Multiple Damage Progression Paths in Model-based Prognostics

Matthew Daigle* and Kai Goebel†
*University of California, Santa Cruz, NASA Ames Research Center, Moffett Field, CA, USA
†NASA Ames Research Center, Moffett Field, CA, USA
{matthew.j.daigle,kai.goebel}@nasa.gov

***Abstract*—Model-based prognostics approaches employ domain knowledge about a system, its components, and how they fail through the use of physics-based models. Component wear is driven by several different degradation phenomena, each resulting in their own damage progression path, overlapping to contribute to the overall degradation of the component. We develop a model-based prognostics methodology using particle filters, in which the problem of characterizing multiple damage progression paths is cast as a joint state-parameter estimation problem. The estimate is represented as a probability distribution, allowing the prediction of end of life and remaining useful life within a probabilistic framework that allows for uncertainty management. We also develop a novel variance control mechanism that maintains an uncertainty bound around the hidden parameters to limit the amount of estimation uncertainty and consequently reduce prediction uncertainty. We construct a detailed physics-based model of a centrifugal pump, to which we apply our model-based prognostics algorithms. We illustrate the operation of the prognostic solution with a number of simulation-based experiments and demonstrate the performance of the chosen approach when multiple damage mechanisms are active. Algorithm performance is evaluated using established prognostics metrics.**

I. Introduction

Model-based prognostics approaches employ domain knowledge about a system, its components, and how they fail through the use of physics-based models, derived from first principles, that capture the underlying physical phenomena [1]–[3]. In contrast, data-driven (model-free) approaches use large amounts of run-to-failure data to train machine learning algorithms to make end of life and remaining useful life predictions [4]. Component wear is driven by several different degradation phenomena. Each of these degradation phenomena results in its own damage progression path, which all combine to contribute to the overall degradation of the component. Due to manufacturing variances and differences in usage and environmental conditions, the damage progression rates for the different damage mechanisms vary among components of the same type. This poses considerable challenges to data-driven approaches lacking vast amounts of training data that cover a significant portion of such cases. In the absence of such data, model-based approaches are better-suited, since they use underlying physical models to help estimate the amount of damage and the rates of damage progression.

Extending previous work in [1], [5], [6], we develop a model-based prognostics methodology using particle filters, in which the problem of characterizing multiple damage progression paths is cast as a joint state-parameter estimation problem. The estimate is represented as a probability distribution, allowing the prediction of end of life and remaining useful life within a probabilistic framework that allows for uncertainty management. To reduce the amount of artificial uncertainty introduced by particle filter-based parameter estimation, we introduce a novel variance control mechanism that maintains an uncertainty bound around an unknown parameter being estimated.

We demonstrate our prognostics methodology on a centrifugal pump application. Centrifugal pumps are used in a wide variety of applications, from water supply to spacecraft fueling systems. Because pumps typically see high usage, they can particularly benefit from prognostics and health management solutions to ensure satisfactory system performance, extended component lifetime, and limited downtime. Model-based diagnosis has been investigated previously with centrifugal pumps [7]–[9]. However, most prognostics approaches for pumps have been data-driven, usually based on pump vibration signals. A principal component analysis method is applied for condition monitoring of a pump using vibration signals in [10]. We illustrate our model-based prognostic approach for centrifugal pumps using a number of simulation-based experiments when multiple damage mechanisms are active. We evaluate algorithm performance using established prognostics metrics.

The paper is organized as follows. Section II formally defines the prognostics problem and describes the prognostics architecture. Section III describes the modeling methodology and develops the centrifugal pump model for prognostics. Section IV describes the particle filter-based damage estimation method and develops the variance control scheme. Section V discusses the prediction methodology. Section VI provides results from a number of simulation-based experiments and evaluates the approach. Section VII concludes the paper.

II. Prognostics Approach

The problem of prognostics is to predict the EOL and/or the RUL of a component. In this section, we first formally define the problem of prognostics. We then describe a general model-based architecture within which a prognostics solution may be implemented.

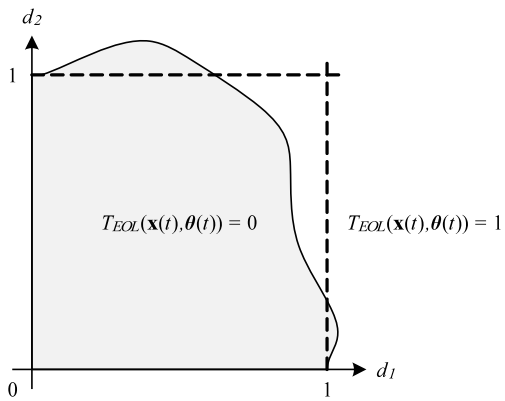

Fig. 1. EOL threshold in the damage space.

A. Problem Formulation

In general, a system model may be defined as

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{f}(t, \mathbf{x}(t), \boldsymbol{\theta}(t), \mathbf{u}(t), \mathbf{v}(t)) \\ \mathbf{y}(t) &= \mathbf{h}(t, \mathbf{x}(t), \boldsymbol{\theta}(t), \mathbf{u}(t), \mathbf{n}(t))\end{aligned}$$

where $\mathbf{x}(t) \in \mathbb{R}^{n_x}$ is the state vector, $\boldsymbol{\theta}(t) \in \mathbb{R}^{n_\theta}$ is the parameter vector, $\mathbf{u}(t) \in \mathbb{R}^{n_u}$ is the input vector, $\mathbf{v}(t) \in \mathbb{R}^{n_v}$ is the process noise vector, $\mathbf{f}$ is the state equation, $\mathbf{y}(t) \in \mathbb{R}^{n_y}$ is the output vector, $\mathbf{n}(t) \in \mathbb{R}^{n_n}$ is the measurement noise vector, and $\mathbf{h}$ is the output equation. This form represents a general nonlinear model with no restrictions on the functional forms of $\mathbf{f}$ or $\mathbf{h}$. Further, the noise terms may be coupled in a nonlinear way with the states and parameters. The parameters $\boldsymbol{\theta}(t)$ evolve in an unknown way. In practice, they are typically considered to be constant, but may, in fact, be time-varying.

Our goal is to predict EOL (and/or RUL) at a given time point t_P using the discrete sequence of observations up to time t_P, denoted as $\mathbf{y}_{0:t_P}$. EOL is defined as the time point at which the component no longer meets a functional requirement (e.g., a pump is overheated). This point is often linked to a damage threshold, beyond which the component fails to function properly. In general, we may express this threshold as a function of the system state and parameters, $T_{EOL}(\mathbf{x}(t), \boldsymbol{\theta}(t))$, which determines whether EOL has been reached, where

$$T_{EOL}(\mathbf{x}(t), \boldsymbol{\theta}(t)) = \begin{cases} 1, & \text{if EOL is reached} \\ 0, & \text{otherwise.} \end{cases}$$

The threshold is linked to a boundary in the multi-dimensional damage space. Inside the boundary, $T_{EOL}(\mathbf{x}(t), \boldsymbol{\theta}(t)) = 0$, and outside the boundary, $T_{EOL}(\mathbf{x}(t), \boldsymbol{\theta}(t)) = 1$. Fig. 1 illustrates this concept with a two-dimensional example, where the damage dimensions are normalized such that $d_i = 1$ corresponds to the maximum allowable damage for d_i, given all other $d_j = 0$. Note that, in general, the different damage mechanisms cannot be considered independently in defining EOL, because increased damage along one dimension may either allow a greater amount of damage or restrict the allowable amount of damage along another damage dimension. For example, in a valve, where EOL is defined by valve opening and closing times, friction damage will cause the valve to open more slowly, but a weakening of the return spring will allow the valve to open more quickly. If this were not the case, then the boundary would form a hypercube in the damage space (denoted by the dashed lines in the figure). The regions of the boundary that extend beyond this hypercube are where more damage is allowed, and the regions that fall within this hypercube are where damage is restricted further.

Using the threshold function, we can formally define EOL with

$$EOL(t_P) \triangleq \arg\min_{t \geq t_P} T_{EOL}(\mathbf{x}(t), \boldsymbol{\theta}(t)) = 1,$$

i.e., EOL is the earliest time point at which the damage threshold is met. RUL may then be defined with

$$RUL(t_P) \triangleq EOL(t_P) - t_P.$$

Note that we are interested in the EOL formed by the combined effects of all damage progressions paths. Therefore, they must be considered simultaneously, rather than independently.

In practice, many sources of uncertainty exist that affect the prediction. Noise is inherent in the process and the measurements, represented by the noise terms $\mathbf{v}(t)$ and $\mathbf{n}(t)$, respectively. Further, the future inputs of the system, which affect the evolution of the state, and therefore the progression of damage, are not always known. Certain input profiles may also excite some damage mechanisms more than others. Thus, it is much more useful to compute a probability distribution of the EOL or RUL, rather than a single prediction point. The goal, then, is to compute, at time t_P, $p(EOL(t_p)|\mathbf{y}_{0:t_P})$ or $p(RUL(t_P)|\mathbf{y}_{0:t_P})$.

B. Prognostics Architecture

In our model-based approach, we develop detailed physics-based models of components and systems that include descriptions of how fault parameters evolve in time. These models depend on unknown and possibly time-varying wear parameters, $\boldsymbol{\theta}(t)$. Therefore, our solution to the prognostics problem takes the perspective of joint state-parameter estimation. In discrete time k, we estimate $\mathbf{x}_k$ and $\boldsymbol{\theta}_k$, and use these estimates to predict EOL and RUL at desired time points.

We employ the prognostics architecture in Fig. 2. The system is provided with inputs $\mathbf{u}_k$ and provides measured outputs $\mathbf{y}_k$. Prognostics may begin at $t = 0$, with the damage estimation module determining estimates of the states and unknown parameters, represented as a probability distribution $p(\mathbf{x}_k, \boldsymbol{\theta}_k|\mathbf{y}_{0:k})$. In parallel, a fault detection, isolation, and identification (FDII) module may be used to determine which damage mechanisms are active, represented as a fault set $\mathbf{F}$. The damage estimation module may then use this to limit the space of parameters that must be estimated. Alternatively, prognostics may begin only when diagnostics has completed. The prediction module uses the joint state-parameter distribution, along with hypothesized future inputs, to compute

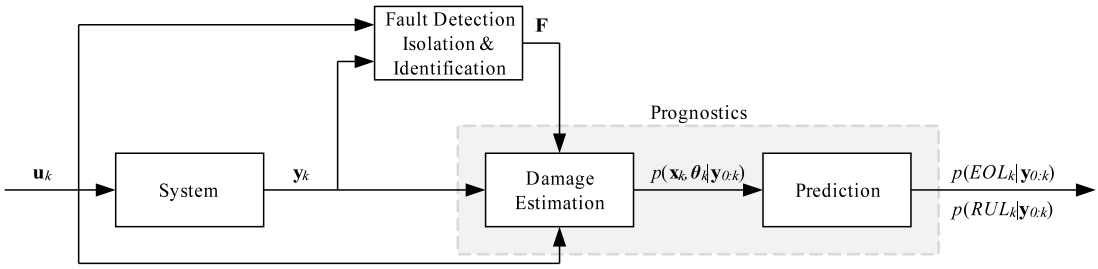

Fig. 2. Prognostics architecture.

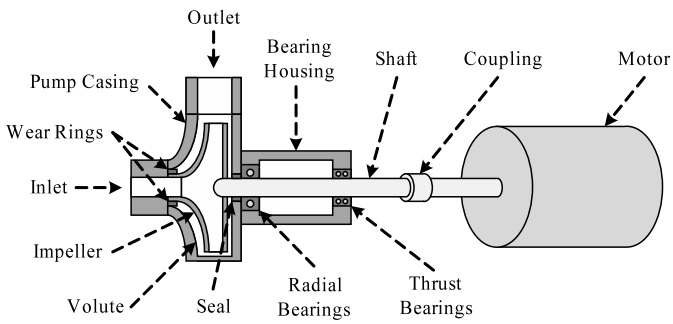

Fig. 3. Centrifugal pump.

EOL and RUL as probability distributions $p(EOL_{k_P}|\mathbf{y}_{0:k_P})$ and $p(RUL_{k_P}|\mathbf{y}_{0:k_P})$ at given prediction times k_P. In this paper, we focus on the damage estimation and prediction modules, and assume that the FDII module does not inform the prognostics, i.e., all possible damage progression paths must be tracked starting from $t = 0$.

III. Pump Modeling

We apply our prognostics approach to a centrifugal pump, and develop a physics-based model of its nominal and faulty behavior. Centrifugal pumps are used in a variety of domains for fluid delivery. A schematic of a typical centrifugal pump is shown in Fig. 3. Fluid enters the inlet, and the rotation of the impeller forces fluid through the outlet. The impeller is driven by an electric motor, typically a three-phase alternating-current induction motor. The radial and thrust bearings help to minimize friction along the pump shaft. The bearing housing contains lubricating oil which bathes the bearings. A seal prevents fluid flow into the bearing housing. Wear rings prevent internal pump leakage from the outlet to the inlet side of the impeller, but a small clearance is typically allowed (a small amount of internal leakage is normal).

The state $\mathbf{x}$ of the pump is given by

$$\mathbf{x}(t) = \begin{bmatrix} \omega(t) \\ T_t(t) \\ T_r(t) \\ T_o(t) \end{bmatrix},$$

where $\omega(t)$ is the rotational velocity of the pump, $T_t(t)$ is the thrust bearing temperature, $T_r(t)$ is the radial bearing temperature, and $T_o(t)$ is the oil temperature.

The rotational velocity of the pump is described using a torque balance,

$$\dot{\omega} = \frac{1}{J}\left(\tau_e(t) - r\omega(t) - \tau_L(t)\right),$$

where J is the lumped motor/pump inertia, τ_e is the electromagnetic torque provided by the motor, r is the lumped friction parameter, and τ_L is the load torque. In an induction motor, a voltage is applied to the stationary part, the stator, which creates a current through the stator coils. With a polyphase supply, this creates a rotating magnetic field which induces a current in the rotating part, the rotor, causing it to turn. A torque is produced on the rotor only when there is a difference between the synchronous speed of the supply voltage, ω_s and the mechanical rotation, ω. This difference, called *slip*, is defined as

$$s = \frac{\omega_s - \omega}{\omega_s}.$$

The expression for the torque τ_e is derived from an equivalent circuit representation for the three-phase induction motor, shown in Fig. 4, based on rotor and stator resistances and inductances, and the slip s [11]:

$$\tau_e = \frac{npR_2}{s\omega_s}\frac{V_{rms}^2}{(R_1 + R_2/s)^2 + (\omega_s L_1 + \omega_s L_2)^2},$$

where R_1 is the stator resistance, L_1 is the stator inductance, R_2 is the rotor resistance, L_2 is the rotor inductance, n is the number of phases (typically 3), and p is the number of magnetic pole pairs. For a 3600 rpm motor, $p = 1$. The dependence of torque on slip creates a feedback loop that causes the rotor to follow the rotation of the magnetic field. The rotor speed may be controlled by changing the input frequency ω_s, e.g., through the use of a variable-frequency drive.

The load torque τ_L is a polynomial function of the flow rate through the pump and the impeller rotational velocity [7], [8]:

$$\tau_L = a_0\omega^2 + a_1\omega Q - a_2 Q^2,$$

where Q is the flow, and a_0, a_1, and a_2 are coefficients derived from the pump geometry [8].

The rotation of the impeller creates a pressure difference from the inlet to the outlet of the pump, which drives the pump flow, Q. The pump pressure is computed as

$$p_p = A\omega^2 + b_1\omega Q - b_2 Q^2,$$

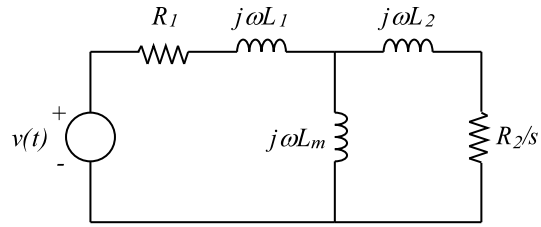

Fig. 4. Induction motor equivalent circuit.

where A is the impeller area, and b_1 and b_2 are coefficients derived from the pump geometry. Flow through the impeller, Q_i, is computed using the pressure differences:

$$Q_i = c\sqrt{|p_s + p_p - p_d|}sign(p_s + p_p - p_d),$$

where c is a flow coefficient, p_s is the suction pressure, and p_d is the discharge pressure. The small (normal) leakage flow from the discharge end to the suction end due to the clearance between the wear rings and the impeller is described by

$$Q_l = c_l A_l \sqrt{|p_d - p_s|}sign(p_d - p_s),$$

where c_l is a flow coefficient and A_l is the leakage flow equivalent area. The discharge flow, Q, is then

$$Q = Q_i - Q_l.$$

Pump temperatures are often monitored as indicators of pump condition. The oil temperature heats up due to the radial and thrust bearings and cools to the environment:

$$\dot{T}_o = \frac{1}{J_o}(H_{o,1}(T_t - T_o) + H_{o,2}(T_r - T_o) - H_{o,3}(T_o - T_a)),$$

where J_o is the thermal inertia of the oil, and the $H_{o,i}$ terms are heat transfer coefficients. Thrust bearing temperature heats up due to the friction between the pump shaft and the bearings, and cools to the oil and the environment:

$$\dot{T}_t = \frac{1}{J_t}(r_t\omega^2 - H_{t,1}(T_t - T_o) - H_{t,2}(T_t - T_a)),$$

where J_t is the thermal inertia of the thrust bearings, r_t is the friction coefficient for the thrust bearings, and the $H_{t,i}$ terms are heat transfer coefficients. The radial bearings behave similarly:

$$\dot{T}_r = \frac{1}{J_r}(r_r\omega^2 - H_{r,1}(T_r - T_o) - H_{r,2}(T_r - T_a))$$

where J_r is the thermal inertia of the radial bearings, r_r is the friction coefficient for the radial bearings, and the $H_{r,i}$ terms are heat transfer coefficients. Note that r_t ande r_r contribute to the overall friction coefficient r.

The overall input vector $\mathbf{u}$ is given by

$$\mathbf{u}(t) = \begin{bmatrix} p_s(t) \\ p_d(t) \\ T_a(t) \\ V(t) \\ \omega_s(t) \end{bmatrix}.$$

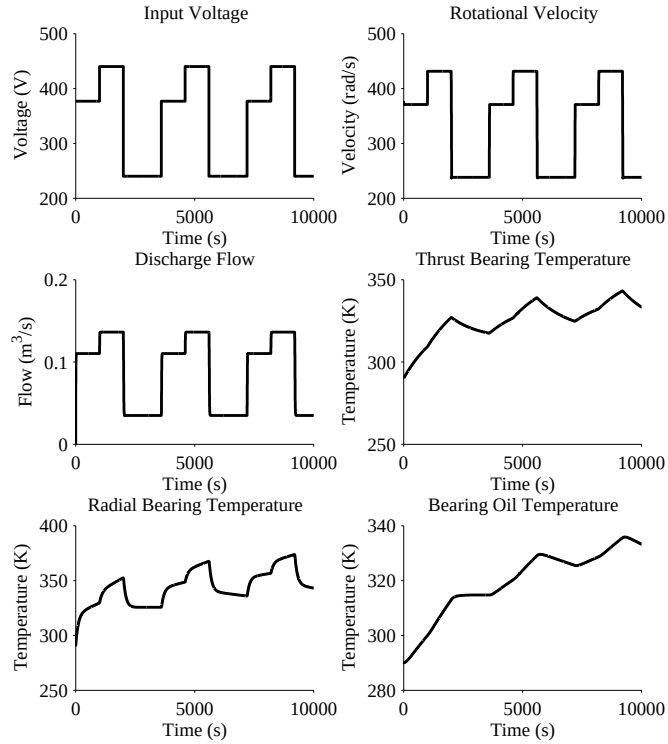

Fig. 5. Nominal pump operation.

The available complete measurement vector $\mathbf{y}$ is given by

$$\mathbf{y}(t) = \begin{bmatrix} \omega(t) \\ Q(t) \\ T_t(t) \\ T_r(t) \\ T_o(t) \end{bmatrix}.$$

Fig. 5 shows nominal pump operation. The input voltage (and frequency) are varied to control the pump speed. The electromagnetic torque is produced initially as slip is 1. This causes a rotation of the motor to match the rotation of the magnetic field, with a small amount of slip remaining, depending on how large the load torque is. As the pump rotates, fluid flow is created. The bearings heat up as the pump rotates and cool when the pump rotation slows.

A. Damage Modeling

The most significant forms of damage for pumps are impeller wear, caused by cavitation and erosion by the flow, and bearing failure, caused by friction-induced wear of the bearings. In each case, we map the form of damage to a particular parameter in the nominal model, and this parameter becomes a state variable that evolves by a damage progression function. This function is parameterized by a set of unknown wear parameters.

Impeller wear is represented as a decrease in impeller area A [9]. We use the erosive wear equation! [12]. Erosive wear rate is proportional to fluid velocity times friction force. Fluid velocity is proportional to fluid volumetric flow rate, and

friction force is proportional to fluid velocity. We lump the proportionality constants into the wear coefficient w_A to obtain

$$\dot{A} = -w_A Q_i^2.$$

A decrease in the impeller area will decrease the pump pressure reduced, which, in turn, reduces the delivered flow, and, therefore, pump efficiency. The pump must operate at a certain minimal efficiency. This requirement defines an EOL criteria. We define A^- as the minimum value of the impeller area at which this requirement is met, hence, $T_{EOL} = 1$ if $A(t) < A^-$.

Bearing wear is captured as an increase in friction. Sliding and rolling friction generate wear of material which increases the coefficient of friction [1], [5], [12]:

$$\dot{r}_t(t) = w_t r_t \omega^2$$
$$\dot{r}_r(t) = w_r r_r \omega^2,$$

where w_t and w_r are the wear coefficients. The slip compensation provided by the electromagnetic torque generation masks small changes in friction, so it is only with very large increases that a change in ω will be observed. These changes can be observed much more readily through the bearing temperatures. Limits on the maximum values of these temperatures define EOL for bearing wear. We define r_t^+ and r_r^+ as the maximum permissible values of the friction coefficients, before the temperature limits are exceeded over a typical usage cycle. So, $T_{EOL} = 1$ if $r_t(t) > r_t^+$ or $r_r(t) > r_r^+$. Vibration and acceleration sensors have also been used in pumps for bearing monitoring, e.g., in [10], however, when using such methods it is difficult to map changes in vibration back to changes in the thrust bearings, radial bearings, or both, while also quantifying the amount of damage.

IV. Damage Estimation

In model-based prognostics, damage estimation reduces to joint state-parameter estimation, i.e., computation of $p(\mathbf{x}_k, \boldsymbol{\theta}_k | \mathbf{y}_{0:k})$. A general solution to this problem is the *particle filter*, which may be directly applied to nonlinear systems with non-Gaussian noise terms [13], [14]. In particle filters, the state distribution is approximated by a set of discrete weighted samples, called *particles*.

With particle filters, the particle approximation to the state distribution is given by

$$\{(\mathbf{x}_k^i, \boldsymbol{\theta}_k^i), w_k^i\}_{i=1}^N,$$

where N denotes the number of particles, and for particle i, $\mathbf{x}_k^i$ denotes the state vector estimate, $\boldsymbol{\theta}_k^i$ denotes the parameter vector estimate, and w_k^i denotes the weight. The posterior density is approximated by

$$p(\mathbf{x}_k, \boldsymbol{\theta}_k | \mathbf{y}_{0:k}) \approx \sum_{i=1}^N w_k^i \delta_{(\mathbf{x}_k^i, \boldsymbol{\theta}_k^i)}(d\mathbf{x}_k d\boldsymbol{\theta}_k),$$

where $\delta_{(\mathbf{x}_k^i, \boldsymbol{\theta}_k^i)}(d\mathbf{x}_k d\boldsymbol{\theta}_k)$ denotes the Dirac delta function located at $(\mathbf{x}_k^i, \boldsymbol{\theta}_k^i)$.

Algorithm 1 SIR Filter

Inputs: $\{(\mathbf{x}_{k-1}^i, \boldsymbol{\theta}_{k-1}^i), w_{k-1}^i\}_{i=1}^N, \mathbf{u}_{k-1:k}, \mathbf{y}_k$
Outputs: $\{(\mathbf{x}_k^i, \boldsymbol{\theta}_k^i), w_k^i\}_{i=1}^N$
for $i = 1$ **to** N **do**
  $\boldsymbol{\theta}_k^i \sim p(\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1}^i)$
  $\mathbf{x}_k^i \sim p(\mathbf{x}_k | \mathbf{x}_{k-1}^i, \boldsymbol{\theta}_{k-1}^i, \mathbf{u}_{k-1})$
  $w_k^i \leftarrow p(\mathbf{y}_k | \mathbf{x}_k^i, \boldsymbol{\theta}_k^i, \mathbf{u}_k)$
end for
$W \leftarrow \sum_{i=1}^N w_k^i$
for $i = 1$ to N **do**
  $w_k^i \leftarrow w_k^i / W$
end for
$\{(\mathbf{x}_k^i, \boldsymbol{\theta}_k^i), w_k^i\}_{i=1}^N \leftarrow \texttt{Resample}(\{(\mathbf{x}_k^i, \boldsymbol{\theta}_k^i), w_k^i\}_{i=1}^N)$

We use the sampling importance resampling (SIR) particle filter, using systematic resampling [15]. The pseudocode for a single step of the SIR filter is shown as Algorithm 1. Each particle is propagated forward to time k by first sampling new parameter values, and then sampling new states using the model. The particle weight is assigned using $\mathbf{y}_k$. The weights are then normalized, followed by the resampling step (see [13]).

Here, the parameters $\boldsymbol{\theta}_k$ evolve by some unknown process that is independent of the state $\mathbf{x}_k$. However, we need to assign some type of evolution to the parameters. The typical solution is to use a random walk, i.e., for parameter θ, $\theta_k = \theta_{k-1} + \xi_{k-1}$, where ξ_{k-1} is sampled from some distribution (e.g., zero-mean Gaussian). With this type of evolution, the particles generated with parameter values closest to the true values should be assigned higher weight, thus allowing the particle filter to converge to the true values.

The selected variance of the random walk noise determines both the rate of this convergence and the estimation performance once convergence is achieved. Therefore, it is very desirable to tune this parameter to obtain the best possible performance. Too large a shock may converge quickly but track with too wide a variance, while too small a shock will take very long to converge, if at all, but once it does, will track with a very small variance. One approach is to use kernel shrinkage, in which the random walk noise is diminished over time [16]. This approach assumes that the parameter is constant, but in reality, this may not be the case, so some amount of noise should still be included to account for unmodeled deviations in the parameter value over time. In [17], [18], this noise (viewed as a model hyper-parameter) is tuned using outer correction loops based on prediction error. In this case, the underlying prognostic model is assumed to contain only a single fault dimension, therefore it cannot be applied in our case.

We develop a $\boldsymbol{\xi}$ adaptation method similar to [17], but with some key distinguishing features. First, we consider a multi-dimensional damage space, therefore, we must simultaneously adapt the random walk noise for multiple parameter values. Second, we cannot use prediction error to drive the adaptation, because we cannot, in general, map errors in prediction to

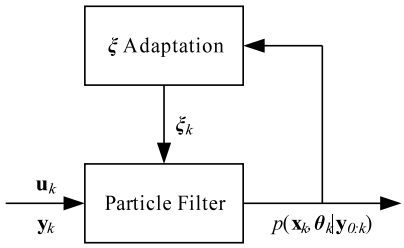

Fig. 6. $\boldsymbol{\xi}$ adaptation scheme.

Algorithm 2 $\boldsymbol{\xi}$ Adaptation

Inputs: $\{(\mathbf{x}_k^i, \boldsymbol{\theta}_k^i), w_k^i\}_{i=1}^N, \boldsymbol{\xi}_{k-1}$
State: $\mathbf{a}$
Outputs: $\boldsymbol{\xi}_k$
if $k = 0$ **then**
  $\mathbf{a} \leftarrow \mathbf{0}$
end if
for all $j \in \{1, 2, \ldots, n_\theta\}$ **do**
  $v_j \leftarrow \text{RMAD}(\{\theta_k^i(j)\}_{i=1}^N)$
  if $\mathbf{a}(j) = 0$ **and** $v_j < T$ **then**
    $\mathbf{a}(j) \leftarrow 1$
  end if
  if $\mathbf{a}(j) = 0$ **then**
    $v_j^* \leftarrow v_{j0}^*$
  else
    $v_j^* \leftarrow v_{j\infty}^*$
  end if
  $\boldsymbol{\xi}_k(j) \leftarrow \boldsymbol{\xi}_{k-1}(j)\left(1 + P\frac{v_j - v_j^*}{v_j^*}\right)$
end for

specific wear parameters, since each output is dependent on multiple damage mechanisms. Instead, we try to control the variance of the hidden wear parameter estimate to a user-specified range by modifying the random walk noise variance. Since the random walk noise is artificial, we should reduce it as much as possible, because this uncertainty propagates into the EOL predictions. So, controlling this uncertainty helps to control the uncertainty of the EOL prediction.

The algorithm for the adaptation of the $\boldsymbol{\xi}$ vector is given as Algorithm 2, and Fig. 6 shows how it interacts with the particle filter. We assume that the $\boldsymbol{\xi}$ values are tuned initially based on the maximum expected wear rates, e.g., if the pump is expected to fail no earlier than 100 hours, then this corresponds to particular maximum wear rate values. The initial wear rate estimate values may start at 0. We use the relative median absolute deviation (RMAD) as the measure of variance:

$$\text{RMAD}(X) = 100\frac{\text{Median}_i\left(|X_i - \text{Median}_j(X_j)|\right)}{\text{Median}_j(X_j)},$$

where X is a data set and X_i is an element of that set. We use RMAD because it is statistically robust, and, because it is a *relative* measure of spread, it can be treated equally for any wear parameter value. The adaptation scheme is essentially a proportional control law, where the error between the actual RMAD of a parameter $\boldsymbol{\theta}(j)$, denoted as v_j in the algorithm, and the desired RMAD value (e.g., 10%), denoted as v_j^* in the algorithm, is normalized by v_j. The error is then multiplied by a factor P (e.g., 1×10^{-3}), and the corresponding variance $\boldsymbol{\xi}(j)$ is increased or decreased by that percentage. We utilize two different setpoints. First, we allow for a convergence period, with setpoint v_{j0}^* (e.g., 50%). Once the v_j reaches T (e.g., $1.2v_{j0}^*$), we mark it using the $\mathbf{a}(j)$ flag, and begin to control it to a new setpoint $v_{j\infty}^*$ (e.g., 10%).

Because there is some inertia to the process of v_j changing in response to a new value of $\boldsymbol{\xi}(j)$, the gain P cannot be too large, otherwise v_j will not converge to the desired value, instead, it will continually shrink and expand. This is illustrated in Fig. 7, where the value of P is varied for estimation of w_A for the pump. For $P = 1 \times 10^{-2}$, this oscillatory behavior occurs because P is too large. In contrast, if P is too small, such as when $P = 1\times10^{-5}$, v_j will converge to v_j^* much more slowly. In our experiments, $P = 1 \times 10^{-3}$ worked well over the entire range of values considered for each wear parameter. Ideally, the wear parameter variance would be zero, but the particle filter needs some amount of noise to accurately track the parameter. So, v_j^* cannot be too small, and we have found that controlling to an RMAD of 10% introduces an acceptable amount of uncertainty while allowing for accurate tracking.

V. Prediction

Prediction is initiated at a given time k_P. Using the current joint state-parameter estimate, $p(\mathbf{x}_{k_P}, \boldsymbol{\theta}_{k_P}|\mathbf{y}_{0:k_P})$, which represents the most up-to-date knowledge of the system at time k_P, the goal is to compute $p(EOL_{k_P}|\mathbf{y}_{0:k_P})$ and $p(RUL_{k_P}|\mathbf{y}_{0:k_P})$. As discussed in Section IV, the particle filter computes

$$p(\mathbf{x}_{k_P}, \boldsymbol{\theta}_{k_P}|\mathbf{y}_{0:k_P}) \approx \sum_{i=1}^N w_{k_P}^i \delta_{(\mathbf{x}_{k_P}^i, \boldsymbol{\theta}_{k_P}^i)}(d\mathbf{x}_{k_P} d\boldsymbol{\theta}_{k_P}).$$

We can approximate a prediction distribution n steps forward as [19]

$$p(\mathbf{x}_{k_P+n}, \boldsymbol{\theta}_{k_P+n}|\mathbf{y}_{0:k_P}) \approx \sum_{i=1}^N w_{k_P}^i \delta_{(\mathbf{x}_{k_P+n}^i, \boldsymbol{\theta}_{k_P+n}^i)}(d\mathbf{x}_{k_P+n} d\boldsymbol{\theta}_{k_P+n}).$$

So, for a particle i propagated n steps forward without new data, we may take its weight as $w_{k_P}^i$. Similarly, we can approximate the EOL as

$$p(EOL_{k_P}|\mathbf{y}_{0:k_P}) \approx \sum_{i=1}^N w_{k_P}^i \delta_{EOL_{k_P}^i}(dEOL_{k_P}).$$

To compute EOL, then, we propagate each particle forward to its own EOL and use that particle's weight at k_P for the weight of its EOL prediction.

If an analytic solution exists for the prediction, this may be directly used to obtain the prediction from the state-parameter distribution. An analytical solution is rarely available, so the general approach to solving the prediction problem is through simulation. Each particle is simulated forward to EOL to obtain the complete EOL distribution. The pseudocode for the

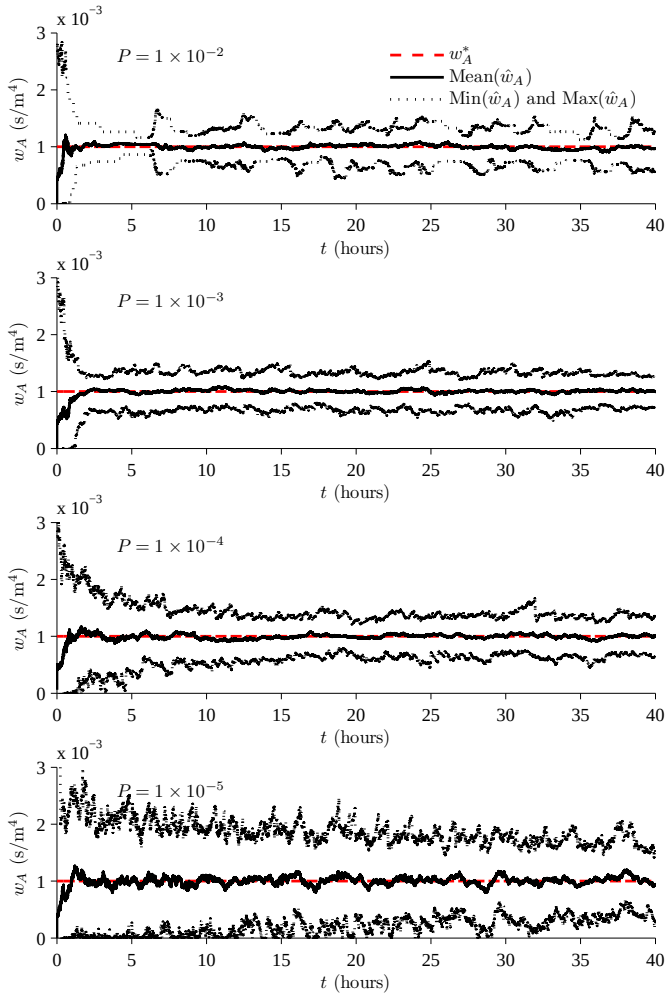

Fig. 7. Estimation of w_A with different values of P.

Algorithm 3 EOL Prediction

Inputs: $\{(\mathbf{x}^i_{k_P}, \boldsymbol{\theta}^i_{k_P}), w^i_{k_P}\}_{i=1}^N$
Outputs: $\{EOL^i_{k_P}, w^i_{k_P}\}_{i=1}^N$
for $i = 1$ **to** N **do**
  $k \leftarrow k_P$
  $\mathbf{x}^i_k \leftarrow \mathbf{x}^i_{k_P}$
  $\boldsymbol{\theta}^i_k \leftarrow \boldsymbol{\theta}^i_{k_P}$
  while $T_{EOL}(\mathbf{x}^i_k, \boldsymbol{\theta}^i_k) = 0$ **do**
    Predict $\hat{\mathbf{u}}_k$
    $\boldsymbol{\theta}^i_{k+1} \sim p(\boldsymbol{\theta}_{k+1}|\boldsymbol{\theta}^i_k)$
    $\mathbf{x}^i_{k+1} \sim p(\mathbf{x}_{k+1}|\mathbf{x}^i_k, \boldsymbol{\theta}^i_k, \hat{\mathbf{u}}_k)$
    $k \leftarrow k + 1$
    $\mathbf{x}^i_k \leftarrow \mathbf{x}^i_{k+1}$
    $\boldsymbol{\theta}^i_k \leftarrow \boldsymbol{\theta}^i_{k+1}$
  end while
  $EOL^i_{k_P} \leftarrow k$
end for

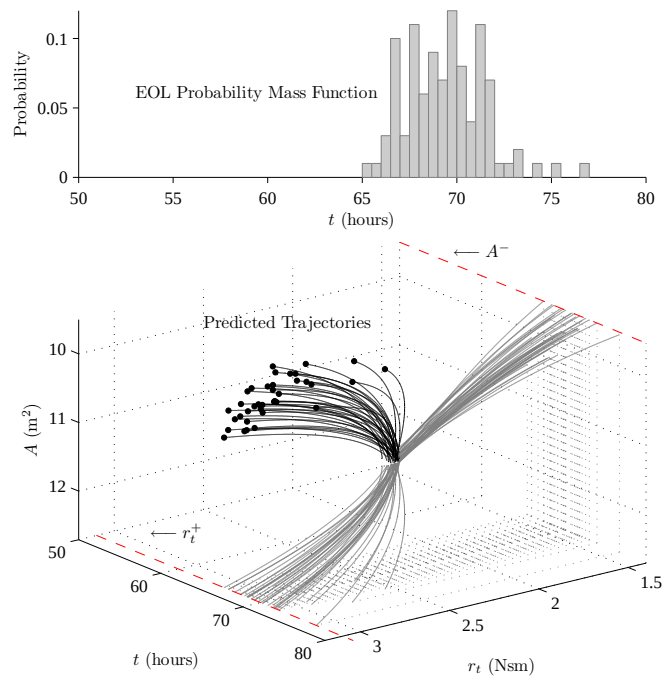

Fig. 8. Simultaneous prediction of impeller wear and thrust bearing wear in the pump.

prediction procedure is given as Algorithm 3 [1]. Each particle i is propagated forward until $T_{EOL}(\mathbf{x}^i_k, \boldsymbol{\theta}^i_k)$ evaluates to 1; at this point EOL has been reached for this particle.

Note that prediction requires hypothesizing future inputs of the system, $\hat{\mathbf{u}}_k$, because damage progression is rarely independent of the system inputs. For the pump, an increased rotation speed will cause bearing friction to increase at a faster rate, and will cause an increased pump flow, which, in turn, will cause impeller wear to increase at a faster rate. The choice of future expected inputs depends on what kind of information the user wants, e.g., for a worst-case scenario one would consider the pump running at its maximum rotation.

Fig. 8 shows example results from the simultaneous prediction of impeller wear and thrust bearing wear for $N = 100$ (not all of the trajectories are shown in the lower plot) for a single future input trajoectory. Initially, the particles have a very tight distribution of friction and impeller area damage values, but the distribution of the wear parameters, w_A and w_{r_t}, is relatively large. As a result, the individual trajectories for the different wear parameter values are easily distinguishable as EOL is approached. Because the the damage threshold is multi-dimensional, we show also the projections of the trajectories onto the damage-time planes. The dotted lines from the ends of those projections onto the time axis distinguish the EOLs. For some particles, A^- is reached first, while for others, r_t^+ is reached first. The different EOL values along with particle weights form an EOL distribution approximated by the probability density function shown in the upper plot.

VI. Results

In this section, we present simulation-based experiments to analyze the performance of the prognostics algorithm in the case of multiple damage progression paths. We first define the metrics used to evaluate the algorithm performance. We then provide detailed results for a single experiment to demonstrate

the approach, followed by results summarized over a large number of experiments.

A. Evaluation Metrics

We evaluate the performance of the wear parameter estimation by quantifying estimation accuracy and spread. Accuracy is calculated using the percentage root mean square error (PRMSE), which expresses relative estimation accuracy of w as a percentage:

$$\text{PRMSE}_w = 100\sqrt{\text{Mean}_k\left[\left(\frac{\hat{w}_k - w_k^*}{w_k^*}\right)^2\right]},$$

where $\hat{w}_k$ denotes the estimated wear parameter value at time k, w_k^* denotes the true wear parameter value at k, and Mean_k denotes the mean over all values of k. In computing PRMSE, we ignore the initial time frame associated with convergence of the wear parameter estimate.

We calculate the spread using RMAD as defined in Section IV. For estimation spread, for time k we, compute for wear parameter w, $\text{RMAD}_{w,k}$ using the distribution of wear parameter values given by the particle set at k as the data set. We denote the average RMAD over multiple k using $\overline{RMAD}_w$:

$$\overline{\text{RMAD}}_w = \text{Mean}_k(\text{RMAD}_{w,k}).$$

In computing estimation spread, we also ignore the initial time frame associated with convergence of the wear parameter estimate.

For a particular prediction point k_P, we compute measures of accuracy and spread for the prediction. For accuracy, we use the relative accuracy (RA) metric [20]:

$$\text{RA}_{k_P} = 100\left(1 - \frac{|RUL_{k_P}^* - \text{Mean}_i(RUL_{k_P}^i)|}{RUL_{k_P}^*}\right).$$

RA is averaged over each prediction point to obtain a single value that characterizes the overall accuracy, denoted as $\overline{\text{RA}}$.

We calculate prediction spread using RMAD, which we denote as RMAD_{RUL} for the RUL prediction. To obtain a single value for overall spread, RSD and RMAD are averaged over all prediction points starting from the prediction at which a prognostics horizon (where RA is within a specified bound) is first reached, denoted using $\overline{\text{RMAD}}_{RUL}$. Prognostics performance is summarized using the α-λ metric which requires that for a given prediction time λ, at least β of the probability mass of the RUL prediction lies within α of the true value [20].

B. Demonstration of Approach

We first provide an example scenario to illustrate the approach. Fig. 9 shows the estimation results for the hidden wear parameters. Estimation error for the sensor signals was negligible, so those plots are not shown. In this case, $w_A^* = 2 \times 10^{-3}$, $w_t^* = 4 \times 10^{-11}$, and $w_r^* = 2 \times 10^{-11}$. Initially, the estimate bounds are significantly above the true value, however, as the estimates begin to converge the RMAD of each is reduced to 50% through the adaptation scheme,

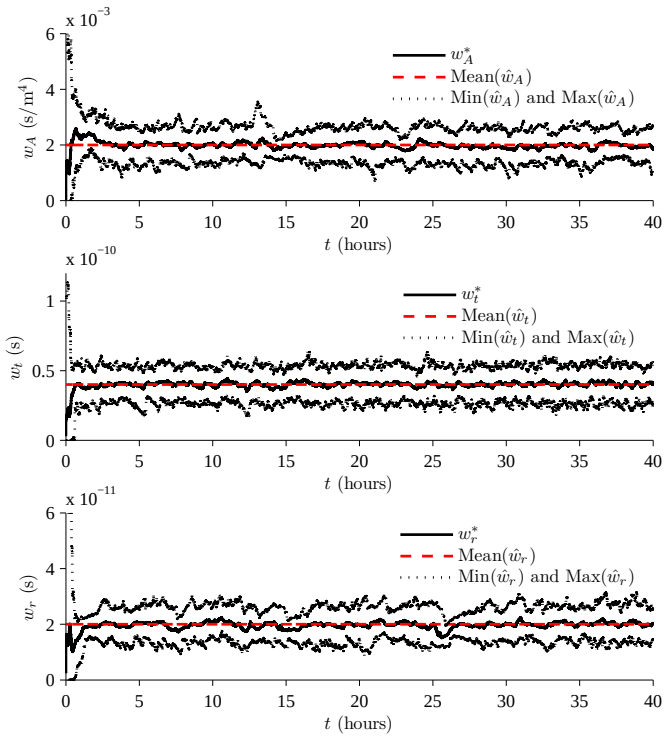

Fig. 9. Simultaneous estimation of pump wear parameters for $N = 500$, $T = 60\%$, $v_0^* = 50\%$, $v_\infty^* = 10\%$, and $P = 1 \times 10^{-3}$.

and then to 10%. Once convergence has occurred, tracking proceeds very well. The RMAD is maintained around 10% to the end of the experiment. The PRMSE of the different wear parameters are correspondingly low, with $\text{PRMSE}_{w_A} = 4.36$ $\text{PRMSE}_{w_t} = 3.60$, and $\text{PRMSE}_{w_r} = 5.51$. The mean RMADs of the wear parameters are $\overline{\text{RMAD}}_{w_A} = 8.60$, $\overline{\text{RMAD}}_{w_t} = 8.42$, and $\overline{\text{RMAD}}_{w_r} = 8.29$, which are less than the controlled value of 10%.

Prediction performance is shown by the α-λ of Fig. 10. Here, the impeller wear damage dominates the EOL prediction. The accurate and precise wear parameter estimates yield correspondingly accurate and precise RUL predictions. Here, $\alpha = 0.1$ and $\beta = 0.5$, so the α-λ test requires that 50% of the probability mass lies within 10% of the true value at each prediction point. The test succeeds at all but the last prediction point, although the probability mass contained within the α-bounds, 49.6%, is very close to the requirement of 50%. The average relative accuracy, using the mean of the RUL distribution, is 97.16%. The average RMAD of the RUL distribution is 9.14%. Maintaining the variance of the wear parameter estimates maintains also the RMAD of the RUL (though not necessarily to the same setpoint).

C. Simulation Results

We performed a number of simulation experiments in which combinations of wear parameter values were selected randomly within a range, with $N = 500$. We selected values in $[0.5 \times 10^{-3}, 4 \times 10^{-3}]$ at increments of 0.5×10^{-3} for w_A,

TABLE I
ESTIMATION AND PREDICTION PERFORMANCE

n	PRMSE_{w_A}	PRMSE_{w_t}	PRMSE_{w_r}	$\overline{\text{RMAD}}_{w_A}$	$\overline{\text{RMAD}}_{w_t}$	$\overline{\text{RMAD}}_{w_r}$	$\overline{\text{RA}}$	$\overline{\text{RMAD}}_{RUL}$
1	6.10	7.06	4.58	8.48	8.39	8.29	96.11	10.21
10	4.56	2.70	3.30	8.59	8.80	8.53	96.74	10.62
100	4.50	2.84	2.49	9.08	8.84	8.87	95.77	11.63

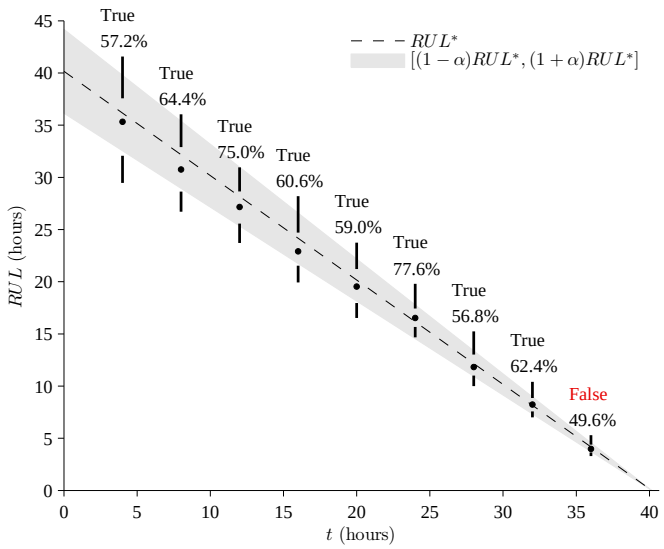

Fig. 10. α-λ performance with $\alpha = 0.1$ and $\beta = 0.5$ for $N = 500$, $T = 60\%$, $v_0^* = 50\%$, $v_\infty^* = 10\%$, and $P = 1 \times 10^{-3}$.

in $[0.5\times10^{-11}, 7\times10^{-11}]$ at increments of 0.5×10^{-11} for w_t, and in $[0.5 \times 10^{-11}, 7 \times 10^{-11}]$ at increments of 0.5×10^{-11} for w_r, such that the maximum wear rates corresponded to a minimum EOL of 20 hours. In order to confirm that the wear parameter variance could still be maintained with additional sensor noise, we varied the sensor noise variance by factors of 1, 10, and 100, and performed 20 experiments for each case. We considered the case where future input of the pump is known, and it is always operated at a constant RPM. This way, no additional uncertainty is introduced and the remaining uncertainty reduces to that involved in the process and sensor noise terms and that introduced by the particle filtering algorithm.

The averaged estimation and prediction performance results are shown in Table I. In all experiments, we used $T = 60\%$, $v_0^* = 50\%$, $v_\infty^* = 10\%$, and $P = 1 \times 10^{-3}$. In each of the cases, the PRMSE for the different wear parameter estimates remained at most around 7% for the normal amount of noise, and under 5% for increased noise. We attribute the higher PRMSE of the normal noise cases to a couple outlier scenarios where convergence was slower, throwing off the estimate early on. In these cases the median PRMSEs were under 5%. The PRMSE for w_A is on average higher than that for the bearing wear parameters because the flow measurement Q is relatively more noisy than the temperature measurements T_t and T_r.

The RMAD of each wear parameter was successfully controlled to 10%, averaging around 8 to 9%. This translated to good prediction performance, with the RA averaging around 96% and the RMAD of the RUL prediction averaging around 11%. Even as the noise increases, the variance control scheme was able to maintain the RMAD setpoint, and so the spread of the RUL prediction increased only slightly as the amount of sensor noise increased.

Fig. 11 shows the RMAD of the wear parameters as a function of wear parameter value. Here, it is clear that the RMAD can be controlled well independently of the wear parameter value. Performance is similar across different wear parameters and their values, which then translates to the similar prediction performance observed across different wear parameter values in Table I.

VII. CONCLUSIONS

We investigated the issues of multiple damage progression paths and developed a model-based prognostics methodology to accommodate them. Damage progression paths are characterized by a fault or damage variable and a set of wear parameters that describe how they evolve in time. Particle filters perform joint state-parameter estimation in order to estimate the health state of the component. The state-parameter distribution is then extrapolated to the EOL threshold to compute RUL predictions in the presence of multiple damage progression paths. A novel variance control mechanism keeps the uncertainty necessary for proper functioning of the particle filter in check, in order to maintain the uncertainty of the unknown wear parameters at a certain level. The framework was applied to a centrifugal pump, and the results demonstrated good performance over a range of wear parameter values and sensor noise levels.

In higher dimensional systems, the particle filter requires a very large number of particles to track successfully. Using only 500 particles was sufficient enough for good results here, but as the number of states or damage mechanisms needed to be tracked increases, the number of particles must increase also. For large enough N, the particle filter approach may not be efficient enough. In future work, we would like to investigate alternative approaches with reduced computational burden for high-dimensional state spaces. Also, the model-based approach presented here could possibly be complemented by data-driven methods that utilize pump vibration or acceleration sensors, in a hybrid model-based and data-driven approach.

ACKNOWLEDGMENTS

The funding for this work was provided by the NASA Fault Detection, Isolation, and Recovery (FDIR) project under the

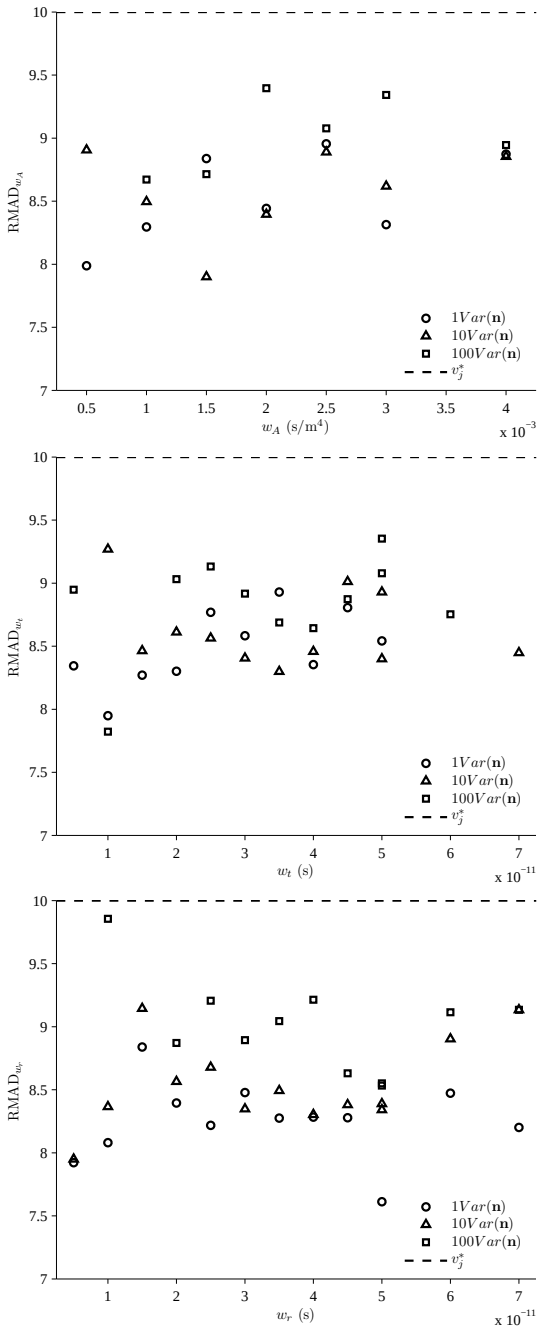

Fig. 11. RMAD of the wear parameter as a function of wear parameter value.

Exploration Technology and Development Program (ETDP) of the Exploration Systems Mission Directorate (ESMD).

REFERENCES

[1] M. Daigle and K. Goebel, "Model-based prognostics under limited sensing," in *2010 IEEE Aerospace Conference*, Mar. 2010.

[2] B. Saha and K. Goebel, "Modeling Li-ion battery capacity depletion in a particle filtering framework," in *Proceedings of the Annual Conference of the Prognostics and Health Management Society 2009*, Sept. 2009.

[3] M. Roemer, C. Byington, G. Kacprzynski, and G. Vachtsevanos, "An overview of selected prognostic technologies with reference to an integrated PHM architecture," in *Proceedings of the First International Forum on Integrated System Health Engineering and Management in Aerospace*, 2005.

[4] M. Schwabacher, "A survey of data-driven prognostics," in *Proceedings of the AIAA Infotech@Aerospace Conference*, 2005.

[5] M. Daigle and K. Goebel, "Improving computational efficiency of prediction in model-based prognostics using the unscented transform," in *Annual Conference of the Prognostics and Health Management Society 2010*, Oct. 2010.

[6] ——, "Model-based prognostics with fixed-lag particle filters," in *Proceedings of the Annual Conference of the Prognostics and Health Management Society 2009*, Sept. 2009.

[7] A. Wolfram, D. Fussel, T. Brune, and R. Isermann, "Component-based multi-model approach for fault detection and diagnosis of a centrifugal pump," in *Proceedings of the 2001 American Control Conference*, vol. 6, 2001, pp. 4443–4448.

[8] C. Kallesøe, "Fault detection and isolation in centrifugal pumps," Ph.D. dissertation, Aalborg University, 2005.

[9] G. Biswas and S. Mahadevan, "A hierarchical model-based approach to systems health management," in *2007 IEEE Aerospace Conference*, Mar. 2007.

[10] S. Zhang, M. Hodkiewicz, L. Ma, and J. Mathew, "Machinery condition prognosis using multivariate analysis," *Engineering Asset Management*, pp. 847–854, 2006.

[11] S. E. Lyshevski, *Electromechanical Systems, Electric Machines, and Applied Mechatronics*. CRC, 1999.

[12] I. M. Hutchings, *Tribology: friction and wear of engineering materials*. CRC Press, 1992.

[13] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[14] O. Cappe, S. J. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential Monte Carlo," *Proceedings of the IEEE*, vol. 95, no. 5, p. 899, 2007.

[15] G. Kitagawa, "Monte Carlo filter and smoother for non-Gaussian non-linear state space models," *Journal of Computational and Graphical Statistics*, vol. 5, no. 1, pp. 1–25, 1996.

[16] J. Liu and M. West, "Combined parameter and state estimation in simulation-based filtering," *Sequential Monte Carlo methods in Practice*, pp. 197–223, 2001.

[17] M. Orchard, G. Kacprzynski, K. Goebel, B. Saha, and G. Vachtsevanos, "Advances in uncertainty representation and management for particle filtering applied to prognostics," in *Proceedings of International Conference on Prognostics and Health Management*, Oct. 2008.

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

[19] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, pp. 197–208, 2000.

[20] A. Saxena, J. Celaya, B. Saha, S. Saha, and K. Goebel, "Metrics for offline evaluation of prognostic performance," *International Journal of Prognostics and Health Management (IJPHM)*, vol. 1, 2010.