2010 Conference on Control and Fault Tolerant Systems
Nice, France, October 6-8, 2010

Particle Filters for Remaining Useful Life Estimation of Abatement Equipment used in Semiconductor Manufacturing

Shane Butler and John Ringwood

Abstract— **Prognostics is the ability to predict the remaining useful life of a specific system, or component, and represents a key enabler of any effective condition-based-maintenance strategy. Among methods for performing prognostics such as regression and artificial neural networks, particle filters are emerging as a technique with considerable potential. Particle filters employ both a state dynamic model and a measurement model, which are used together to predict the evolution of the state probability distribution function. The approach has similarities to Kalman filtering, however, particle filters make no assumptions that the state dynamic model be linear or that Gaussian noise assumptions must hold true.**

The technique is applied in predicting the degradation of thermal processing units used in the treatment of waste gases from semiconductor processing chambers. The performance of the technique demonstrates the potential of particle filters as a robust method for accurately predicting system failure.

In addition to the use of particle filters, Gaussian Mixture Models (GMM) are employed to extract signals associated with the different operating modes from a multi-modal signal generated by the operating characteristics of the thermal processing unit.

I. INTRODUCTION

Within the semiconductor manufacturing industry there is an increasing focus on process yields, tool uptime and wafer throughput. Furthermore, as the size of wafers increases alongside the introduction of new material technologies, the value of wafers is also increasing [1]. As a result, the impact of equipment failures resulting in tool downtime and loss of wafers is a major concern.

Regular Preventative Maintenance (PM) has long been a standard approach to maintenance in semiconductor manufacturing. However, as condition monitoring technology improves, Condition-Based Maintenance (CBM) approaches are being introduced alongside the more traditional PM approaches. A key component of any effective CBM program is prognostics.

Prognostics is the ability to estimate the Remaining-Useful-Life (RUL) of equipment and provide maintainers with notification to take corrective action in a timely and organised manner, significantly reducing both maintenance costs and impacts on tool uptime and availability.

Within semiconductor manufacturing, regular PM activities on process tool chambers will continue to be standard practice as the chambers require regular maintenance to maintain performance characteristics. Ideally, tool operators would like for any necessary maintenance on support equipment such as vacuum pumps, and abatement equipment be performed at the same time, to maximise overall tool availability [2]. Prognostics represents a key component in achieving this goal.

In this study, we address the issue of degradation of equipment used in the abatement of exhaust emissions from semiconductor processing chambers. Gaussian mixture models are used for feature extraction and multimode signal tracking. Particle filters are then employed to generate RUL estimates using the extracted features from the GMM as inputs.

The use of particle filters as a tool for prognostics has been increasing in recent years, and they have been applied to a range of applications, including Lithium-ion battery capacity depletion [3], turbine engine blade and gearbox plate crack growth prediction [4]. The attraction of particle filters is the framework provided for handling the significant levels of uncertainty inherent in the generation of long-term predictions.

II. PROBLEM DESCRIPTION

Typical semiconductor fabrication processes utilise a wide range of dangerous chemicals, which are often corrosive, toxic, and flammable. The fabrication processes will generally only consume a small proportion of the chemicals used within the processing chamber, which results in large quantities of chemicals and environmentally damaging greenhouse gases being discharged from the process chamber into the exhaust system.

A standard method for treating chamber effluent streams is to employ scrubbers for the treatment of effluent gases [5]. However, on particularly harsh processes, such methods are often not sufficient to remove the more toxic elements of the exhaust stream and additional treatment is required. One method is to incinerate the more toxic effluent process gases to oxidise the toxic materials and reduce their toxicity. The oxidised effluent streams are then forwarded to the large scrubbers for further treatment. The equipment used to incinerate the exhaust gas streams is commonly known as a Thermal Processing Unit (TPU).

A issue that can occur with the use of TPUs on certain processes is the generation of silicon oxides in the combustion process, which can deposit on the walls of the combustion chamber. The deposits formed can be relatively large and, as the deposition gradually increases, this can result in reduced

This work was supported by Enterprise Ireland under grant EI/CTFD/05/IT/323, and the NUI Maynooth Postgraduate Travel Fund.

Shane Butler and John Ringwood are with the Department of Electronic Engineering, National University of Ireland (Maynooth), Maynooth, Co. Kildare, Ireland. shane.butler@eeng.nuim.ie, john.ringwood@eeng.nuim.ie

The authors would like to acknowledge the support of Edwards Vacuum (formerly BOC Edwards) in carrying out this work.

combustion of the effluent stream and eventually clog the combustion chamber.

Shown in Fig. 1(a) is a plot of the Combustor Temperature (CT) signal from a TPU which suffered from clogging of the combustion chamber. Shown below it in Fig. 1(b) is a zoomed in section from the first three days of data shown in Fig. 1(a).

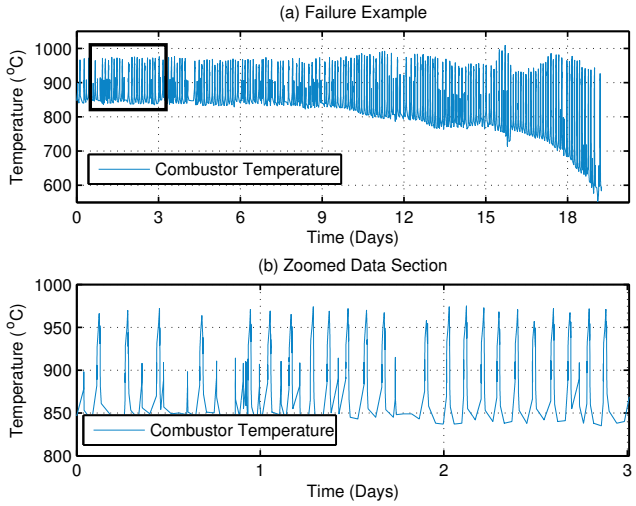

Fig. 1. TPU Failure Example

The clogging of the combustion chamber is reflected in the CT signal by a gradual loss of temperature as the level of deposits increase. Another feature of the CT signal is the large fluctuations in temperature observed over certain periods as seen in Fig. 1(b). This is as a result of how the TPU system is operated. Wafer processing generally occurs intermittently as batches of wafers arrive and leave each tool. In a typical TPU combustion chamber, free of deposits, the temperature is maintained using methane gas at approximately 850^o C. However, during processing of wafers when large quantities of gas are being discharged to the TPU, a signal is sent to the TPU unit which results in oxygen or oxygen-enriched gas being injected into the combustion chamber to increase the temperature in order to ensure complete combustion of the effluent stream. This results in the large fluctuations in CT observed in the data.

In the following sections, the issue of TPU combustion chamber clogging is addressed using both Gaussian mixture models to extract relevant features from the multi-modal CT signal, and particle filters to estimate the RUL of the system using the extracted features as inputs.

A. Data Collection

In this study, a data set from several TPUs installed in a large semiconductor manufacturing facility was employed to develop and test a particle filter approach to RUL estimation. Each of the TPUs in the facility are connected to a networked monitoring system, which also monitors all of the mechanical dry pumps within the facility. Each piece of equipment on the monitoring network sends updates on sensor values and status to a central database for further processing and storage [1].

In addition to the sensor data, a status signal indicating those times at which oxygen is being injected into the TPU combustion chamber (Processing Mode) or otherwise (Idling Mode) is recorded by the networked monitoring system.

III. MULTI-MODE SIGNAL TRACKING

Within a typical TPU system, only the CT signal, provides any indication that deposits are forming within the combustion chamber. The method in which the TPUs are operated results in a the generation of a multi-modal CT signal with the different signal modes corresponding to the different operating modes of the TPU. Fig. 2 plots the distribution of CT values over a three day period from a normally operating TPU.

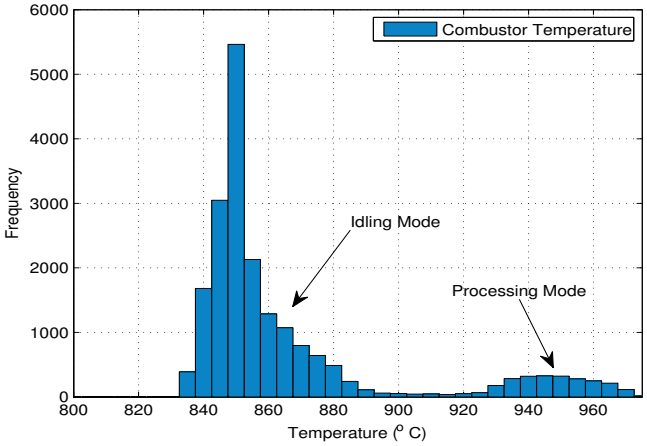

Fig. 2. Distribution of CT Values

The distribution of the CT signal reflects the different operating modes of the TPU system associated with the injection of oxygen gas into the combustion chamber. It is proposed to model the underlying distribution as a mixture of univariate Gaussian distributions. By employing a sliding window and iterating through the data, it will allow for changes in the distribution of the CT signal in each of the TPU operating modes to be tracked over time. At each iteration, the best fit mixture of Gaussian distributions is fit to the CT signal within each window. The length of the window is considered in section III-C.

A. Gaussian Mixture Models

To track the CT signal in each of the modes of the TPU system, it is proposed to use a sliding window, where at each iteration the best-fit mixture of Gaussian distributions is determined which best describes the underlying distribution of the CT signal within each window.

In modelling the underlying distribution of the data we consider it as a superposition of K Gaussian densities such that [6],

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|\theta_k) \quad (1)$$

where the k values refer the individual densities, π_k is the discrete probability that a point is sampled from density k, and is commonly referred to as the mixture parameter. The

individual component densities are given by $p(x|\theta_k)$ where [6],

$$p(x|\theta_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} \quad (2)$$

is the probability that x takes on a certain value given that is from density k, and mu_k and σ_k^2 are the mean and variance of the individual densities respectively. The probabilities are also subject to a number of constraints [6] such that,

$$\sum_{k=1}^{K} \pi_k = 1 \quad where \quad 0 \leq \pi_k \leq 1 \quad (3)$$

The problem is now to estimate the parameters of the model Θ, which include the mixture parameters, mean, and variance of each of the individual densities.

$$\Theta = \{\{\pi_1, \mu_1, \sigma_1\}, ...\{\pi_k, \mu_k, \sigma_k\}\} \quad (4)$$

The most common approach to estimating parameters of the model (4) is the Expectation-Maximisation (EM) algorithm [6]. It is an iterative procedure to find the maximum likelihood estimates of the model parameters. However, the EM algorithm suffers from the requirement that the user must specify the number of components and does not automatically adjust this number to fit the data. In this work, the Figueiredo-Jain (F-J) algorithm was employed which automatically optimises the number of components to fit the data, and estimates the statistical parameters of the model using a modified version of the EM algorithm [7]. This method was used to determine the parameters of the Gaussian mixture model at each iteration of the sliding window.

B. Feature Extraction

The TPU systems are operated in two distinct modes, idling and processing. This would suggest that we would expect the underlying distribution of the CT signal to be a mixture of two Gaussian densities. However, analysis of the distribution using the F-J algorithm, identifies the underlying distribution as a mixture of three Gaussian densities. This is illustrated in Fig. 3(a), which shows the distribution of the CT signal, and Fig. 3(b) which shows the best fit mixture of three Gaussian densities identified using the F-J algorithm, which best model the underlying distribution.

This additional mode is as a result of two principal factors:

1) periods where oxygen gas is injected for short periods of time and the times when no gas is being injected (idling) and,
2) the temperature signal is falling from an immediately preceding period of high values (processing).

This additional mode will be referred to as the transition mode.

The availability of a status signal indicating the injection of oxygen into the combustion chamber is of significant benefit as it allows the data to be first partitioned by mode in a supervised manner. If the CT signal were simply bi-modal then we could use the partitioned data to track the CT values by mode. However, analysis has shown that the data is in fact tri-modal as seen in Fig. 3.

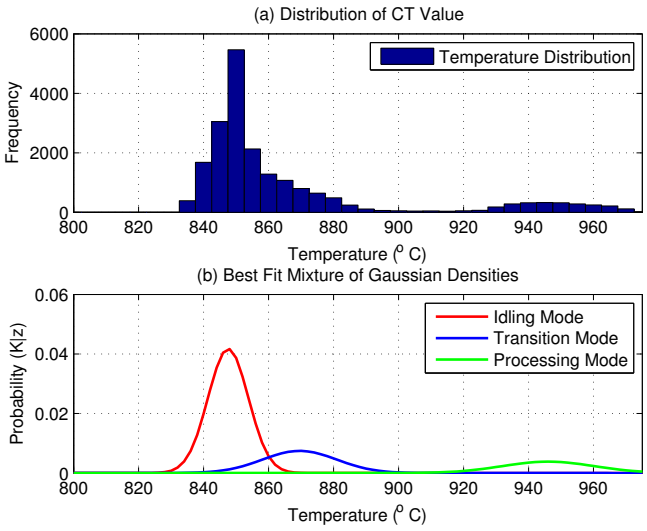

Fig. 3. Distribution of CT Values by Mode

An additional benefit of the status signal, which allows supervised partition of the data, is that wafer production does not occur in a regular predictable manner. In the absence of a status signal, the ability to identify different distributions associated with the different operating modes within each window would be influenced by the rate of wafer processing within each respective window which can vary widely, affecting the generation of associated distributions.

By using the status signal, we separate the CT signal into values associated with idling and processing. Within each of these modes we then model the associated data as a mixture of two Gaussian densities. This is illustrated for the idling mode in Fig. 4 below. Fig. 4(a) shows the distribution of CT values over a three day period at which no gas was being injected into the chamber and 4(b) shows the best fit mixture of two Gaussians as identified by the F-J algorithm, representing the idling and transition modes in the underlying distribution.

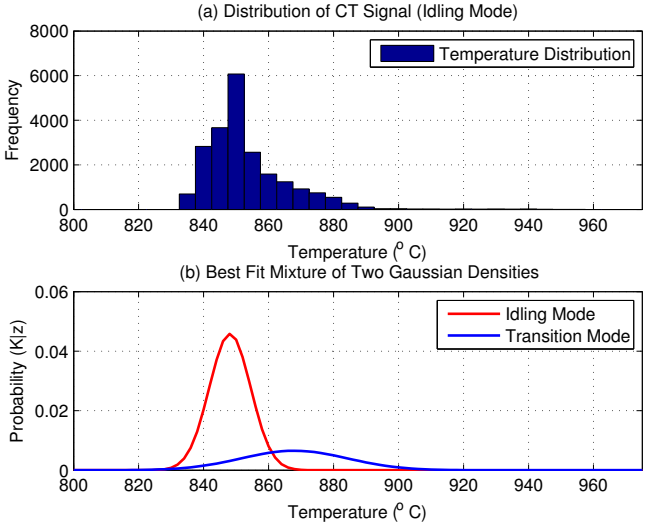

Fig. 4. Distribution of CT Values (Idling)

The distribution of the data within the processing mode for

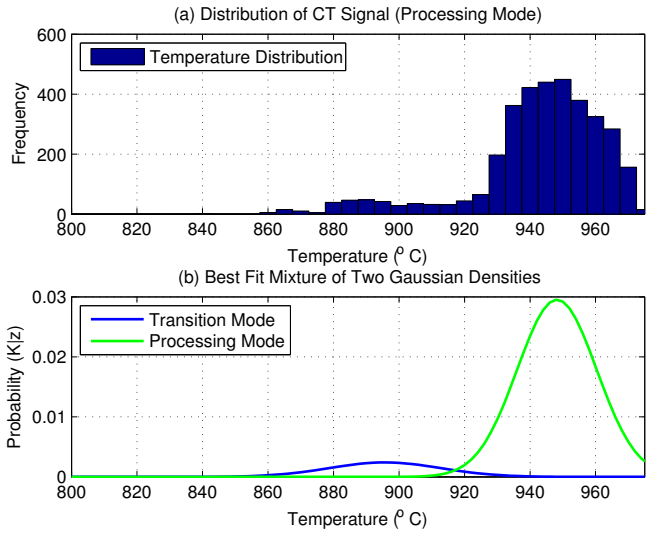

Fig. 5. Distribution of CT Values (Processing)

data from the same period is shown in Fig. 5. Once again it is a combination of two Gaussian densities, generated by the processing and transition modes.

C. TPU Mode Tracking

Analysis of the failure mode of TPUs, suffering from clogging of the combustion chamber, shows that the temperature values associated with the idling mode gradually fall as the level of clogging increases. To track how the CT signal values associated with the idling mode change over time we first identify all those samples at which no gas is being injected into the chamber. We then employ a sliding window, whereby at each iteration, the best-fit mixture of two Gaussian densities is determined and the mean value of the lower density represents the value of the CT signal at that instant. The window iterates through the data, resulting in a moving signal which tracks the changes in the distribution of values in the idling mode. A number of factors which affect the performance of the algorithm have to be considered.

The length of window over which the underlying distribution is a major factor to consider. The longer the window, the greater the number of samples present and the greater the likelihood that we are identifying the true underlying distribution. However, if the window is too long, it will not respond quickly enough to the changes in the underlying distribution as the TPU starts to clog. A range of values were considered and a sliding window with a length of 4.5 hours was chosen. This window length is also consistent with existing TPU signal tracking techniques using window methods [8]. The sliding window iterated through the data in steps of 10 minutes.

The performance of the algorithm in tracking the CT values associated with the idling mode is shown in Fig. 6. Each value in the mode tracking signal represents the mean of the lower Gaussian density estimated by the sliding window up to that point. The tracking of this signal represents a robust and reliable method for tracking the clogging of the TPU combustion chamber, from the raw, multi-modal CT signal. In the following section, we develop a prognostic approach

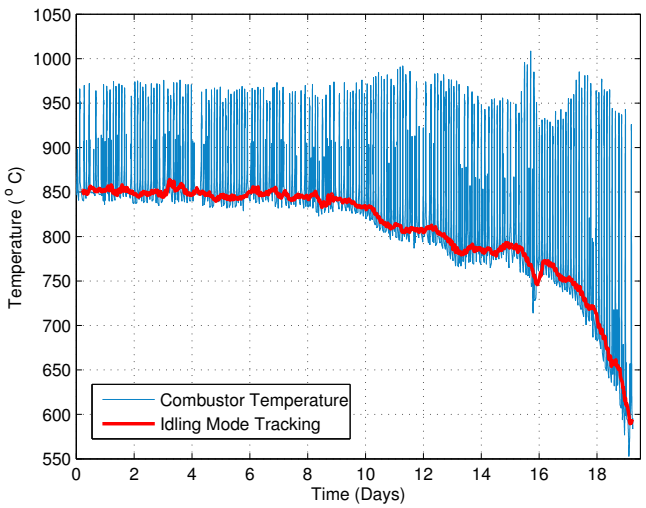

Fig. 6. TPU Mode Tracking (Idling)

to estimating the RUL of a TPU using the extracted idling mode signal as input.

IV. PARTICLE FILTERS FOR PROGNOSTICS

Prognostics is the ability to predict accurately and precisely the RUL of a failing system. It involves the generation of long-term predictions describing the evolution of a particular fault mode or condition. Inherent in the generation of such predictions is a large element of uncertainty which must be handled appropriately.

To perform accurate prognostics, and provide a long term prediction of equipment failure, two important conditions must be satisfied: A model which describes the progression of the fault condition and an accurate estimate of the current state. In this study, we consider the use of particle filters to estimate the RUL of a degrading TPU. Particle filters employ a state dynamic model and a measurement model to predict the posterior Probability Distribution Function (PDF) of the system state. The appeal of particle filters is that they avoid the linearity and Gaussian noise assumptions associated with Kalman filtering and provide a robust framework for long-term prognosis while accounting effectively for uncertainties [4].

Particle filters are a class of Sequential Monte Carlo (SMC) methods that use both information available from system measurements, but also incorporate any system models available which describe the system behaviour. The use and application of particle filters for prognostics has developed in recent years and is growing in acceptance as a providing an appropriate framework for handling the various sources of uncertainty which arises in estimating the RUL of a system [4], [9], [3].

Using particle filters, the system state PDF is approximated by a set of particles which represent sampled values from the unknown state space, and an associated set of weights which represent the discrete probability mass for each particle. The set of particles are recursively updated using a non-linear process model, a measurement model, measurement updates and an *a priori* estimate of the state

PDF. In this study, we use particle filters to estimate the RUL of a degrading TPU. The approach taken comprises two steps; state estimation and RUL prediction.

A. State Estimation

The principle of particle filtering is the approximation of the conditional state probability distribution, $p(x_k|z_k)$ by a set of samples or "particles" with a set of corresponding weights, representing the discrete probability masses. Particles are generated from an initial estimate of the state PDF $p(x_0)$ and are recursively updated using a nonlinear process model (5) which describes the evolution of the system under investigation, and a measurement model (6) which uses a set of available measurements $z_{1:k} = (z_1, ...z_k)$.

$$x_k = f_k(x_{k-1}, \omega_k) \leftrightarrow p(x_k|x_{k-1}) \tag{5}$$

$$z_k = h_k(x_k, \upsilon_k) \leftrightarrow p(z_k|x_k) \tag{6}$$

where f_k is a possibly nonlinear function describing the state evolution, h_k is a possibly nonlinaer function and ω and υ represent the process and measurement noise sequences respectively.

Considering the problem from a Bayesian perspective, the objective is to recursively calculate some degree of belief in the state x_k at time k, taking different values, given the data $z_{1:k}$ up to time k, and to construct a conditional state PDF $p(x_k|z_{1:k})$. As in any Bayesian estimation problem, the estimation process comprises two main steps; prediction and update.

In the prediction step, the knowledge of both the previous state estimate and the process model in (5) is used to generate an *a priori* estimate of the state PDF for the next time instant [10] as shown in (7).

$$p(x_k|z_{1:k-1}) = \int p(x_k|x_{k-1})p(x_{k-1}|z_{1:k-1})dx_{k-1} \tag{7}$$

The next stage is the filtering step. At time k, a measurement z_k becomes available, and this is combined with the *a priori* state estimate to generate the *a posteriori* state PDF using Bayes' rule [10], as

$$p(x_k|z_{1:k}) = \frac{p(z_k|x_k p(x_k|z_{1:k-1}))}{p(z_k|z_{1:k-1})} \tag{8}$$

The actual distributions are approximated by a set of samples and a set of corresponding normalised weights. Consider the dataset $\{x_k^i, w_k^i\}_{i=1}^{N_s}$ which characterises the posterior PDF $p(x_k|z_{1:k})$, where x_k is the set of sample points with associated weights w_k. The weights are normalised such that $\sum_i w_k^i = 1$. The posterior density at time k can then be approximated [10], by

$$p(x_k|z_{1:k}) \approx \sum_{i=1}^{N} \tilde{w_k}(x_{0:k}^i).\delta(x_{0:k} - x_{0:k}^i) \tag{9}$$

The weights are chosen using the principle of *importance sampling* [10], such that the weight update equation is given by,

$$w_k = w_{k-1} \frac{p(z_k|x_k)p(x_k|x_{k-1})}{q(x_k|x_{0:k-1}, z_{1:k})} \tag{10}$$

where the importance density function $q(x_k|x_{0:k-1}, z_{1:k})$ can be approximated by the *a priori* PDF for the state [9], $p(x_k|x_{k-1})$, such that w_k becomes,

$$w_k = w_{k-1} p(z_k|x_k) \tag{11}$$

A final issue which must be considered is that of particle degeneracy. As the algorithm iterates, the variance of the weights continually increases so that after a few iterations all but one of the particles will have negligible weights. To overcome this problem, a common method is to perform *resampling* of the weights.

Resampling is performed when the effective number of particles $P_{eff} < P_{threshold}$, where P_{eff} is computed as the inverse of the sum of the squared normalised particle weights in (10). The resampling operation is carried out by selecting a new set of P particles from the current set of particles, where the probability of selecting a particle is proportional to its current weight. The old set of particles is then replaced with the new resampled set, and each particle is assigned an equal weighting, given by $1/P$.

B. Prediction

To estimate the RUL of a system using particle filters, a number of approaches are possible. A thorough description of the different approaches is presented by Orchard [4]. The simplest approach is to extend the trajectories of the individual particles $\hat{x}_{0:t+p}^i$ for the required number of p steps into the future as in (12), where the current state value estimate associated with each particle is used as the initial condition,

$$\tilde{p}_{t+p}^{(i)} = E[f_{t+p}(\tilde{x}_{t+p-1}^{(i)}, \omega_{t+p})]; \tag{12}$$

where $\tilde{p}_{t+p}^{(i)}$ is the predicted value of particle i at time p, E is the expectation and f_{t+p} is the process model in (5) which is recursively calculated to estimate particle values at time p. The current particle weights are propagated in time without any change.

Presented in [4] are two alternative methods which involve applying update equations to the particle weights or *resampling* of the weights. However, it is noted that the error generated by considering the particle weights invariant for future time instants is negligible with respect to other sources of error which may appear in practical applications, such as model inaccuracies or measurement and process noise assumptions [4], and that the simplified method provides satisfactory performance.

To generate an estimate of the RUL of the TPU system, a specific hazard zone must be specified within which it is expected that the probability of equipment failure is very high. For the current application, the clogging of the TPU combustion chamber is well reflected by the loss of temperature. A specific temperature range is specified with upper (H_{ub}) and lower bounds (H_{ub}) which represents the hazard zone. This range is defined from both historical failure analysis and input from equipment maintainers.

Once the hazard zone has been specified, it is possible to combine the predicted particle trajectories, their weights and

the specified hazard zone to generate the system RUL PDF as in (13),

$$p_{TTF}(TTF) = \sum_{i=1}^{N} Pr\{H_{lb} \leq x_{TTF}^{(i)} \leq H_{ub}\} \tilde{w}_{TTF}^{(i)} \qquad (13)$$

where $p_{TTF}(TTF)$ is the probabiity of system failure at time TTF, $\tilde{w}_{TTF}^{(i)}$ is the weight of particle i at time TTF, and $x_{TTF}^{(i)}$ is the predicted value of particle i at time TTF.

V. RESULTS

The algorithm was tested on a number of historical failures. For this study, the state transition model described by 14 was used to model the evolution of the chamber clogging process. The model was adapted from an ageing model for battery Lithium-ion battery cells [11] which exhibit similar degradation characteristics to the TPU clogging process.

$$x_{k+1} = x_k - \beta_1 \frac{exp(\frac{\beta_2}{t_k})}{t_k^2} - \beta_3 exp(\beta_4 t_k) \qquad (14)$$

where x is the combustor temperature signal value in the idling mode and β_1,β_2, β_3 and β_4 are model parameters which were estimated by performing a curve fitting exercise on historical failure examples. Both the measurement and process noise variance ω_k and υ_k respectively were modeled as Gaussian densities.

An example of the performance of the particle filter approach is illustrated in Fig. 7(a). The hazard zone for the TPU systems was specified as between 640^oC and 660^oC. The figure shows the mean value of the state estimate and the upper and lower limits of the estimated state PDF at each sample time. A prediction of the RUL of the TPU system was performed once the state PDF reached 775^o. The upper and lower bounds on the predicted particle trajectories is shown by the shaded region. Shown in Fig. 7(b) is the predicted RUL PDF of the TPU system calculated from (13).

VI. CONCLUSIONS

The use of particle filters for prognostics is continually growing in acceptance as a suitable technique for handling the significant levels of uncertainty associated with the generation of long-term predictions. In addition, the generation of a RUL PDF and associated confidence intervals is a natural component of the approach, which is lacking in the use of alternate prognostic approaches such as artificial neural networks. In this study we have illustrated how the particle filter approach can provides accurate and actionable estimates of the RUL, with sufficient lead time provided to maintenance personnel to take corrective action in an organised and timely manner, reducing instances of equipment failure. In addition, by simply altering the model parameters and hazard zone specifications, the approach can be applied to TPU systems operating on different processes with different failure characteristics.

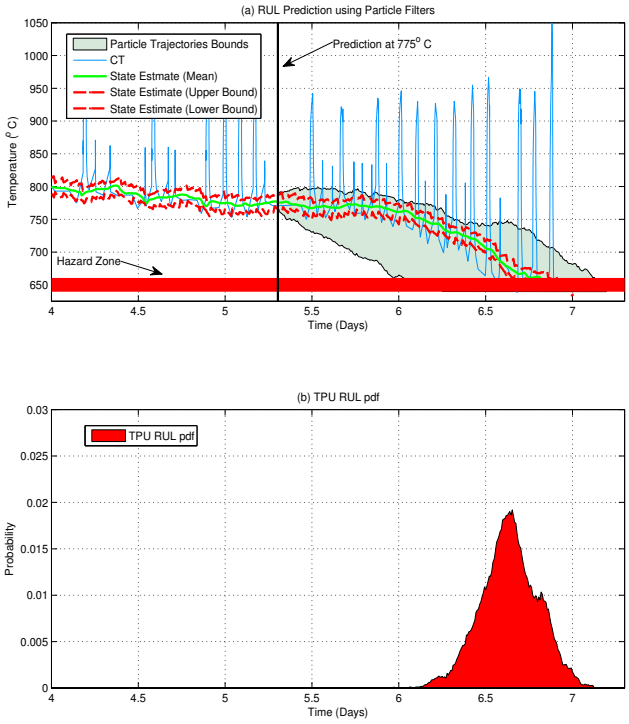

Fig. 7. Estimation of TPU RUL

REFERENCES

[1] M. Mooney and G. Shelley, "Data collection and networking capabilities enable pump predictive diagnostics," *Solid State Technology*, vol. 48, pp. 49–63, 2005.

[2] X. Yao, E. Fernandez-Gaucherand, M. Fu, and S. Marcus, "Optimal preventive maintenance scheduling in semiconductor manufacturing," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 17, no. 3, pp. 345 – 356, Aug. 2004.

[3] B. Saha, K. Goebel, S. Poll, and J. Christophersen, "Prognostics methods for battery health monitoring using a Bayesian framework," *Instrumentation and Measurement, IEEE Transactions on*, vol. 58, no. 2, pp. 291 –296, Feb 2009.

[4] M. Orchard, "A particle filtering-based framework for on-line fault diagnosis and failure prognosis," Ph.D. dissertation, Georgia Institute of Technology, 2007.

[5] C. Chang and S. Sze, *ULSI Technology*. Secaucus, NJ, USA: Mcgraw-Hill College, 1996.

[6] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[7] M. Figueiredo and A. Jain, "Unsupervised learning of finite mixture models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 3, pp. 381 –396, Mar 2002.

[8] M. Mooney, "EADS abatement model requirement specification," Edwards Vacuum, Tech. Rep., 2008.

[9] 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*, vol. 31, no. 3-4, pp. 221–246, 2009.

[10] M. 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*, vol. 50, no. 2, pp. 174 –188, Feb 2002.

[11] R. L. H. II, "An aging model for lithium-ion cells," Ph.D. dissertation, University of Akron, 2008.