

CHAPTER 5

Prognostic Performance Metrics

Kai Goebel, Abhinav Saxena, Sankalita Saha, Bhaskar Saha, and Jose Celaya

5.1 Introduction.....	150
5.2 Background.....	151
5.2.1 Prediction Categorization.....	151
5.2.1.1 Forecasting.....	151
5.2.1.2 Prognostics.....	151
5.2.2 Prediction Methods.....	152
5.2.3 Performance Evaluation Methods.....	152
5.3 Metrics for Prognostic Applications.....	154
5.3.1 Certification Metrics.....	154
5.3.2 Cost Benefit Metrics.....	154
5.3.3 Metrics for Computational Performance.....	155
5.3.4 Metrics for Reliability Analysis.....	155
5.3.5 Metrics for Prognostics Algorithm Performance.....	155
5.3.5.1 Challenges.....	156
5.3.6 Error-based Metrics.....	157
5.3.6.1 FP, FN, and ROC.....	157
5.3.6.2 Spread-Based Metrics.....	158
5.3.6.3 Anomaly Correlation Coefficient (ACC).....	159
5.3.6.4 Prognostic Horizon.....	159
5.3.6.5 α - λ Performance.....	160
5.3.6.6 Relative Accuracy.....	162
5.3.6.7 Cumulative Relative Accuracy.....	162
5.3.6.8 Convergence.....	163
5.3.6.9 Robustness.....	164
5.3.6.10 RUL Online Precision Index (RUL-OPI).....	164
5.3.7 Incorporating Uncertainty Estimates.....	165
5.3.8 Guidelines for Applying Prognostics Metrics.....	167
5.3.8.1 Guidelines on choosing performance parameters.....	167
5.3.8.2 Guidelines for dealing with uncertainties.....	167
5.3.8.3 Guidelines to resolve ambiguities.....	168
5.4 Summary.....	169
Acknowledgments.....	169
References.....	169

5.1 Introduction

We define prognostics here strictly as the predicting of Remaining Useful Life (RUL) of a component or system. The prediction is typically performed only after the “health” of the component or system deteriorates beyond a certain threshold. Often times, that threshold is tripped because a fault occurs. A fault is a state of a component or system that deviates from the normal state such that the integrity of the component is outside of its required specification. A fault does not necessarily imply that the overall system does not operate anymore; however, the damage that characterizes the fault often grows under the influence of operations to a failure. The latter is the state at which the component or system does not meet its desired function anymore. It is the task of prognostics to estimate the time that it takes from the current time to the failed state, conditional on anticipated future usage. This would give operators access to information that has significant implications on system safety or cost of operations. Where safety is impacted, the ability to predict failure allows operators to take action that preserves the assets either through rescue operation or through remedial action that avert failure altogether. Where minimizing cost of operations is the primary objective, predictive information allows operators to avert secondary damage, or to perform maintenance in the most cost-effective fashion. Often times, there is a mix of objectives that need to be optimized together, sometimes weighted by different preferences.

Predicting remaining component or system life can be accomplished in several ways. Where sufficient historical run-to-failure trajectories are available, data mining techniques can be employed to perform the predictions. Traditionally reliability based predictions have been used widely in the manufacturing industry to schedule preventive maintenance. In contrast, the focus of this chapter is mainly on condition based prognostic systems for a particular monitored unit under test (UUT). Instead of considering the entire population for a statistical life estimates, one can employ physics-based models to perform the predictions or a combination of models and history data. In either case, predictions are conditional on future conditions and are subject to significant amounts of uncertainty. Methods for prognostics ideally express their confidence of their own prediction based on an assessment of the various uncertainty sources. Besides uncertainty of future usage, uncertainty also comes from the current state assessment, the models used, measurement noise, etc.

Metrics can be understood as a standardized language by which technology developers and users communicate their findings and compare results. This aids in allowing the proper expression of requirements as well as the dissemination of scientific information. Two surveys on methods for prognostics, one on data-driven methods (Schwabacher, 2005) and one on artificial-intelligence-based methods (Schwabacher & Goebel, 2007) reveal a lack of standardized methodologies for performance evaluation or a lack of performance methods altogether. The most recent ISO standard by International Organization for Standards (ISO, 2004) for prognostics in condition monitoring and diagnostics of machines does not even provide a firm definition of any such method. Nonetheless, there has been recently a significant push towards crafting suitable metrics to evaluate prognostic performance (Saxena et al., 2010). These metrics address primarily evaluation of algorithmic performance for prognostics applications. They are mostly focused on tackling offline performance evaluation methods for applications where run-to-failure data are available and true End-of-Life (EoL) is known a priori. They are therefore particularly useful for the algorithm development phase where feedback from the metrics can be used to fine-tune prognostic algorithms. It needs to be appreciated that these

metrics are continuously evolving. Efforts are also underway towards designing on-line performance metrics although they have not reached a significant level of maturity.

This chapter presents a discussion on prognostics metrics. After a review of performance assessment for prediction/forecasting applications a categorization of prognostic metrics into several classes is performed. This categorization suggests that there can be various different objectives that drive improvements in prognostic performance and correspondingly different set of metrics may be used to obtain performance feedback.

5.2 Background

As more diverse research communities and practitioners start adopting PHM techniques it becomes imperative to use standardized prognostic methodologies (Uckun et al., 2008) as well as to use metrics to measure performances. However, since prognostics is a developing field, the challenges in developing standards and metrics are numerous (Engel, 2008; Engel et al., 2000). We start out by providing an overview of prognostic concepts that are used in a variety of domains.

5.2.1 Prediction Categorization

Prior to delineating the methods to assess prognostic performance it may be useful to provide a brief discussion about different types of applications in which predictions are employed. Based on an analysis of aerospace, medicine, nuclear energy, finance, weather and automotive domains, it was found that one can distinguish roughly between forecasting and prognostics.

5.2.1.1 Forecasting

Forecasting is found in applications where predictions are made to describe expected future behavior without predicting a fixed target. That is, there is no notion of “end of life” and consequently there is no concept of RUL. Example application areas are weather or finance domains. The prediction format can be either quantitative (e.g., prediction of exact numbers) or qualitative (e.g., high or low demands) in nature. Furthermore, the data trends are generally non-monotonic in such applications. Predictions may be discrete (e.g., forecasting market demand for a particular month) or continuous (e.g., variation of temperature over the period of next week). Details and more references to such applications in various domains can be found in (Saxena et al., 2008).

5.2.1.2 Prognostics

The other class of applications makes use of critical thresholds such that if the system under test crosses this threshold it is declared to have failed or lost its functional capability. This class of applications – e.g. medicine, nuclear, mechanical and electrical industrial systems, – involves predictions of RUL and involves decay or fault propagation models to capture the behavior of the system.

Predictions can be made in two forms: 1.) an *event prediction* where the time for EoL is estimated; and 2.) a *decay prediction* where the complete future trajectory is predicted until EoL is reached. It must be noted, however, that EoL criteria need not always be a complete loss or failure. In safety critical applications EoL is often a degraded state where performance level has deteriorated to cross a predetermined safety margin even though the component may still retain partial functionality. For example, in the electronics domain, EoL of a switching device (such as

a MOSFET) is not necessarily the complete loss of the switching functionality. Instead, it could be a decrease in the switching frequency below a certain threshold level.

There are two main types of applications where predictions for system health are made: These include, *predicting wear* of a system or *predicting failure* in the event of a fault.

Failure Predictions: An otherwise healthy system may encounter a fault that grows due to continued usage (or exposure to adverse conditions) that may result into a failure. In such cases it is critical to detect the presence of a fault (ideally shortly after it happens), the particular fault mode, its severity and rate of growth, so that appropriate decisions may be taken to avoid undesired, possibly catastrophic, events. Here the task of prognosis is to estimate expected End-of-Life (EoL), i.e., determine when the system will no longer operate under specifications. In some cases it is not only important to know when the system will break but also how it will approach the failure. In those cases, instead of predicting just the event of EoL, a complete trajectory may be predicted, where the end point of the trajectory also determines the EoL. Examples of such applications include structural faults like cracks in metallic structures or die-attach degradation in power semiconductor devices.

Wear Predictions: There are many situations where systems undergo expected normal wear and need to be maintained or replaced whenever the wear levels impact functionality. In these cases, the system does not experience a fault condition even under the degraded performance. Therefore, the health of the system is tracked from the very beginning of system deployment and detection and diagnosis are not predecessors for prognostics. As stated earlier, the end point of these trajectories can be used to determine the EoL point so appropriate decisions may be taken. Examples of such applications include battery capacity degradation and valve leakage due to wear.

5.2.2 Prediction Methods

There are several ways to carry out prognostics. In some cases a detailed physical model of the unit under observation can be used. The model captures the unit's behavior under operational and environmental conditions and provides an expected response that describes the current and (given the proper input) future states. Alternative to a physics-based model, historical data can be utilized to estimate expected time to failure. The key is to either have access to a sufficient amount of existing historical data (e.g., medicine) or to be able to experimentally generate run-to-failure trajectories (e.g., for some mechanical systems). Then, a variety of data-driven or statistical techniques can be applied.

The availability of run-to-failure data allows the straightforward evaluation of prediction performance by comparing the predicted EoL to the actual EoL. However, there are many applications where run-to-failure experiments cannot be afforded or where very little failure history data are available (e.g., aerospace). It becomes somewhat more difficult to assess the performance in such cases due to the absence of knowledge about the future outcomes. Methods are tested on experimental or simulated data and, when fielded, are expected to perform similarly on real systems. However, algorithm functionality does rarely translate without loss of performance from simulation environment or lab to the field. Indeed, validation and verification of prognostic methods remains a thorny issue.

5.2.3 Performance Evaluation Methods

Techniques employed for prediction or forecasting in the application areas enumerated above use metrics that are based on accuracy and precision with several slight variations (Saxena, et al., 2008). Mostly, they are customized to better serve a particular domain. In

medicine and finance, for example, several statistical measures are used to benefit from the availability of large datasets. In contrast, predictions in medicine are commonly evaluated based on hypothesis testing methodologies. And in the finance domain, errors are calculated based on reference prediction models. The precision and accuracy metrics include, for example, mean squared error (MSE), standard deviation (SD), mean absolute deviation (MAD), median absolute deviation (MdAD), mean absolute percentage error (MAPE) and similar variants. Other domains, like aerospace, electronics, and nuclear are less mature with respect to fielded prognostics applications. There, metrics from other system health techniques, like diagnostics, have been used with the goal to capture the characteristics of prognostics (with varied success). Metrics used include false positives, false negatives and receiver operator characteristics (ROC) curves (Goebel & Bonissone, 2005). Other metrics include those from the reliability domain like mean time between failures (MTBF) or the ratio mean time between unit replacements (MTBF/MTBUR). Adaptations include for example the augmentation with business metrics such as Return on Investment (ROI) (Sandborn, 2005), Technical Value (TV) (Byington et al., 2005), Net Present Value (NPV) (Kacprzyński et al., 2001) and Life Cycle Cost (LCC) (Luna, 2009).

It becomes apparent that there are several types of metrics for prognostics based on the purpose of prognostics and the end user. A categorization with these objectives in mind allows a more targeted choice of appropriate metrics. Coble & Hines (Coble & Hines, 2008) categorized prognostic algorithms into three categories based on type of models/information used for predictions. Wheeler et al. (Wheeler et al., 2009) categorized end users from a health management stakeholder's point of view. The top-level user groups were: Operations, Regulatory, and Engineering. We combine and expand on these notions and categorize prognostic metrics based both on their goal as well as end users (see Table 1).

Table 1 - Categorization of Prognostic Metrics based on End Usage

Metrics	Assessment Goals	Operations				Engineering		Regulatory
		Program Manager	Plant Manager	Operator	Maintainer	Designer	Researcher	Policy Maker
Certification Metrics	Assess conformance to safety assurance and certification requirements	X						X
Cost-Benefit Metrics	Assess the economic viability for specific applications before It can be approved or funded	X						X
Reliability Based Metrics	Assess probabilities of failures based on statistical evidence from multiple systems			X		X		X
Algorithm performance Metrics	Assess performance of prediction algorithms in predicting EoL	X	X	X	X		X	X
Computational Performance Metrics	Assess computational requirements					X	X	

A favorable Cost-Benefit case is the hinge pin of a successfully fielded prognostic solution and

cost-benefit metrics allow a quantification of the degree of fulfillment. Similarly, verifiability and certification metrics determine the degree to which a prognostic solution conforms with safety assurance and certification requirements. Both these top level metrics categories require prognostic estimates to satisfy stringent performance metrics, which are often derived from reliability analysis or condition-based prognostic methods. Computational performance metrics are important for implementation and figure in the trade space of cost-benefit analysis and algorithmic performance. Often, most of the metrics mentioned are connected through a requirement specification process.

5.3 Metrics for Prognostic Applications

In this section, the various prognostic metrics are presented by categories in more detail. Particular emphasis is given to performance metrics.

5.3.1 Certification Metrics

Regulatory bodies (such as the FAA) are concerned with whether a fielded system might negatively impact (directly or indirectly) overall system safety. Associated metrics can be expressed as the logical conjunction of an exhaustive set of safety related use cases.

5.3.2 Cost Benefit Metrics

Establishing cost benefits of prognostics is an important step in integrating the health management practices into fielded applications. Thus, metrics that measure economic viability of prognostics have started gaining in importance. Some of the most common Cost-Benefit metrics include:

5.3.2.1 MTBF to MTBUR Ratio: This reliability statistics-based method expresses the efficiency of a maintenance operation by measuring the ratio between the lengths of time a component is expected to last and the length of time for which it was used before it was replaced (Teal & Larsen, 2003).

5.3.2.2 Life Cycle Cost (LCC): LCC is fundamentally the sum of acquisition cost and cost of operations. To assess the value of prognostics, LCC is compared with and without prognostics (Wilkinson et al., 2004).

5.3.2.3 Return on Investment (ROI): In an ROI calculation the difference between return and investment (the gain) is divided by the investment. It is one of the most commonly used metrics (not just in the context of prognostics) that assesses the benefits of deploying a PHM system.

5.3.2.4 Technical Value: The benefits achieved through accurate detection, fault isolation and prediction of critical failure modes are weighed against the costs associated with false alarms, inaccurate diagnoses/prognoses, and resource requirements of implementing and operating specific techniques. (Dzakowic & Valentine, 2006; Vachtsevanos et al., 2006).

5.3.2.5 Total Value: Given the coverage a PHM system provides for multiple fault modes in a system, total value quantifies the usefulness of a PHM technology in a particular application. Total value is defined as the summation of the benefits prognostics provides over all the failure modes that it can diagnose or prognose, minus the implementation cost, operation and maintenance cost, and consequential cost of incorrect assessments. This metric connects the algorithm performance in a PHM system to the management and operational performance.

5.3.3 Metrics for Computational Performance

Computational performance is one of the most closely related factors to actual implementation of a system. It provides a mechanism to negotiate between computational and time resources that are demanded and the required fidelity and accuracy of the PHM system. Most of the metrics that can be used to quantify computational performance come from theoretical computer science and computer engineering. These are not specific to prognostics and are just mentioned here for completeness. Computational performance metrics include: Computational Complexity, CPU Time, Memory Size, and Data Rate. Depending on the length of the prediction horizon in a prognostic application, data processing capabilities are of greater significance from the design and implementation point of view. All the above metrics help specify the hardware requirements or otherwise specify constraints within which a software must work and still satisfy algorithmic performance requirements.

5.3.4 Metrics for Reliability Analysis

Referring again to table 1, reliability analysis metrics are chiefly used by operators, designers, and policy makers. Reliability analysis stems from statistical evidences aggregated from historical data. Failure rates and distributions are extracted from history data or experimental data, which are then used to make failure predictions for a system under test. While prognostics is the science of prediction based on condition and usage of the monitored UUT, the reliability analysis predicts failures based on expected outcome from the observed statistic over a population of the UUT. It has been the traditional way to use these metrics to assess the costs and the risks of using a system. Used correctly, reliability metrics are connected to system improvement due to prognostic performance.

Broadly classifying there are two types of reliability metrics as described below (Wood, 2001):

5.3.4.1 Constant Rate Reliability Metrics: these are the most popular reliability metrics in the electronics industry as they represent a good approximation of the flat region of the reliability bathtub curve. Mean life metrics usually assume an exponential distribution, which makes them equivalent to constant rate metrics. These rates are typically measured from field data and are simple and intuitive to explain. Some common examples of these metrics are: Mean Time Between Failure (MTBF), Mean Time to Failure (MTTF), Part Return/Repair Rate, Part Replacement Rate, Mean Time Between Service Call (MTBSC) and Mean Time Between Maintenance Action (MTBMA).

5.3.4.2 Probability of Success Metrics: When systems do not show constant failure rates, specifying mean times does not suffice. In such cases a better way is to specify probability of success or, in other words, the probability that a system performs a required function under stated condition for a stated period of time. Another way to specify probability of success is to measure the percentage of population that survives a specific duration. Therefore, these metrics are usually time dependent, i.e., the probability of success will depend on the length of the mission. These may be specified as the percentiles of the distributions. A common example used in mechanical systems domain is *L_x Life* which specifies the number of hours after which at least *x*% of the population would have failed. Other metrics commonly used are Failure Free Operating Time, Maintenance Free Operating Time, Mean Mission Duration, etc. More discussion on reliability based metrics may be found in (IEEE-Standard1413, 2002).

5.3.5 Metrics for Prognostics Algorithm Performance

Before using performance metrics, an a priori analysis should be conducted to identify the relevant factors in a given application and address them appropriately. A good set of metrics

should accommodate all or most of these factors. In this context the challenges surrounding prognostics (as compared to, say, diagnostics) should be discussed. It should also be noted that the metrics detailed in this chapter are continuously evolving as the field matures further.

5.3.5.1 Challenges

Prognostics requires special consideration in a number of areas. These include the acausality phenomenon, need for run-to-failure data, online performance evaluation, and expression of estimation uncertainty.

Acausality: An acausal system is defined as a system with outputs and internal states that depend on future input values. Prognostics has acausal properties. It requires input from future events (for instance knowledge about operational conditions and load profiles) in order to make accurate predictions. To accurately assess the performance (both accuracy or precision), one must also know the true EoL to compare with the predicted EoL estimates. In some cases, future operating conditions are well known. This is the case for example for stationary applications with constant operating conditions. However, in non-stationary applications and where knowledge about future events is not available, estimates may be derived based on past usage history, the expected mission profile, and predictions for future operating and environmental conditions that are not controllable (e.g., weather conditions). This however, adds uncertainty to the overall process and complicates prognostic performance evaluation.

Run-to-Failure Data from Real Applications: Assessing the correctness of prognostics benefits greatly from allowing the system to fail such that the prediction can be confirmed. For many systems, this is not feasible because it may be too expensive or because it negatively impacts system safety. But if a corrective action (such as maintenance or repair) is taken, one has just removed the ability to assess how early the prediction was. This is sometimes referred to as the “paradox of prognostics”.

Online Performance Evaluation: The aforementioned considerations lead to an argument in favor of controlled run-to-failure (RtF) experiments for the algorithm development phase. While this may allow offline performance evaluation, some issues remain: First, it is difficult to extend the results of offline conditions to a real-time scenario; Second, a RtF experiment needs often times frequent disassemblies to gather ground truth data. This assembly-disassembly process creates variations in the system performance and the EoL point shifts from what it may have been in the beginning of the experiment. Since actual EoL is observed only at the end there is no guarantee that a prediction made based on initial part of data will be very accurate. Whereas, this does not necessarily mean that prognostic algorithm is poorly trained, it is difficult to prove otherwise. Therefore, one must be careful while interpreting the performance assessment results. Third, even controlled subscale RtF experiments can be very expensive and time consuming, in particular if one seeks to conduct statistically significant number of experiments for an exhaustive set of components and fault modes.

Uncertainty in Prognostics: The quantification of prediction confidence is indispensable in prognostics. Consider a remaining life estimate of, say, 5 hours. If one knew that the confidence bounds for a given risk acceptance level are also 5 hours, then reactive action has to be taken immediately. If, however, the confidence bounds are at plus/minus 1 hour, then a completely different set of action can be taken. Without such information any prognostic estimate is of limited use and cannot be incorporated in mission critical applications (Uckun, et al., 2008). Uncertainties arise from various sources in a PHM system (Coppe et al., 2009; Hastings & McManus, 2004; Orchard et al., 2008). Some of these sources include:

- Model uncertainties (errors in the representation and parameters of both the system model and fault propagation model),
- Measurement uncertainties (these arise from sensor noise, ability of sensor to detect and disambiguate between various fault modes, loss of information due to data preprocessing, approximations and simplifications),
- Operating environment uncertainties,
- Future load profile uncertainties (arising from unforeseen future and variability in usage history data),
- Input data uncertainties (estimate of initial state of the system, variability in material properties, manufacturing variability), etc.

Assessing the levels and characteristics of uncertainties arising from each of these sources is often times not a trivial task. It is even more difficult to determine how these uncertainties combine at different stages of the prognostic process and propagate through the – possibly complex and non-linear – system. On top of that, statistical properties may not follow any known parametric distributions, therefore complicating analytical solutions.

Owing to all of these challenges Uncertainty Representation and Management has become an active area of research in the field of PHM (DeNeufville, 2004; Ng & Abramson, 1990; Orchard, et al., 2008; Sankararaman et al., 2009; Tang et al., 2009). Methods for prognostic performance evaluation, then, must be able to incorporate various expressions of uncertainties.

Performance metrics for prognostics can be classified into accuracy, precision, and robustness. We use the working definition for accuracy as those that assess the degree of closeness of predictions to the actual failure time. Precision is defined as the spread of predictions performed at the same time. Robustness is defined as the sensitivity of the predictions with changes of algorithm parameter variations or external disturbances. There are a large number of prognostic performance metrics that have been used. However, as discussed earlier, most of these metrics do not take into consideration the particular challenges of prognostics. Hence, we feature here only a subset of general metrics especially suitable for prognostics. For a comprehensive list of performance metrics, the reader is referred to (Saxena et al., 2008).

5.3.6 Error-based Metrics

Many metrics are based on the assessment of the error, i.e., the deviation of the actual output from the target. One example of such an error metric is the average scale independent error. This metric provides an exponential weight of the errors in RUL predictions and averages over several UUTs (Vachtsevanos et al. 2006, Hyndman and Koehler 2006);

$$A(i) = \frac{1}{L} \sum_{l=1}^L \exp \left\{ - \frac{|\Delta^l(i)|}{D_0} \right\} \quad (1)$$

Where, Δ is the error and D_0 is a normalizing constant whose value depends on the magnitudes in the application. The range of $A(i)$ varies between 0 and 1 where 1 represents perfect score. Other error-based metrics include Root Mean Squared Error (RMSPE), and Mean Absolute Percentage Error (MAPE).

5.3.6.1 FP, FN, and ROC

FP and FN are at heart also error-based metrics, but they deserve special consideration. A

common way to assess performance is to treat predictions as dichotomous forecasts by means of categorizing them into false positives (FP), false negatives (FN), true positives (TP) and true negatives (TN) (Goebel & Bonissone, 2005). FP assesses unacceptable early predictions and FN assesses unacceptable late predictions at specified time instances. User must set acceptable ranges (t_{FN} and t_{FP}) for prediction. Early predictions result in excessive lead time, which may lead to unnecessary corrections. Also note that, a prediction that is late more than a critical threshold time units (t_c) is equivalent to not making any prediction and having the failure occurring. Mathematically, FP is defined as

$$FP(r_*^l(i)) = \begin{cases} 1 & \text{if } \Delta^l(i) > t_{FP} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where t_{FP} = user defined acceptable early prediction. While false negative (FN) is defined as

$$FN(r_*^l(i)) = \begin{cases} 1 & \text{if } -\Delta^l(i) > t_{FN} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where t_{FN} = user defined acceptable late prediction.

FP and FN both can vary between values is 0 and 1 where 1 represents perfect score. FP and FN can then be compiled into Receiver Operating Characteristic (ROC) curve. The ROC allows to assess the tradeoff between false positives (FP) and false negatives (FN) (Ebert 2007, Palmer et al. 2004) in a comprehensive fashion by plotting (1 - false negatives) over the false positives (see Figure 1). The ideal curve would have zero false positives and zero false negatives, but such a curve cannot realistically be achieved for real-world problems. Use of Time dependent ROC has been suggested that depicts ROC obtained for forecasts made for different time horizons. Also each point on the ROC curve may be associated with a point wise fixed width confidence bounds to indicate confidence in predictions. Tuning the prognostic algorithm such that a ROC can be generated may prove difficult in practice (e.g., due to lack of data or lack of tuning “parameters”).

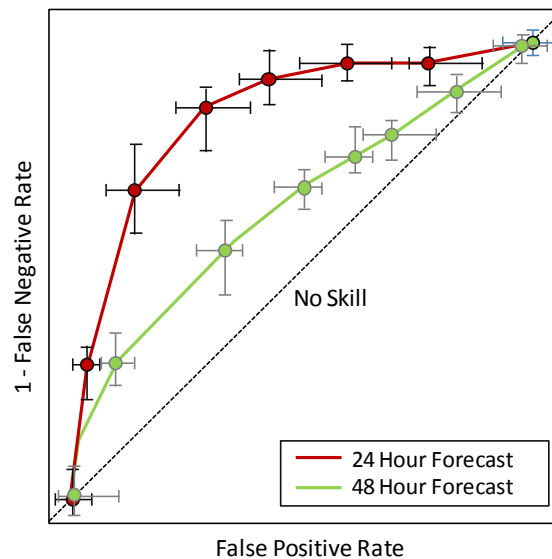


Figure 1 - ROC curve.

5.3.6.2 Spread-Based Metrics

Spread-based metrics measure the dispersion/spread of the error. The most basic spread-based

metric is the sample deviation which considers the error dispersion with respect to the error sample mean (Hoaglin et al., 1983; Vachtsevanos, et al., 2006).

$$S(i) = \sqrt{\frac{\sum_{l=1}^n (\Delta^l(i) - M)^2}{n-1}} \quad (4)$$

where M is the sample mean of the error. This metric is restricted to the assumption of normal distribution of the error. It is, therefore, recommended to carry out a visual inspection of the error plots. SD can vary between 0 and ∞ , where the perfect score is 0. Other spread-based metrics include the Mean Absolute Deviation from Sample Median (MAD) and the Median Absolute Deviation from the sample median (MdAD).

5.3.6.3 Anomaly Correlation Coefficient (ACC)

This metric is used to measure correspondence or phase difference between prediction and observations, subtracting out the historical mean at each point and is frequently used to verify output from numerical prediction models (Ebert, 2007). ACC is not sensitive to error or bias, so a good anomaly correlation does not guarantee accurate predictions. In the PHM context, ACC computed over a few time-steps after t_p can be used to modify long term predictions. However, the method requires computing a baseline from history data which may be difficult to come by. Mathematically, ACC can be represented as follows:

$$ACC = \frac{\sum (\pi^l(i|j) - z_{\#}(i))(z_*(i) - z_{\#}(i))}{\sqrt{\sum (\pi^l(i|j) - z_{\#}(i))^2 \sum (z_*(i) - z_{\#}(i))^2}} \quad (5)$$

where, $z_*(i)$ is a prediction variable (e.g. $f_n^l(i)$ or $h_n^l(i)$), and $z_{\#}(i)$ is the corresponding history data value. ACC can vary between -1 and 1 where 1 represents perfect score.

Thus, ACC averages the absolute percentage errors in the predictions of multiple UUTs at the same prediction horizon. The percentage is computed based on the mean value of the prediction and ground truth. This prevents the percentage error from being too large for the cases where the ground truth is close to zero. This metric is computed at a particular time and does not capture performance variation with time.

5.3.6.4 Prognostic Horizon

The Prognostic Horizon (PH) can be formally defined as the difference between the time index i when the predictions first meet the specified performance criteria (based on data accumulated until time index i) and the time index for EoL. PH can be considered as a robustness metric. The basic notion behind the metric is that a longer PH implies more time to act based on a prediction that has some credibility. The performance requirement is specified in terms of an allowable error bound (α) around the true EoL where the choice of α depends on the estimate of time required to take a corrective action. PHs are typically determined offline during the validation phase for an algorithm-application pairing. PH performance is then used as a guideline for algorithm deployment where actual EoL is not known in advance.

$$PH = EoL - i \quad (6)$$

where:

$$i = \min\{j | (j \in P) \wedge ((r_* - \alpha \cdot EoL) \leq r(j) \leq (r_* + \alpha \cdot EoL))\}$$

P is a set of all time indices for which a prediction is made

r_* is the true RUL

$r(j)$ is the predicted RUL at time index j

α is the accuracy modifier

EOL is the ground truth End of Life point

PH output is a score that is characterized by both the length of remaining life of a system and the time scales in the problem at hand. As shown in Figure 2, the desired level of accuracy with respect to the *EoL* ground truth is specified as $\pm\alpha$ -bounds. A remaining life estimate within those bounds has sufficient utility to a user (it is not too far off from the target to be actionable). The PH for an algorithm is declared as soon the corresponding predictions enter the α -bounds. RUL values can be superimposed for various algorithms, thus providing an easy aid in their comparison. As evident from Figure 2, the algorithm A1 has a longer PH than algorithm A2.

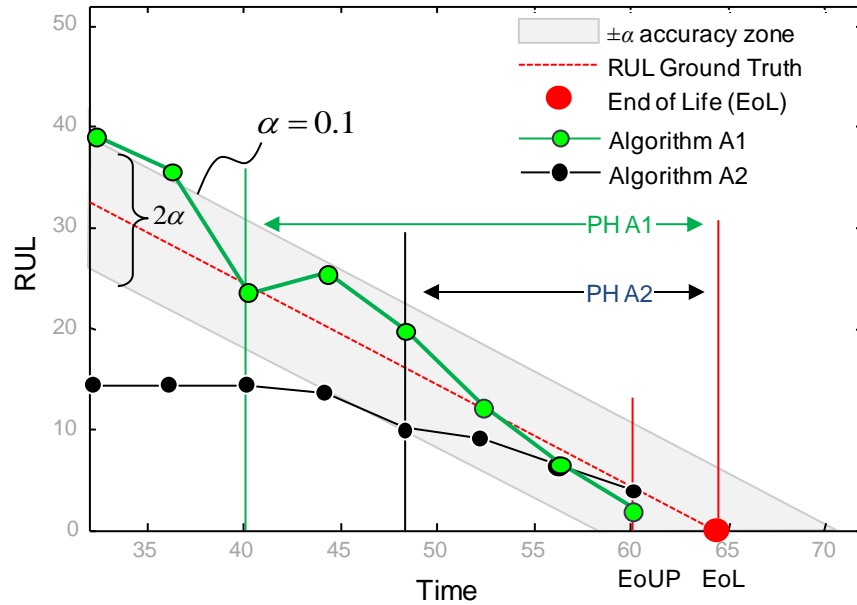


Figure 2 – Prognostic Horizon

5.3.6.5 α - λ Performance

α - λ metric quantifies the prediction quality by determining whether the prediction falls within specified limits at particular times with respect to a performance measure. The evaluation times may be specified either as a ratio to the total remaining life from the point the first prediction is made or it may be specified as a given absolute time interval before *EoL* is reached. α - λ performance could be expressed either as α - λ accuracy, as α - λ precision, or as α - λ robustness metric. In the discussion below we delineate α - λ performance without loss of generality as an accuracy performance measure.

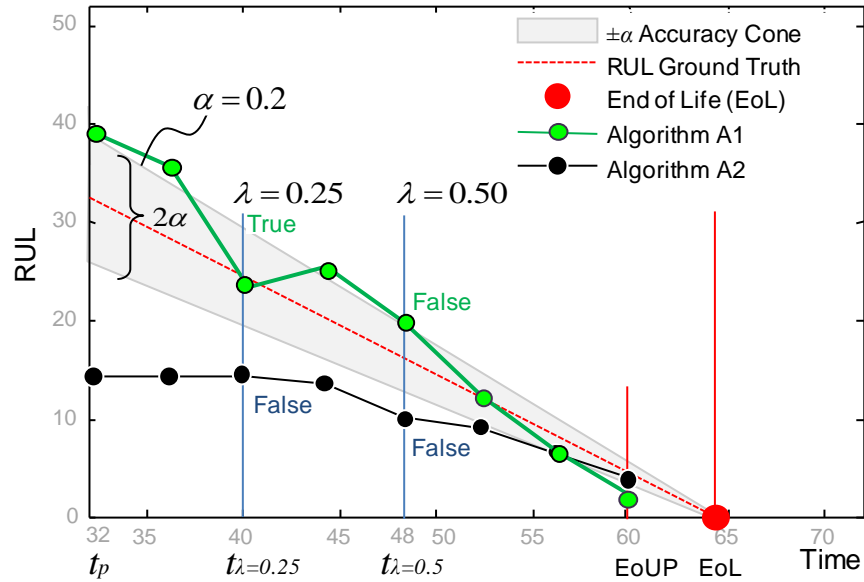


Figure 3 - α - λ Performance

Here we define α - λ accuracy as the prediction accuracy to be within α % of the actual RUL at specific time instance expressed as a fraction of time between the point when an algorithm starts predicting and the actual failure (Figure 3). Consider an example case where this metric determines whether a prediction falls within 20% accuracy (i.e., $\alpha=0.2$) halfway to failure from the time the first prediction is made, (i.e., $\lambda=0.5$). The α - λ accuracy metric is defined as,

$$\alpha - \lambda \text{ Accuracy} = \begin{cases} 1 & \text{if } (1 - \alpha) \cdot r_*(i) \leq r'(i_\lambda) \leq (1 + \alpha) \cdot r_*(i) \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Where α : accuracy modifier

λ : time window modifier such that $t_\lambda = t_p + \lambda(t_{EoL} - t_p)$

For illustrating the usage of this performance measure several prediction algorithms employed in (Goebel et al., 2008; Liu et al., 2010; Saxena et al., 2009) are compared in Figure 4. Here this metric evaluates whether predictions made by various algorithms lie within 10% error when evaluated at halfway to the EoL.

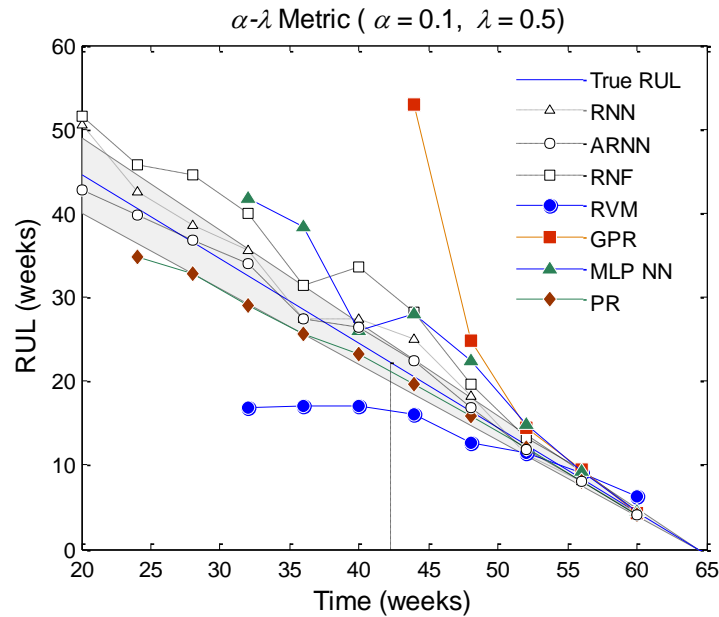


Figure 4 - α - λ Performance Comparison

5.3.6.6 Relative Accuracy

RA is defined as a measure of error in RUL prediction relative to the actual RUL.

$$RA_{\lambda} = 1 - \frac{|r_{*}(i_{\lambda}) - \langle r^l(i_{\lambda}) \rangle|}{r_{*}(i_{\lambda})} \quad (8)$$

where

λ is the time window modifier such that $t_{\lambda} = t_p + \lambda(t_{EoL} - t_p)$,

l is the index for l^{th} unit under test (UUT),

$r_{*}(i_{\lambda})$ is the ground truth RUL at time index i_{λ} ,

$\langle r(i_{\lambda}) \rangle$ is an appropriate point estimate of the predicted RUL distribution at time index i_{λ} .

5.3.6.7 Cumulative Relative Accuracy

RA conveys information at a specific time. In order to estimate the general behavior of the algorithm, RA can be evaluated at multiple time instances to provide an aggregate accuracy level or the Cumulative Relative Accuracy (CRA).

$$CRA_{\lambda} = \frac{1}{|\ell_{\lambda}|} \sum_{i \in \ell_{\lambda}} w(r(i)) RA_{\lambda} \quad (9)$$

where:

$w(r(i))$ is a weight factor as a function of RUL at all time indices,

ℓ_{λ} is the set of all time indexes when a prediction was made

It may be desirable to give more weight to RA evaluated at times closer to EoL since good performance close to EoL is important for condition-based decision making. Therefore, one would expect that λ is chosen in a meaningful fashion, e.g., the time required to apply a corrective action. RA evaluated at $\lambda = 0.5$ indicates the time when a system is expected to have consumed half of its remaining life. Alternatively, RA could be evaluated at time instances where the damage magnitude has reached 50% of the failure threshold. This metric is also useful

in comparing different algorithms for a given λ in order to get an idea on how well a particular algorithm does at significant times.

5.3.6.8 Convergence

Convergence expresses the rate at which any metric (M) – like accuracy or precision – improves with time. The error of different algorithm metric evaluation is connected into a curve. Convergence is then defined as the distance between the origin and the centroid of the area under the curve for a given metric.

$$C_M = \sqrt{(x_c - t_p)^2 + y_c^2}, \quad (10)$$

where:

C_M is the Euclidean distance between the center of mass (x_c, y_c) and $(t_p, 0)$

$M(i)$ is a non-negative prediction accuracy or precision metric with a time varying value

(x_c, y_c) is the center of mass of the area under the curve $M(i)$ between t_p and t_{EoUP} , defined as following

$$x_c = \frac{\frac{1}{2} \sum_{i=P}^{EoUP} (t_{i+1}^2 - t_i^2) M(i)}{\sum_{i=P}^{EoUP} (t_{i+1} - t_i) M(i)},$$

$$y_c = \frac{\frac{1}{2} \sum_{i=P}^{EoUP} (t_{i+1} - t_i) M(i)^2}{\sum_{i=P}^{EoUP} (t_{i+1} - t_i) M(i)} \quad (11)$$

EoUP is the time index for last useful prediction made. Alternatively one may use EoP, but EoUP makes sure that performance is evaluated only based on those predictions that are useful from a practical view point since any prediction made after EoUP does not leave enough time to carry out any corrective measure. (Saxena et al., 2009).

As stated earlier, convergence banks on the implicit assumption that algorithm performance should improve with time. For illustration of the concept, consider three cases that converge at different rates in Figure 5. Lower distance implies a faster convergence.

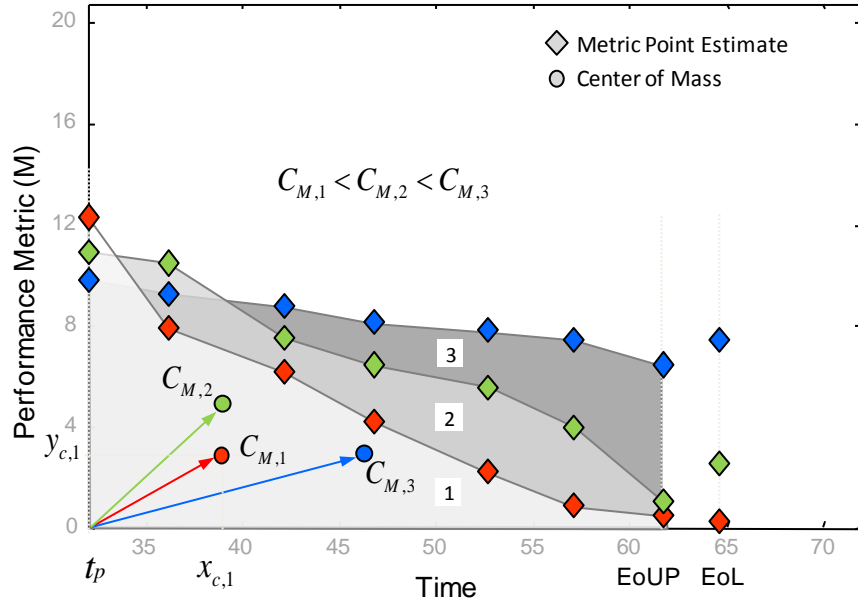


Figure 5 - Convergence Metric

5.3.6.9 Robustness

A robustness metric has the task of quantifying the sensitivity of an algorithm with respect to its parameters, such as those found in expressing prior distribution, initial conditions, and training data size. Confidence bounds of a robust algorithm are not expected to change much with variations of algorithm parameters. Mathematically, the robustness metric R_b can be defined as

$$R_b = \frac{\int_{x_{mean}-\eta}^{x_{mean}+\eta} f(x) dx}{4\eta\delta} \quad (12)$$

where x is the investigated algorithm parameter and $f(x)$ is the confidence bound variation function with respect to x .

The assessment of algorithm robustness is of high value in particular since most of the time an accurate prior is difficult to obtain with limited data source and extensive experiments on the actual engineering system are often prohibitive due to time and cost constraints (Guan et al., 2011).

5.3.6.10 RUL Online Precision Index (RUL-OPI)

This index quantifies the length of 95% confidence bounds relative to the predicted RUL at any given time instant (Orchard et al., 2009). The index is normalized between 0 and 1. It can be used as an online performance metric to ensure if I_1 remains close to 1 as system deteriorates, i.e., EoL approaches:

$$I_1(i) = e^{-\left(\frac{\sup\{CI(i)\} - \inf\{CI(i)\}}{r(i)}\right)} \quad (13)$$

where $0 < I_1 \leq 1, \forall i \in [1, \langle EoL \rangle], i \in I^+$

5.3.7 Incorporating Uncertainty Estimates

Prognostics algorithms typically involve estimating the probability distribution function (pdf) of the EoL and RUL rather than single point predictions which enables them to handle uncertainties arising from various sources such as noise, loading conditions and so on. They also allow for propagation of uncertainties for subsequent predictions (Orchard & Vachtsevanos, 2009). Thus, it is necessary to ensure that prognostic performance metrics include these factors. The most common form of assessing a pdf output is through estimates of mean and variance of the distribution owing to their simplicity and easy interpretation (Goebel et al., 2008). However, in reality, these distributions are rarely smooth or symmetric and hence mean and variance are not robust enough to evaluate the performance. A combination of mean as the measure of location and Quartiles or Inter Quartile Range (IQR) as a measure of spread can provide better estimates of the distribution (Hoaglin, et al., 1983).

The metrics shown in the previous sections do not explicitly accommodate for uncertainty estimating capability of the prognostic algorithms. However, a fairly straightforward way to do so is to specify an allowable error bound for a given metric. This error bound could be asymmetric as shown in Figure 6. In case of prognostics, typically a wider error margin to the “left” of the prediction (that is, an early prediction) may be preferred because early predictions have lower cost and safety consequences than late ones.

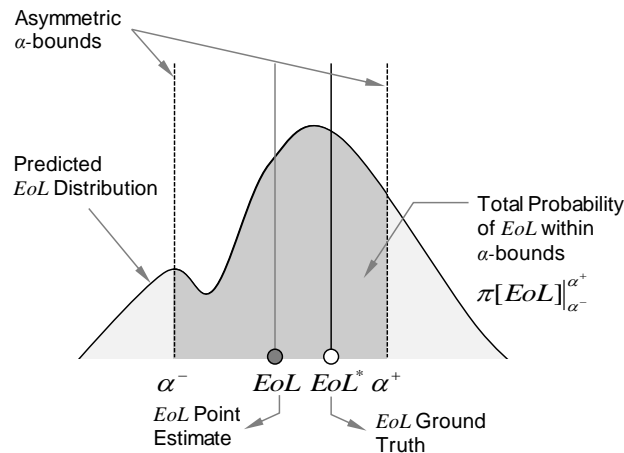


Figure 6 - Concepts for incorporating uncertainties (Saxena et al., 2010 CC 3.0)

These concepts can be analytically incorporated into the metrics by calculating the probability mass of a prediction falling within the specified α -bounds. As an illustrative example, consider again the α - λ accuracy. The α -bounds are expressed as a percentage of actual RUL $r(i_\lambda)$ at t_λ .

$$\alpha - \lambda \text{ Accuracy} = \begin{cases} 1 & \text{if } \pi[r(i_\lambda)]_{-\alpha}^{+\alpha} \geq \beta \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

where:

$r(i_\lambda)$ is the predicted RUL at time index i_λ

$\pi[r(i_\lambda)]_{-\alpha}^{+\alpha}$ is the probability mass of the prediction PDF within the α -bounds that are given by

$\alpha^+ = r_*(i_\lambda) + \alpha \cdot r(i_\lambda)$ and $\alpha^- = r_*(i_\lambda) - \alpha \cdot r(i_\lambda)$

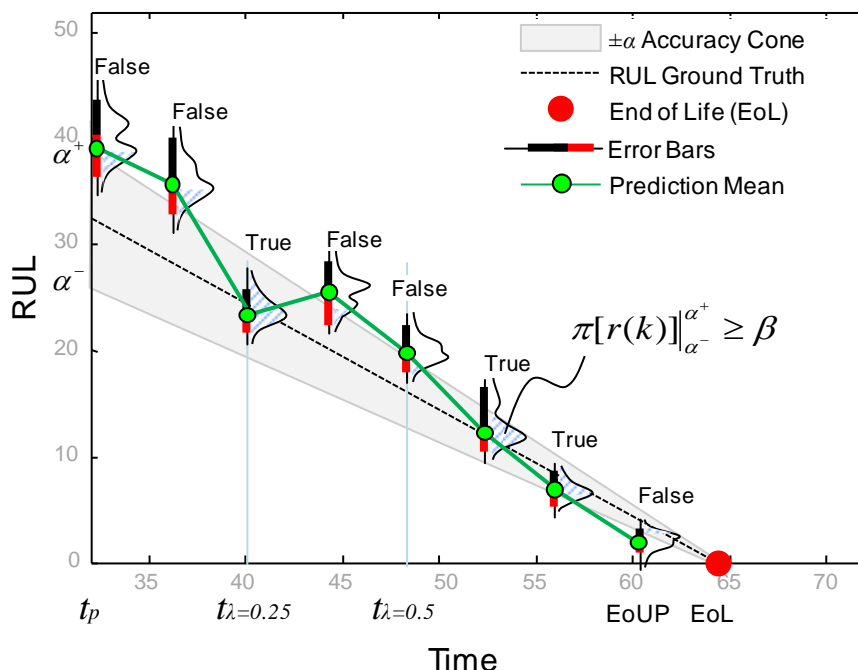


Figure 7 - α - λ accuracy with the accuracy cone shrinking with time on RUL vs. time plot
 With $\alpha = 0.1$ and $\lambda = 0.5$, the criteria for matching the metric requirement is determined by assessing the intersection of the uncertainty with the α -cone as shown in Figure 7. Note that there may be no prediction assessed at time t_λ for a given λ and the corresponding $i_\lambda \notin p$ because the set of time indexes (p) where a prediction is made is determined by the frequency of prediction step in a prognostic algorithm. In such cases one can make choose λ' closest to λ such that $i_{\lambda'} \in p$. To illustrate the application of α - λ accuracy further, refer to Figure 8 where the performance of a Recurrent Neural Network algorithm is plotted for every time instant when a prediction was made. The figure indicates at any point with either “True” or “False” (written above the upper quartiles) whether α - λ accuracy metric is satisfied or not based on β -criterion.

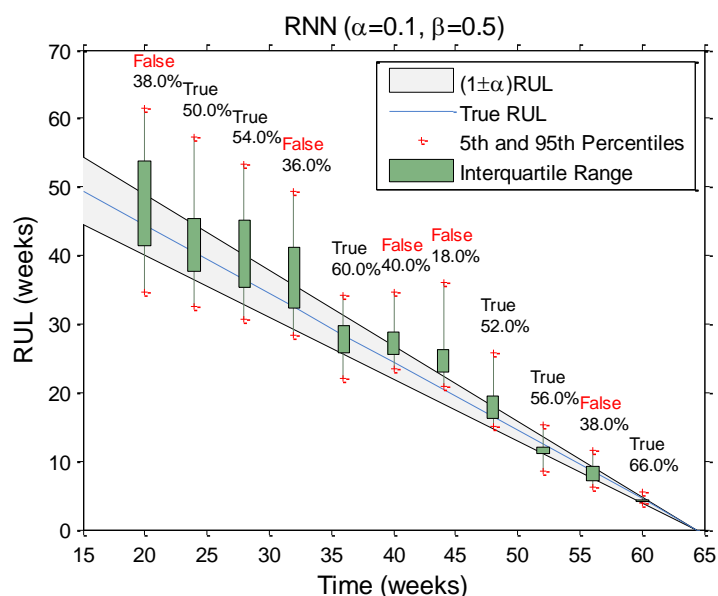


Figure 8 - α - λ accuracy for a particular algorithm with distribution information (Liu, et al., 2010, CC 3.0)

5.3.8 Guidelines for Applying Prognostics Metrics

Given the structure of the prognostics metrics described in this chapter, one can observe a progression in the manner how these metrics characterize the algorithm performance. The first metric, Prognostic Horizon, identifies whether an algorithm predicts within a specified error limits around the actual EoL and, if it does, how much time it allows for any corrective action. Therefore if an algorithm does not allow a sufficient PH it may not be meaningful to continue on computing other metrics. On the other hand if an algorithm passes the PH test, the next metric, α - λ Performance, identifies whether the algorithm performs within desired error margins of the actual RUL at any given time instant that may be of interest to in a particular application. This is a more stringent requirement of staying within a converging cone of the error margins as a system nears EoL. If this criterion is also met, the next step is to quantify the accuracy levels relative to the actual RUL. This is accomplished by the metrics Relative Accuracy and Cumulative Relative Accuracy. These metrics assume that prognostic performance improves as more information becomes available with time and hence, by design, an algorithm will satisfy these metrics criteria if it converges to true RULs. Therefore, the fourth metric, Convergence, quantifies how fast the algorithm converges if it does satisfy all previous metrics. These metrics can be considered as a hierarchical test that provides several levels of comparison among different algorithms in addition to the specific information these metrics individually provide regarding algorithm performance. Of course the use of other metrics such as robustness stands by itself to assess sensitivity of any of these or even other metrics with respect to a key system parameter.

5.3.8.1 Guidelines on choosing performance parameters

Time critical nature of prognostic application resulted in metrics for which the performance evolves with time and needs to be tracked. This required several special parameters that must be specified to define time criticality (λ), confidence level (β), or acceptable error bounds (α). The choice of α depends on the estimate of time required to take a corrective action. Depending on the situation this corrective action may correspond to performing maintenance (manufacturing plants) or bringing the system to a safe operating mode (operations in a combat zone). Adjustments to these parameters may translate into significant changes in the cost-benefit-risk equation in a process. Therefore, it is suggested that these parameters be chosen carefully to clearly specify prognostic requirements (Saxena et al., 2010). Requirements engineering is a discipline that provides guidelines to obtain these requirements in a systematic manner. For instance, in a safety critical military application first a Failure Modes Affects and Criticality Analysis (FMECA) or Hazard and Operability Analysis (HAZOP) must be conducted to identify most critical failures. Then based on available sensors, measurement quality, noise levels, etc. desired confidence levels must be derived. For safety critical systems a more conservative failure threshold may be chosen, while for commercial applications a less conservative but more cost effective threshold is preferred. It must be noted that the choice of metrics and performance specifications is an iterative process that negotiates between user requirements and constraints originating from performance needs, available resources, established maturity level of PHM, and time criticality for that application.

5.3.8.2 Guidelines for dealing with uncertainties

A Prognostic system models a stochastic process and hence the behavior observed from a particular run (single realization of the stochastic process) does not represent the complete behavior of the predicted trajectories. Assuming that all measures practically possible for

uncertainty reduction have been taken during the algorithm development phase, such observations should be treated only as isolated realization of the process. A level of confidence or probability of occurrence should be attached to such predictions. Otherwise, multiple trajectories should be aggregated from several runs to achieve statistical significance and more sophisticated stochastic analyses may be carried out. Another aspect dealing with uncertainties is related to prognostic algorithm output. Different algorithms represent uncertainties in different ways. Some specify parametric distribution and other as non parametric ones. Furthermore, some result in a closed form analytical equation for these distributions and other only result in discretized histograms. It is very important to carefully treat these distributions and not lose critical information by approximating these by known simpler forms such as normal distribution or by computing their statistical moments (Saxena et al., 2009; Saxena, et al., 2010). A common practice has been to compute mean and variance for all types of distributions whereas they may not be very meaningful for non-normal distributions. Use of more robust estimators like median, L-estimator, or M-estimator for expressing central tendency and Inter Quartile Range (IQR), Mean Absolute Deviation (MAD), or Median Absolute Deviation (MdAD) for expressing the spread is suggested (Devore, 2004).

5.3.8.3 Guidelines to resolve ambiguities

In practice, there can be several situations where the definitions discussed above result in ambiguity. Some of such situations are very briefly discussed here with suggested resolutions.

While applying the PH metric, a common situation encountered is when the RUL trajectory jumps out of the $\pm\alpha$ accuracy bounds temporarily. Situations like this result in multiple time indexes where RUL trajectory enters the accuracy zone to satisfy the metric criteria. A simple and conservative approach to deal with this situation is to declare a PH at the latest time instant the predictions enter accuracy zone and never comes out thereafter. Another option is to use the original PH definition and further evaluate other metrics to determine whether the algorithm satisfies all other requirements. As discussed in (Saxena et al., 2009) situations like these can occur due to a variety of reasons.

Inadequate system model: Real systems often exhibit inherent transients at different stages during their life cycles. These transients get reflected as deviations in computed RUL estimates from the true value if the underlying model assumed for the system does not account for these behaviors. In such cases, one must step back and refine the respective models to incorporate such dynamics.

Operational transients: Another source of such behaviors can be due to sudden changes in operational profiles under which a system is operating. Prognostic algorithms may show a time lag in adapting to such changes and hence resulting in temporary deviation from the real values. Therefore, whenever inconsistent behavior of PH metric is observed, one must identify the root cause of it and accordingly interpret the results. The situations discussed here are more common typically towards the end when a system nears EoL. This is because in most cases the fault evolution dynamics are too fast and complex to model or learn from data as the system nears EoL. Therefore, RUL curve deviates from the error band near t_{EoL} . To determine whether such deviations are critical for post-prognostic decision making, the concept of t_{EoUP} or End-of-Useful-Predictions (EoUP) is introduced. This index represents the minimum allowable PH that is required to take a corrective measure. Any predictions made beyond EoUP are of little or no use from a practical viewpoint.

5.4 Summary

This chapter presents several performance metrics for offline evaluation of prognostics algorithms. A brief overview of different methods employed for performance evaluation is also included. Because metrics developed in the context of forecasting differ from prognostics in the systems health management context and to account for the additional considerations, metrics specialized for prognostics (but not necessarily for the application) are needed. These metrics were introduced and their use was illustrated with recommendations.

Acknowledgments

The authors would like to express their gratitude to colleagues at the Prognostic Center of Excellence (NASA Ames Research Center) and external partners at Impact Technologies, Georgia Tech, and Clarkson University for participating in research discussions, evaluating metrics in their respective applications, and providing a valuable feedback. This work was funded by NASA Aviation Safety Program-IVHM Project.

References

- Byington, C. S., M. J. Roemer and P. W. Kalgren. 2005. *Verification and Validation of Diagnostic/Prognostic Algorithms*. Machinery Failure Prevention Technology Conference (MFPT 59), Virginia Beach, VA.
- Coble, J. B. and J. W. Hines. 2008. *Prognostic Algorithm Categorization with PHM Challenge Application*. 1st International Conference on Prognostics and Health Management (PHM08), Denver, CO.
- Coppe, A., R. T. Haftka, N. Kim, N. and F. Yuan. 2009. *Reducing Uncertainty in Damage Growth Properties by Structural Health Monitoring*. Annual Conference of the Prognostics and Health Management Society (PHM09) San Diego, CA.
- DeNeufville, R. 2004. *Uncertainty Management for Engineering Systems Planning and Design*. Engineering Systems Symposium MIT, Cambridge, MA.
- Devore, J. L. 2004. *Probability and Statistics for Engineering and the Sciences* (6th ed.): Thomson.
- Dzakowic, J. E. and G. S. Valentine. 2006. *Advanced Techniques For The Verification And Validation Of Prognostics & Health Management Capabilities*. Machinery Failure Prevention Technologies Conference (MFPT 60), Virginia Beach, VA.
- Ebert, B. 2007. Forecast Verification - Issues, Methods and FAQ, 2007, from http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html.
- Engel, S. J. 2008. *Prognosis Requirements and V&V: Panel Discussion on PHM Capabilities: Verification, Validation, and Certification Issues*. International Conference on Prognostics and Health Management (PHM08), Denver, CO.
- Engel, S. J., B. Gilmartin, K. Bongort and A. Hess. 2000. *Prognostics, the Real Issues Involved with Predicting Life Remaining*. IEEE Aerospace Conference, Big Sky, MT.
- Goebel, K., and P. P. Bonissone. 2005. *Prognostic Information Fusion for Constant Load Systems*. 7th Annual Conference on Information Fusion.
- Goebel, K., B. Saha, A. Saxena, J. Celaya and J. P. Christopherson. 2008. Prognostics in Battery

- Health Management. *IEEE Instrumentation and Measurement Magazine*, 11, 33 - 40.
- Guan, X., R. Jha, Y. Liu, A. Saxena, J. Celaya and K. Goebel. 2010. Comparison of Two Probabilistic Fatigue Damage Assessment Approaches Using Prognostic Performance Metrics. *International Journal of Prognostics and Health Management*, 1-13.
- Hastings, D. and H. McManus. 2004. *A Framework for Understanding Uncertainty and its Mitigation and Exploitation in Complex Systems*. Engineering Systems Symposium MIT, Cambridge MA.
- Hoaglin, D. C., F. Mosteller and J. W. Tukey (Eds.). 1983. *Understanding Robust and Exploratory Data Analysis*: John Wiley & Sons.
- Hyndman, R. J. and A. B. Koehler. 2006. Another look at measures of forecast accuracy, *International Journal of Forecasting*, Elsevier, vol. 22(4), 679-688.
- IEEE-Standard1413. 2002. IEEE Guide for Selecting and Using Reliability Predictions Based on IEEE 1413™ 1413.1TM (pp. 97): IEEE Standards Coordinating Committee 37.
- Condition Monitoring and Diagnostics of Machines - Prognostics part 1: General Guidelines, ISO/IEC Directives Part 2 C.F.R. 2004.
- Kacprzynski, G. J., M. Gumina, M. J. Roemer, D. E. Caguiat, T. R. Galie and J. J. McGroarty. 2001. *A Prognostic Modeling Approach for Predicting Recurring Maintenance for Shipboard Propulsion System*. 55th Meeting of the Society for Machinery Failure Prevention Technology, Virginia Beach, VA.
- Liu, J., A. Saxena, K. Goebel, B. Saha and W. Wang, 2010. *An Adaptive Recurrent Neural Network for Remaining Useful Life Prediction of Lithium-ion Batteries*. 2nd Annual Conference of the Prognostics and Health Management Society (PHM10), Portland OR.
- Luna, J. J. 2009. *Metrics, models and scenarios for evaluating PHM effects on logistics support*. Annual Conference of the Prognostics and Health Management Society, San Diego, CA.
- Ng, K.-C. and B. Abramson. 1990. Uncertainty Management in Expert Systems. *IEEE Expert Systems*, 5, 20.
- Orchard, M., G. Kacprzynski, K. Goebel, B. Saha and G. Vachtsevanos, G. 2008. *Advances in Uncertainty Representation and Management for Particle Filtering Applied to Prognostics*. International Conference on Prognostics and Health Management (PHM08), Denver, CO.
- Orchard, M. E., L. Tang, K. Goebel and G. Vachtsevanos, G. 2009. *A Novel RSPF Approach to Prediction of High-Risk, Low-Probability Failure Events*. Annual Conference of the Prognostics and Health Management Society (PHM09), San Diego, CA.
- Orchard, M. E. and G. Vachtsevanos. 2009. A Particle-Filtering Approach for On-line Fault Diagnosis and Failure Prognosis. *Transactions of the Institute of Measurement and Control*, 31(3-4), 221-246.
- Palmer, T. N., A. Alessandri, U. Andersen et al. 2004. Development of A European Multimodel System for Seasonal-to-Interannual Prediction (Demeter)," *Bulletin of the American Meteorological Society*, vol. 85, pp. 853-872, June 01, 2004.
- Sandborn, P. 2005. *A Decision Support Model for Determining the Applicability of Prognostic Health Management (PHM) Approaches to Electronic Systems*. Reliability and Maintainability Symposium (RAMS), Arlington, VA.
- Sankararaman, S., Y. Ling, C. Shantz, S. Mahadevan. 2009. *Uncertainty Quantification in Fatigue Damage Prognosis*. Annual Conference of the Prognostics and Health Management Society (PHM09), San Diego, CA.
- Saxena, A., J. Celaya, E. Balaban, K. Goebel, B. Saha and S. Saha. 2008. *Metrics for Evaluating*

- Performance of Prognostics Techniques*. 1st International Conference on Prognostics and Health Management (PHM08), Denver, CO.
- Saxena, A., J. Celaya, B. Saha, S. Saha and K. Goebel. 2009. *Evaluating Algorithmic Performance Metrics Tailored for Prognostics*. IEEE Aerospace Conference, Big Sky, MT.
- Saxena, A., J. Celaya, B. Saha, S. Saha and K. Goebel. 2009. *On Applying the Prognostics Performance Metrics*. Annual Conference of the Prognostics and Health Management Society (PHM09) San Diego, CA.
- Saxena, A., J. Celaya, B. Saha, S. Saha and K. Goebel. 2010. Metrics for Offline Evaluation of Prognostic Performance. *International Journal of Prognostics and Health Management*, 1(1), 21.
- Saxena, A., J. Celaya, S. Saha, B. Saha, I. Roychoudhury and K. Goebel. 2010. *Requirements Specification for Prognostics Performance – An Overview*. AIAA Infotech @ Aerospace, Atlanta GA.
- Schwabacher, M. 2005. *A Survey of Data Driven Prognostics*. AIAA Infotech@Aerospace Conference, Arlington, VA.
- Schwabacher, M. and K. Goebel. 2007. *A Survey of Artificial Intelligence for Prognostics*. AAAI Fall Symposium, Arlington, VA.
- Tang, L., G. Kacprzyński, K. Goebel and G. Vachtsevanos. 2009. *Methodologies for Uncertainty Management in Prognostics*. IEEE Aerospace Conference, Big Sky, MT.
- Teal, C. and B. Larsen. 2003. *Technology Update II: Wire Systems Diagnostics & Prognostics*. Digital Avionics Systems Conference.
- Uckun, S., K. Goebel and P. J. F. Lucas. 2008. *Standardizing Research Methods for Prognostics*. International Conference on Prognostics and Health Management (PHM08), Denver, CO.
- Vachtsevanos, G., F. L. Lewis, M. Roemer, A. Hess and B. Wu. 2006. *Intelligent Fault Diagnosis and Prognosis for Engineering Systems* (1st ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.
- Wheeler, K. R., T. Kurtoglu and S. Poll. 2009. *A Survey of Health Management User Objectives Related to Diagnostic and Prognostic Metrics*. ASME 2009 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (IDETC/CIE), San Diego, CA.
- Wilkinson, C., D. Humphrey, B. Vermeire and J. Houston, J. 2004. *Prognostic and Health Management for Avionics*. IEEE Aerospace Conference, Big Sky, MT.
- Wood, A. P. 2001. *Reliability-Metric Varieties and their Relationships*. Reliability and Maintainability Symposium, 2001, Philadelphia, PA.