Availability Model of a PHM-Equipped Component

Michele Compare, Luca Bellani, Enrico Zio

Availability Model of a PHM-Equipped Component

Michele Compare, Luca Bellani, and Enrico Zio

***Abstract*—A variety of prognostic and health management (PHM) algorithms have been developed in the last years and some metrics have been proposed to evaluate their performances. However, a general framework that allows us to quantify the benefit of PHM depending on these metrics is still lacking. We propose a general, time-variant, analytical model that conservatively evaluates the increase in system availability achievable when a component is equipped with a PHM system of known performance metrics. The availability model builds on metrics of literature and is applicable to different contexts. A simulated case study is presented concerning crack propagation in a mechanical component. A simplified cost model is used to compare the performance of predictive maintenance based on PHM with corrective and scheduled maintenance.**

***Index Terms*—Availability, cost-benefit analysis, Monte Carlo (MC) simulation, prognostics and health Management (PHM) metrics.**

NOMENCLATURE

Δt	Time interval between two successive remaining useful life (RUL) predictions.
λ	Time window modifier, such that $t_\lambda = T_{\mathrm{pr}} + \lambda(T_f - T_{\mathrm{pr}})$; $\lambda \in [0, 1]$.
$\lfloor x \rfloor$	Integer part of x; that is, $n \leq x < n + 1$, $x \in \mathbb{R}$, $n \in \mathbb{N}$.
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2.
$\mathcal{U}(a, b)$	Uniform distribution between a and b.
$\mathcal{W}(a, b)$	Weibull distribution with shape parameter a and scale parameter b; the cumulative density function (cdf) is $F_\mathcal{X}(x) = 1 - e^{(\frac{x}{a})^b}$.
$\mu_\mathcal{X}$	Mean of random variable $\mathcal{X}$.
$\mu_{\overline{\Upsilon_\lambda}}$	Mean of Υ_λ conditional on RUL is overestimated.
$\mu_{\underline{\Upsilon_\lambda}}$	Mean of Υ_λ conditional on RUL is underestimated.
$\overline{RA}$	Relative accuracy overestimation, which quantifies the percentage error of the RUL predictions, only when the RUL is overestimated.
$\overline{ra}$	Mean of $\overline{\mathrm{RA}}$.
$\sigma_\mathcal{X}$	Standard deviation of random variable $\mathcal{X}$.
$\sigma_{\overline{\Upsilon_\lambda}}$	Standard deviation of Υ_λ conditional on RUL is overestimated.
$\sigma_{\underline{\Upsilon_\lambda}}$	Standard deviation of Υ_λ conditional on RUL is underestimated.
$\underline{RA}$	Relative accuracy underestimation, which quantifies the percentage error of the RUL predictions, only when the RUL is underestimated.
$\underline{ra}$	Mean of $\underline{\mathrm{RA}}$.
Υ_λ	Point value summarizing the uncertainty in R_λ (e.g., mean, median, 10th percentile, etc.).
$A(t)$	Component availability at time t.
$Be(p)$	Bernoulli distribution with parameter p: $\mathcal{X} \sim Be(p) \Rightarrow \mathcal{X} \in \{0, 1\}$ and $\mathbb{P}(X = 1) = p$.
C	Cost of the PHM-driven maintenance.
C_{cor}	Cost of the corrective maintenance action.
c_{cor}	Cost of a single corrective maintenance action.
c_{DT}	Cost of component operation interruption.
c_{prev}	Cost of a single preventive maintenance action.
C_{sched}	Cost of the scheduled maintenance action.
DT	Component downtime over the time horizon T $\mathrm{DT} = \int_0^T (1 - A(t))dt$.
DTD	Detection time delay, $T_{\mathrm{det}} - T_d$.
f_{DTD}	Probability density function (pdf) of DTD.
f_{R_λ}	pdf of the predicted RUL at the time window indicated by λ.
$f_{T_d}(t)$	pdf of time T_d.
$f_{T_f}(t)$	pdf of failure time T_f.
fn	False negatives.
fp	False positives.
$h \cdot \Delta t$	Time required to maintenance decision, if $\Upsilon_\lambda < h \cdot \Delta t$.
$j \cdot \Delta t$	Time required to arrange predictive maintenance activity.
$MTTF$	Mean time to failure, i.e., $\mathbb{E}[T_f]$.
N	Number of maximum RUL predictions before failure.
N_{cor}	Number of corrective maintenance actions during the whole life-cycle of the component.
N_{prev}	Number of predictive maintenance actions during the whole life-cycle of the component.
$PS(M)$	Prediction spread of metric M.
R_λ	Uncertain predicted RUL at the time indicated by λ.
RA	Relative accuracy.
ra	Expected value of relative accuracy.
RUL_λ^*	Actual RUL at the time indicated by λ.
T	Component time-horizon.
T_d	Time instant at which the system reaches the detection threshold.

Manuscript received July 31, 2016; revised December 15, 2016; accepted February 11, 2017. Date of publication March 16, 2017; date of current version May 31, 2017. The work of E. Zio was supported in part by the China NSFC under Grant 71231001. Associate Editor: H. Waeselynck.

M. Compare is with the Energy Department, Politecnico di Milano, Milano 20133, Italy, and also with the Aramis S.r.l., Milano 20124, Italy (e-mail: michele.compare@aramis3d.com).

L. Bellani is with the Energy Department, Politecnico di Milano, Milano 20133, Italy (e-mail: luca.bellani1991@gmail.com).

E. Zio is with the Energy Department, Politecnico di Milano, Milano 20133, Italy, the Aramis S.r.l., Milano 20124, Italy, and also with the Foundation Electricité de France, CentraleSupélec, Châtenay-Malabry 92290, France (e-mail: enrico.zio@polimi.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TR.2017.2669400

T_f	Time instant at which the system reaches the failure threshold.
T_{cor}	Time required to perform corrective maintenance.
T_{det}	Actual detection time.
T_{prev}	Time required to perform predictive maintenance.
T_{pr}	Time of the first RUL prediction.
T_s	Prediction time at which the decision to remove the component from operation is taken.
$W(t)$	Binary indicator variable: $W(t) = 1$ if the component is working at time t, $W(t) = 0$, otherwise.

I. Introduction

PROGNOSTICS and health management (PHM) focuses on detection (i.e., the recognition of a deviation from the normal operating condition), diagnostics (i.e., the characterization of the abnormal state), and prognostics (i.e., the prediction of the evolution of the abnormal state up to failure) [1]–[5].

PHM is very important for industry because it allows identifying problems at an early stage and timely performing the necessary maintenance actions to anticipate failures ([6], [7]). The estimation of the component remaining useful life (RUL) enables setting an efficient and agile maintenance management, capable of providing the right part to the right place at the right time, together with the necessary resources to perform the maintenance task. This reduces the interruption of business operations and possible additional malfunctions introduced by errors deriving from maintenance.

Boosted by the appealing potential of PHM for industry, a large number of algorithms have been developed in recent years (see [8]–[10] for overviews). PHM engineers look at the available alternative PHM solutions to find the best combination for their problem. For this, performances must be compared to make a decision about the best portfolio of solutions to invest in ([11]–[13]).

A variety of performance metrics and indicators have been introduced for detection (e.g., [2], [14], [15]), diagnostics (e.g., [16]), and prognostics (e.g., [10], [17]–[19]). Based on a thorough survey of the literature on prediction metrics in different engineering fields, a classification of prognostic metrics into three main groups is proposed in [17], driven by the functional need that the information provided by the metrics relates to the following.

1) Algorithmic performance metrics, which look at the capability of the PHM algorithms of predicting the future evolution of the component degradation. The metrics of this class are further divided into the following four groups.
   a) Accuracy-based, i.e., metrics evaluating the closeness between the estimated and the corresponding true values of the RUL.
   b) Precision-based, i.e., metrics evaluating the variability of the RUL estimations.
   c) Robustness-based metrics, i.e., metrics related to the ability of the algorithm to provide RUL estimations that tolerate perturbations.
   d) Trajectory-based, i.e., metrics evaluating the capability of trajectory prediction.
2) Computational metrics, which refer to the computational burden of the PHM algorithms. These metrics are particularly important when selecting algorithms for online prognostic applications, where the time to provide the estimation becomes a fundamental decision driver.
3) Cost-benefit metrics, which are intended to trade of the benefit gained with the prognostic capability against its cost.

The metrics of the latter class are fundamental for companies that have to decide about investing to purchase the necessary instrumentation, software, and specialized knowledge to yield benefits from PHM. Obviously, both PHM costs and benefits are expected to depend on the algorithmic performance metrics: roughly speaking, the stronger the detection, diagnostic, and prognostic capabilities, the larger the benefit brought by PHM, the larger its cost. For this, linking the cost-benefit metrics to the algorithmic performance metrics is fundamental for conveying investments to PHM and, thus, for the development of the PHM technology.

In spite of its relevance, this issue has been addressed in a few works, only. For example, return on investment has been used as a cost-benefit metric in [19], [20], and [21], where, however, the predicted RUL is assumed to obey a known distribution, which is not soundly related to the algorithm performance metrics. Also, the decisions about when to remove the system from operation are considered not dependent on time. Another cost-benefit metric proposed in the PHM literature is the technical value (TV, [15]), which depends on the performance in detection, diagnostics, and prognostics of critical failure modes and on the costs associated with false alarms. However, TV relies on cost terms that are difficult to estimate (e.g., the savings realized by isolating a fault in advance) and assumes, again, constant performance metrics, whereas, in practice, they depend on time (e.g., the probability of a failure mode). Moreover, TV does not fully account for the scenarios stemming from erroneous detection, diagnosis, and prognosis.

The aim of this work is to build a general, time-dependent, modeling framework to link a set of selected PHM algorithmic performance metrics to the component availability, the metric that enters most of the cost-benefit models used by decision makers (DMs) to select the best option to invest in. The assumption underlying this framework is that any PHM system can be summarized by a set of parameters (i.e., performance metrics), which are the input variables of the mathematical model that estimates the PHM costs and benefits, accounting for the related uncertainties. This entails that we do not need to run any PHM algorithm to estimate the component availability; rather, we only need to know the values of its algorithmic performance metrics, whichever the PHM system is.

It turns out that one of the main advantages of the proposed modeling framework is that it enables estimating the system availability also before the PHM system is developed. Indeed, the mapping between availability and algorithmic performance metrics can also be used to identify the PHM system design specifications, which the PHM developers have to fulfill to guarantee the economic viability of the PHM investment.

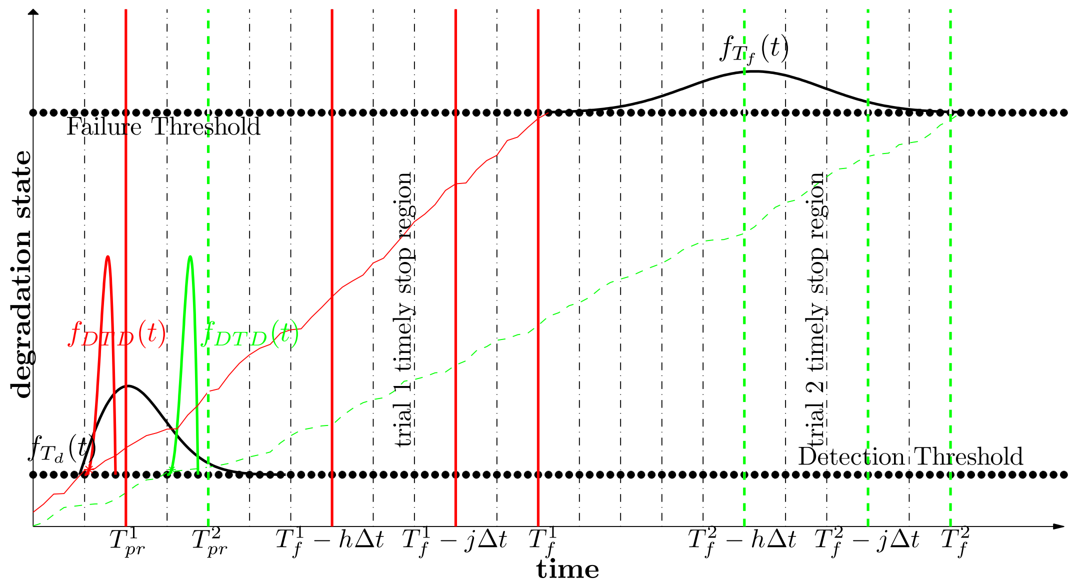

Fig. 1. Examples of degradation evolutions over time.

The remainder of the paper is organized as follows. Section II briefly introduces the model setting. Section III considers the impact of PHM on system availability. Section IV illustrates a simulated case study of single crack propagation and performs a sensitivity analysis of the availability model. Section V compares the operational costs of PHM-driven maintenance and corrective and scheduled maintenance. Section VI concludes the paper.

II. Model Setting

Consider a degrading component, which is monitored every Δt units of time with respect to a continuous indicator variable of the degradation state (see Fig. 1). The degradation process is stochastic and the monitored degradation state variable is characterized by two thresholds: the detection threshold, which mainly depends on the characteristics of the instrumentation used to measure the degradation variable (i.e., for values below this threshold it is not possible to detect the degradation state), and the failure threshold, above which the component function is lost.

The uncertainty in the time instant T_d at which the component reaches the first threshold is described by the probability density function (pdf) $f_{T_d}(t)$. If no action is taken, the component continues to degrade up to failure time T_f, whose uncertainty is described by the pdf $f_{T_f}(t)$.

We realistically assume that detection is not perfect and, thus, we use metrics of literature to characterize the detection performance. In this respect, two detection metrics are widely used in practice: false positive probability (i.e., the probability of triggering undue alarms) and false negative probability (i.e., the probability of missing due alarms) ([14], [15]). An additional detection metric is the detection time delay (DTD, [2]), which measures the interval from the time when the component reaches the detectable degradation state and its detection. We use this latter performance metric, motivated by a twofold justification: on one side, DTD can be regarded as a time-dependent false negative indicator (i.e., alarms are certainly missing up to DTD); on the other side, the DTD values depend on the detection algorithm settings, which can be adjusted such that the false positive probability is negligible in the early phases of the component life ([2]). This introduces a simplification in the model development. To be realistic, we assume that DTD is affected by uncertainty, which is described by the pdf $f_{\mathrm{DTD}}(t)$.

In this setting, the PHM system starts to predict the RUL at time $T_{\mathrm{pr}} = (\lfloor \frac{T_d + \mathrm{DTD}}{\Delta t} \rfloor + 1)\Delta t$, where $\lfloor \Box \rfloor$ indicates the integer part of its argument. The number of predictions that the PHM system can perform before failure is $N = \lfloor \frac{T_f - T_{\mathrm{pr}}}{\Delta t} \rfloor$. From now on, we assume that the system actually fails at time $T_{\mathrm{pr}} + N\Delta t$, instead of T_f; the smaller Δt, the smaller the approximation.

Finally, notice that in the modeling framework developed in this paper we assume, for simplicity, that the PHM-equipped component is affected by a single failure mode. This assumption prevents us from tackling the complex issue of embedding diagnostic metrics into the availability model, and considering all consequent scenarios that originate from decisions based on erroneous diagnoses of the failure mode. This diagnostic issue will be investigated in future research work.

III. Availability Model of a Component Equipped With a PHM System

In this section, we build the mathematical model of the availability of a component equipped with a PHM system. The availability at time t, $A(t)$, is defined as [22]

$$A(t) = \mathbb{P}(W(t) = 1) \tag{1}$$

where $W(t)$ is the component working indicator function and $W(t) = 1$ if the component is working and $W(t) = 0$, otherwise. Notice that this definition is different from that used in [23], where the probability in (1) is integrated up to t, and divided by t, which gives the average availability up to t [22].

To compute $A(t)$ for the PHM-equipped component, we need to link PHM metrics and maintenance decisions with the probability to remove the component from operation during its life. As already highlighted, the benefit of the PHM system mainly lies in that the knowledge of the predicted RUL allows timely arranging the maintenance actions: the time required to perform the PHM-driven preventive maintenance action, $T_{\rm prev}$ (i.e., replace the degraded component before failure), is smaller than the time $T_{\rm cor}$ required to reset the system into operation upon failure.

For generality, we define T_s as the prediction time at which the decision to remove the component from operation is taken, provided that the Decision Maker (DM) always removes the component upon receiving the alarm from the PHM system. We also assume that maintenance decisions are based on point-wise estimations, Υ_λ, of the component RUL predicted at time $t_\lambda = T_{\rm pr} + \lambda(T_f - T_{\rm pr})$, $\lambda \in [0, 1]$. This allows applying the developed model to both situations where the prognostic algorithm gives only the point estimate of the expected RUL value (e.g., based on similarity measures [24], on regression techniques [25]–[27], etc.) and when the prognostic algorithm gives also the uncertainty in the RUL estimation (e.g., based on particle filtering [1], [2], [12], [13]), in which case Υ_λ can be the mean, median, or some percentile of the predicted RUL distribution.

We assume that the PHM-equipped component is stopped when Υ_λ is smaller than $h \cdot \Delta t$ and, also, that $j \cdot \Delta t$ time is required to arrange its maintenance activity. Thus, the RUL predictions are useful only if they allow us to stop the system at least $j \le h$ time intervals Δt before failure. This modeling assumption is also beneficial for the flexibility of the model, which can be applied to more stringent $(j = h)$ or relaxed $(j = 1)$ situations. Roughly speaking, if PHM suggests to remove the component at $T_s \le T_f - j \cdot \Delta t$, then $W(t) = 0 \ \forall t \in [T_s + j \cdot \Delta t, T_s + j \cdot \Delta t + T_{\rm prev}]$. Otherwise (i.e., if PHM fails to trigger an alarm at all prediction times before $T_f - j \cdot \Delta t$), $W(t) = 0 \ \forall t \in [T_f, T_f + T_{\rm cor}]$.

The proposed availability model relies on the false positive (fp) metric, which is defined as ([18])

$$fp_\lambda = \mathbb{E}[\Phi P_\lambda],\ \Phi P_\lambda = \begin{cases} 1, \text{ if } \Upsilon_\lambda - \mathrm{RUL}^*_\lambda < -d^{\rm threshold}_\lambda \\ 0, \text{ otherwise} \end{cases} \tag{2}$$

where RUL^*_λ is the actual component RUL at time t_λ and $d^{\rm threshold}_\lambda$ is a user-defined parameter, which we set to 0, so that fp becomes an estimator of the probability of having RUL predictions smaller than the real ones (i.e., $fp_\lambda = \mathbb{P}(\Upsilon_\lambda < \mathrm{RUL}^*_\lambda)$).

To compute the probability $\mathbb{P}(T_s = t_\lambda)$ to remove the component from operation at prediction instant $\lambda = 0, \frac{1}{N}, \ldots, \frac{N-1}{N}$, we divide the life cycle of the component into two different intervals:

1) early stop region: $t < T_f - h \cdot \Delta t$;
2) timely stop region: $T_f - h \cdot \Delta t \le t \le T_f - j \cdot \Delta t$.

Before going into model details, we remind that our objective is to build an availability model whose input parameters are the values of the algorithm performance metrics, whichever is the PHM system. This model can be embedded into more or less refined cost-benefit models to either support DMs in defining the PHM system design specifications that guarantee the economic viability of the PHM investment, or to quantify the expected profit of an existing PHM system of given performance metrics (i.e., of known values of the performance). In this latter case, the metrics values are estimated through a test campaign, as specified in [11].

A. Early Stop Region

In the considered maintenance setting, $\mathbb{P}(T_s = t_\lambda) = \mathbb{P}(\Upsilon_\lambda < h \cdot \Delta t)$. Then, for the early stop region, we can write

$$\mathbb{P}(T_s = t_\lambda) = \mathbb{P}(\Upsilon_\lambda < h \cdot \Delta t) \le \mathbb{P}(\Upsilon_\lambda < \mathrm{RUL}^*_\lambda). \tag{3}$$

This inequality is justified by the fact that event $\{\Upsilon_\lambda < h \cdot \Delta t\}$ is a subset of $\{\Upsilon_\lambda < \mathrm{RUL}^*_\lambda\}$; this allows conditioning the stop probability on the occurrence of underestimated predictions

$$\mathbb{P}(T_s = t_\lambda) = \mathbb{P}(T_s = t_\lambda | \Upsilon_\lambda < \mathrm{RUL}^*_\lambda) \times fp_\lambda. \tag{4}$$

To further develop (4), we need to characterize the uncertainty on predictions Υ_λ. For this, we consider the relative accuracy metric (RA_λ, [14], [17], [18]), which is a time variant index quantifying the percentage error between the actual RUL^*_λ and its estimation at a time t_λ

$$\mathrm{RA}_\lambda = 1 - \left| \frac{\Upsilon_\lambda - \mathrm{RUL}^*_\lambda}{\mathrm{RUL}^*_\lambda} \right|. \tag{5}$$

RA_λ is a random variable because it is a function of two dependent stochastic quantities, i.e., Υ_λ, which represents on the uncertainty in the PHM RUL predictions, and RUL^*_λ, which represents the uncertainty in the failure time.

For simplicity, we develop the availability model by first assuming that we know the actual failure times and, thus, the value of RUL^*. Then, we remove the dependence of the PHM-equipped component availability on RUL^*_λ by integrating on all its possible values. This is done through the Monte Carlo (MC) procedure given in the Appendix.

To characterize the uncertainty in RA_λ, we rely on ra_λ, which is defined as

$$ra_\lambda = \mathbb{E}[\mathrm{RA}_\lambda] = \mathbb{E}\left[1 - \left| \frac{\Upsilon_\lambda - \mathrm{RUL}^*_\lambda}{\mathrm{RUL}^*_\lambda} \right| \right] = 1 - \frac{\mathbb{E}[|\Upsilon_\lambda - \mathrm{RUL}^*_\lambda|]}{\mathrm{RUL}^*_\lambda}. \tag{6}$$

By the total probability theorem [28], the numerator on the right hand side can be further developed as

$$\begin{aligned} &\mathbb{E}[|\Upsilon_\lambda - \mathrm{RUL}^*_\lambda|] \\ &\quad = \mathbb{E}[\Upsilon_\lambda - \mathrm{RUL}^*_\lambda | \Upsilon_\lambda \ge \mathrm{RUL}^*_\lambda] \times \mathbb{P}(\Upsilon_\lambda \ge \mathrm{RUL}^*_\lambda) \\ &\qquad + \mathbb{E}[\mathrm{RUL}^*_\lambda - \Upsilon_\lambda | \Upsilon_\lambda < \mathrm{RUL}^*_\lambda] \times \mathbb{P}(\Upsilon_\lambda < \mathrm{RUL}^*_\lambda) \\ &\quad = \mathbb{E}[\Upsilon_\lambda - \mathrm{RUL}^*_\lambda | \Upsilon_\lambda \ge \mathrm{RUL}^*_\lambda] \times (1 - fp_\lambda) \\ &\qquad + \mathbb{E}[\mathrm{RUL}^*_\lambda - \Upsilon_\lambda | \Upsilon_\lambda < \mathrm{RUL}^*_\lambda] \times fp_\lambda \end{aligned}$$

From this, it emerges that due to the modulus in the definition of RA_λ, if we used only ra_λ and fp_λ to describe the uncertainty on Υ_λ, we would get one equation encoding two variables (i.e., $\mathbb{E}[\Upsilon_\lambda - \mathrm{RUL}^*_\lambda | \Upsilon_\lambda \ge \mathrm{RUL}^*_\lambda]$ and $\mathbb{E}[\mathrm{RUL}^*_\lambda - \Upsilon_\lambda | \Upsilon_\lambda < \mathrm{RUL}^*_\lambda]$)

and, thus, we could not establish a useful link between ra_λ and $\mathbb{P}(T_s = t_\lambda | \Upsilon_\lambda < \text{RUL}^*_\lambda)$ [see (4)].

To solve this issue, we can observe that in the early stop region the component can be stopped if the RUL is underestimated, only. Then, we can slightly modify RA_λ into $\underline{\text{RA}}_\lambda$ by applying the same definition in (5) only to RUL underestimations (i.e., when $\Phi P_\lambda = 1$)

$$\underline{\text{RA}}_\lambda = 1 - \frac{\text{RUL}^*_\lambda - \Upsilon_\lambda}{\text{RUL}^*_\lambda} = \frac{\Upsilon_\lambda}{\text{RUL}^*_\lambda}, \qquad \text{if } \Upsilon_\lambda \le \text{RUL}^*_\lambda. \quad (7)$$

Thus

$$\underline{ra}_\lambda = 1 - \frac{\mathbb{E}[\text{RUL}^*_\lambda - \Upsilon_\lambda | \text{RUL}^*_\lambda \ge \Upsilon_\lambda]}{\text{RUL}^*_\lambda} = \frac{\mathbb{E}[\Upsilon_\lambda | \text{RUL}^*_\lambda \ge \Upsilon_\lambda]}{\text{RUL}^*_\lambda} = \frac{\mu_{\underline{\Upsilon}_\lambda}}{\text{RUL}^*_\lambda}. \quad (8)$$

From (8), the mean $\mu_{\underline{\Upsilon}_\lambda}$ of Υ_λ conditional on that the RUL is underestimated, can be derived as $\underline{ra}_\lambda (N-k)\Delta t$, where $(N-k)\Delta t = \text{RUL}^*_\lambda$ and $k = \lfloor N\lambda \rfloor$. Moreover, the standard deviation $\sigma_{\underline{\Upsilon}_\lambda}$ of Υ_λ conditional on that the RUL is underestimated, can be derived from the prediction spread (PS) ([11], [13]) of $\underline{\text{RA}}$, $\sigma_{\underline{\text{RA}}_\lambda}$, which is defined as

$$\sigma_{\underline{\text{RA}}_\lambda} = \sqrt{Var\left[\frac{\Upsilon_\lambda}{\text{RUL}^*_\lambda} \middle| \text{RUL}^*_\lambda \ge \Upsilon_\lambda\right]} \quad (9)$$

Thus, $\sigma_{\underline{\Upsilon}_\lambda} = \sigma^2_{\underline{\text{RA}}_\lambda}[(N-k)\Delta t]^2$, using the known quadratic property $\text{Var}[\alpha X] = \alpha^2 \text{Var}[X]$, which is valid for any random variable X (e.g., [28]) and $\alpha \in \mathbb{R}$.

For the sake of generality, we do not make any parametric assumption about the distribution of Υ_λ conditional on that the RUL is underestimated. Rather, we exploit the well-known one-sided Chebyshev's inequalities (e.g., [29]), which can be applied to find probability upper and lower bounds of any random variable that is known in terms of its first moments (i.e., mean μ and variance σ^2)

$$\mathbb{P}(X \ge \mu + a) \le \frac{\sigma^2}{\sigma^2 + a^2} \quad (10)$$

$$\mathbb{P}(X \le \mu - a) \le \frac{\sigma^2}{\sigma^2 + a^2} \quad (11)$$

where $a > 0$.

To guarantee that $a > 0$, we distinguish the following two cases that can occur at time t_λ.

1) The expected RUL underestimate $\mu_{\underline{\Upsilon}_\lambda} = \underline{ra}_\lambda (N-k)\Delta t$ is larger than $h \cdot \Delta t$ (time t_{λ_1} in Fig. 2). In this case, we will stop the component with probability

$$\begin{aligned}\mathbb{P}(T_s = t_\lambda | \Upsilon_\lambda < \text{RUL}^*_\lambda) &= \mathbb{P}(\Upsilon_\lambda \le h \cdot \Delta t | \Upsilon_\lambda < \text{RUL}^*_\lambda)\\ &= \mathbb{P}(\Upsilon_\lambda - [\underline{ra}_\lambda (N-k)\Delta t] \le h \cdot \Delta t\\ &\quad - [\underline{ra}_\lambda (N-k)\Delta t] | \Upsilon_\lambda < \text{RUL}^*_\lambda)\\ &\mathbb{P}(\Upsilon_\lambda - \mu_{\underline{\Upsilon}_\lambda} \le -a | \Upsilon_\lambda < \text{RUL}^*_\lambda)\\ &\le \frac{\sigma^2_{\underline{\text{RA}}_\lambda}[(N-k)\Delta t]^2}{\sigma^2_{\underline{\text{RA}}_\lambda}[(N-k)\Delta t]^2 + a^2}\end{aligned} \quad (12)$$

where $a = \underline{ra}_\lambda (N-k)\Delta t - h \cdot \Delta t$.
When (12) is taken as equality, it provides an estimate of the minimum benefit achievable from a PHM system with known prognostic metrics, as it gives the maximum stop probability when we would not like to stop the component, to exploit all its useful lifetime.

2) The RUL underestimate $\mu_{\underline{\Upsilon}_\lambda} = \underline{ra}_\lambda (N-k)\Delta t$ is smaller than $h \cdot \Delta t$ (time t_{λ_2} in Fig. 2). In this case, we expect to stop the component. The probability that we will not stop the component is

$$\begin{aligned}\mathbb{P}(T_s > t_\lambda | \Upsilon_\lambda < \text{RUL}^*_\lambda) &= \mathbb{P}(\Upsilon_\lambda \ge h \cdot \Delta t | \Upsilon_\lambda < \text{RUL}^*_\lambda)\\ &= \mathbb{P}(\Upsilon_\lambda - [\underline{ra}_\lambda (N-k)\Delta t] \ge h \cdot \Delta t\\ &\quad - [\underline{ra}_\lambda (N-k)\Delta t] | \Upsilon_\lambda < \text{RUL}^*_\lambda)\end{aligned}$$

where we added the same quantity to both sides of the inequality. This entails that

$$\mathbb{P}(\Upsilon_\lambda - \mu_{\underline{\Upsilon}_\lambda} \ge a | \Upsilon_\lambda < \text{RUL}^*_\lambda) \le \frac{\sigma^2_{\underline{\text{RA}}_\lambda}[(N-k)\Delta t]^2}{\sigma^2_{\underline{\text{RA}}_\lambda}[(N-k)\Delta t]^2 + a^2} \quad (13)$$

where $a = h \cdot \Delta t - \underline{ra}_\lambda (N-k)\Delta t$.
Equation (13) gives an upper bound for the probability of not stopping the component and, thus, a lower bound of the stop probability. Given that in the early stop region a missed stop is beneficial for the availability of the model, to conservatively estimate the stop probability given an underestimation of RUL^*, we assume that it is equal to 1.

To sum up, the bounds of the stop probability for prediction times in $[T_{\text{pr}}, T_f - h\Delta t)$, i.e., $\lambda \in \{0, \frac{1}{N}, \ldots, \frac{N-h-1}{N}\}$, are [see (4)]:

1) If $\mu_{\underline{\Upsilon}_\lambda} = \underline{ra}_\lambda (N-k)\Delta t > h \cdot \Delta t$

$$\mathbb{P}(T_s = t_\lambda) = \frac{\sigma^2_{\underline{\text{RA}}_\lambda}[(N-k)\Delta t]^2}{\sigma^2_{\underline{\text{RA}}_\lambda}[(N-k)\Delta t]^2 + a^2} \times fp_\lambda \quad (14)$$

where $a = \underline{ra}_\lambda (N-k)\Delta t - h \cdot \Delta t$.

2) Otherwise

$$\mathbb{P}(T_s = t_\lambda) = fp_\lambda. \quad (15)$$

B. Timely Stop Region

In this section, we develop the analytical model of the stop probability for time prediction instants between $T_f - h \cdot \Delta t$ and $T_f - j \cdot \Delta t$, which then enters the availability model in (1). The major difference with respect to the predictions before $T_f - j \cdot \Delta t$ lies in that in the present case the component can be stopped even if the current RUL estimate Υ_λ is larger than RUL^*, provided that $\Upsilon_\lambda \le h \cdot \Delta t$

$$\begin{aligned}\mathbb{P}(T_s = t_\lambda) &= \mathbb{P}(\Upsilon_\lambda \le h \cdot \Delta t)\\ &= \mathbb{P}(\Upsilon_\lambda \le h \cdot \Delta t | \Upsilon_\lambda \le \text{RUL}^*_\lambda)\mathbb{P}(\Upsilon_\lambda \le \text{RUL}^*_\lambda)\\ &\quad + \mathbb{P}(\Upsilon_\lambda \le h \cdot \Delta t | \Upsilon_\lambda > \text{RUL}^*_\lambda)\mathbb{P}(\Upsilon_\lambda > \text{RUL}^*_\lambda)\\ &= \mathbb{P}(\Upsilon_\lambda \le h \cdot \Delta t | \Upsilon_\lambda \le \text{RUL}^*_\lambda) fp_\lambda\\ &\quad + \mathbb{P}(\Upsilon_\lambda \le h \cdot \Delta t | \Upsilon_\lambda > \text{RUL}^*_\lambda)(1 - fp_\lambda).\end{aligned} \quad (16)$$

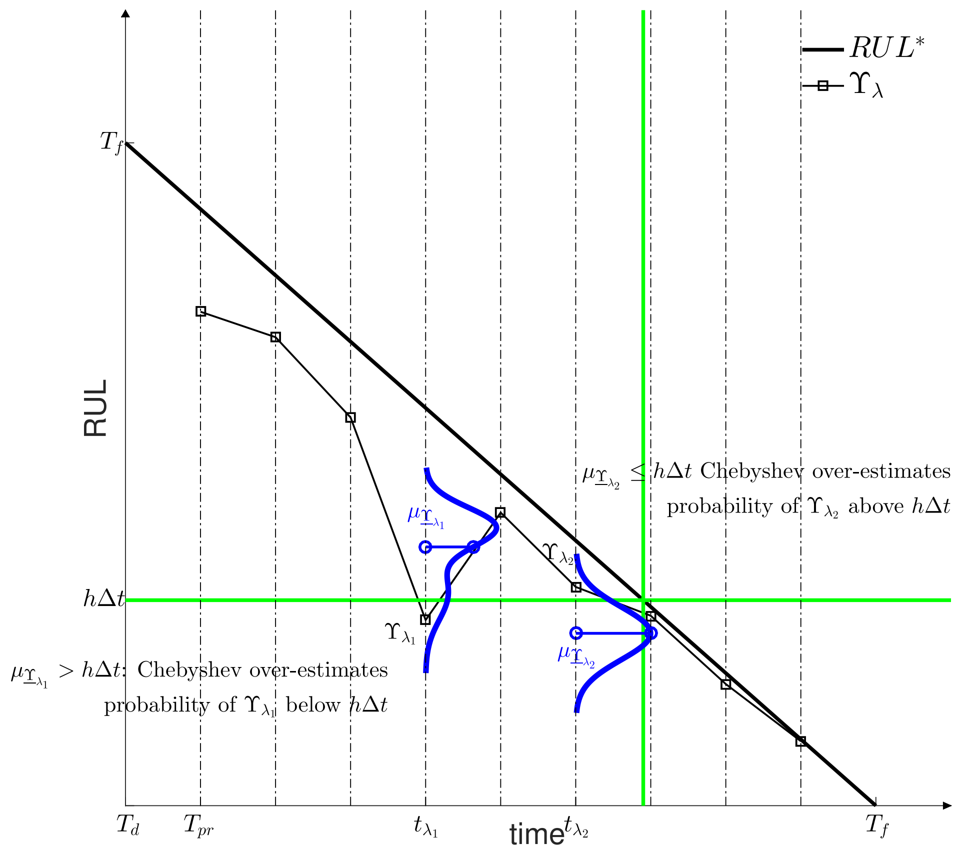

Fig. 2. Time t_{λ_1}, $\mu_{\underline{\Upsilon}_\lambda} = \underline{ra}_\lambda(N-k)\Delta t > h \cdot \Delta t$: the Chebyshev inequality (11) gives an upper bound for the stop probability. Time t_{λ_2} $\mu_{\underline{\Upsilon}_\lambda} = \underline{ra}_\lambda(N-k)\Delta t < h \cdot \Delta t$, the Chebyshev inequality (11) gives an upper bound for the no-stop probability.

The first addendum reduces to fp_λ: since we are at time instants after $T_f - h \cdot \Delta t$, then $\mathrm{RUL}^*_\lambda \leq h \cdot \Delta t$ for all prediction instants and $\mathbb{P}(\Upsilon_\lambda \leq h \cdot \Delta t | \Upsilon_\lambda \leq \mathrm{RUL}^*_\lambda) = 1$.

Now, we need to model the second addendum, (i.e., the case $\Phi P_\lambda = 0$). Proceeding similarly as before, we can define $\overline{\mathrm{RA}}_\lambda$ as

$$\overline{\mathrm{RA}}_\lambda = 1 - \frac{\Upsilon_\lambda - \mathrm{RUL}^*_\lambda}{\mathrm{RUL}^*_\lambda} = 2 - \frac{\Upsilon_\lambda}{\mathrm{RUL}^*_\lambda}, \qquad \text{if } \Upsilon_\lambda > \mathrm{RUL}^*_\lambda.$$

The mean and standard of $\overline{\mathrm{RA}}_\lambda$ deviation are

$$\overline{ra}_\lambda = 1 - \frac{\mathbb{E}[\Upsilon_\lambda - \mathrm{RUL}^*_\lambda | \Upsilon_\lambda > \mathrm{RUL}^*_\lambda]}{\mathrm{RUL}^*_\lambda} = 2 - \frac{\mathbb{E}[\Upsilon_\lambda]}{\mathrm{RUL}^*_\lambda}$$

$$\sigma_{\overline{\mathrm{RA}}_\lambda} = \sqrt{Var\left[\frac{\Upsilon_\lambda}{\mathrm{RUL}^*_\lambda}\middle| \Upsilon_\lambda > \mathrm{RUL}^*_\lambda\right]}.$$

Similarly as before, $\mu_{\overline{\Upsilon}_\lambda} = (2 - \overline{ra}_\lambda)(N-k)\Delta t$ and $\sigma^2_{\overline{\Upsilon}_\lambda} = \sigma^2_{\overline{\mathrm{RA}}_\lambda}[(N-k)\Delta t]^2$; this allows us exploiting the Chebyshev's inequalities to estimate the upper and lower bounds of $\mathbb{P}(\Upsilon_\lambda \leq h \cdot \Delta t | \Upsilon_\lambda > \mathrm{RUL}^*_\lambda)$ as follows.

1) If at time t_λ the RUL overestimate $\mu_{\overline{\Upsilon}_\lambda} = (2 - \overline{ra}_\lambda)(N - k)\Delta t$ is smaller than $h \cdot \Delta t$, then, we have

$$\mathbb{P}(\Upsilon_\lambda \geq h \cdot \Delta t | \Upsilon_\lambda > \mathrm{RUL}^*_\lambda) = \mathbb{P}(\Upsilon_\lambda - [(2-\overline{ra}_\lambda)(N-k)\Delta t] \geq h \cdot \Delta t - [(2 - \overline{ra}_\lambda)(N-k)\Delta t] | \Upsilon_\lambda > \mathrm{RUL}^*_\lambda)$$

where we have added the same quantity to both hands of the inequality. This entails that

$$\mathbb{P}(\Upsilon_\lambda - \mu_{\overline{\Upsilon}_\lambda} \geq a | \Upsilon_\lambda > \mathrm{RUL}^*_\lambda) \leq \frac{\sigma^2_{\overline{\mathrm{RA}}_\lambda}[(N-k)\Delta t]^2}{\sigma^2_{\overline{\mathrm{RA}}_\lambda}[(N-k)\Delta t]^2 + a^2}$$

where $a = h \cdot \Delta t - (2 - \overline{ra}_\lambda)(N-k)\Delta t$. Since $\mathbb{P}(\Upsilon_\lambda < h \cdot \Delta t | \Upsilon_\lambda > \mathrm{RUL}^*_\lambda) = 1 - \mathbb{P}(\Upsilon_\lambda \geq h \cdot \Delta t | \Upsilon_\lambda > \mathrm{RUL}^*_\lambda)$, we can write

$$\mathbb{P}(\Upsilon_\lambda \leq h \cdot \Delta t | \Upsilon_\lambda > \mathrm{RUL}^*_\lambda) \geq 1 - \frac{\sigma^2_{\overline{\mathrm{RA}}_\lambda}[(N-k)\Delta t]^2}{\sigma^2_{\overline{\mathrm{RA}}_\lambda}[(N-k)\Delta t]^2 + a^2}. \tag{17}$$

We take (17) as an equality, which guarantees that we are underestimating the benefit derived from the PHM system, as we are underestimating the probability of triggering alarms in the time instants where the component should be removed from operation to avoid failures.

2) If at time t_λ the mean RUL overestimate $\mu_{\overline{\Upsilon}_\lambda} = (2 - \overline{ra}_\lambda)(N-k)\Delta t$ is larger than $h \cdot \Delta t$, we assume that

$$\mathbb{P}(\Upsilon_\lambda \leq h \cdot \Delta t | \Upsilon_\lambda > \mathrm{RUL}^*_\lambda) = 0. \tag{18}$$

This assumption is conservative, because according to the Chebyshev's inequality, 0 is always smaller than the upper bound of the stop probability.

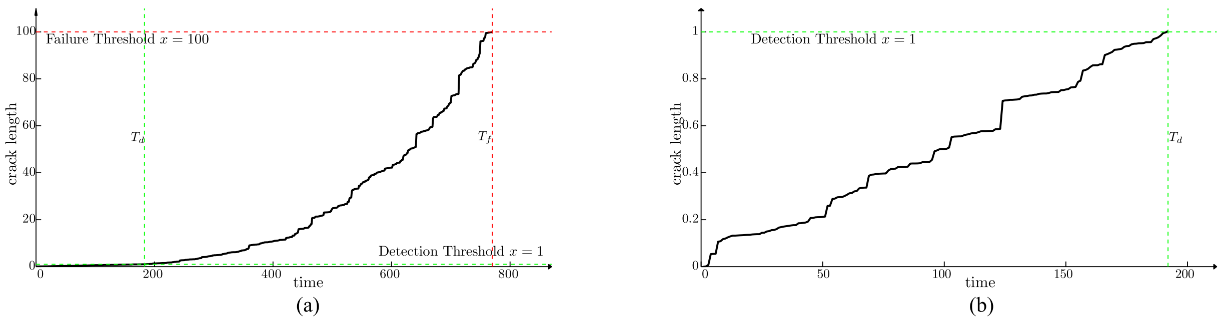

Fig. 3. Crack propagation process: example. (a) Behavior in the whole life cycle. (b) Zoom in the part below the detection threshold.

To sum up, from (16) the lower bound of the stop probability for prediction times in $[T_f - h\cdot\Delta t, T_f - j\cdot\Delta t]$, i.e., $\lambda \in \{\frac{N-h}{N},\dots,\frac{N-j}{N}\}$, reads

1) If $(2-\overline{ra}_\lambda)(N-k)\Delta t \le h\cdot\Delta t$

$$\mathbb{P}(T_s=t_\lambda)=fp_\lambda+\left(1-\frac{\sigma^2_{\overline{\mathrm{RA}}_\lambda}[(N-k)\Delta t]^2}{\sigma^2_{\overline{\mathrm{RA}}_\lambda}[(N-k)\Delta t]^2+a^2}\right)(1-fp_\lambda) \tag{19}$$

where $a = h\cdot\Delta t - (2-\overline{ra}_\lambda)(N-k)\Delta t$.

2) Otherwise

$$\mathbb{P}(T_s = t_\lambda) = fp_\lambda. \tag{20}$$

Finally notice that for the time prediction instants in $[T_f - j\cdot\Delta t, T_f]$, we assume that the component is never removed and it will fail at time T_f.

To remove the dependence of $A(t)$ from T_f and, thus, from RUL*, which is represented by N in all equations, it is not possible to find out an analytical formula. Nonetheless, $A(t)$ can be estimated using the MC approach in the Appendix, which integrates the results of different failure times.

IV. Case Study

In this section, we apply the developed modeling framework to a component affected by a fatigue degradation mechanism, which is simulated according to the Paris Erdogan (PE) model ([2], [31], see Fig. 3):

1) The crack length x_i reaches the first threshold, $x = 1\,\mathrm{mm}$, according to the following equation:

$$x_{i+1} = x_i + a\times e^{\omega_i^1}$$

where $a = 0.003$ is the growth speed parameter and $\omega_i^1 \sim \mathcal{N}(-0.625, 1.5)$ models the uncertainty in the speed values. The uncertainty in the arrival time at $x = 1$ is described by $f_{T_d}(t)$.

2) The crack length reaches the failure threshold $x = 100\,\mathrm{mm}$ according to the following equation:

$$x_{i+1} = x_i + C\times e^{\omega_i^2}(\eta\sqrt{x_i})^n$$

where $C = 0.005$ and $n = 1.3$ are parameters related to the component material properties and are determined by experimental tests; $\eta = 1$ is a constant related to the characteristics of the load and the position of the crack and $\omega_i^2 \sim \mathcal{N}(0,1)$ is used to describe the uncertainty in the crack growth speed values. The uncertainty in the arrival time at $x = 100$ is described by $f_{T_f}(t)$.

The numerical values are taken from [2].

We assume that $T_{\mathrm{cor}} = \chi T_{\mathrm{pred}}$, with $\chi > 1$.

TABLE I
PHM and Maintenance Policy Data

Parameter	Value
h	5
j	3
χ	2
T_{pred}	100
Δt	10
T (horizon)	2900

(a) Maintenance Policy Data

Parameter	Value			
interval	$[0, 0.25)$	$[0.25, 0.5)$	$[0.5, 0.75)$	$[0.75, 1]$
$\underline{ra}$	0.75	0.8	0.85	0.9
fp	0.3	0.4	0.5	0.6
$PS\ (\sigma^2_{\underline{\mathrm{RA}}})$	0.04	0.0225	0.01	0.0025
$\overline{ra}$	0.75	0.8	0.85	0.9
$PS\ (\sigma^2_{\overline{\mathrm{RA}}})$	0.04	0.0225	0.01	0.0025

(b) PHM Data

A. Availability Model: Application

In this section, we apply the availability model developed in Section III to the case study of the component equipped with a PHM system for the crack propagation mechanism. We use the MC simulation [31] scheme reported in the Appendix with stop probabilities equal to the bounds given by Chebyshev's inequalities introduced earlier. As already pointed out, this allows computing the minimum benefit achievable with a PHM system of given characteristics and performance.

The availability performance is evaluated on a time horizon corresponding to approximately 3/4 component life cycles, whereas the maintenance policy data and PHM data are summarized in Tables I a and I b, respectively.

Fig. 4 shows the values of $\underline{ra}$, $\overline{ra}$, fp, $\sigma_{\underline{\mathrm{RA}}}$, and $\sigma_{\overline{\mathrm{RA}}}$ as a function of λ; for simplicity, we have assumed $\underline{ra} = \overline{ra} = ra^*$ and $\sigma_{\underline{\mathrm{RA}}} = \sigma_{\overline{\mathrm{RA}}} = \sigma_{\mathrm{RA}}$.

Fig. 5 compares the instantaneous unavailability over time (i.e., $1 - A(t)$) of the PHM-equipped component with that of a component undergoing corrective maintenance. The bars in Fig. 5 represent the 68th two-sided confidence interval of the MC simulation error.

From Fig. 5, it emerges that the mean unavailability in the whole time horizon T (i.e., $\frac{1}{T}\int_0^T (1-A(t))dt$, [23]) of the PHM-equipped component is significantly smaller than that of the component with the corrective maintenance (0.1326 versus

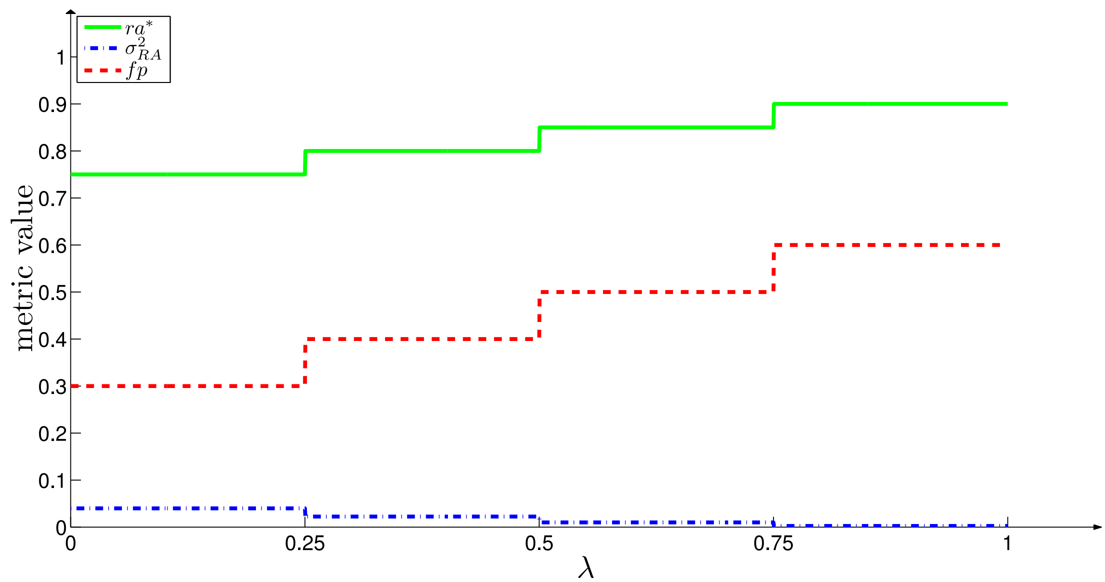

Fig. 4. PHM metric data for the availability model.

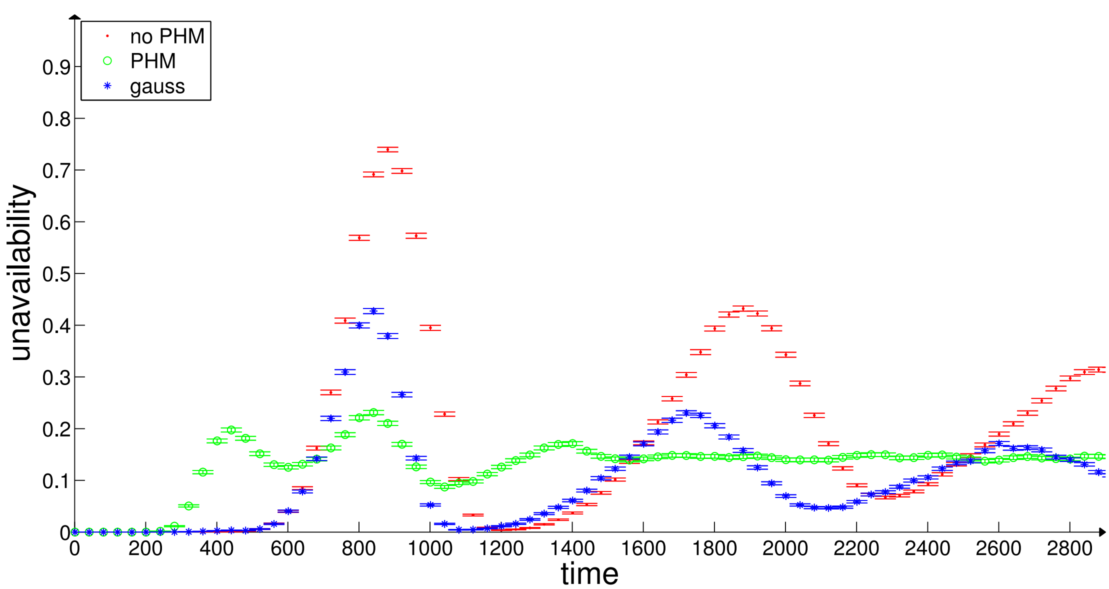

Fig. 5. Unavailability of the component under corrective maintenance and of the PHM-equipped maintenance, both using the developed model and assuming normal predictions.

0.1856). The point-wise unavailability $(1 - A(t))$ of the PHM-equipped component oscillates less than that of the component under corrective maintenance, and reaches a stable steady state value much earlier. This indicates that the PHM system reduces the variability in the population of similar components affected by the same degradation process and equipped with the same PHM system.

Notice that the first peak of the unavailability of the PHM-equipped component precedes that of the component undergoing corrective maintenance: this is due to the large value of σ_{RA} in the early stop region, which strongly increases the probability of triggering early alarms. The second peaks are close for the two components: this is indicative of the PHM capability of correctly identifying the component failure time. After the second peak, the positions of the unavailability peaks of the component under corrective maintenance follows the typical almost-periodic behavior, with period $MTTF + T_{\mathrm{cor}}$.

Fig. 5 also compares the point-wise unavailability curves of the two components (equipped and not equipped with PHM) to that obtained assuming that the uncertainty in Υ_λ is normally distributed. The mean unavailability is 0.1040: the additional information about the uncertainty distribution of the RUL predictions allows us to not use the Chebyshev's inequalities in (13)–(17) and, thus, to have smaller mean unavailability values. This is due to the fact that, as already pointed out, the Chebyshev's inequalities overestimate the stop probability when the

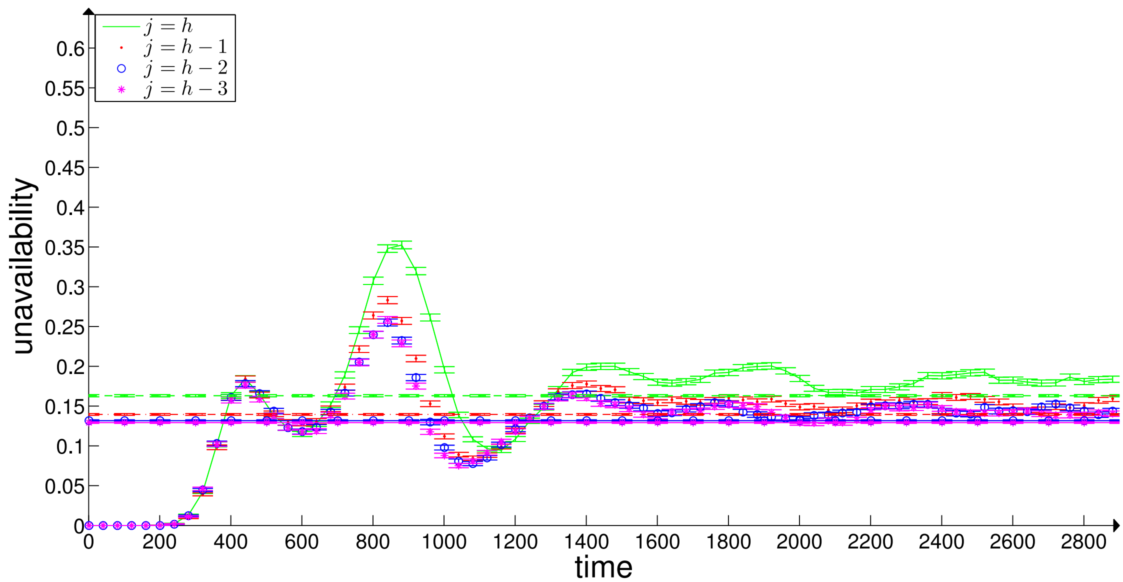

Fig. 6. Sensitivity of unavailability to j.

prediction time is far from failure, as it can be seen from the anticipated increase of the unavailability curve. Notice that this result (i.e., the mean unavailability estimated when the distribution of Υ_λ is known is smaller than that estimated through the Chebyshev's inequalities) does not depend on the particular distribution of Υ_λ. Rather, it is due to the fact that the inequalities are inherently upper bounds of the true probabilities of stopping the component in the early stop region or leaving it working in the timely stop region.

B. Availability Model Sensitivity Analysis

In this section, we perform a sensitivity analysis to investigate how the considered PHM metrics and maintenance policy data affect the availability of the system. To do this, we exploit the one-at-a-time approach ([32]), i.e., we change one parameter at a time to analyze the corresponding changes of both the unavailability $1 - A(t)$ and its mean, the reference setting being that summarized in Table I a and I b.

To reduce the possible cases to be investigated, we assume that the ra^*, fp, and σ_{RA} metrics are nondecreasing stepwise functions in the intervals $[0, 0.25)$, $[0.25, 0.5)$, $[0.5, 0.75)$, and $[0.75, 1]$. In particular, ra^* takes values ra_1^*, $ra_1^* + \frac{1}{3}(ra_4^* - ra_1^*)$, $ra_1^* + \frac{2}{3}(ra_4^* - ra_1^*)$, and ra_4^*, respectively; fp values are fp_1, $fp_1 + \frac{1}{3}(fp_4 - fp_1)$, $fp_1 + \frac{2}{3}(fp_4 - fp_1)$, and fp_4 and those of σ_{RA} are σ_{RA_1}, $\sigma_{\mathrm{RA}_1} - \frac{1}{3}(\sigma_{\mathrm{RA}_1} - \sigma_{\mathrm{RA}_4})$, $\sigma_{\mathrm{RA}_1} - \frac{2}{3}(\sigma_{\mathrm{RA}_1} - \sigma_{\mathrm{RA}_4})$ and σ_{RA_4}.

Fig. 6 summarizes the impact of parameter j on the component unavailability. From the analysis of this figure, it can be inferred that the closer the value of j to h, the larger the component unavailability. This is justified by the fact that we have fewer instants to take action. In particular, when $j = h$ (i.e., when the component can be removed from operation only in the early stop region), the value of the mean unavailability is much larger than the values of the other cases, as there is no possibility to avoid failures in the timely stop region. Notice also that the first peak does not change for the different curves: it is related to the early stops, only, which do not depend on j. On the contrary, the larger the value of j, the larger the second peak, which relates to the component failures: the larger the value of j, the larger the portion of the simulated component undergoing failures and, thus, the wider the peak area.

Fig. 7 evaluates the impact of Δt on the component unavailability. The worst case is represented by $\Delta t = 1$ (i.e., the "continuously" monitored component). This is due to the fact that our model overestimates the stop probability at time instants far from failure: the smaller the Δt, the larger the number of predictions, the larger the probability of early removing the component from operation, which does not allow exploiting all the useful life of the component. As it can be seen later, we need to improve the accuracy in this region to avoid early stops.

With regards to the sensitivity to other Δt values, the mean unavailability values for $\Delta t = 10$ and $\Delta t = 25$ are very similar to each other, though the case with $\Delta t = 10$ is slightly better. When $\Delta t = 50$, we can see a nonnegligible increase in the mean unavailability; this is due to the reduction of predictions in the timely stop area, which leads to an increase in the number of failures, as it can be seen from the wider second peak.

Fig. 8 shows the impact of fp_1 and fp_4 on the component unavailability: different figures refer to different values of fp_1 and show the unavailability for different values of fp_4. Notice that only the cases where $fp_4 \geq fp_1$ are considered, being the performance metrics assumed not decreasing over time.

From Fig. 8, it emerges that unavailability is more sensitive to fp_1 than fp_4: there is a progressive increase in the mean unavailability from Fig. 8 top-left to Fig. 8 bottom-right, which is due to the early alarms that become more frequent as fp_1 increases. In fact, the value of fp_1 is directly linked to the number of early predictions after T_{pr}: the larger its value, the larger the number of early alarms. This can also be seen by the increasing value of the first peak of the point-wise unavailability curves, which is due to false alarms, rather than failures.

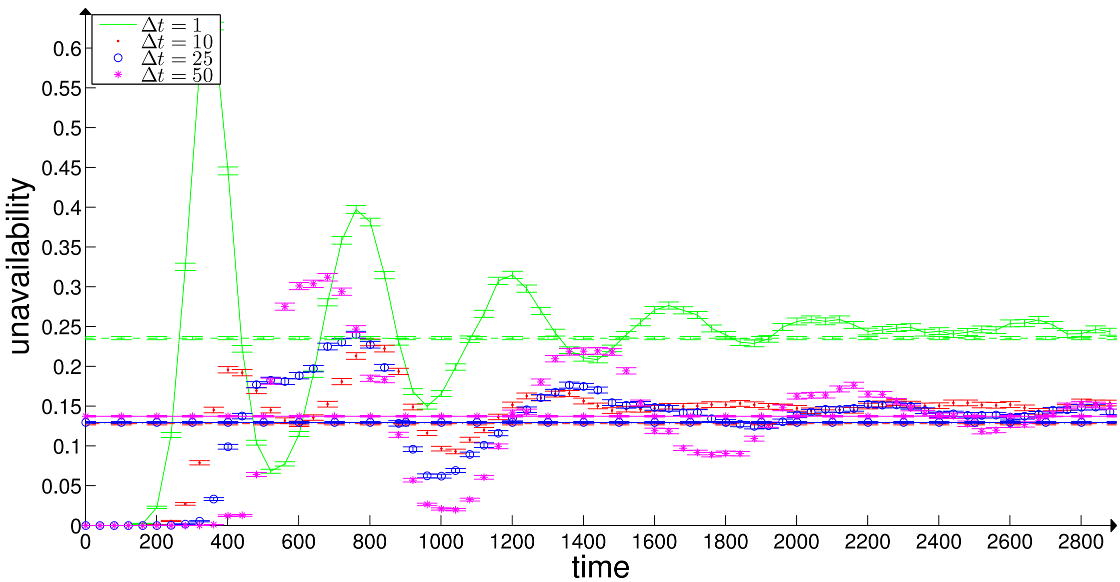

Fig. 7. Sensitivity of unavailability to Δt.

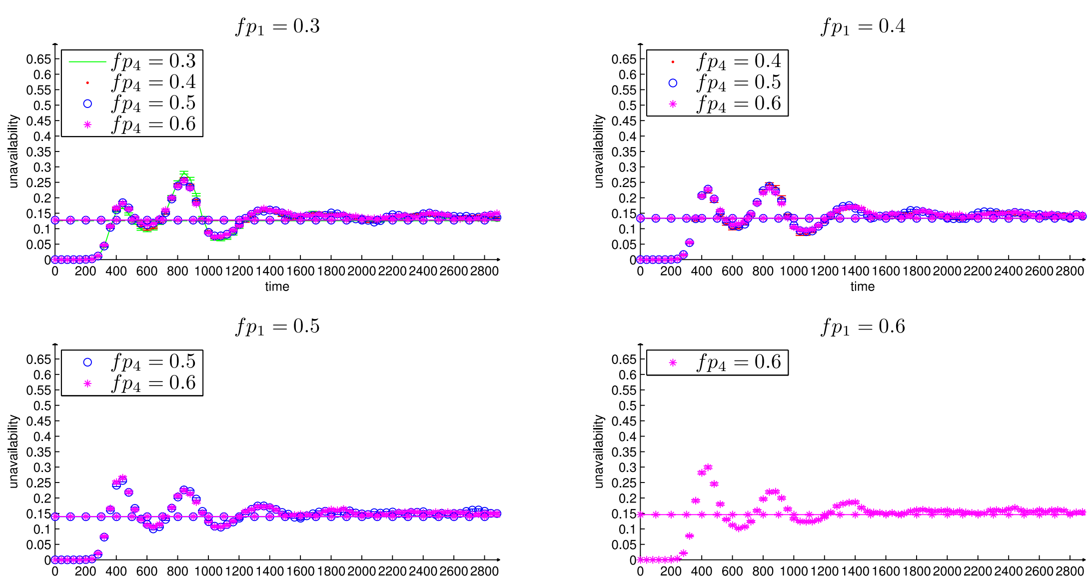

Fig. 8. Sensitivity of unavailability to fp.

The component unavailability is not sensitive to fp_4. This result is due to the following two opposite trends arising with increasing values of fp_4.

1) On one side, there are a smaller number of failures, which result in a reduction of the component unavailability. This can be seen from Table II, columns 3 and 6, which report the total number N_{cor} of corrective maintenance actions performed over the 15 000 MC trials.
2) On the other side, predictions related to the first instants yield a larger number of early stops as fp_4 increases. In this respect, Table II, columns 4 and 8 report the total number N_{prev} of preventive maintenance actions performed over the 15 000 MC trials: the larger fp_4, the larger N_{prev}. These early stops do not allow exploiting the entire life time of the system and, thus, increase the component unavailability.

TABLE II
NUMBER OF FAILURES VERSUS fp VALUES

fp_1	fp_4	N_{cor}	N_{prev}	fp_1	fp_4	N_{cor}	N_{prev}
0.3	0.3	8034	37 105	0.4	0.5	7756	39 726
0.3	0.4	7953	37 395	0.4	0.6	7732	40 021
0.3	0.5	7850	37 799	0.5	0.5	7642	41 456
0.3	0.6	7839	38 072	0.5	0.6	7639	41 756
0.4	0.4	7829	39 394	0.6	0.6	7449	43 491

Fig. 9 shows the impact of $\underline{ra}_1$ and $\underline{ra}_4$ on the unavailability curves. Different figures refer to different values of $\underline{ra}_1$ and

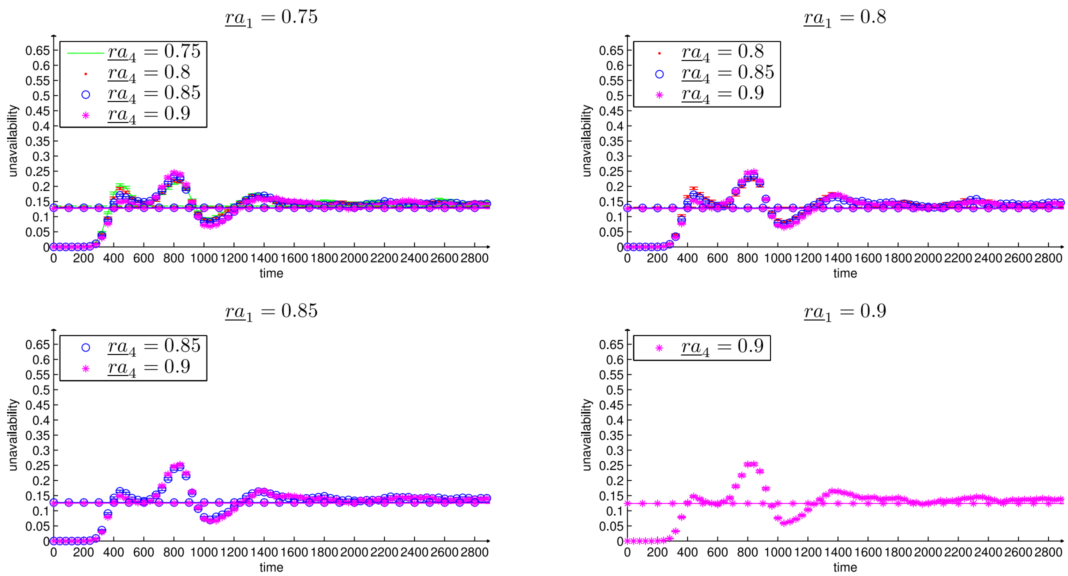

Fig. 9. Sensitivity of Unavailability to $\underline{ra}$.

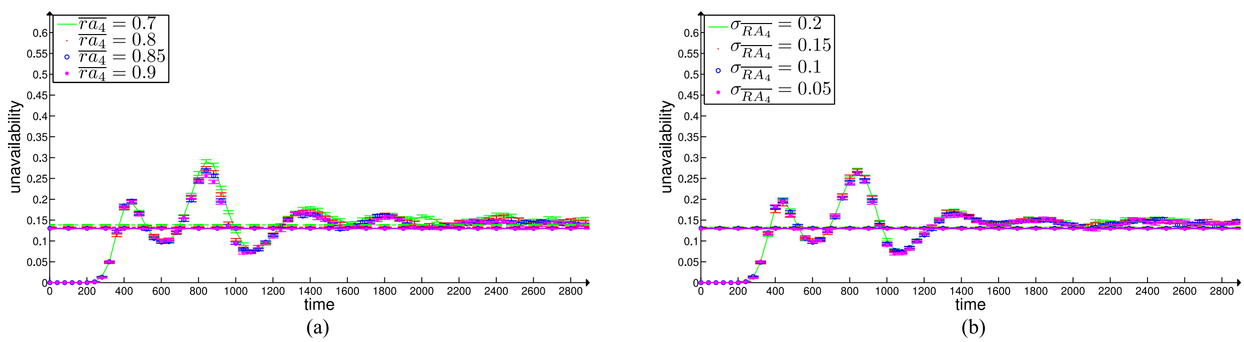

Fig. 10. Sensitivity of unavalability to $\overline{ra_4}$ and $\sigma_{\overline{\mathrm{RA}_4}}$. (a)Sensitivity of unavalability to $\sigma_{\overline{\mathrm{ra}_4}}$, (a)Sensitivity of unavalability to $\sigma_{\overline{\mathrm{RA}_4}}$

show the component unavailability for different values of $\underline{ra}_4$, $\underline{ra}_4 \geq \underline{ra}_1$.

From the analysis of Fig. 9, we can conclude that the component unavailability is not very sensitive to both $\underline{ra}_1$ and $\underline{ra}_4$. This is due to the relatively weak dependency of the stop probability on ra values, which directly derives from the Chebyshev's inequalities.

With regards to the sensitivity to $\underline{ra}_4$, considerations similar to those made for fp_4 about the two competing phenomena can be drawn to explain the poor sensitivity.

The sensitivity with respect to $\overline{ra}$ and $\sigma_{\overline{\mathrm{RA}}}$ is summarized in Fig. 10(a) and (b), respectively. The decision to focus only on the last values of the two metrics is due to the consideration that $\overline{ra}$ and $\sigma_{\overline{\mathrm{RA}}}$ affect the stop probability only for time instants between $T_f - h \cdot \Delta t$ and $T_f - j \cdot \Delta t$. Fig. 10(a) shows that there is little sensitivity to $\overline{ra_4}$, because these values affect the unavailability behavior only for few prediction times. For the same reasons, there is little sensitivity also to $\sigma_{\overline{\mathrm{RA}_4}}$ [see Fig. 10(b)] . To appreciate a larger sensitivity to these parameters, the values of j should be increased, so as to increase the time interval in which these values enter the computation of unavailability.

Finally, Fig. 11 shows the impact of $\sigma_{\underline{\mathrm{RA}_1}}$ and $\sigma_{\underline{\mathrm{RA}_4}}$ on unavailability, the organization of the subplots being similar to that of Figs. 8 and 9.

From the analysis of Fig. 11, it can be inferred that both $\sigma_{\underline{\mathrm{RA}_1}}$ and $\sigma_{\underline{\mathrm{RA}_4}}$ have a strong impact on the mean unavailability of the PHM-equipped component: the smaller their values, the smaller the mean unavailability. In this respect, we can notice that smaller values of $\sigma_{\underline{\mathrm{RA}_1}}$ entail smaller values of the first peak value, which completely disappears if $\sigma_{\underline{\mathrm{RA}_1}} \leq 0.1$, meaning that most early stops are prevented. Moreover, as $\sigma_{\underline{\mathrm{RA}_4}}$ increases, there is a larger stop probability in the last instants before failure, which leads the mean unavailability to decrease according to Chebyshev's inequalities. The very large sensitivity to σ_{RA} allows us to conclude that the precision of the predictions is, indeed, one of the most important driver of a PHM system, even more than the accuracy.

V. Reverse Engineering Approach for Maintenance Decision Making

To the authors' best knowledge, there is no simple link between the metric values of a PHM system and its development

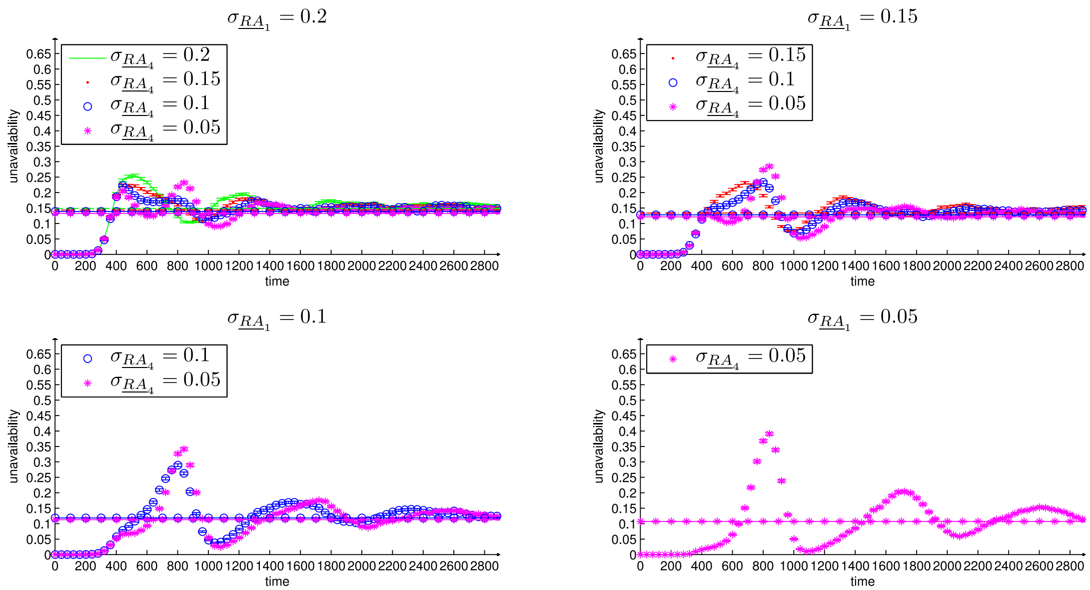

Fig. 11. Sensitivity of Unavailability to $\sigma_{\underline{\mathrm{RA}}}$.

cost. Then, to give the DM the possibility of trading-off the benefit arising from PHM against its possible initial development costs, we propose a reverse engineering approach. That is, we develop a cost analysis to compare the operational undiscounted cost C of a PHM-equipped component (i.e., without considering the initial development cost) with those of corrective and scheduled maintenance (i.e., C_{cor} and C_{sched}) [1]. This provides the DM with a rough estimate of the economic benefit achievable by PHM and, thus, gives him/her a basis to decide whether to invest in PHM or not.

Notice that in the scheduled maintenance policy, the repair actions are performed every 650 units of time: this corresponds to the 10th percentile of the component failure time, which in [1] has been proved to be the optimal maintenance interval for the same case study considered in this paper.

To compare the operational costs of the different maintenance policies, we consider the following simplified cost model ([33], [34]):

$$C = N_{\mathrm{cor}} \times c_{\mathrm{cor}} + N_{\mathrm{prev}} \times c_{\mathrm{prev}} + c_{\mathrm{DT}} \times \mathrm{DT} \tag{21}$$

where c_{cor} and c_{prev} are the costs related to corrective and preventive maintenance, respectively, whereas c_{DT} represents the cost due to component operation interruption. Maintenance costs are set to $c_{\mathrm{cor}} = 800$, $c_{\mathrm{prev}} = 500$, and $c_{\mathrm{DT}} = 50$, in arbitrary units; $\mathrm{DT} = \int_0^T (1 - A(t))dt$ is the total expected downtime over the whole time horizon T, in arbitrary units; N_{cor} and N_{prev} are the number of corrective and preventive maintenance actions performed over the component life cycle. Obviously, in case of corrective maintenance policy $N_{\mathrm{prev}} = 0$.

Notice that the considered cost values are illustrative and the operational cost model may be refined, for example, by including possible costs for monitoring and predicting, savings due to the possibility of properly arranging preventive maintenance, discounting rates, etc.

The cost-benefit analysis is developed for different values of the considered PHM metrics: ra^*, σ_{RA} (which are assumed to be equal for both the underestimation and over-estimation cases), and fp. In details, fp takes values in the set $\{0.3, 0.45, 0.6\}$, ra^* in the set $\{0.8, 0.85, 0.9, 0.95\}$; σ_{RA} ranges in $[0.01, 0.16]$ at discrete points equally spaced by 0.015.

The remaining data are set as in the general case study presented in Section IV-A. For simplicity, ra^*, σ_{RA} are all considered constant functions of λ.

Fig. 12 shows an estimate of the operational cost C for the PHM-driven maintenance, which is computed for each combination of the three considered metrics. These cost values are compared with those of the scheduled and corrective maintenance approaches ($C_{\mathrm{sched}} \simeq 25600$ $C_{\mathrm{cor}} \simeq 30690$), as derived from MC simulation.

When $C < C_{\mathrm{sched}}$, it is worthwhile considering the development of a PHM system; on the contrary, when $C \geq C_{\mathrm{cor}}$, PHM is always unfavorable. Additional information to develop more refined analysis (e.g., a more accurate cost model and/or more information about the PHM system to be analyzed) is required in the in-between case ($C_{\mathrm{sched}} \leq C < C_{\mathrm{cor}}$). From Fig. 12, it can also be seen that, as expected, the metric value which mostly affects C is σ_{RA}: the smaller this value, the smaller the operational cost C with PHM. When $\sigma_{\mathrm{RA}} \leq 0.01$, the PHM-driven maintenance is always cheaper than the scheduled maintenance, whereas if $\sigma_{\mathrm{RA}} \geq 0.16$, PHM-driven maintenance is always more expensive of both the scheduled and corrective approaches, unless the other two metrics take their best values. Moreover, when $\sigma_{\mathrm{RA}} \geq 0.07$ the maintenance based on PHM is always more expensive than the scheduled one, whichever the values of the other metrics. Although this result seems discouraging, we have to keep in mind that this is the least achievable benefit derived from a PHM tool and that here we are assuming a constant value of σ_{RA} over the whole λ, which is a quite conservative hypothesis as we may expect that the predictions will

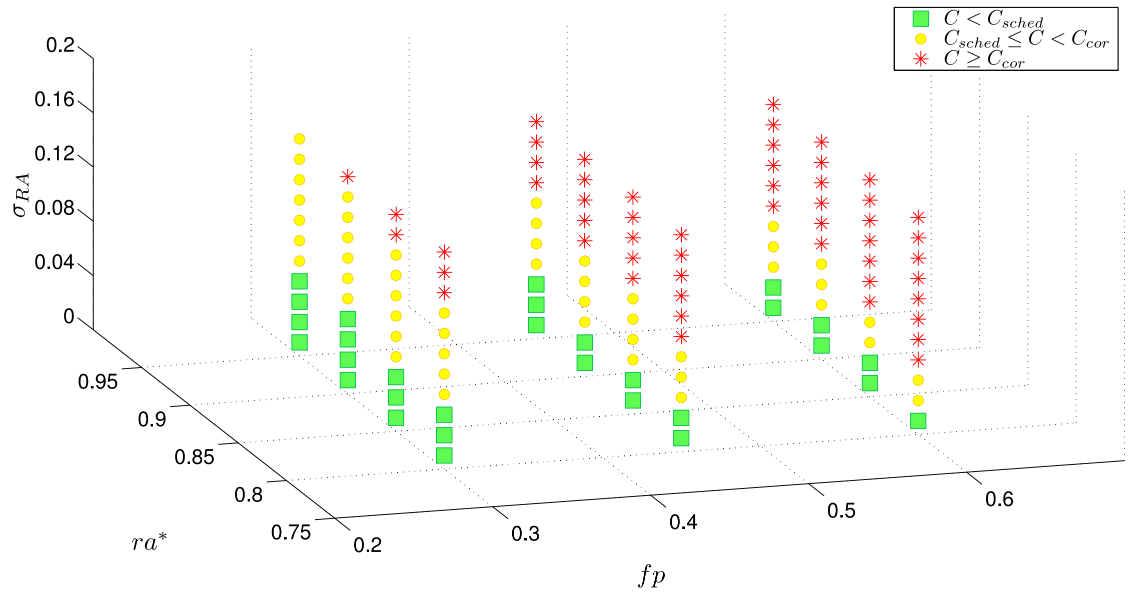

Fig. 12. Cost Analyzes as function of fp, ra, and σ_{RA}: $C < C_{\mathrm{sched}}$ (best case) is indicated by "□,' '$C_{\mathrm{sched}} \le C < C_{\mathrm{cor}}$ (medium case) by "o", $C \ge C_{\mathrm{cor}}$ by "'∗" (worst case) .

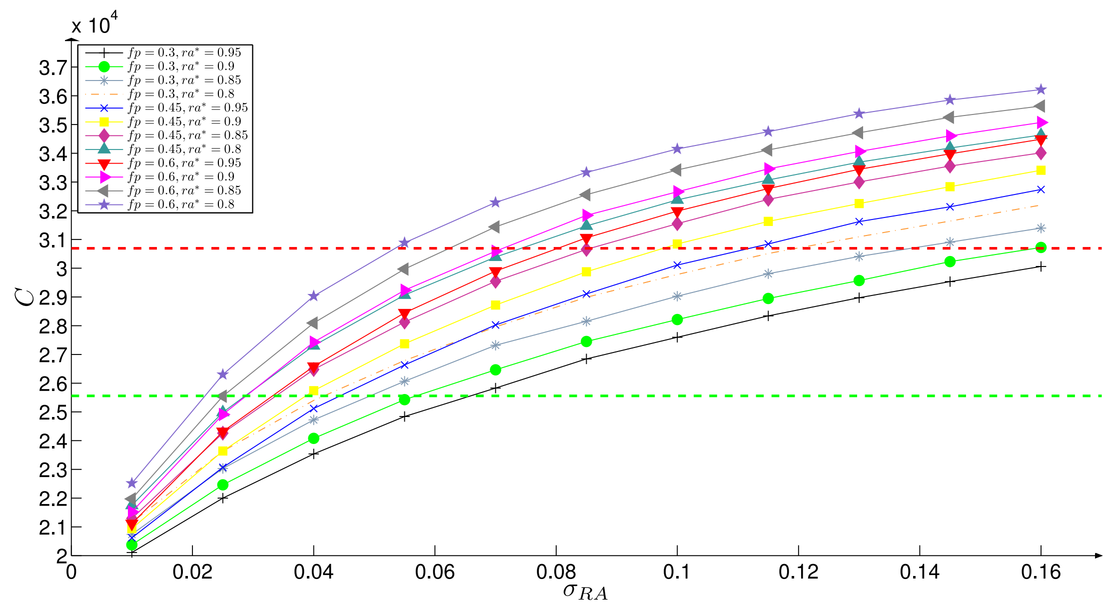

Fig. 13. Sensitivity of C to the metrics.

be more accurate as the component approaches failure. With respect to ra^*, the larger its value, the smaller the cost, whereas larger fp values entail larger cost. As already pointed out in the previous section, these results depend on the values of j, h, and Δt, which lead the availability to be more sensitive to fp at small values of λ.

Finally, Fig. 13 plots the operational cost C over σ_{RA} for different combinations of ra^* and fp. We can notice that C ranges in $[20000, 37000]$, i.e., the best combination of metrics guarantees a minimum achievable benefit of about $\frac{25000-20000}{25000} = 20\%$ of the value of the scheduled maintenance. From the elbow shape of the curves in Fig. 13, we can also infer that for smaller values of σ_{RA}, ra^* has more impact on cost than fp, whereas as σ_{RA} increases, improvements in the value of ra^* are less relevant than fp: when $\sigma_{\mathrm{RA}} = 0.16$, all cost values that correspond to smaller values of fp are smaller than those corresponding to larger fp values, for any value of ra^*.

VI. Conclusion

In this paper, we have presented a general framework to compute the availability of a PHM-equipped component, which is based on time-variant prognostic metrics proposed in the literature (fp, $\underline{\mathrm{RA}}$, and $\overline{\mathrm{RA}}$). The modeling framework proposed

Data: $\underline{ra}_\lambda, \overline{ra}_\lambda, \sigma_{\underline{RA}_\lambda}, \sigma_{\overline{RA}_\lambda}, fp_\lambda \ \forall\lambda \in [0;1]; T_{cor}, T_{pred}, f_{T_f}, f_{T_d}, f_{DTD}, \Delta t, h, j, T$

Result: $U(t) = 1 - A(t)$

Repeat B times the following procedure:

Initialize $T_{es} \leftarrow 0$, MC counter $W_b(0:T) \leftarrow 0$

while $T_{es} < T$ **do**

$\quad stop \leftarrow 0$, $T_d \sim f_{T_d}(t) + T_{es}$, $DTD \sim f_{DTD}(t)$, $T_{pr} \leftarrow (\lfloor \frac{T_d + DTD}{\Delta t} \rfloor + 1)\Delta t$, $T_f \sim f_{T_f}(t) + T_d$, $N \leftarrow \lfloor \frac{T_f - T_{pr}}{\Delta t} \rfloor$, $T_{main} \leftarrow T_{cor}$, $T_s \leftarrow T_f$

$\quad$ **if** $N > j$ **then**

$\quad\quad$ **for** k *in* $0 : N - h - 1$ **do**

$\quad\quad\quad \lambda \leftarrow \frac{k}{N}$

$\quad\quad\quad$ **if** $\underline{ra}_\lambda (N-k)\Delta t \leq h \cdot \Delta t$ **then**

$\quad\quad\quad\quad stop \sim Be(fp_\lambda)$

$\quad\quad\quad$ **else**

$$a \leftarrow h \cdot \Delta t - \underline{ra}_\lambda (N-k)\Delta t, \quad stop \sim Be\left(\frac{\sigma^2_{\underline{RA}_\lambda}[(N-k)\Delta t]^2}{\sigma^2_{\underline{RA}_\lambda}[(N-k)\Delta t]^2 + a^2} \times fp_\lambda\right)$$

$\quad\quad\quad$ **end**

$\quad\quad\quad$ **if** $stop = 1$ **then**

$\quad\quad\quad\quad t_\lambda \leftarrow T_{pr} + k \cdot \Delta t$, $T_{main} \leftarrow T_{pred}$, $T_s \leftarrow t_\lambda + j \cdot \Delta t$, go to $\star$

$\quad\quad\quad$ **end**

$\quad\quad$ **end**

$\quad\quad$ **for** k *in* $N - h : N - j$ **do**

$\quad\quad\quad \lambda \leftarrow \frac{k}{N}$

$\quad\quad\quad$ **if** $(2 - \overline{ra}_\lambda)(N-k)\Delta t \leq h \cdot \Delta t$ **then**

$$a \leftarrow h \cdot \Delta t - (2 - \overline{ra}_\lambda)(N-k)\Delta t, \quad stop \sim Be\left(fp_\lambda + (1 - fp_\lambda) \times \frac{\sigma^2_{\overline{RA}_\lambda}[(N-k)\Delta t]^2}{\sigma^2_{\overline{RA}_\lambda}[(N-k)\Delta t]^2 + a^2}\right)$$

$\quad\quad\quad$ **else**

$\quad\quad\quad\quad stop \sim Be(fp_\lambda)$

$\quad\quad\quad$ **end**

$\quad\quad\quad$ **if** $stop = 1$ **then**

$\quad\quad\quad\quad t_\lambda \leftarrow T_{pr} + k \cdot \Delta t$, $T_{main} \leftarrow T_{pred}$, $T_s \leftarrow t_\lambda + j \cdot \Delta t$, go to $\star$

$\quad\quad\quad$ **end**

$\quad\quad$ **end**

$\quad$ **end**

$\quad \star$: $T_{es} \leftarrow T_s + T_{main}$, $W_b(T_s : T_{es}) \leftarrow 1$

end

After B times

$$U(t) = 1 - A(t) \simeq \frac{\sum_{b=1}^{B} W_b(t)}{B}$$

allows estimating the probability of removing the component from operation at different time instants. To get this estimation, we have not made any parametric assumption to characterize the uncertainty in the predicted RUL; rather, we have exploited the one-sided Chebyshev's inequalities, which encode the Prediction Spread of RA. These inequalities provide an estimation of the least achievable availability benefit of a given PHM system, as they give an upper bound of the probability of removing the component from operation at time instants far from failure (i.e, when stopping the component does not allow exploiting all its useful life time) and an upper bound for the probability of not removing the component from operation at time instants near failure (i.e, when not stopping the component causes failure).

In the developed model, the component can be removed from operation at all time instants except the last $j < h$ instants: this allows applying the model to settings in which some time is needed to organize the preventive maintenance action, where if the failure is predicted too late, then there is no advantage in preventive maintenance.

The model has been applied to a simulated case study and a sensitivity analysis has been performed to identify the performance indicators which the unavailability is more sensitive to.

A simplified cost model has finally been developed in support to DMs, who have to decide whether to invest in PHM.

Further research work will investigate the application of the developed availability model to real engineering situations, to identify when PHM can really bring advantages to the industry business.

Other possible developments concern further improvements of the availability model, e.g., for relaxing some conservative assumptions or approximations and encoding diagnostics.

Appendix
Availability Simulation Algorithms

The pseudocode of the algorithm for estimating the component unavailability is as shown at the top of the previous page.

References

[1] E. Zio and M. Compare, “Evaluating maintenance policies by quantitative modeling and analysis,” *Rel. Eng. Syst. Safety*, vol. 109, pp. 53–65, 2013.

[2] M. Compare, P. Baraldi, P. Turati, and E. Zio, “Interacting multiple-models, state augmented Particle Filtering for fault diagnostics,” *Probab. Eng. Mech.*, vol. 40, pp. 12–24, 2015.

[3] M. M. Simões, J. M. Gomes, and C. F. Yasin, “A literature review of maintenance performance measurement: A conceptual framework and directions for future research,” *J. Qual. Maintenance Eng.*, vol. 17, pp. 116–137, 2011.

[4] H. Wang, “A survey of maintenance policies of deteriorating systems,” *Eur. J. Oper. Res.*, vol. 139, no. 3, pp. 469–489, 2002.

[5] E. Zio, “Challenges and opportunities in reliability engineering: The big KID (Knowledge, Information and Data),” *IEEE, Trans. Rel.*, to be published.

[6] M. S. Kan, A. C. C. Tan, and J. Mathew, “A review on prognostic techniques for non-stationary and non-linear rotating systems,” *Mech. Syst. Signal Process.*, vol. 62, pp. 1–20, 2015.

[7] K. Pipe, “Practical prognostics for condition based maintenance,” in *Proc. 2008 Int. Conf. Prognostics Health Manage.*, 2008, pp. 1–10.

[8] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, “Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications,” *Mech. Syst. Signal Process.*, vol. 42, no. 1/2, pp. 314–334, 2014.

[9] X. S. Si, W. Wang, C. H. Hu, and D. H. Zhou,“Remaining useful life estimation—A review on the statistical data driven approaches,” *Eur. J. Oper. Res.*, vol. 213, no. 1, pp. 1–14, 2011.

[10] E. Zio, “Prognostics and health management of industrial equipment,” *Diagnostics Prognostics Eng. Syst., Methods Tech., IGI Global*, pp. 333–356, 2012.

[11] A. Saxena, J. Celaya, B. Saha, S. Saha, and K. Goebel, “Evaluating algorithm performance metrics tailored for prognostics,” in *Proc. IEEE Aerosp. Conf.*, pp. 1–13, 2009.

[12] B. Saha and K. Goebel, “Modeling Li-ion battery capacity depletion in a particle filtering framework,” in *Proc. Annu. Conf. Prognostics Health Manage. Soc.*, pp. 2909–2924, 2009.

[13] A. Saxena, K. Goebel, D. Simon, and N. Eklund, “Damage propagation modeling for aircraft engine run-to-failure simulation,” in *Proc. IEEE Int. Conf. Prognostics Health Manage.*, pp. 1–9, 2008.

[14] Z. Zeng, F. Di Maio, E. Zio, and R. Kang, “A hierarchical decision-making framework for the assessment of the prediction capability of prognostic methods.” in *Proc. Inst. Mech. Eng., Part O: J. Risk Rel.*, 1748006X16683321, 2016.

[15] J. E. Dzakowic and G. S. Valentine, “Advanced techniques for the verification and validation of prognostics and health management capabilities,” *Proc. Machinery Failure Prevention Technologies*, Virginia Beach, VA, USA, pp. 1–11, 2007.

[16] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, “A review of process fault detection and diagnosis: Part I: Quantitative model-based methods,” *Comput. Chem. Eng.*, vol. 27, no. 3, pp. 293–311, 2003.

[17] A. Saxena, J. Celaya, B. Saha, S. Saha, and K. Goebel, “Metrics for offline evaluation of prognostic performance,” *Int. J. Prognostics Health Manage.*, vol. 1, no .1, pp. 4–23, 2010.

[18] A. Saxena, J. Celaya, B. Saha, S. Saha, K. Goebel, M. Schwabacher, and E. Balaban, “Metrics for evaluating performance of prognostic techniques,” in *Proc. IEEE Int. Conf. Prognostics Health Manage.*, pp. 1–17, 2008.

[19] L. Goodman, S. Wood, and A. Turner, “Return-on-investment (ROI) for electronic prognostics in mil/aero systems,” in *Proc. IEEE Autotestcon Orlando*, 2005, pp. 73–75.

[20] S. Wood and D. Goodman, “ Return-on-investment (ROI) for electronic prognostics in high reliability telecom applications,” in *Proc. Annu. Int. Telecommun. Energy Conf.*, 2006, pp. 1–3.

[21] K. Feldman, T. Jazouli, and P. A. Sandborn, “A methodology for determining the return on investment associated with prognostics and health management,” *IEEE Trans. Rel.*, vol. 58, no. 2, pp. 305–316, Jun. 2009.

[22] E. Zio, “An introduction to the basics of reliability and risk analysis,” *World Sci.*, vol. 13, pp. 89–91, 2007.

[23] T. Jazouli, P. Sandborn, and A. Kashani-Pour , “A direct method for determining design and support parameters to meet an availability requirement,” *Int. J. Perform. Eng.*, vol. 10, no. 2, pp. 211–225, 2014.

[24] E. Zio and F. Di Maio , “A data-driven fuzzy approach for predicting the remaining useful life in dynamic failure scenarios of a nuclear system,” *Rel. Eng. Syst. Safety*, vol. 95, no. 1, pp. 49–57, 2010.

[25] W. Zhao, T. Tao, and E. Zio, “System reliability prediction by support vector regression with analytic selection and genetic algorithm parameters selection,” *Appl. Soft Comput.*, vol. 30, pp. 792–802, 2015.

[26] W. Zhao, T. Tao, E. Zio, Y. F. Li, and J. Sun, “Condition-based component replacement of the pneumatic valve with the unscented particle filter,” in *Proc. IEEE Prognostics Syst. Health Manage. Conf.*, 2014, pp. 290–296.

[27] W. Zhao, T. Tao, and E. Zio, “Parameters tuning in support vector regression for reliability forecasting,” *Chem. Eng.*, vol. 33, pp. 523–528, 2013.

[28] J. Jacod and P. Protter, *Probability Essentials*. New York, NY, USA: Springer, 2004, pp. 197–211.

[29] I. R. Savage, “Probability inequalities of the tchebycheff type,” *J. Res. Nat. Bur. Stand., B. Math. Math. Phy.*, vol. 65B, pp. 211–222, 1961.

[30] P. Paris and F. Erdogan, “ A critical analysis of crack propagation laws,” *J. Basic Eng.*, vol. 85, no. 4, pp. 528–533, 1963.

[31] E. Zio, *The Monte Carlo Simulation Method for System Reliability and Risk Analysis*. New York, NY, USA: Springer, 2013.

[32] E. Zio, *Computational methods for Reliability and Risk Analysis*. New York, NY, USA: Springer, 2009.

[33] G. Curcurú, G. Galante, and A. Lombardo, “A predictive maintenance policy with imperfect monitoring,” *Rel. Eng. Syst. Safety*, vol. 95, no. 9, pp. 989–997, 2010.

[34] A. Ponchet, M. Fouladirad, and A. Grall, “Imperfect condition-based maintenance assessment on a finite time span,” in *Proc. IEEE Int. Conf. Qual., Rel., Risk, Maint. Safety Eng.*, 2012, pp. 390–395.

Authors’ photographs and biographies not available at the time of publication.