Transactions of the Institute of Measurement and Control 31, 3/4 (2009) pp. 221–246

A particle-filtering approach for on-line fault diagnosis and failure prognosis

Marcos E. Orchard[1,2] and George J. Vachtsevanos[1]

[1]School of Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0250, USA
[2]Electrical Engineering Department, University of Chile, Av. Tupper 2007, Santiago, Chile

This paper introduces an on-line particle-filtering (PF)-based framework for fault diagnosis and failure prognosis in non-linear, non-Gaussian systems. This framework considers the implementation of two autonomous modules. A fault detection and identification (FDI) module uses a hybrid state-space model of the plant and a PF algorithm to estimate the state probability density function (pdf) of the system and calculates the probability of a fault condition in real-time. Once the anomalous condition is detected, the available state pdf estimates are used as initial conditions in prognostic routines. The failure prognostic module, on the other hand, predicts the evolution in time of the fault indicator and computes the pdf of the remaining useful life (RUL) of the faulty subsystem, using a non-linear state-space model (with unknown time-varying parameters) and a PF algorithm that updates the current state estimate. The outcome of the prognosis module provides information about the precision and accuracy of long-term predictions, RUL expectations and 95% confidence intervals for the condition under study. Data from a seeded fault test for a UH-60 planetary gear plate are used to validate the proposed approach.

Key words: failure prognosis; fault detection; fault identification; particle filtering.

1. Introduction

Critical aircraft assets (exhibiting attributes of reliability, robustness and high confidence under a variety of flight regimes) are required to be available when needed, and maintained on the basis of their current condition rather than on the basis

Address for correspondence: Marcos E. Orchard, School of Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0250, USA.
E-mail: marcos.orchard@gmail.com
Figures 1–8 appear in colour online: http://tim.sagepub.com

 10.1177/0142331208092026

of scheduled maintenance practices. Moreover, condition-based maintenance (CBM) requires that the health of critical components/systems be monitored and diagnostic/prognostic strategies be developed to detect and identify incipient failures – fault detection and identification (FDI) – and predict the remaining useful life (RUL) of the failing component. New and innovative technologies must be developed and implemented to address these concerns.

The complexity of the problem indicates that it is appropriate to combine model-based and state-estimation techniques to implement on-line FDI/prognostic approaches. In this sense, recursive Bayesian algorithms are well suited to solve the problem of real-time estimation since they incorporate process data (in the form of sequential observations) into the *a priori* state estimate by considering the likelihood of measured values (Doucet *et al*., 2001). Particularly, sequential Monte Carlo (SMC) methods – also referred to as particle filtering (PF) – provide a solid and consistent theoretical framework to handle model non-linearities or non-Gaussian process/observation noise. Founded on the concept of sequential importance sampling (SIS) and Bayesian theory, PF has been the subject of an intensive amount of research over the past years in many diverse disciplines including economics, biostatistics and statistical signal processing problems in the engineering domain such as time series analysis, target tracking and communications (Arulampalam *et al*., 2002).

The underlying principle of the methodology is the approximation of the conditional state probability distribution $p(x_{0:t} \mid y_{0:t})$ by a swarm of points called 'particles'. These particles contain samples from the state-space and a set of weights – associated with them – representing discrete probability masses. Particles can be easily generated and recursively updated given a non-linear process model (which describes the evolution in time of the system under analysis), a measurement model, a set of available measurements $Y = \{Y_t, t \in \mathbb{N}\}$, and an initial estimation for the state probability density function (pdf), $p(x_0)$. Furthermore, PF allows information from multiple measurement sources to be fused in a principled manner, which is an attribute of decisive significance for fault detection/diagnosis purposes.

Although several applications of PF for FDI may already be found in literature (de Freitas, 2002; Kadirkamanathan *et al*., 2002; Koutsoukos *et al*., 2002; Li and Kadirkamanathan, 2001; Verma *et al*., 2003, 2004), little work has been done in the prognosis arena. In this sense, this paper introduces a general framework where hybrid-state dynamic models are used to represent the behaviour of the system under no-fault and faulty operating conditions, and real-time PF algorithms are utilized to estimate the state pdf. These pdf estimates directly indicate the probability of each faulty mode and are also used as initial conditions in a two-level procedure for failure prognosis.

The organization of the paper is as follows. Section 2 provides the theoretical background for Bayesian estimation and PF and it also indicates the *state-of-the-art* for the application of these methods in FDI and prognosis. Section 3 introduces the proposed approach for on-line FDI and presents the results obtained for a real application example. Section 4 focuses on the prognosis issue, provides the theoretical

foundation, shows an illustrative example and analyses the results obtained for the prediction of axial crack growth in an UH-60 planetary carrier plate. Main conclusions and final remarks are stated in Section 5.

2. Theoretical background

2.1 Bayesian estimation and PF

Non-linear filtering is the process of using noisy observation data to estimate at least the first two moments of a state vector governed by a dynamic non-linear, non-Gaussian state-space model (Haug, 2005). Although in principle the estimation procedure may be implemented on continuous-time systems, the present paper is solely focused on discrete-time systems since the streaming measurement data is sent (and received) through digital devices in most of the applications relevant to FDI and prognosis.

Mathematically speaking, let $X = \{X_t, t \in \mathbb{R}\}$ be a $\mathbb{R}^{n_x}$-valued Markov process characterized both by its initial distribution $p(x_0)$ and the transition probability $p(x_t \mid x_{t-1})$. Moreover, let $p(x_t \mid x_{t-1})$ be defined by (1), where $\{\omega_t\}_{t\geq 0}$ is a sequence of independent random variables, not necessarily Gaussian.

$$x_t = f_t(x_{t-1}, \omega_t) \tag{1}$$

Noisy observations $Y = \{Y_t, t \in \mathbb{N}\}$ are assumed to be conditionally independent, given $X = \{X_t, t \in \mathbb{N}\}$. Equation (2) defines the marginal distribution $p(y_t \mid x_t)$, where $\{\upsilon_t\}_{t\geq 0}$ is a sequence of independent random variables.

$$y_t = g_t(x_t, \upsilon_t) \tag{2}$$

Let $x_{0:t} \triangleq \{x_0, \ldots, x_t\}$ and $y_{1:t} \triangleq \{y_1, \ldots, y_t\}$ denote, respectively, the signal and the observations up to time t. It is of interest to estimate the *posterior distribution* $p(x_{0:t} \mid y_{1:t})$, the marginal distribution $p(x_t \mid y_{1:t})$, and the expectations (3) for any function $f_t : \mathbb{R}^{n_x} \to \mathbb{R}^{n_{f_t}}$ integrable with respect to $p(x_{0:t} \mid y_{1:t})$, (Doucet *et al.*, 2001).

$$I(f_t) = \mathrm{E}_{p(x_{0:t}|y_{1:t})}[f_t(x_{0:t})] \triangleq \int f_t(x_{0:t}) p(x_{0:t}|y_{1:t}) dx_{0:t} \tag{3}$$

This task can be basically achieved by performing two sequential steps, namely *prediction* and *filtering* (Arulampalam *et al.*, 2002). On one hand, *prediction* uses both the knowledge of the previous state estimate and the process model to generate the *a priori* state pdf estimate for the next time instant:

$$p(x_{0:t}|y_{1:t-1}) = \int p(x_t|x_{t-1}) p(x_{0:t-1}|y_{1:t-1}) dx_{0:t-1} \tag{4}$$

On the other hand, the *filtering* step, which can be implemented by using the recursion formula (5), generates the *posterior* state pdf, by using Bayes formula:

$$p(x_{0:t}|y_{1:t}) \propto p(y_t|x_t) \cdot p(x_t|x_{0:t-1}) \cdot p(x_{0:t-1}|y_{1:t-1}) \tag{5}$$

Expressions (3), (4) and (5) do not have analytical solution in most cases. In this sense, SMC algorithms (particle filters) make feasible their evaluation through the use of efficient sampling strategies (Arulampalam *et al.*, 2002; Doucet *et al.*, 2000).

2.2 SMC methods: PF

Consider a sequence of probability distributions $\{\pi_t(x_{0:t})\}_{t\geq 1}$, where it is assumed that $\pi_t(x_{0:t})$can be evaluated pointwise up to a normalizing constant. SMC methods, also referred to as particle filters, are a class of algorithms designed to approximately obtain samples sequentially from $\{\pi_t\}$, ie, to generate a collection of $N \gg 1$ weighted random samples $\{w_t^{(i)}, x_{0:t}^{(i)}\}_{i=1,\ldots,N}$, $w_t^{(i)} \geq 0, \forall t \geq 1$, satisfying (Andrieu *et al.*, 2001):

$$\sum_{i=1}^{N} w_t^{(i)} \varphi_t(x_{0:t}^{(i)}) \underset{N\to\infty}{\longrightarrow} \int \varphi_t(x_{0:t}) \pi_t(x_{0:t}) dx_{0:t}, \tag{6}$$

where φ_t is any π_t–integrable function.

In the particular case of the Bayesian Filtering problem, the *target distribution* $\pi_t(x_{0:t}) = p(x_{0:t} \mid y_{1:t})$ is the *posterior* pdf of $X_{0:t}$, given a realization of noisy observations $Y_{1:t} = y_{1:t}$. Using (1) and (2), $\pi_t(x_{0:t})$ may be written as (Doucet *et al.*, 2000)

$$\pi_t(x_{0:t}) = p(x_0) \prod_{k=1}^{t} f_k(x_k|x_{k-1}) g_k(y_k|x_k) \tag{7}$$

Let a set of N paths $\{x_{0:t-1}^{(i)}\}_{i=1,\ldots,N}$ be available at time $t-1$. Furthermore, let these paths distribute according to $q_{t-1}(x_{0:t-1})$, also referred to as the *importance* density function at time $t-1$. Then, the objective is to efficiently obtain a set of N new paths (particles) $\{\tilde{x}_{0:t}^{(i)}\}_{i=1,\ldots,N}$ approximately distributed according to $\pi_t(\tilde{x}_{0:t})$ (Andrieu *et al.*, 2001).

For this purpose, the current paths $x_{0:t-1}^{(i)}$ are extended by using the kernel $q_t(\tilde{x}_{0:t}|x_{0:t-1}) = \delta(\tilde{x}_{0:t-1} - x_{0:t-1}) \cdot q_t(\tilde{x}_t|x_{0:t-1})$, ie, $\tilde{x}_{0:t} = (x_{0:t-1}, \tilde{x}_t)$. The *importance sampling* procedure generates consistent estimates for (3), by approximating (7) with the empirical distribution (Andrieu *et al.*, 2001)

$$\tilde{\pi}_t^N(x_{0:t}) = \sum_{i=1}^{N} w_{0:t}^{(i)} \delta(x_{0:t} - \tilde{x}_{0:t}^{(i)}) \tag{8}$$

where $w_{0:t}^{(i)} \propto w_{0:t}(\tilde{x}_{0:t}^{(i)})$ and $\sum_{i=1}^{N} w_{0:t}^{(i)} = 1$.

The most basic SMC implementation – the SIS particle filter – computes the value of the particle weights $w_{0:t}^{(i)}$, by setting the importance density function equal to the

a priori pdf for the state, ie, $q_t(\tilde{x}_{0:t}|x_{0:t-1}) = p(\tilde{x}_t|x_{t-1}) = f_t(\tilde{x}_t|x_{t-1})$. In that manner, the weights for the newly generated particles are evaluated from the likelihood of new observations. The efficiency of the procedure improves as the variance of the importance weights is minimized. The choice of the importance density function is critical for the performance of the particle filter scheme and hence, it should be considered in the filter design.

2.3 Resampling step: SIR particle filter

One of the main difficulties that must be addressed in the implementation of SIS particle filters is the degeneracy problem (Doucet, 1998) since, after a few iterations, all but one particle have a negligible weight (Andrieu *et al*., 2001; Arulampalam *et al*., 2002; Doucet *et al*., 2000). Several authors have proposed methods to overcome this problem (Kong *et al*., 1994; Liu, 1996) measuring the degeneracy in the particle population with $\hat{N}_{eff}$, an estimate of the effective sample size N_{eff} (Doucet *et al*., 2000).

$$N_{eff} = N \cdot \left(1 + \mathrm{var}_{\pi(\cdot|y_{0:t})}(w_{0:t})\right)^{-1}, \quad \hat{N}_{eff} = \left(\sum_{i=1}^{N} (w_t^{(i)})^2\right)^{-1} \tag{9}$$

Whenever $\hat{N}_{eff} \leq N_{thres}$, a fixed threshold, a resampling algorithm (Arulampalam *et al*., 2002; Doucet *et al*., 2000; Pitt and Shephard, 1999; Van der Merwe *et al*., 2006) is performed to eliminate particles with small weights, concentrating the computational efforts in those having large ones. Considering the latter, the algorithm for the sampling importance resampling (SIR) particle filter is as follows (Doucet *et al*., 2000):

Sequential Importance Sampling Resampling (SIR) Particle Filter

1. *Importance sampling*

- For $i = 1,\ldots,N$, sample $\tilde{x}_t^{(i)} \sim \pi(x_t|\tilde{x}_{0:t-1}^{(i)}, y_{0:t})$ and set $\tilde{x}_{0:t}^{(i)} \triangleq (x_{0:t-1}^{(i)}, \tilde{x}_t^{(i)})$.
- Evaluate the importance weights

$$w(\tilde{x}_{0:t}^{(i)}) = w_{0:t-1}^{(i)} \cdot \frac{p(y_t|\tilde{x}_t^{(i)})p(\tilde{x}_t^{(i)}|x_{0:t-1}^{(i)})}{q_t(\tilde{x}_t^{(i)}|x_{0:t-1}^{(i)})} \tag{10}$$

$$w_{0:t}^{(i)} = w(\tilde{x}_{0:t}^{(i)}) \cdot \left(\sum_{i=1}^{N} w(\tilde{x}_{0:t}^{(i)})\right)^{-1} \tag{11}$$

2. *Resampling algorithm*

If $\hat{N}_{eff} \geq N_{thres}$

- $\breve{x}_{0:t}^{(i)} = \tilde{x}_{0:t}^{(i)}$ for $i = 1,\ldots,N$; otherwise
- For $i = 1,\ldots,N$, sample an index $j(i)$ distributed according to a discrete distribution satisfying $P(j(i) = l) = w_t^{(l)}$ for $l = 1,\ldots,N$.
- For $i = 1,\ldots,N$, $\breve{x}_{0:t}^{(i)} = \tilde{x}_{0:t}^{j(i)}$ and $\breve{w}_t^{(i)} = N^{-1}$

After the resampling procedure, the new particle population $\{\breve{x}_{0:t}^{(i)}\}_{i=1,\ldots,N}$ is an independent and identically distributed (i.i.d.) sample of the empirical distribution (12), and thus the weights are reset to $\breve{w}_t^{(j)} = N^{-1}$.

$$\breve{\pi}_t^N(x_{0:t}) = \frac{1}{N}\sum_{i=1}^{N} N_t^{(i)}\delta(x_{0:t} - \tilde{x}_{0:t}^{(i)}) = \frac{1}{N}\sum_{i=1}^{N}\delta(x_{0:t} - \breve{x}_{0:t}^{(i)}) \tag{12}$$

2.4 PF in real-time diagnosis applications

PF has a direct application in the arena of FDI. Indeed, once the current state of the system is known, it is natural to implement FDI procedures by comparing the process behaviour with patterns regarding normal or faulty operating conditions.

Approaches introduced in Koutsoukos *et al*. (2002) and de Freitas (2002) make use of PF not only as a tool for state estimation, but also as a means of obtaining the probability of a determined fault mode in a system. This attribute is also found in other interesting results published in the literature, and it is of paramount importance for the present work, since it sets the foundations for a procedure aimed at including customer specifications in the design.

In this sense, two applications of PF algorithms for FDI purposes are of particular interest. These approaches are based on the concept of hybrid dynamic models and the inclusion of risk functions for the allocation of particles among discrete states. The variable resolution particle filter (VRPF; Verma *et al*., 2003, 2004) incorporates the concept of 'abstract particles' in Markov Chain processes, where each particle may represent a single state or a set of similar states. This algorithm has the advantage that only a limited amount of particles is needed to represent large portions of the state-space, when measurements indicate that the likelihood is low. Moreover, once the likelihood of an abstract particle increases, it is possible to specialize the state-space representation to include more specific states, considering a bias–variance trade-off.

The risk sensitive particle filter (RSPF; Thrun *et al*., 2001), on the other hand, incorporates a cost model in the importance distribution to generate more particles in high-risk regions of the state-space (Verma *et al*., 2003). This methodology has proven to be very helpful in improving the tracking of states that are critical to the performance of a six-wheel robot (Verma *et al*., 2003). An important drawback of this approach, though, is the fact that it needs the inclusion of exogenous models to evaluate the risk associated with every fault mode, task that may prove to be difficult to implement.

2.5 PF in real-time prognosis applications

Prognosis may be understood as the result of the procedure where long-term (multi-step) predictions – describing the evolution in time of a fault indicator – are generated with the purpose of estimating the RUL of a failing component/subsystem. Several approaches related to prognosis may be found in the literature. Few of them, however, offer appropriate tools for real-time estimation of the RUL as a continuous function of time.

The most comprehensive effort in establishing an on-line prognosis framework can be found in applications associated with the use of filtering techniques for the study of fatigue crack dynamics (Ray and Tangirala, 1996). The filtering concept enhances the deterministic crack growth modelling standpoint, based on the application of Paris' Equation (Patrick *et al*., 2007; Ray and Tangirala, 1996), and keeps a close relationship with the physics of the problem. Efforts have been made to employ Markov processes and extended Kalman filters (EKF) to estimate the first two moments of a Gaussian state pdf of the system, also assuming independence between measurement noise and uncertainties in material properties. In this case, the obtained Gaussian pdf is afterwards projected in time and used to test M disjoint statistical hypothesis, which divide the feasible range for crack length values.

Regarding particle filters, most authors have visualized this technique as a tool for detection, but not for prognosis. This is mainly because there are no clear indications about how to project the particle population in time, when model non-linearities and non-Gaussian noise structures are assumed. In specific applications, such as chaos prediction, the absence of both process and measurement noise is assumed for prediction purposes (Gustafsson and Hriljac, 2003), thus obtaining a long-term prediction with minimum variance. Each particle is then used as an initial condition for deterministic models to be used for decision theory, risk calculations and other statistical approaches. The implications of these assumptions, though, could be significant in real processes, especially in the presence of vibration signals and, therefore, they must be evaluated with care.

3. Particle filter-based fault diagnosis

A fault diagnosis procedure involves the tasks of FDI and fault identification (assessment of the severity of the fault). In this sense, the proposed particle-filter-based diagnosis framework aims to accomplish these tasks, under general assumptions of non-Gaussian noise structures and non-linearities in process dynamic models, using a reduced particle population to represent the state pdf. The method also allows fusing and utilizing information present in a feature vector (measurements) to determine not only the operating condition (mode) of a system, but also the causes for deviations from desired behavioural patterns. This compromise between model-based and data-driven techniques is accomplished by the use of a particle filter-based module built upon the non-linear dynamic state model (13):

$$\begin{cases} x_d(t+1) = f_b(x_d(t) + n(t)) \\ x_c(t+1) = f_t(x_d(t), x_c(t), \omega(t)) \\ \text{Features}(t) = h_t(x_d(t), x_c(t), v(t)) \end{cases} \tag{13}$$

where f_b, f_t and h_t are non-linear mappings, $x_d(t)$ is a collection of Boolean states associated with the presence of a particular operating condition in the system (normal operation, fault type #1, #2, etc.), $x_c(t)$ is a set of continuous-valued states that describe the evolution of the system given those operating conditions, $\omega(t)$ and $v(t)$ are non-Gaussian distributions that characterize the process and feature noise signals respectively. Since the noise signal $n(t)$ is a measure of uncertainty associated with Boolean states, it is recommendable to define its probability density through a random variable with bounded domain. For simplicity, $n(t)$ may be assumed to be zero-mean i.i.d. uniform white noise.

A PF approach based on model (13) allows statistical characterization of both Boolean and continuous-valued states, as new feature data are received. As a result, at any given instant of time, this framework provides an estimate of the probability masses associated with each fault mode, as well as a pdf estimate for meaningful physical variables in the system. Once this information is available within the FDI module, it is conveniently processed to generate proper fault alarms and to inform about the statistical confidence of the detection routine. Furthermore, pdf estimates for the system continuous-valued states (computed at the moment of fault detection) may be also used as initial conditions in failure prognostic routines, giving an excellent insight about the inherent uncertainty in the prediction problem. As a result, a swift transition between the two modules (FDI and prognosis) may be performed, and more-over, reliable prognosis can be achieved within a few cycles of operation after the fault is declared. This characteristic is, in fact, one of the main advantages of the proposed particle-filter-based diagnosis framework. The following application example helps to illustrate most of the implementation aspects that must be taken into account when applying this methodology, as well as the type of results that can be achieved.

3.1 Detection of crack growth in a UH-60 planetary gear carrier plate

Consider the case of a seeded fault test on a carrier plate, a critical component of the planetary gear transmission system that transmits mechanical power from the engines to the main rotor blades of the helicopter. During this test, a cyclic load profile is applied to the plate to analyse how it affects the growth of an axial crack. Given the existence of a fault condition is known in a seeded fault test, the main objective of this case study is to determine when this crack increases along its axis. Customer specifications include early detection of changes in the growth rate, and a desired statistical confidence level. It is important to note that in this application, it is possible to use features based on the ratio between the fundamental harmonic and the sidebands of the vibration signal spectrum to compute a noisy estimate of the crack length (Patrick *et al*., 2007).

Thus, two main operating conditions are distinguished: the *normal* condition reflects the fact that the crack is growing very slowly or not growing at all, meanwhile the *faulty* condition indicates an abrupt change in the growth rate. In this case,

a PF-based FDI module is implemented using non-linear model (14) to describe the expected rate of growth in the crack, where $x_{d,1}$ and $x_{d,2}$ are Boolean states that indicate *normal* and *faulty* conditions respectively, x_c is the continuous-valued state that represents the crack length, β is a time-varying model parameter dependent on the loading profile that is being applied to the gearbox, and where $\omega(t)$ and $v(t)$ have been selected as zero mean Gaussian noises for simplicity. The initial crack length in the data set used for this analysis is 3.4'', which determines the initial condition of (14).

Besides detecting the faulty condition, it is desired to obtain some measure of the statistical confidence of the alarm signal. For this reason, two outputs will be extracted from the FDI module. The first out is the expectation of the Boolean state $x_{d,2}$, which constitutes an estimate of the probability of fault. The second output is the statistical confidence needed to declare the fault via hypothesis testing (H0: 'the crack is not growing' vs H1: 'The crack is rapidly growing'). The latter output needs another pdf to be considered as the baseline. In this case, historical data has been collected to define this baseline as a Normal distribution $N(\mu, \sigma^2)$.

$$\begin{aligned}
\begin{bmatrix} x_{d,1}(t+1) \\ x_{d,2}(t+1) \end{bmatrix} &= f_b\left(\begin{bmatrix} x_{d,1}(t) \\ x_{d,2}(t) \end{bmatrix} + n(t)\right) \\
x_c(t+1) &= x_c(t) + \beta \cdot x_c(t) \cdot x_{d,2}(t) + \omega(t) \\
y(t) &= x_c(t) + v(t) \\
f_b(x) &= \begin{cases} \begin{bmatrix} 1 & 0 \end{bmatrix}^T, & \text{if} \left\| x - \begin{bmatrix} 1 & 0 \end{bmatrix}^T \right\| \leq \left\| x - \begin{bmatrix} 0 & 1 \end{bmatrix}^T \right\| \\ \begin{bmatrix} 0 & 1 \end{bmatrix}^T, & \text{else} \end{cases} \\
\begin{bmatrix} x_{d,1}(0) & x_{d,2}(0) & x_c(0) \end{bmatrix}^T &= \begin{bmatrix} 1 & 0 & 3.4 \end{bmatrix}^T
\end{aligned} \tag{14}$$

One way to generate an indicator of statistical confidence for the detection procedure is to consider the sum of the weights of all particles i such that $x_c^{(i)}(T) \geq z_{1-\alpha,\mu,\sigma^2}$, where α is the desired test confidence and T is the detection time, which is essentially equivalent to an estimate of (1–*type II error*), or equivalently the probability of detection. If additional information is required, it is possible to compute the value of the Fisher's Discriminant Ratio, as in (15).

$$F_{index}(T) = \left| \mu - \sum_{i=1}^{N} w_T^{(i)} \cdot x_c^{(i)}(T) \right|^2 \cdot \left(\sigma^2 + \sum_{i=1}^{N} w_T^{(i)} \cdot \left(x_c^{(i)}(T) - \sum_{j=1}^{N} w_T^{(j)} \cdot x_c^{(j)}(T) \right)^2 \right)^{-1} \tag{15}$$

Figure 1 shows the results obtained when the proposed FDI approach was applied to the problem of crack growth detection in the planetary gear plate, using the state-space

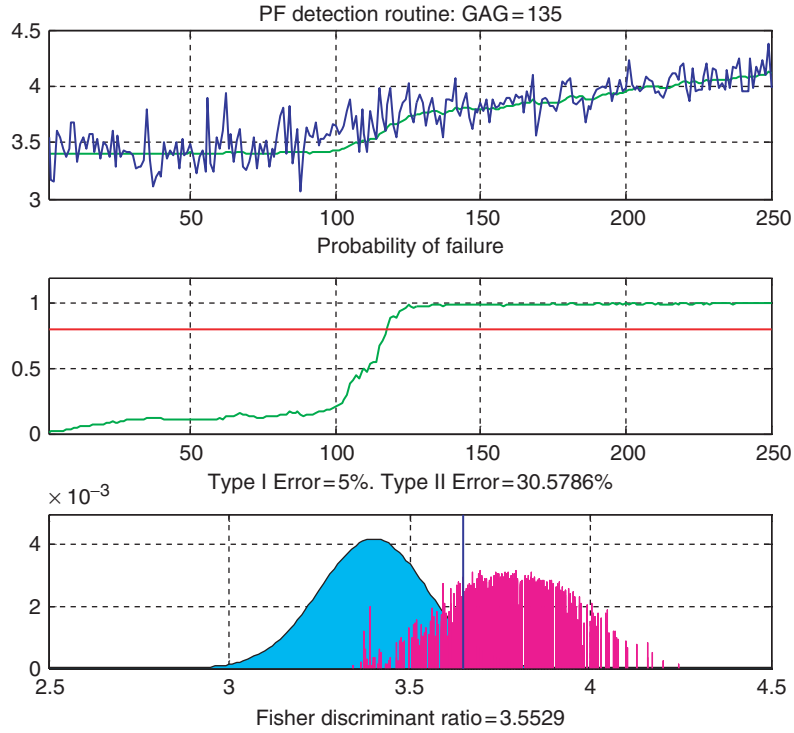

Figure 1 Particle filter-based FDI module. Changes in growth rate in a UH-60 planetary carrier plate

model (14) and 500 particles to describe crack evolution in time. By comparing the trend of the vibration-based crack estimate over time, it is clear that no significant increment in the crack length happened before the 100th ground–air–ground (GAG) cycle. The FDI algorithm only needs 35 additional GAG cycles from this point to detect a change in the crack growth rate, with a confidence level of nearly 70% (type II error ≈30%).

It must be noted that, in this approach, no particular specification about the detection threshold has to be made prior to the actual experiment. Customer specifications are translated into acceptable margins for the type I and II errors in the detection routine. The algorithm itself will indicate when the type II error (false negatives) has decreased to the desired level.

Figure 1 shows three indicators that are simultaneously computed. The first indicator, depicted as a function of time, shows the probability of a determined failure mode, and it is based on the estimate of the Boolean state $x_{d,2}$ in model (14). FDI alarms may be triggered whenever this indicator reaches a pre-determined threshold. If more information is needed, the value of the Fisher's Discriminant Ratio or the type II detection error (second and third indicators, respectively) may be considered. The vertical line that discriminates between the two pdfs in Figure 1 is fixed by the desired

type I detection error (probability of false positives), considering the data used as a baseline for detection purposes.

4. PF for prognosis in stochastic non-linear systems

Prognosis may be essentially understood as the generation of long-term predictions for a fault indicator, made with the purpose of estimating the RUL of a failing component. This paper presents a two-level procedure that has been developed, and subsequently tested, to address the issue of failure prognosis. This procedure reduces the uncertainty associated with long-term predictions by using the current state pdf estimate, a process noise model and a record of corrections made to previously computed predictions. In a first prognosis level, p-step ahead predictions are generated based on an *a priori* estimate, adjusting their associated probabilities according to the noise model structure. A second prognosis level uses these predictions and the definition of critical thresholds to estimate the RUL pdf, also referred to as the time-to-failure (TTF) pdf, and simultaneously implements a correction model (*outer correction loop*) to compensate for all main error sources. A detailed description of each level is now presented.

4.1 First prognosis level: generation of long-term predictions

The first prognosis level is related to the generation of a p-step ahead long-term prediction for the state pdf, which can be obtained in a recursive manner using both the model update Equation (1) and the current state estimate, as shown in (16).

$$\begin{aligned}\tilde{p}(x_{t+p}|y_{1:t}) &= \int \tilde{p}(x_t|y_{1:t}) \prod_{j=t+1}^{t+p} p(x_j|x_{j-1}) dx_{t:t+p-1} \\ &\approx \sum_{i=1}^{N} w_t^{(i)} \int \cdots \int p(x_{t+1}|x_t^{(i)}) \prod_{j=t+2}^{t+p} p(x_j|x_{j-1}) dx_{t+1:t+p-1}\end{aligned} \tag{16}$$

The evaluation of these integrals, though, may be difficult and/or may require significant computational effort. To illustrate the latter, consider the predicted conditional state pdf $\hat{p}(x_{t+k}^{(i)}|\hat{x}_{t+k-1}^{(i)})$, which describes the state distribution at the future time instant $t+k$ ($k=1,\ldots,p$) when the particle $\hat{x}_{t+k-1}^{(i)}$ is used as initial condition. Assuming that the current weights $\{w_t^{(i)}\}_{i=1,\ldots,N}$ are a good representation of the state pdf at time t, then it is possible to approximate the predicted state pdf at time $t+k$, by using the law of total probabilities and the particle weights at time $t+k-1$, as shown in (17):

$$\hat{p}(x_{t+k}|\hat{x}_{1:t+k-1}) \approx \sum_{i=1}^{N} w_{t+k-1}^{(i)} \cdot \hat{p}(x_{t+k}^{(i)}|\hat{x}_{t+k-1}^{(i)}); \; \hat{x}_t^{(i)} = \tilde{x}_t^{(i)}; \; k = 1,\ldots,p \tag{17}$$

To evaluate (17), the weight of every particle should be modified (at each prediction step) to take into account the fact that noise and process non-linearities could change the shape of the state pdf as time passes. However, since the weight update procedure is needed as part of a prediction problem, it cannot depend on the acquisition of new measurements. Additionally, before proceeding with the next prediction step, it is necessary to allocate a new set of particles within the domain of the probability distribution (17). To overcome most of these difficulties, two main approaches are proposed.

4.1.1 P-step ahead long-term predictions – first approach: This first approach predicts the evolution in time of each particle by successively taking the expectation of the model update Equation (1) for every future time instant, considering the state value associated to that particle as initial condition, as shown in (18).

$$\hat{x}_{t+p}^{(i)} = E[f_{t+p}(\tilde{x}_{t+p-1}^{(i)}, \omega_{t+p})]; \hat{x}_t^{(i)} = \tilde{x}_t^{(i)} \tag{18}$$

In this sense, the first approach for long-term prediction is the simplest in terms of computational effort. Basically, it states that the error that can be generated by considering the particle weights invariant for future time instants is negligible with respect to other sources of error that may appear in practical applications, such as model inaccuracies or even in the assumptions made for process and measurement noise parameters.

Therefore – from this standpoint – (18) is considered sufficient to extend the trajectories $\hat{x}_{0:t+k}^{(i)}$, while the current particle weights are propagated in time without changes. The computational burden of this method is significantly smaller and, as it will be shown in simulation results, the method still offers a satisfactory view about how the system behaves in practical applications.

4.1.2 P-step ahead long-term predictions – second approach: The second approach for long-term prediction proposes a solution for the problem of uncertainty representation at future time instants, which is especially useful if the prediction time horizon is large. Instead of recalculating the particle weights, it proposes that uncertainty for future transitions may be incorporated by simply resampling the predicted state pdf (17).

Thus, the information about the distribution of the state for future time instants is now given by the position of the particles, not by the particle weight value. The implementation of this methodology, however, must ensure that the resampled population is representative of (17). A computationally affordable solution for this predicament is proposed, based on the assumption of uncorrelated process noise (diagonal covariance matrix for $\omega(t)$) and the use of kernel transitions to describe the state pdf before the resampling step, as it is also done in the case of the regularized particle filter (RPF; Musso *et al*., 2001).

Consider, in this sense, a discrete approximation (19) for the predicted state pdf (17), where $K(\cdot)$ is a kernel density function, which may correspond to the process noise pdf, a Gaussian kernel or a rescaled version of the Epanechnikov kernel (21).

$$\hat{p}(x_{t+k}|\hat{x}_{1:t+k-1}) \approx \sum_{i=1}^{N} w_{t+k-1}^{(i)} K_h\left(x_{t+k} - E\left[x_{t+k}^{(i)}|\hat{x}_{t+k-1}^{(i)}\right]\right) \tag{19}$$

$$K_h = \frac{1}{h^{n_x}} K\left(\frac{x}{h}\right), h_{opt} = A \cdot N^{-\frac{1}{n_x+4}}, A = \left(8c_{n_x}^{-1} \cdot (n_x+4) \cdot \left(2\sqrt{\pi}\right)^{n_x}\right)^{\frac{1}{n_x+4}} \tag{20}$$

$$K(x) = \begin{cases} \frac{n_x+2}{2c_{n_x}}\left(1 - \|x\|^2\right) & \text{if} \|x\| < 1 \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

where c_{n_x} is the volume of the unit sphere in $\mathbb{R}^{n_x}$. It is reasonable to try to represent the uncertainty present in (19), instead of just projecting the conditional expectations of the state variables. One way to achieve this task is to generate a new population of equally weighted particles for the time instant $t+k$, $1 \le k \le p$, performing an inverse transform resampling procedure for the particle population. This method obtains samples distributing according to (19), selecting N realizations of $u(i) \sim U(0,1)$ and interpolating a value for $\hat{x}_{t+k}^{(i)}$ from the cumulative state distribution $F(X_{t+k} \le x_{t+k}) = \int_{-\infty}^{x_{t+k}} \hat{p}(x_{t+k}|\hat{x}_{1:t+k-1})dx_{t+k}$ in accordance with $\hat{x}_{t+k}^{(i)} = F^{-1}(u^{(i)})$.

The inherent randomness present in the inverse transform resampling method, however, may lead to unrepresented areas in the domain of the cumulative state distribution function, situation that is difficult to correct in long-term predictions, since there are no measurements available that may be used for this purpose. To overcome this difficulty, a two-step procedure is proposed.

The first step in the resampling strategy performs a simplified version of the inverse transform resampling procedure, which will focus in representing the growth of uncertainty present in (19). In this sense, samples distributing according to (19) are obtained by selecting $u^{(i)} = i \cdot (N+1)^{-1}$ $(i: 1, \ldots, N)$ and interpolating a value for $\hat{x}_{t+k}^{(i)}$ from the cumulative state distribution $F(X_{t+k} \le x_{t+k}) = \int_{-\infty}^{x_{t+k}} \hat{p}(x_{t+k}|\hat{x}_{1:t+k-1})dx_{t+k}$ in accordance with $\hat{x}_{t+k}^{(i)} = F^{-1}(u^{(i)})$.

To avoid loss of diversity among particles, an additional step inspired by the RPF is performed. In this sense, it is assumed that the state covariance matrix $\hat{S}_{t+k}$ equal to the empirical covariance matrix of $\hat{x}_{t+k}$ and that a set of equally weighted samples for $\hat{x}_{t+k-1}$ is available, in such a way that the efficiency in the use of Epanechnikov kernels for pdf approximation is maximized.

In consequence, considering all of the above, the regularization algorithm (Musso *et al*., 2001) when applied for long-term predictions is as follows:

Long-term predictions: second approach

- Apply modified inverse transform resampling procedure. For $i = 1, \ldots, N$ $w_{t+k}^{(i)} = N^{-1}$
- Calculate $\hat{S}_{t+k}$, the empirical covariance matrix of $\left\{E\left[x_{t+k}^{(i)}|\hat{x}_{t+k-1}^{(i)}\right], w_{t+k}^{(i)}\right\}_{i=1}^{N}$
- Compute $\hat{D}_{t+k}$ such that $\hat{D}_{t+k}\hat{D}_{t+k}^{T} = \hat{S}_{t+k}$
- For $i = 1, \ldots, N$, draw $\varepsilon^{i} \sim K$, the Epanechnikov kernel and assign $\hat{x}_{t+k}^{(i)*} = \hat{x}_{t+k}^{(i)} + h_{t+k}^{opt}\hat{D}_{t+k}\varepsilon^{i}$, where h_{t+k}^{opt} is computed as in (20)

It is important to notice that the assumption of uncorrelated process noise is only included for the sake of reducing the computational effort of the resampling procedure. In fact, there are no theoretical restrictions for the application of this methodology in the presence of correlated process noise.

4.2 Second prognosis level: estimation and statistical characterization of the RUL of equipment

The final outcome for any prognosis algorithm is an estimate for the system RUL pdf, which is intrinsically entangled with the probability of failure at future time instants. This probability can be obtained from long-term predictions, when the empirical knowledge about critical conditions for the system is included in the form of thresholds for main fault indicators, also referred to as the hazard zones.

In real applications, it is expected for the hazard zones to be statistically determined on the basis of historical failure data, defining a critical pdf with lower and upper bounds for the fault indicator (H_{lb} and H_{up}, respectively).

Since the hazard zone specifies the probability of failure for a fixed value of the fault indicator, and the weights $\{w_{t+k}^{(i)}\}_{i=1,\ldots,N}$ represent the predicted probability for the set of predicted paths, then it is possible to compute the probability of failure at any future time instant (namely the RUL pdf) by applying the law of total probabilities, as shown in (22). Once the RUL pdf is computed, combining the weights of predicted trajectories with the hazard zone specifications, it is well known how to obtain prognosis confidence intervals, as well as the RUL expectation.

$$\hat{p}_{TTF}(ttf) = \sum_{i=1}^{N} \Pr\left(Failure|X = \hat{x}_{ttf}^{(i)}, H_{lb}, H_{up}\right) \cdot w_{ttf}^{(i)} \tag{22}$$

Expression (22) provides a solution for the RUL pdf estimation problem that is suitable for on-line applications. As it depends on the predicted trajectory weights, though, it is subject to uncertainty and it may be sensitive to modelling errors. Moreover, uncertainty inherent to RUL expectations increases as the prediction horizon grows. This issue is of special interest in prognosis, since the estimation of the

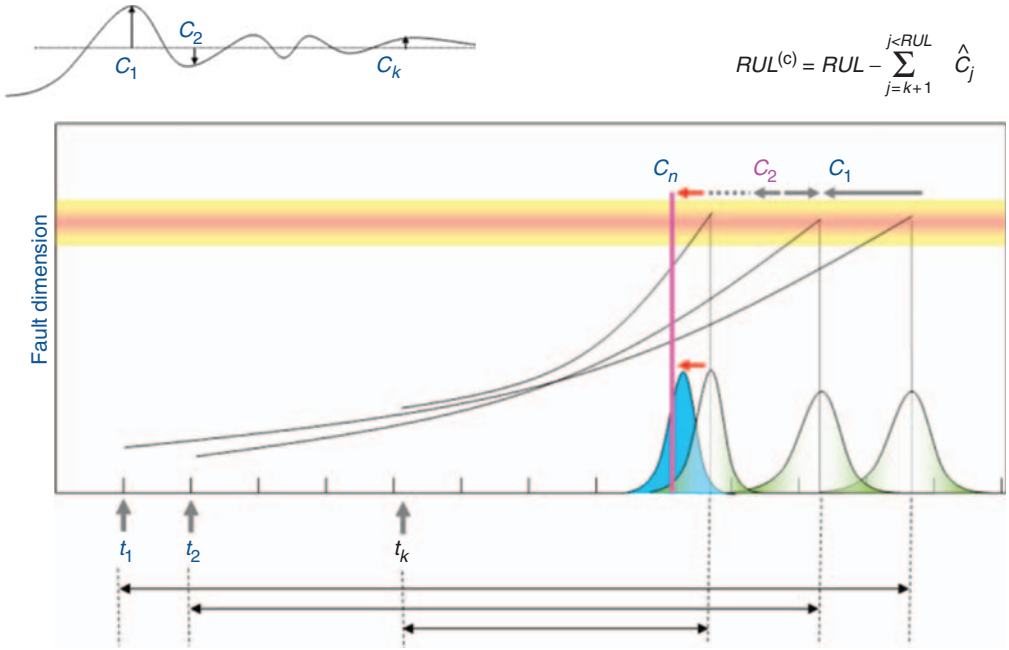

Figure 2 Illustration of outer correction algorithm for RUL expectation

RUL must be done immediately after the fault condition has been detected, and hence most of the prediction horizons involve long time periods.

In particular, to reduce the uncertainty inherent to a particle filter-based failure prognosis and improve the accuracy of the RUL expectation, an additional *outer correction loop* has been included as part of the proposed second prognosis level, see Figure 2.

This *outer loop* is basically a data-driven learning paradigm. It computes a series of correction terms C_j $(j=1,\ldots,k)$ that measure the difference between the RUL expectation computed at the current time $t=j$ and the one that was computed in the previous iteration of the prognosis algorithm. Once k correction terms are obtained, a linear autoregressive model is built to establish a relationship between all past correction terms. The obtained linear autoregressive model is then used to generate an estimate for all future corrections $\hat{C}_{k+1},\ldots,\hat{C}_{RUL}$ that would be applied to the current RUL expectation if measurement data were to be acquired until the failure time, assuming that both process and measurement noises are wide sense stationary (WSS).

Finally, the current RUL expectation is corrected, obtaining $RUL^{(c)}$. In simple words, the proposed *outer correction loop* intends to capture the pattern of past measurement-driven prediction updates inside a simple model, which can be used afterwards to estimate and correct for the accuracy of the current prediction.

The learning scheme proposed here is just an example about how the combination of model-based and data driven techniques in an *outer correction loop* can significantly improve the prognosis algorithm accuracy. Other approaches may be also implemented as a manner of incorporating information from past instances: additional *outer correction loops* may also help to reduce prediction uncertainty by modifying both the structure and parameters of process noise in the dynamic model. These topics will be considered for future research efforts.

4.3 Illustrative example: RUL statistical characterization

Consider the problem of RUL estimation in a process for which the evolution in time of a known failure condition (for instance, a crack in a material) is described by the model (23), where $\omega_2(t)$ is zero mean Gaussian noise.

$$\begin{aligned}
&\begin{cases} x_1(t+1) = x_1(t) + 3 \cdot 10^{-4}(0.05 + 0.1 \cdot x_2(t))^3 + \omega_1(t) \\ x_2(t+1) = x_2(t) + \omega_2(t) \end{cases} \\
&y(t) = x_1(t) + v(t) \\
&\omega_1(t) \sim \mathit{Gamma}\,(0.15, 0.3) \\
&v(t) \sim \frac{1}{4}\mathcal{N}(-0.5, 0.25) + \frac{3}{4}\mathcal{N}(0.5, 0.25)
\end{aligned} \tag{23}$$

Furthermore, to analyse the effect that inaccuracies and model errors imply in RUL estimates, let us assume that noise is believed to be Gaussian. In that case, the first two moments of both process and observation noises may be estimated using historical data, obtaining:

$$\omega_1(t) \sim \mathcal{N}(0.045, 0.1162), \quad v(t) \sim \mathcal{N}(0.25, 0.5). \tag{24}$$

The hazard zone, which in real applications must be defined on the basis of customer specifications or ground truth failure data, is defined here as a normal pdf with parameters $\mu = 9.0$ and $\sigma = 0.3$. The main objective is to generate a 95% confidence interval for the RUL of the process, 40 cycles after the fault condition is detected. In addition to the techniques described in Sections 4.1 and 4.2, an Extended Kalman Filter (EKF)-based prognosis procedure has also been considered as a means for both comparison and performance evaluation for the proposed PF-based techniques.

Results are summarized in Figure 3, where the light–dark and black lines represent, respectively, the noisy measurements and the process output estimation obtained from an SIR particle filter, and where the dotted line shows the actual evolution of the failure condition for future time instants (information that is unknown when the RUL estimation is performed).

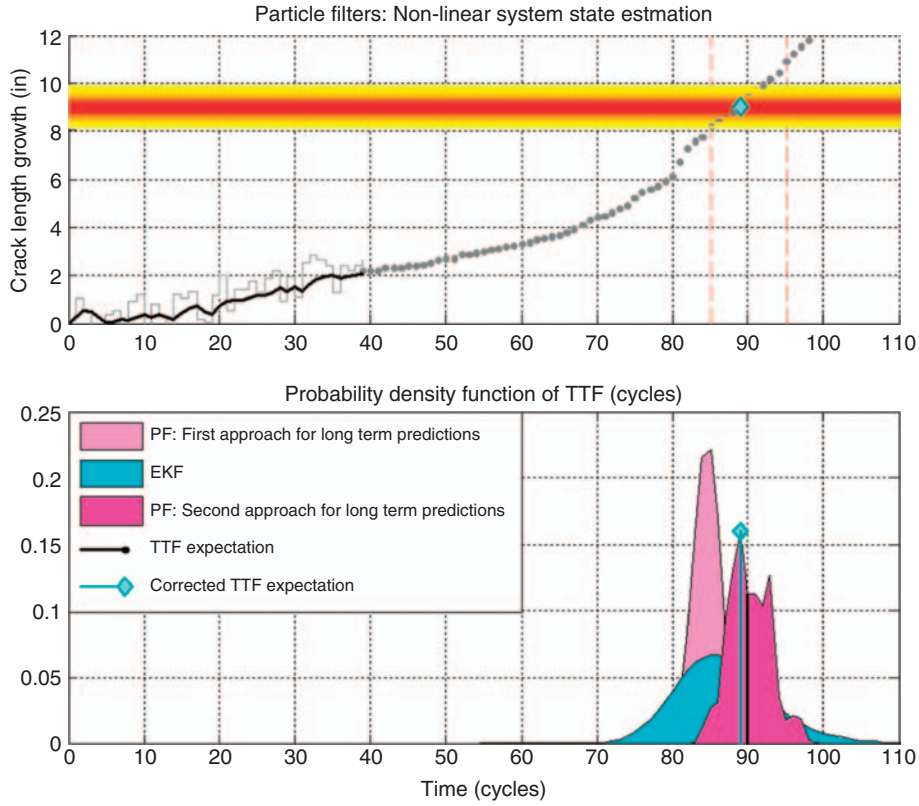

Figure 3 Result comparison for RUL statistical characterization

Figure 3 provides valuable information that may be used to evaluate the capability of the algorithm to predict the evolution in time of the state probability distribution, particularly when some metrics (such as precision and accuracy) are invoked to assess the algorithm performance.

Results show that the PF–based second approach for long-term predictions is capable of overcoming the bias introduced by model errors, because of its ability to represent the state probability space. The combination of resampling techniques and Epanechnikov kernels for pdf approximation in long-term predictions is able to simultaneously reduce the impact of model inaccuracies and provide a balanced result in terms of accuracy and precision in the RUL estimate. Furthermore, the actual fault indicator (unknown when the long-term predictions were performed) reaches the previously defined hazard zone inside the 95% confidence interval, validating the RUL pdf estimate.

Finally, it must be noted that when the *outer loop* correction scheme – introduced as part of the second prognosis level – is applied to the PF-based second approach for long-term prediction generation, it allows improving the estimate of the RUL

expectation to the extent that the corrected estimate time-to-failure (TTF) coincides with the time instant where the actual failure growth reaches the mean of the hazard zone (9′′) (Figure 3).

4.4 Case study: UH-60 planetary carrier plate. Analysis of axial crack growth

Consider the case of prognosis for the evolution of an axial crack on the plate of the UH-60 planetary gearbox, shown in Figure 4.

Although this fault mode can lead to a critical failure condition in the aircraft, there was no certain way to determine its existence save by a detailed off-line inspection of this piece of equipment – a procedure that obviously involves large financial cost. Under this scenario, the use of algorithms capable of estimating the RUL by only analysing vibration-based features becomes extremely attractive and would help to dramatically decrease operational and maintenance costs as well as avoid catastrophic events.

With the purpose of testing the feasibility and efficiency of such techniques, a seeded fault test was conducted to collect fault data under a fixed known loading profile. In this test, the crack was artificially grown until it reached a total length of

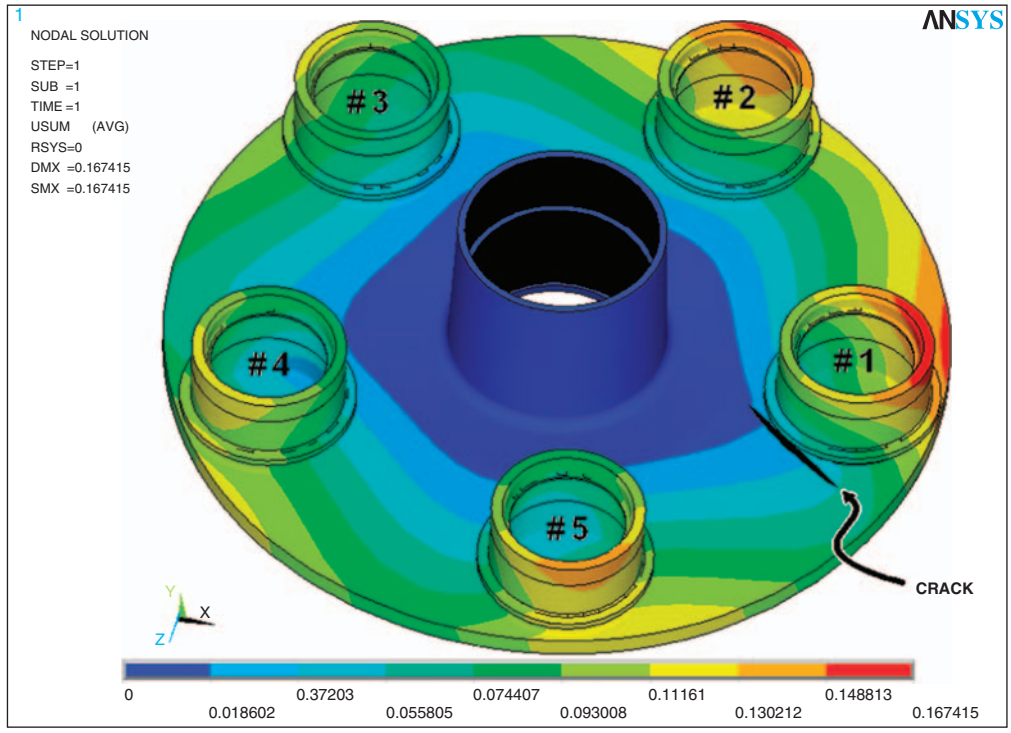

Figure 4 ANSYS model of the planetary gear plate, showing crack location

1.34'', after that the gearbox was forced to operate emulating load changes that vary from 20% to 120% in a three (min) GAG cycle (Figure 5). Given the fact that the initial crack length was perfectly known in this case, a deterministic prognosis approach was considered at first to estimate bounds for the failure time.

From material structure theory, it is well known that the crack growth evolution may be explained by using an empirical model such as the Paris' Equation (25), given the proper set of coefficients (Patrick *et al.*, 2007):

$$\frac{dL}{dn} = C \cdot (U(n) \cdot \Delta K(n))^m \tag{25}$$

where L is the total crack length, C and m are material related coefficients, n is the cycle index, $U(n)$ is a parameter that models the effect of crack closure during cycle n and $\Delta K(n)$ is the crack tip stress variation during the cycle n, measured in $(\mathrm{MN/m^{3/2}})$. Although simple, model (25) requires the computation of two critical parameters to be used in any prognosis routine: $\Delta K(n)$ and $U(n)$. The stress $K(n)$ may be estimated for a constant load (usually 100%) by using finite element analysis (FEA) tools such as ANSYS, for different crack lengths and crack orientations.

Considering a proportional relationship between the stress in the tip of the crack and the load percentage, it is in fact possible to construct a mapping relating both the current crack length and load variation per cycle with $\Delta K(n)$.

Albeit the former piece of information is helpful, it is insufficient to estimate the evolution of the crack length. On one hand, the closure effect parameter $U(n)$ cannot be efficiently measured and only empirical approximations exist for certain materials,

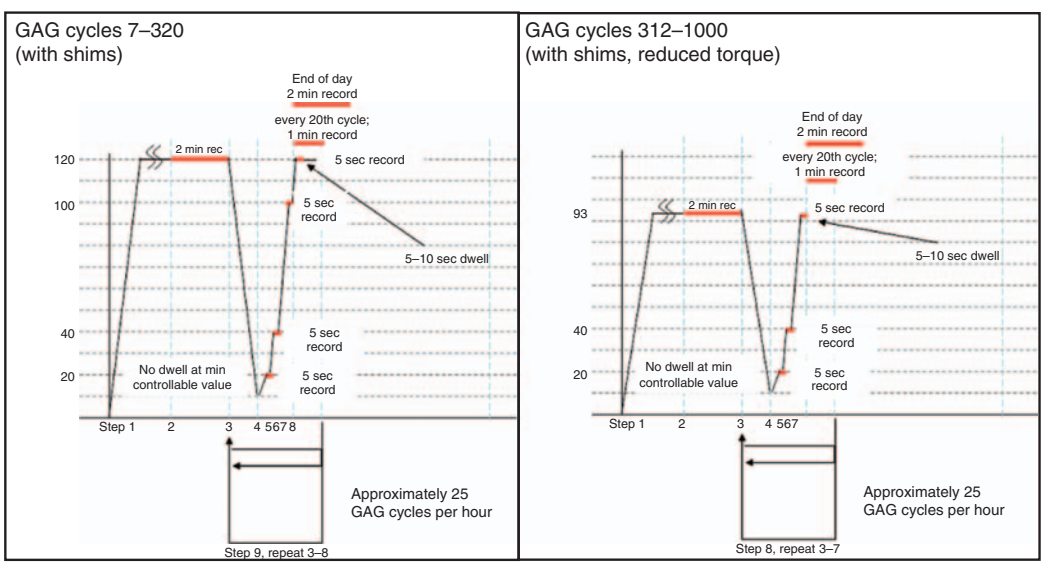

Figure 5 Loading profile (%) vs GAG cycles

such as Ti-6A1-4V. Even in the case of that material, only upper and lower bounds may be computed, and thus it is impossible to compute expectations and/or determine statistically the validity of confidence intervals. On the other hand, the crack length has to be first estimated to come up with an approximate value for $\Delta K(n)$ and therefore any estimation error will affect tremendously the accuracy of the long-term prediction. In this sense, long-term predictions and bounds generated by means of a deterministic model are reasonably good for regular maintenance scheduling, though insufficient for the on-line determination of confidence intervals and on-flight corrective actions.

The inclusion of process data, measured and pre-processed in an on-line fashion, improves tremendously the prospect of what can be achieved in terms of RUL estimation and prognosis in general. Indeed, the use of features based on the ratio between the fundamental harmonic and the sidebands in the vibration data spectrum (Patrick *et al.*, 2007) gives the basis for the implementation of any of the PF-based prognosis methodologies introduced in Section 4. Consequently, under this new approach, not only it is possible to estimate the expected growth of the crack, but also the unknown closure parameter in the crack growth model (25) and the RUL pdf, enabling the computation of any statistics such as expectations, confidence intervals, etc.

The following crack growth state model (based on Paris' Equation) has been implemented for purposes of on-line state and model parameter estimation:

$$\begin{cases} L(t+1) = L(t) + C \cdot \alpha(t) \cdot \left\{ (\Delta K_{inboard}(t))^m + (\Delta K_{outboard}(t))^m \right\} + \omega_1(t) \\ \alpha(t+1) = \alpha(t) + \omega_2(t) \\ \Delta K_{inboard}(t) = f_{inboard}(\text{Load}(t), L(t)) \\ \Delta K_{outboard}(t) = f_{outboard}(\text{Load}(t), L(t)) \end{cases} \tag{26}$$
$$\text{Feature}(t) = h(L(t)) + v(t)$$

where $L(t)$ is the total crack length estimation at GAG cycle t, $\alpha(t)$ is an unknown time-varying model parameter to be estimated (unitary initial condition), C and m are model constants related to material properties, ΔK is the variation in crack tips stress related to the load profile and the current crack length (estimated through off-line analysis of the system with ANSYS) and $\omega_1(t)$, $\omega_2(t)$ and $v(t)$ are non-Gaussian white noises.

Process model (26) necessitates a noisy estimate of the crack length based on the value of the feature data point to be used in on-line applications. This requirement is easily satisfied via a non-linear mapping $h(\cdot)$, which is corrected or improved according to the ground truth crack length data that is acquired (at specific and very limited time instants) from strain gages sensors allocated on the surface of the planetary gear plate.

As a result, in the proposed scheme, two update loops run in parallel. The first one, referred to as the *inner loop*, basically uses the feature data and the previous state pdf

estimate to update the crack length and model parameter estimates and thus, the RUL pdf estimate through the prognosis approaches discussed in Sections 4.1 and 4.2. On the other hand a second loop, namely the *outer loop*, revises the non-linear mapping $h(\cdot)$ between the vibration-based feature value and the crack length every time it gets an update from the strain gages allocated on the plate. It is expected, for future on-line applications, that the non-linear mapping $h(\cdot)$ would be still valid, save for minor adjustments.

At any given time instant, each particle from the current particle population determines both an initial condition for a long-term prediction and a probability associated with that prediction, see Figures 6 and 7 where each plausible long-term prediction is depicted with a different colour.

The time instant when each predicted trajectory reaches a given threshold defines a probable failure time and thus, a realization of the RUL pdf. RUL expectations, 95% confidence interval for long-term predictions and ± 3 sigma intervals may be computed once the RUL pdf is estimated through the described procedure.

Table 1 shows the results for this particular case study, comparing all the statistics for long-term prediction with the ground truth data that was supplied from strain gages allocated on the surface of the plate.

Ground truth data points, ie, strain gages crack length measurements, shown in Table 1 were provided incrementally up to 650 GAG cycles in a 'blind' test format. Thus, for instance, the prediction result of Table 1 for GAG #36 (1.60'') has been obtained at GAG #0 knowing only the initial crack length. Subsequently, the predicted value for GAG #100 (2.40'') has been obtained at GAG #36 after the ground truth data value of 2.00'' was used to adjust the non-linear mapping $h(\cdot)$. Analogously, the prediction for GAG #230 was made at GAG #100. The rest of the table was constructed in the same manner.

Every time a new point of ground truth data is included, a more accurate initial condition for the prediction algorithm is estimated, and hence the overall precision of the algorithm is enhanced. The modularity of the proposed approach allows even modifying the set of thresholds considered in the analysis, every time that it is required to increase the hazard level. Compare, for example, the different thresholds that are shown in both Figures 6 and 7.

To illustrate this fact more clearly, consider that the prediction algorithm is launched at GAG cycle 100. Crack length thresholds at 3.0'', 3.5'' and 4.5'' may be established at that time. Given this scenario, the prediction algorithm provides answers to the question: what are the expected (in a probabilistic sense) times at which the crack will reach the corresponding lengths of 3.0'', 3.5'' and 4.5''?

By estimating the RUL pdf, the algorithm supplies the RUL expectation (mean time) and the 95% confidence interval for each case. As the crack length evolves in time, however, the hazard thresholds can be easily modified to continue the analysis of its growth, eventually reaching the condition of Figures 7 and 8, where only one remaining hazard threshold is of interest (~ 6.2'') with a TTF expected value of 713 GAG cycles,

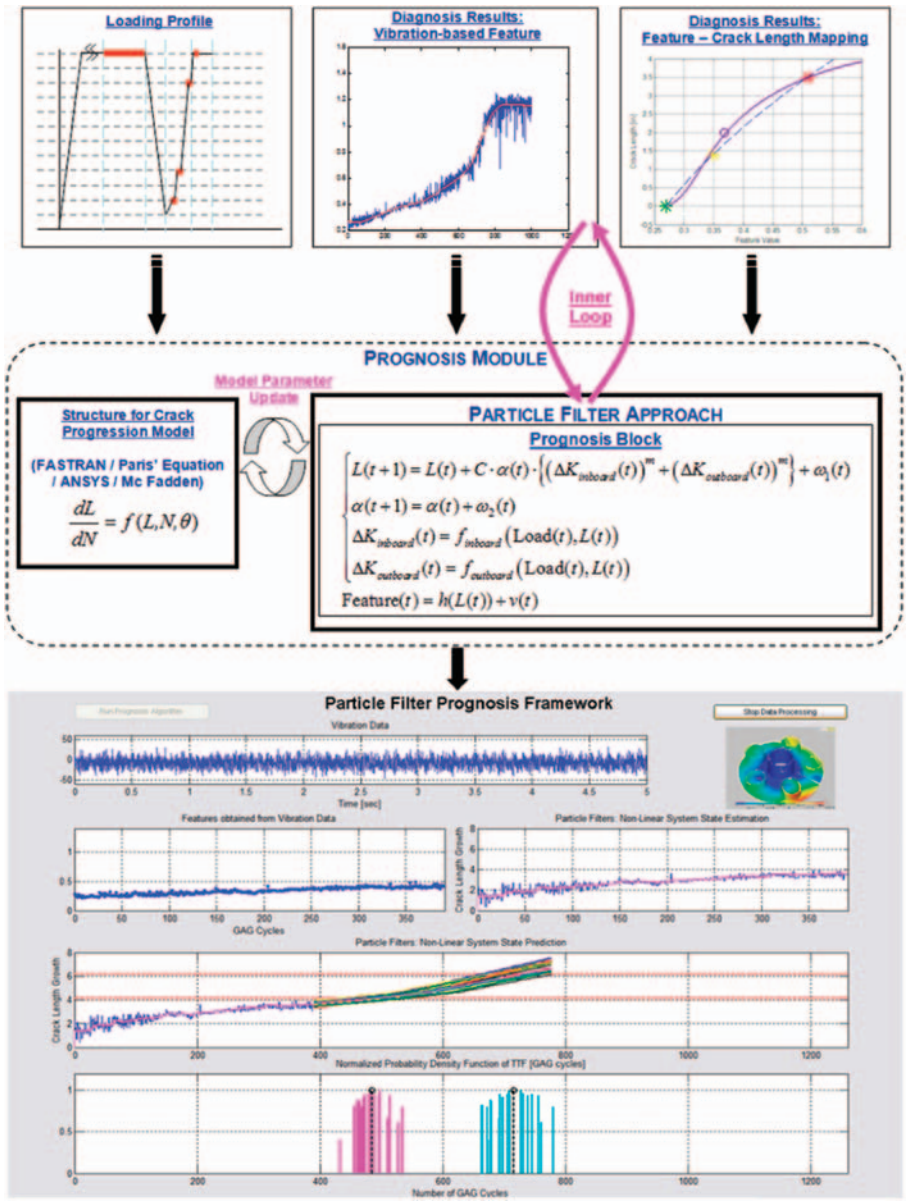

Figure 6 PF-based approach for prognosis; crack growth in a planetary carrier plate

or equivalently an expected RUL of 325 GAG cycles, which is extremely close to the value of 714 GAG cycles that was provided in the ground truth data set for the time of failure. The accuracy of the algorithm has been validated at every step of the *'blind'* test, confirming the robustness of the approach with respect to changes in the load profile

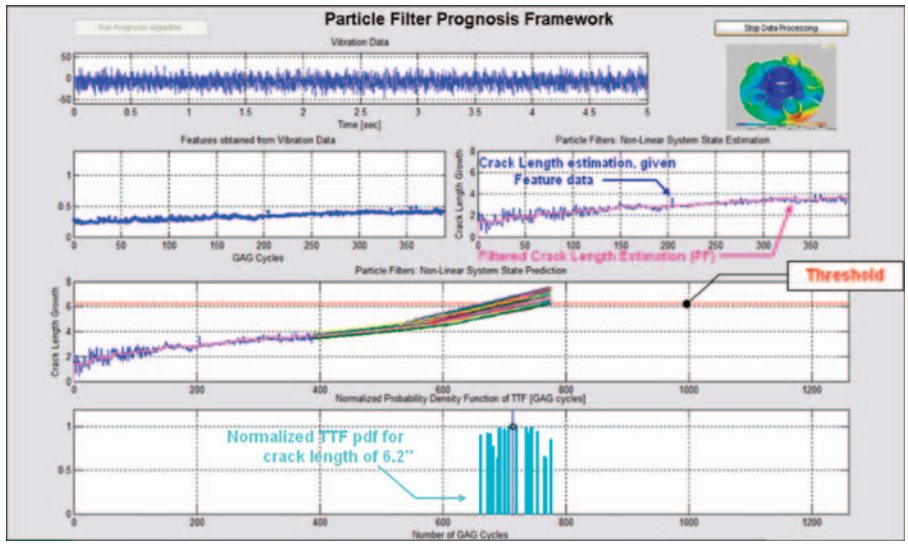

Figure 7 Prediction results for a single hazard threshold

Table 1 Prediction results for PF-based approach for prognosis

Measured crack length		Confidence intervals				
Gag	Crack length (inches)	-3σ	-95%	Mean	$+95\%$	$+3\sigma$
0	1.34	N/A	N/A	1.34	N/A	N/A
36	2.00	0.74	1.03	1.60	2.17	2.46
100	2.50	1.93	2.09	2.40	2.71	2.87
230	3.02	2.73	2.79	2.90	3.01	3.07
400	3.54	3.41	3.54	3.80	4.06	4.19
550	4.07	3.85	4.11	4.30	4.60	4.75
650	4.52	4.20	4.48	4.71	5.08	5.70
750	6.78	6.38	6.42	6.61	6.76	6.84

(depicted in Figure 5) and/or in the signal-to-noise ratio of the feature-based noisy crack length estimate, which steadily improved as the crack length increased.

Given the PF-based pdf state estimate, additional information about the operating conditions of the system may be also extracted. For instance, consider the estimate of the parameter $\alpha(t)$ in model (26) (Figure 9). Sudden changes in the parameter estimate are indicators of changes in the testing operating conditions, as in GAG cycle #320, where the maximum value of the load, applied to the carrier plate, was reduced.

Finally, it is important to mention that the proposed methodology has been compared with an EKF-based approach for long-term prediction. Other approaches

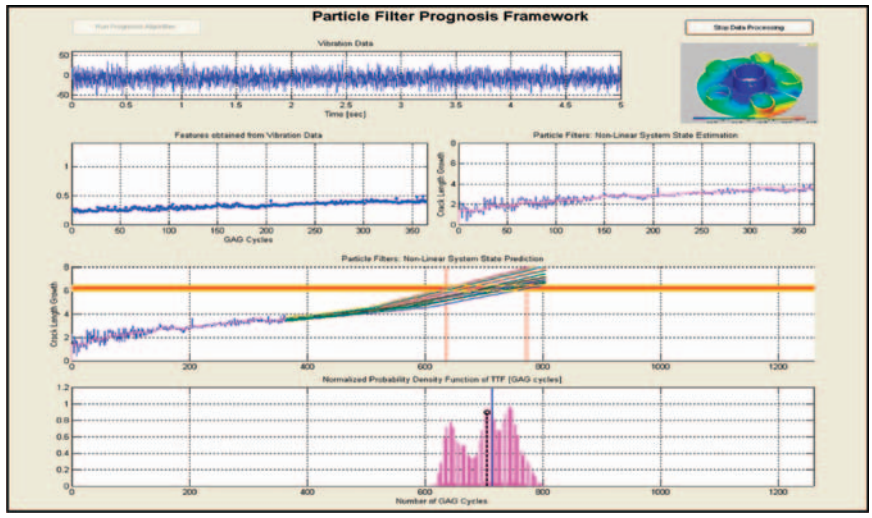

Figure 8 Prediction results for a unique hazard zone at 6.2″

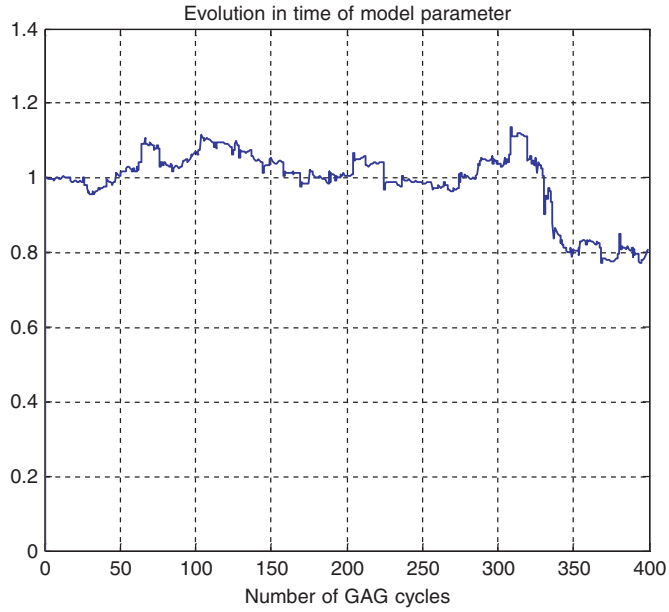

Figure 9 Time-varying model parameter vs GAG cycles

such as IMM or Unscented Kalman Filter were not considered, since they implied a significantly higher computational burden than the proposed approach. Results were always favourable for the proposed PF-based prognosis scheme in terms of accuracy and precision of the RUL pdf estimate.

The PF framework for the prediction of the RUL may be easily implemented in real time on-board a HUMS or other health monitoring platform for on-line applications; in fact, an integrated architecture that combines vibration data processing, feature extraction, fault diagnosis and failure prognosis based on this concept is described in Patrick *et al*. (2007).

5. Conclusions

This paper introduces an architecture for the development, implementation, testing and assessment of a PF-based framework for failure FDI and prognosis. The FDI framework has been successful in pinpointing abnormal conditions, such as changes in the growth rate of an axial crack (UH-60 gear plate). Regarding prognosis, the proposed method was successfully tested in an illustrative example. Furthermore, it was shown that a prediction method based on a combination of a resampling scheme and Epanechnikov kernels (for pdf approximation in long-term predictions) is able to simultaneously reduce the impact of model errors and provide a balanced result in terms of accuracy and precision in the RUL estimates. It was also shown that an approach simply based on the expectation of the long-term prediction provides acceptable results, and that it is suitable for on-line implementation. In particular, a successful case study has been presented to illustrate the performance of a simple implementation (SIR particle filter and an expectation-based long-term prediction generation). This application example used real failure data from a seeded fault test in a UH-60 planetary carrier plate, providing an excellent insight about the effect of model inaccuracies and customer specifications (eg, hazard zone definition, desired prediction window) in the algorithm performance.

References

Andrieu, C., **Doucet, A.** and **Punskaya, E.** 2001: Sequential Monte Carlo methods for optimal filtering. In Doucet, A., de Freitas, N. & Gordon, N., editors. *Sequential Monte Carlo methods in practice*. Springer-Verlag.

Arulampalam, M.S., **Maskell, S.**, **Gordon, N.** and **Clapp, T.** 2002: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing* 50, 174–88.

de Freitas, N. 2002: Rao-Blackwellised particle filtering for fault diagnosis. *IEEE Aerospace Conference Proceedings* (Cat. No. 02TH8593), pt. 4, 1767–72.

Doucet, A. 1998: On sequential Monte Carlo methods for Bayesian filtering. Technical Report, Engineering Department, University of Cambridge.

Doucet, A., **de Freitas, N.** and **Gordon, N.** 2001: An introduction to Sequential Monte Carlo methods. In Doucet, A., de Freitas, N. & Gordon, N., editors. *Sequential Monte Carlo methods in practice*. Springer-Verlag.

Doucet, A., **Godsill, S.** and **Andrieu, C.** 2000: On Sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing* 10, 197–208.

Gustafsson, F. and **Hriljac, P.** 2003: Particle filters for system identification with

application to chaos prediction. *13th IFAC Symposium on System Identification, Rotterdam*. The Netherlands.

Haug, A.J. 2005: A tutorial on Bayesian estimation and tracking techniques applicable to nonlinear and non-Gaussian processes. MITRE Technical Report, MTR 05W0000004, The MITRE Corporation.

Kadirkamanathan, V., **Li, P.**, **Jaward, M.H.** and **Fabri, S.G.** 2002: Particle filtering-based fault detection in non-linear stochastic systems. *International Journal of Systems Science* 33, 259–65.

Kong, A., **Liu, J.S.** and **Wong, W.H.** 1994: Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association* 89, 278–88.

Koutsoukos, X., Kurien, J. and **Zhao, F.** 2002: Monitoring and diagnosis of hybrid systems using particle filtering models. *International Symposium on Mathematical Theory of Networks and Systems*.

Li, P. and **Kadirkamanathan, V.** 2001: Particle filtering based likelihood ratio approach to fault diagnosis in nonlinear stochastic systems. *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews* 31, 3.

Liu, J.S. 1996: Metropolized independent sampling with comparison to rejection sampling and importance sampling. *Statistics and Computing* 6, 113–19.

Musso, C., **Oudjane, N.** and **Le Gland, F.** 2001: Improving regularised particle filters. In Doucet, A., de Freitas, N. & Gordon, N., editors. *Sequential Monte Carlo methods in practice*. Springer-Verlag.

Patrick R., Orchard, M., Zhang, B., Koelemay, M., Kacprzynski, G., Ferri, A. and **Vachtsevanos, G.** 2007: An integrated approach to helicopter planetary gear fault diagnosis and failure prognosis. *42nd Annual Systems Readiness Technology Conference*, AUTOTESTCON 2007, Baltimore, MD, September 2007.

Pitt, M.K. and **Shephard, N.** 1999: Filtering via simulation: auxiliary particle filters. *Journal of the American Statistical Association* 94, 590–99.

Ray, A. and **Tangirala, S.** 1996: Stochastic modeling of fatigue crack dynamics for on-line failure prognosis. *IEEE Transactions on Control Systems Technology* 4, 443–51.

Thrun, S., Langford, J. and **Verma, V.** 2001: Risk sensitive particle filters. *Neural Information Processing Systems* (*NIPS*), December 2001.

Van der Merwe, R., Doucet, A., de Freitas, N. and **Wan, E.** 2006: The unscented particle filter. Technical Report CUED/F-INFENG/TR 380, Cambridge University Engineering Department.

Verma, V., **Gordon, G.**, **Simmons, R.** and **Thrun, S.** 2004: Tractable particle filters for robot fault diagnosis. *IEEE Robotics & Automation Magazine* 11, 56–66.

Verma, V., Thrun, S. and **Simmons, R.** 2003: Variable resolution particle filter. *Proceedings of the International Joint Conference of Artificial Intelligence*.