Engineering*, 28(8):1377 – 1387, 2004.

[100] J. A. Twiddle, S. K. Spurgeon, C. Kitsos, and N. B. Jones. A discrete-time sliding mode observer for estimation of auto-regressive model coefficients with an application in condition monitoring. In *Variable Structure Systems, VSS'06. International Workshop on*, pages 127 –132, June 2006.

[101] L. Tang, J. DeCastro, G. Kacprzynski, K. Goebel, and G. Vachtsevanos. Filtering and prediction techniques for model-based prognosis and uncertainty management. In *Prognostics and Health Management Conference, 2010. PHM '10*, pages 1–10, January 2010.

[102] S. Fulton and M. Kim. Ismi consensus preventive and predictice maintenance vision. Technical report, International Sematech Manufacturing Inititiative (ISMI), 2007.

[103] D. Brown, G. Georgoulas, B. Bole, H. L. Pei, M. Orchard, L. Tang, B. Saha amd A. Saxena, K. Goebel, and G. Vachtsevanos. Prognostics enhanced reconfigurable control of electro-mechanical actuators. In *Annual Conference of the Prognostics and Health Management Society*, 2009.

[104] B. Saha and K. Goebel. Uncertainty management for diagnostics and prognostics of batteries using bayesian techniques. In *Aerospace Conference, 2008 IEEE*, pages 1–8, March 2008.

[105] B. Saha and K. Goebel. Model adaptation for prognostics in a particle filtering framework. *International Journal of Prognostics and Health Management*, 2011.

[106] L. Tang, G. J. Kacprzynski, K. Goebel, and G. Vachtsevanos. Methodologies for uncertainty management in prognostics. In *Aerospace conference, 2009 IEEE*, pages 1–12, March 2009.

[107] World Wind Energy Association (WWEA). World wind energy report 2010. Technical report, World Wind Energy Association (WWEA), 2010.

[108] Z. Hameed, Y.S. Hong, Y.M. Cho, S.H. Ahn, and C.K. Song. Condition monitoring and fault detection of wind turbines and related algorithms: A review. *Renewable and Sustainable Energy Reviews*, 13(1):1 – 39, 2009.

[109] Bin Lu, Yaoyu Li, Xin Wu, and Zhongzhou Yang. A review of recent advances in wind turbine condition monitoring and fault diagnosis. In *Power Electronics and Machines in Wind Applications, 2009. PEMWA 2009. IEEE*, pages 1 –7, june 2009.

[110] R. W. Hyers, J. G. Mcgowan, K. L. Sullivan, J. F. Manwell, and B. C. Syrett. Condition monitoring and prognosis of utility scale wind turbines. *Energy Materials: Materials Science and Engineering for Energy Systems*, 1(3):187–203, 2006-09-01T00:00:00.

[111] R. F. Orshagh, H. Lee, M. Watson, and C. S. Byington. Advanced vibration monitoring for wind turbine health management, 2007.

[112] G. Swiszcz, A. Cruden, C. Booth, and W. Leithead. A data acquisition platform for the development of a wind turbine condition monitoring system. In *Condition Monitoring and Diagnosis, 2008. CMD 2008. International Conference on*, pages 1358 –1361, april 2008.

[113] S. Yang, W. Li, and C. Wang. The intelligent fault diagnosis of wind turbine gearbox based on artificial neural network. In *Condition Monitoring and Diagnosis, 2008. CMD 2008. International Conference on*, pages 1327 –1330, april 2008.

[114] W. Yang, P.J. Tavner, and M.R. Wilkinson. Condition monitoring and fault diagnosis of a wind turbine synchronous generator drive train. *Renewable Power Generation, IET*, 3(1):1 –11, march 2009.

[115] D. J. Lekou, F. Mouzakis, A. Anastasopoulus, and D Kourousis. Emerging techniques for health monitoring of wind turbine gearboxes and bearings. In *Proceedings of European Wind Energy Conference (EWEC)*, Marseille, France, March 2009.

[116] K. Schroeder, W. Ecke, J. Apitz, E. Lembke, and G. Lenschow. A fibre bragg grating sensor system monitors operational load in a wind turbine rotor blade. *Measurement Science and Technology*, 17(5):1167, 2006.

[117] O. Uluyol, G. Parthasarathy, W. Foslien, and K. Kim. Power curve analytic for wind turbine performance monitoring and prognostics. In *Annual Conference of the Prognostics and Health Management Society*, 2011.

[118] P. Caselitz and J. Giebhardt. Rotor condition monitoring for improved operational safety of offshore wind energy converters. *Journal of Solar Energy Engineering*, 127(2):253–261, 2005.

[119] A. Zaher, S.D.J. McArthur, D.G. Infield, and Y. Patel. Online wind turbine fault detection through automated scada data analysis. *Wind Energy*, 12(6):574–593, 2009.

[120] M. C. Garcia, M. A. Sanz-Bobi, and J. del Pico. Simap: Intelligent system for predictive maintenance: Application to the health condition monitoring of a wind-turbine gearbox. *Computers in Industry*, 57(6):552–568, 2006.

[121] P. Guo, D. Infield, and X. Yang. Wind turbine generator condition-monitoring using temperature trend analysis. *Sustainable Energy, IEEE Transactions on*, 3(1):124–133, January 2012.

[122] E. Wiggelinkhuizen, T. Verbruggen, H. Braam, L. Rademakers, M. C. Tipluica, A. Maclean, A. J. Christensen, E. Becker, and D. Scheffler. Conmow: Condition monitoring for offshore wind farms, 2007.

[123] E. Wiggelinkhuizen, T. Verbruggen, H. Braam, L. Rademakers, J. Xiang, and S. Watson. Assessment of condition monitoring techniques for offshore wind farms. *Journal of Solar Energy Engineering*, 130(3):031004, 2008.

[124] Z. I. Botev, J. F. Grotowski, and D. P. Kroese. Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5):2916–2957, 2010.