Metrics for evaluating performance of prognostic techniques

Article · October 2008

DOI: 10.1109/PHM.2008.4711436

Abhinav Saxena
GE Global Research
116 PUBLICATIONS **3,412** CITATIONS
SEE PROFILE

Jose Celaya
NASA
86 PUBLICATIONS **2,354** CITATIONS
SEE PROFILE

Kai Goebel
Palo Alto Research Center
409 PUBLICATIONS **8,286** CITATIONS
SEE PROFILE

Bhaskar Saha
NIO USA
65 PUBLICATIONS **3,242** CITATIONS
SEE PROFILE

Metrics for Evaluating Performance of Prognostic Techniques

Abhinav Saxena, *Member, IEEE,* Jose Celaya, *Member, IEEE*, Edward Balaban, *Member, IEEE*,
Kai Goebel, Bhaskar Saha, *Member, IEEE,* Sankalita Saha, *Member, IEEE*, and Mark Schwabacher

***Abstract*—Prognostics is an emerging concept in condition based maintenance (CBM) of critical systems. Along with developing the fundamentals of being able to confidently predict Remaining Useful Life (RUL), the technology calls for fielded applications as it inches towards maturation. This requires a stringent performance evaluation so that the significance of the concept can be fully exploited. Currently, prognostics concepts lack standard definitions and suffer from ambiguous and inconsistent interpretations. This lack of standards is in part due to the varied end-user requirements for different applications, time scales, available information, domain dynamics, etc. to name a few issues. Instead, the research community has used a variety of metrics based largely on convenience with respect to their respective requirements. Very little attention has been focused on establishing a common ground to compare different efforts. This paper surveys the metrics that are already used for prognostics in a variety of domains including medicine, nuclear, automotive, aerospace, and electronics. It also considers other domains that involve prediction-related tasks, such as weather and finance. Differences and similarities between these domains and health maintenance have been analyzed to help understand what performance evaluation methods may or may not be borrowed. Further, these metrics have been categorized in several ways that may be useful in deciding upon a suitable subset for a specific application. Some important prognostic concepts have been defined using a notational framework that enables interpretation of different metrics coherently. Last, but not the least, a list of metrics has been suggested to assess critical aspects of RUL predictions before they are fielded in real applications.**

***Index Terms*—Prognostics and Health Management, Performance Metrics**

CONTENTS

I. INTRODUCTION

Prognostics is emerging at the forefront of Condition Based Maintenance (CBM) of critical systems giving rise to the term Prognostic Health Management (PHM). We define prognostics to be the detection of a failure precursor followed by the prediction of remaining useful life (RUL). There are major challenges in building a successful prognostics system that can be deployed in field applications [1]. Research efforts are focusing on developing algorithms that can provide a RUL estimate, generate a confidence bound around the predictions, and be integrated with existing diagnostic systems. A key step in successful deployment of a PHM system is prognosis certification. Since prognostics is still considered relatively immature (as compared to diagnostics), more focus so far has been on developing prognostic methods rather than evaluating and comparing their performances. Tests are conducted based on specific requirements to declare the goodness of the algorithms but little or no effort is made to generalize the performance over a variety of other situations. Hence, there is no direct way of comparing different efforts if one needs to identify the most suitable algorithm among several. This calls for a set of general metrics that can be used in a standardized manner. Furthermore, different users of prognosis have different requirements; hence these metrics should be tailored for each end user (*customer based verification*) [2]. This poses a conflicting requirement to the idea of generalization of metrics. This confusion has prevailed for some time in the CBM/PHM community and there is a need to classify various metrics into categories catering to different requirements. We have attempted here to evaluate the verification process such that it can provide a structure for how to choose performance metrics for specific tasks and also compare an algorithm with other competing ones.

In this paper we provide a concise review of a variety of

Manuscript received May 18, 2008. This work was supported in part by the U.S. National Aeronautics and Space Administration (NASA) under the Integrated Vehicle Health Management (IVHM) program.

Abhinav Saxena, Jose Celaya, and Sankalita Saha are with Research Institute for Advanced Computer Science at NASA Ames Research Center, Moffett Field, CA 94035 USA (Phone: 650-604-3208; fax: 650-604-4036; e-mail: asaxena@riacs.edu).

Kai Goebel, Edward Balaban, and Mark Schwabacher are with NASA Ames Research Center, Moffett Field, CA 94035 USA.

Bhaskar Saha is with Mission Critical Technologies at NASA Ames Research Center, Moffett Field, CA 94035 USA.

domains that involve prediction tasks of some kind. All these domains have fielded prognostics or forecasting applications and have, therefore, implemented performance metrics that evaluate and compare one system with another. These metrics have been consolidated and categorized into several categories based on different criteria that will be useful to the CBM/PHM community. For the sake of consistency and clear description, a notational framework has been introduced and included along with basic prognostics-related terms and definitions. The various metrics collected have been briefly explained and discussed with respect to how they can be of use to PHM applications. Several suggestions have been made for possibly useful PHM tailored metrics that may be used to evaluate and compare different algorithms in a standardized manner. Finally, several ideas have been discussed that can lead to newer metrics to evaluate other important prognosis aspects.

II. Motivation

For end-of-life predictions of critical systems, it becomes imperative to establish a fair amount of faith in the prognostic systems before incorporating their predictions into the decision-making process. A maintainer needs to know how good the prognostic estimates are before he/she can optimize the maintenance schedule. Without any reasonable confidence bounds a prediction completely loses its significance. Confidence bounds are a function of uncertainty management capabilities of an algorithm whereas performance metrics provide a means to establish sanity of any claims regarding such confidence bounds. Therefore, these algorithms should be tested rigorously and evaluated on a variety of performance measures before they can be certified. Furthermore, metrics help establish design requirements that must be met. In the absence of standardized metrics it has been difficult to quantify acceptable performance limits and specify crisp and unambiguous requirements to the designers. Standardized metrics will provide a lexicon for a quantitative framework for requirements and specifications.

There are a number of other reasons that make performance evaluation important. In general three broad categories namely, *scientific*, *administrative*, and *economic*, have been identified that include most reasons to carry out performance evaluations [3]. Performance evaluation allows comparing different schemes numerically and provides an objective way to measure how changes in training, equipment or prognostics models (algorithms) affect the quality of predictions. This provides a deeper understanding from the research point of view and yields valuable feedback for further improvements. One can identify bottlenecks in the performance and guide research and development efforts in the required direction. As these methods are further refined, quantitatively measuring improvement in predictions generates scores that can be used to justify for research funding in areas where either PHM has not yet picked up or where better equipment and facilities are needed. These scores can also be translated into costs and benefits to calculate Return-on-Investment (ROI) type indexes to justify their fielded applications.

Performance evaluation is usually the foremost step once a new technique is developed. In many cases benchmark datasets or models are used to evaluate such techniques on a common ground so they can be fairly compared. Prognostic systems, in most cases, have neither of these options. Various research teams have shown how to evaluate their algorithms using a set of performance metrics; there have, however, been inconsistencies in the choice of such metrics. This makes it rather difficult to compare various algorithms even if they have been declared successful based on their respective evaluations. It is true that prognostic methods are application oriented and that it is difficult to develop a generic algorithm useful in every situation. Therefore, the evaluation methods may need to be different as well. Furthermore, the inconsistent use of terminology in different applications has led to confusion in even the most basic definitions. So far very little has been done to identify a common ground when it comes to testing and comparing different algorithms. In two surveys of methods for prognostics (one of data-driven methods and one of artificial-intelligence-based methods) [4, 5], it can be seen that there is a lack of standardized methodology for performance evaluation and in many cases performance evaluation is not even formally addressed. Even the ISO standard [6] for prognostics in condition monitoring and diagnostics of machines lacks a firm definition of such metrics. There must, however, be a way to establish a common ground that can give a fair idea of how an algorithm fares when compared to others. Therefore, in this paper we have attempted to review the various domains where prognostics type applications exist and have matured to a point of being fielded. We also review the state-of-the-art in PHM technology and try to structure the verification methods in a logical fashion. Finally, we suggest several new metrics that can be of use to the prognostic community and present some ideas that, we hope, will serve as starting points for future discussions.

III. Prognostics Terms and Definitions

In this section we describe some commonly used terms in prognostics. Similar terms have been used interchangeably by different researchers and in some cases the same term has been used to represent different notions. This list is provided to reduce ambiguities that may arise by such non-standardized use.

A. Assumptions

- Here prognostics is considered to be the detection of failure precursors and the prediction of RUL based on the current state assessment and expected future operational conditions of the system.
- It is possible to estimate a health index as an aggregate of features and conditions

- RUL estimation is a prediction/ forecasting/ extrapolation process.
- Algorithms under consideration are capable of generating a single RUL value for each prediction. That is, algorithms that produce RUL distributions can be adapted to compress the distribution to a single estimated number for comparison purposes.
- All systems are under continuous monitoring and have the measurement capability that can acquire data as a fault evolves.

B. Glossary

RUL: Remaining Useful Life – amount of time left before system health falls below a defined failure threshold

UUT: Unit Under Test

i : Index for time instant t_i

EOL: End-of-Life - Time index of actual end of life

EOP: End-of-Prediction – earliest time index, i, when prediction has crossed the failure threshold

0: Time index for time of the birth of the system, t_0

F: Time index for the time when fault occurs, t_F

D: Time index at which the fault is detected by diagnostic system, t_D

P: Time index at which the first prediction is made by the prognostic system, t_P

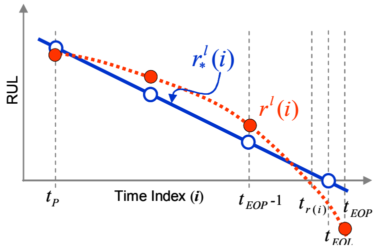


Figure 1. Illustration depicting some important prognostic time definitions and prediction concepts.

$f_n^l(i)$: Value of the n^{th} Feature for the l^{th} UUT at time index i

$c_n^l(i)$: Value of the n^{th} operational condition for the l^{th} UUT at time index i

$r^l(i)$: RUL Estimation at time t_i given that data is available up to time t_i for the l^{th} UUT

$\pi^l(i\,|\,j)$: Prediction at time index i given data up to time t_j for the l^{th} UUT. Prediction may be made in any domain, e.g. feature, health, etc.

$\Pi^l(i)$: Trajectory of predictions at time index i for the l^{th} UUT

$h^l(i)$: Health of system for the l^{th} UUT

Definition 1 - ***Time Index:*** The time in a prognostics application can be discrete or continuous. We will use a time index i instead of the actual time, e.g., i=10 means t_{10}. This takes care of cases where sampling time is not uniform. Furthermore, time indexes are invariant to time-scales.

Definition 2 - ***Time of Detection of Fault:*** Let D be the time index (t_D) at which the diagnostic or fault detection algorithm detected the fault. This process will trigger the prognostics algorithm which should start making RUL predictions shortly after the fault was detected as soon as enough data has been collected. For some applications, there may not be an explicit declaration of fault detection, e.g., applications like battery health management, where prognosis is carried out on a decay process. For such applications t_D can be considered equal to t_0 (time of birth) i.e., we expect to trigger prognosis as soon as enough data has been collected and not wait for an explicit diagnostic flag (Figure 2).

Definition 3 - ***Time to Start Prediction:*** We will differentiate between the time when a fault is detected (t_D) and the time when the system starts predicting (t_P). For certain algorithms $t_D = t_P$ but in general $t_P \geq t_D$ as these algorithms need some time to tune with additional fault progression data before they can start making predictions (Figure 2). In cases where a data collection system is continuously collecting data even before fault detection, enough data is already available to start making predictions right away and hence $t_P = t_D$.

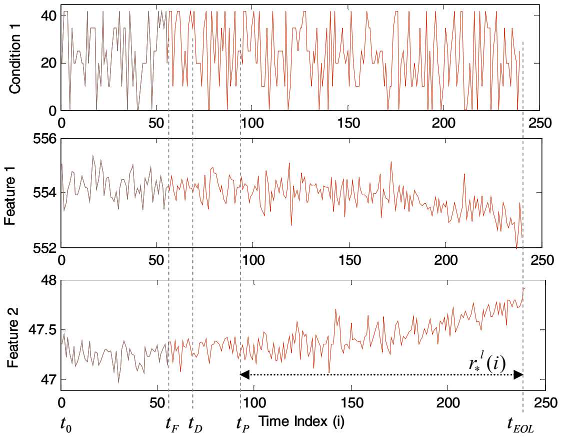


Figure 2. Features and conditions for l^{th} UUT

Definition 4 - ***Prognostics Features:*** Let $f_n^l(i)$ be a feature at time index i, where n = 1, 2, … , N is the feature number, and l = 1, 2, … , L is the UUT index (an index identifying the different units under test). In prognostics, irrespective of the analysis domain, i.e., time, frequency, wavelet, etc., features take the form of time series and they can be physical variables, system parameters or any other quantity that can be computed from measurable variables of the system that provides or aides the prognosis. The features can be also referred to as a feature vector $F^l(i)$ of the l^{th} UUT at time index i.

Definition 5 - ***Operational Conditions:*** Let $c_m^l(i)$ be an operational condition at time index i, where m = 1, 2, … , M is the condition number, and l = 1, 2, … , L is the UUT index. The operational conditions describe how the system is being operated and are sometimes referred to as the load on the system. The conditions can also be referred to as a vector $C^l(i)$ of the l^{th} UUT at time index i.

Definition 6 - ***Health Index:*** Let $h^l(i)$ be a health index at time index i for UUT l = 1, 2, … , L. h can be considered a normalized aggregate of health indicators (relevant features) and operational conditions.

Definition 7 - ***Ground Truth:*** Ground truth, denoted by the subscript $_*$, represents our best belief of the true value of a system variable. In the feature domain $f^l_{*n}(i)$ may be directly or indirectly calculated from measurements. In the health domain, $h^l_*(i)$ is the computed health at time index i for UUT l = 1, 2, … , L after a run to failure test. This health index represents an aggregate of information provided by features and operational conditions up to time index i.

Definition 8 – ***History Data:*** History data, denoted by the subscript $_\#$, encapsulates all the information we know about a system *a priori*. Such information may be of the form of archived measurements or EOL distributions, and can refer to variables in both the feature and health domains represented by $f^l_{\#n}(i)$ and $h^l_\#(i)$ respectively.

Definition 9 - ***Point Prediction:*** Let $\pi^l(i\,|\,j)$ be a point prediction of a variable of interest at time index i given information up to time t_j, where $t_j \le t_i$. $\pi^l(i\,|\,j)$ for $i = EOL$ represents the critical threshold for a given health indicator. Predictions can be made in any domain, features or health. In some cases it is useful to extrapolate features and then aggregate them to compute health and in other cases features are aggregated to a health and then extrapolated to estimate RUL.

Definition 10 - ***Trajectory Prediction:*** Let $\Pi^l(i)$ be the trajectory of predictions at time index i such that $\underline{\Pi}^l(i)$ = $\left\{\pi^l(i\,|\,i), \pi^l(i+1\,|\,i), \ldots, \pi^l(EOL\,|\,i)\right\}$

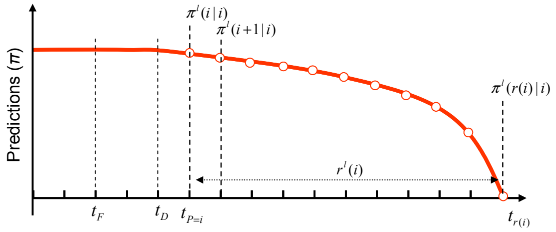


Figure 3. Illustration showing a trajectory prediction. Predictions may modify every time instant and hence the corresponding RUL estimate.

Definition 11 - ***RUL Estimation:*** Let $r^l(i)$ be the remaining useful life estimation at time index i given that the information (features and conditions) up to time index i and an expected operational profile for the future are available. As shown in Figure 4, prediction is made at time t_i and it predicts the RUL given information up to time i for the l^{th} UUT RUL will be estimated as $r^l(i) = \underset{z}{\arg}\left\{h(z)=0\right\} - i$.

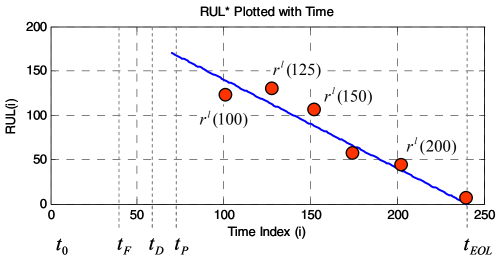


Figure 4. Comparing RUL predictions from ground truth (t_p [70,240], t_{EOL} = 240, t_{EOP} > 240).

IV. Forecasting Applications Classification

Based on our survey of several forecasting application domains, we identified two major classes of forecasting applications (see Figure 5).

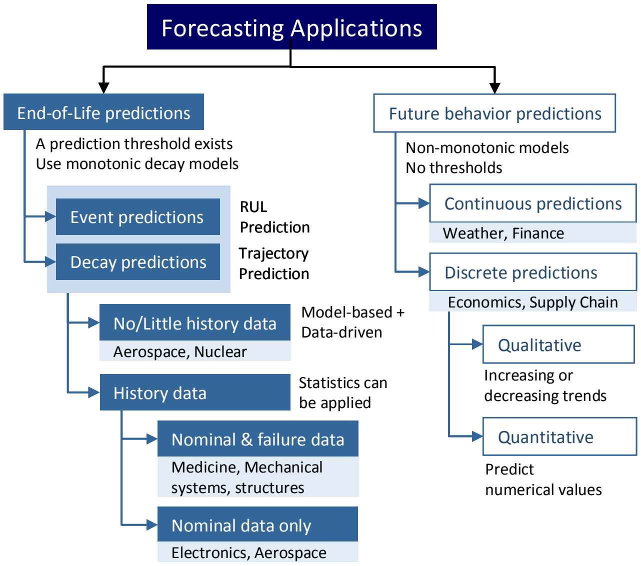


Figure 5. Different categories of the forecasting applications

In one class of applications a prediction is made on a continuous basis, and the trend of data is generally non-monotonic. These preditions 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). Predictions can be quantitative (e.g. prediction of exact numbers) or qualitative (e.g. high or low demands) in nature. Applications like weather and finance have been in existence for quite a while and have matured to a good extent. The other class of applications involves the existence of a critical threshold such that the system under test is declared to have lost a defined degree of functional capability (including complete failure) if it crosses the threshold. These applications usually can be modeled using decay models. Here the task of prognostics is to predict a RUL estimate. In some cases, where enough history data exists (e.g. medicine) or can be experimentally generated (e.g. mechanical systems) for nominal and failure conditions, a variety of data-driven or statistical techniques can be applied. In such situations it is also relatively easy to evaluate the performance by comparing the prediction *a* *posteriori*. However, there are critical applications where run-to-failure experiments can not be afforded and very little failure history data is available (e.g. aerospace). In such cases a variety of methods based on data-driven and model-based techniques have been proposed. It becomes extremely tricky and difficult to assess the performance in such cases due to absence of knowledge about the future outcomes. Methods are tested on experimental or simulated data and are expected to perform on real systems. Unfortunately algorithm performance does not always translate meaningfully from one dataset to another or one domain to another. Therefore, a standard set of metrics independent of application domain would be very desirable.

V. Forecasting Domains Reviewed

In this section we provide a concise assessment of prediction performance evaluation methods in various domains. Specific relevant performance metrics have been listed in the next section.

A. Aerospace

The aerospace industry is likely the field with the most vibrant research and development activity in systems prognostics today. This happened for a good reason – systems health inspections on spacecraft and aircraft are often difficult and costly, and sometimes impossible. The consequences of a premature failure can, however, be dire.

Prognostic algorithms are beginning to be applied to monitoring condition of aircraft structures, avionics, wiring, control actuators, power supplies, and propulsion systems. Prognostic functionality is being incorporated into the health management system of the latest military aircraft (e.g. Joint Strike Fighter [7]) and civilian aircrafts, in order to reduce the overall lifecycle cost and improve flight readiness. Original equipment manufacturers as well as, increasingly, small businesses have established dedicated prognostics research groups. Active work on aerospace prognostics is also being conducted by national governments (including research labs in the armed forces and the aerospace agencies) as well as academic organizations both in the US and elsewhere.

The aerospace industry has also led in developing the metrics to evaluate prognostic algorithms. Most of the metrics have, historically, focused on the technical merits of prognostic techniques, such as accuracy and precision [1], although in recent years more attention have been given to those assessing the business merits (ROI[8-10], Total Value [11], and others). As the prognostic systems make their way into the commercial aerospace sector, they are expected to help with maintenance scheduling, optimal operating mode determination, and asset purchasing decisions.

B. Electronics

Prognostics for electronics is currently less advanced than prognostics for mechanical systems. Many researchers in the field therefore take their inspiration from previous work in mechanical prognostics, and use similar algorithms and metrics, including the usual accuracy metrics [12-14]. Some of the work in electronics prognostics emphasizes the potential cost savings provided by prognostics, and therefore relies on cost/benefit metrics such as ROI [8-10], life cycle cost [15], and MTBF/MTBUR ratio [16]. Methods used for data collection include measuring the temperatures of components [17, 18], installing "canaries" (electronic devices that are designed to fail before the operational devices do) [19], collecting data about operational conditions such as vibration [13], usage hours [17], or ambient temperature, using strain gauges to measure the strain on solder joints [19], and detecting when the performance of a system degrades (for example, when more correctable errors begin to occur) [18].

C. Medicine

Medicine is a field where diagnostics and prognostics have a long tradition. Indeed, medicine has a large body of tests and indicators that are used commonly to aid in decision-making such as blood pressure and cholesterol levels. The field has come to trust these prognostic indicators when they have been subject to the double blind clinical trial. While this test is not perfect, it provides a metric against which other results can be compared. Although prognostics is a common tool in medicine, the most significant constraint is the way in which prognostic results are measured. Typically, survival rates are quantized into increments such that the problem boils down to a classification problem [20, 21]. For example, one would typically measure the number of cancer survivors past, say, 10 years, and then assess whether the prediction was correct or not. Despite that constraint, there are a number of ancillary metrics (e.g., *coverage*, *informativity* [22], *discrimination* and *calibration* [21], *accuracy*, *precision*, *interseparability* and *resemblance* [23], etc.) that have been in use which quantify the quality of a prediction in the context of a regression problem. In addition, important insights are given into how to deal with incomplete data [24], and area that is jointly of interest to other fields like statistics as well.

D. Nuclear

With increasing energy demands nuclear power plants play an important role in the energy sector. Average life of a nuclear reactor being 20-30 years, efforts are underway to extend the life of these reactors using advanced monitoring and maintenance techniques. While advanced diagnostics has been implemented in the US and Europe, prognostics is still at conceptual levels. Most metrics developed so far have been to establish a profitable business case rather than maturing prognostics itself [25, 26]. Data records like overall plant operating efficiency and maintenance, machinery repair records, etc. are used to derive cost-benefit analysis for prognosis. For instance, improved thermal efficiency is translated into gas cost savings and increase in available capacity translated into savings from not using the spare unit, etc. However, the lack of prognostics deployment has resulted in very little research in improving the prognosis itself and hence not many verification schemes.

E. *Finance*

Forecasting techniques are used in finance and economics. These are usually based on statistical methods based on regression or time series analysis techniques. Performance evaluation methods have developed and studied for prediction algorithms within the context of the forecasting research [27]. These methods are focused on *prediction accuracy* and *model/algorithm* selection [28]. The *prediction accuracy* is approached by computing statistics over the prediction error. Such statistics infer parameters like central tendency and variability by either assuming a particular form of the probability distribution of the error or by not making any assumptions about the error distribution using methods like the *median* and MAD (median absolute deviation) [27-30]. On the other hand, in *forecasting model selection* the intention is to select forecasting model/algorithm that performs statistically better than a baseline algorithm. Model selection techniques are also heavily based on prediction errors [28].

A few other well-known model selection techniques in this domain include the Diebold-Mariano test based on average out-of sample error [31], Pearson's Chi-square test on contingency tables, Theil's U statistics [32] and the time-distance criterion which measures the models horizontally as against the usual vertical methods such as mean and mean squared error (MSE). Another set of statistics which is quite relevant in this area is related to the direction/sign of prediction, i.e., whether there will be an increase or decrease in the forecasting variable. This includes confusion rate which is the number of falsely predicted changes divided by the number of observed changes and the Henriksson and Merton test [33] which compares models based on accuracy of direction/sign prediction.

F. *Weather*

Forecasting weather patterns has probably been one of man's earliest attempts at modeling and prediction, and continues to be just as significant today as it was before. Various modeling and forecasting methodologies have evolved from the study of weather as well as a variety of metrics to compare these techniques [34]. However, the essence of the widely used metrics can be grouped in two categories: those that measure bias or error with respect to a baseline [35], and those that measure resolution or the ability of the forecast to distinguish between different outcomes [36]. The baselines to be used as a basis of comparison can also vary between aggregate weather history (over the last 10 years, for example), current measurements or even reference forecasts. This kind of approach is well suited to a field where measurements have improved in accuracy but our understanding of weather patterns is still evolving.

Another related application that drew our attention is wind mill power prediction where the task of prediction matches with prognosis in that it uses different time scales depending on specific applications [37]. For instance, scheduling of power plants based on 3-10 hrs prediction horizon, (2) assessing the value of produced electricity for various end users on a prediction horizon of 0-48 hrs, and (3) Longer time scales for maintenance planning. Performance is usually evaluated using a reference model, often referred to as *persistence* models. Error based metrics like Bias, MAE, RMSE, MSE, SDE, Coefficient of determination (R2), etc. are the most common ones here as well [38]. Also mentioned is a metric called *Surplus* for a given period, which is the sum of all positive prediction errors. All errors are determined for *k*-0 step look ahead. Another metric used is cumulated squared prediction errors for *k*-step ahead prediction (*k* is small for short term and large for long term predictions).

G. *Automotive*

Prognostics has recently become a vital part of on-board diagnostics (OBD) of the latest vehicles. "The goal of this technology is to continually evaluate the diagnostics information over time in order to identify any significant potential degradation of vehicle subsystems that may cause a fault, to predict the remaining useful life of the particular component or subsystem and to alert the driver before such a fault occurs" [39]. Mostly, the approach consists of trending of residuals extracted from diagnostic information [40]. The metrics used are mainly accuracy measures like MSE [41] or Gaussian pdf overlaps [42]. The overall methodology is data-driven and suitable where extensive baseline data is available [43].

VI. Prognostics Metrics Classifications

A variety of prognostics metrics are used in the domains reviewed above. Depending on the end use of the prognostic information, basic accuracy and precision based metrics are transformed into more sophisticated measures. Several factors were identified that classify these metrics into different classes. In this section we attempt to enumerate some of these classifications.

A. *Functional Classification*

The most important classification is based on the information these metrics provide to fulfill specific functions. In general we identified three major categories, namely: (1) Algorithm performance metrics, (2) Computational performance metrics, and (3) Cost-benefit metrics. As evident from their names these metrics measure success based on entirely different criteria. As shown in Figure 6, the algorithmic performance metrics can be further classified into four major subcategories.

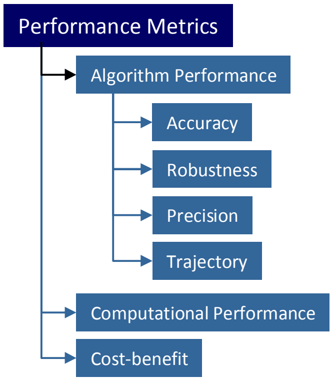


Figure 6. Functional classification of prognostics metrics.

B. *End User Based Classification*

Prognostics information may be used by different people for entirely different purposes. In general, end users of prognosis may be classified into the five categories shown in Table 1.

Table 1. Classification of prognostic metrics based on end user requirements

End User	Goals	Metrics
Program Manager	Assess the economic viability of prognosis technology for specific applications before it can be approved and funded	Cost-benefit type metrics that translate prognostics performance in terms of tangible and intangible cost savings
Plant Manager	Resource allocation and mission planning based on available prognostic information	Accuracy and precision based metrics that compute RUL estimates for specific UUTs. Such predictions are based on degradation or damage accumulation models.
Operator	Take appropriate action and carry out re-planning in the event of contingency during mission	Accuracy and precision based metrics that compute RUL estimates for specific UUTs. These predictions are based on fault growth models for critical failures.
Maintainer	Plan maintenance in advance to reduce UUT downtime and maximize availability	Accuracy and precision based metrics that compute RUL estimates based on damage accumulation models.
Designer	Implement the prognostic system within the constraints of user specifications. Improve performance by modifying design.	Reliability based metrics to evaluate a design and identify performance bottlenecks. Computational performance metrics to meet resource constraints.

VII. Prognostics Metrics

A. *Algorithmic Performance Metrics*

Most metrics found in our survey fall into the category of algorithmic performance evaluators. A concise list of such metrics has been included in Table 2. The table is further divided into subcategories as identified in Section VI.A. As one can see, accuracy and precision based metrics dominate the table. The notion of robustness has been talked about in several cases but formal definitions were not found. Similarly, trajectory prediction metrics were not explicitly defined. Some mention of metrics like *similarity measure* [1] and *prediction behavior error* [43] have been mentioned that may be adapted for trajectory prediction performance evaluation. A more detailed description of various metrics can be found in the table.

In general algorithmic performance can be measured by evaluating the *errors* between the predicted and the actual RULs. Other metrics use *error* to quantify various other characteristics such as statistical moments, robustness, convergence, etc. calculation of error requires availability of ground truth data, which is rarely available in many situations. In that case history data, if accessible, may be utilized to make corresponding inferences. Of course this assumes that the current process draws from the history data distribution.

B. *Classification Based on Predicted Entity*

Within PHM applications, we identified three major classes of the forms of prediction outputs and hence the corresponding metrics. Prognostics performance can be established based on different forms of the prediction outputs, e.g. future health index trajectory at t_P, an RUL estimate at t_P, or a RUL trajectory as it evolves with time. Some algorithms provide a distribution over predicted entities to establish confidence in predictions. Metrics to evaluate such outputs differ in form from those required for single value predictions. In other cases such a distribution is obtained from multiple UUTs, e.g., from fleet applications. The basic form of the metrics used for various categories may be similar, but the underlying information conveyed is usually different in a statistical sense. Figures 7-9 illustrate some representative examples.

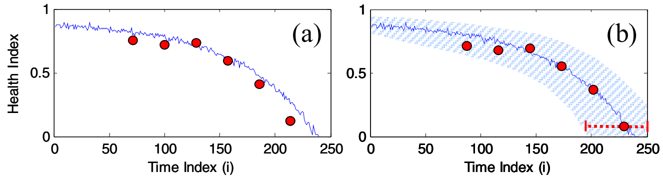


Figure 7. (a) Predictions are made in the health domain for a single UUT. A health trajectory is predicted to consider evolution of fault in the system. (b) Predictions can be in the form of distributions with associated confidence bounds.

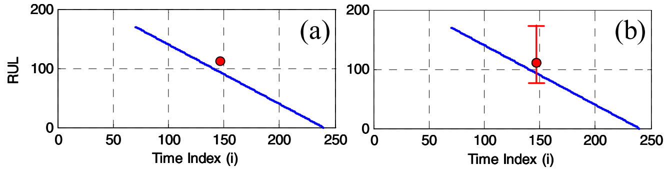


Figure 8. (a) Each prediction in the health domain appears as a point prediction in the RUL domain, which then may be compared with ground truth (b) RUL predictions may be obtained with corresponding confidence limits.

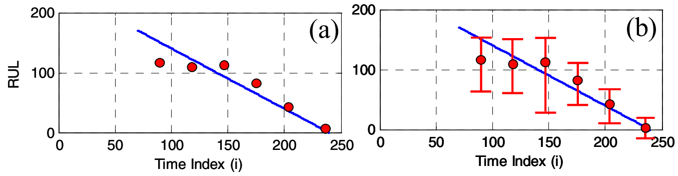
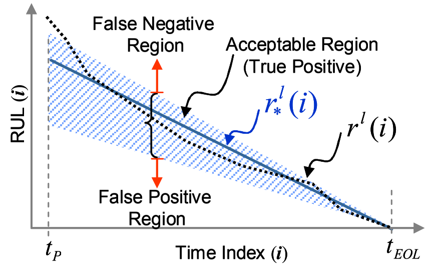


Figure 9. (a) A further assessment can be made on how well an algorithm's RUL estimate evolves over time and converges to the true value as more data becomes available. (b) Such RUL trajectories may be accompanied by corresponding error bars as well.

Table 2. List of metrics for algorithm performance evaluation[*]

Metric Name	Definition	Description	Range	Selected References
Accuracy Based Metrics				
Error	$\Delta^l(i) = r_*^l(i) - r^l(i)$	Error defines the basic notion of deviation from desired output. Most accuracy based metrics are derived directly or indirectly from error.	(-∞,∞) Perfect score = 0	
Average scale independent error	$A(i) = \frac{1}{L}\sum_{l=1}^{L}\exp\left\{-\frac{\lvert\Delta^l(i)\rvert}{D_0}\right\}$	Weighs exponentially the errors in RUL predictions and averages over several UUTs; where D_0 is a normalizing constant whose value depends on the magnitudes in the application.	(0,1] Perfect score =1	[1, 28]
Average bias	$B_l = \frac{\sum_{i=P}^{EOP}\left\{\Delta^l(i)\right\}}{(EOP - P + 1)}$	Averages the errors in predictions made at all subsequent times after prediction starts for the l^{th} UUT. This metric can be extended to average biases over all UUTs to establish overall bias.	(-∞,∞) Perfect score =0	[1]
Timeliness	$A(i) = \frac{1}{L}\sum_{l=1}^{L}\varphi\{\Delta^l(i)\}$ where, $\varphi(z) = \begin{cases}\exp\{\lvert z\rvert/a_1\}-1, & if \quad z<0\\ \exp\{\lvert z\rvert/a_2\}-1, & if \quad z\geq 0\end{cases}$, and $a_1 > a_2 > 0$	Exponentially weighs RUL prediction errors through an asymmetric weighting function. Penalizes the late predictions more than early predictions.	(0,∞) Perfect score =0	[1]
False Positives (FP)	$FP\left(r_*^l(i)\right) = \begin{cases}1 & \text{if} \quad \Delta^l(i) > t_{FP}\\ 0 & \text{otherwise}\end{cases}$ where t_{FP} = user defined acceptable early prediction	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.	[0,1] Perfect score = 0	[44]
False Negatives (FN)	$FN\left(r_*^l(i)\right) = \begin{cases}1 & \text{if} \quad -\Delta^l(i) > t_{FN}\\ 0 & \text{otherwise}\end{cases}$ where t_{FN} = user defined acceptable late prediction	 False Negative Region Acceptable Region (True Positive) $r_*^l(i)$ $r^l(i)$ RUL (i) False Positive Region t_P Time Index (i) t_{EOL}	[0,1] Perfect score = 0	[44]
Mean absolute percentage error (MAPE)	$MAPE(i) = \frac{1}{L}\sum_{l=1}^{L}\left\lvert\frac{100\Delta^l(i)}{r_*^l(i)}\right\rvert$	Averages the absolute percentage errors in the predictions of multiple UUTs at the same prediction horizon. Instead of the mean, median can be used to compute Median absolute percentage error (MdAPE) in a similar fashion.	[0,∞) Perfect score = 0	[28, 29]

[*] For the sake of conciseness, the references cited here are the ones that may be considered representative of the general field and can be grouped as a comprehensive source for these metrics.

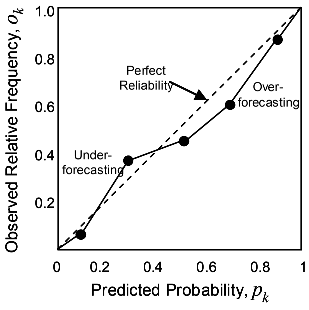
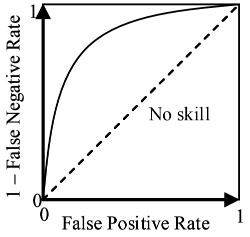
Anomaly correlation coefficient (ACC)	$ACC = \frac{\sum\left(\pi^l(i\mid j) - z_\#(i)\right)\left(z_*(i) - z_\#(i)\right)}{\sqrt{\sum\left(\pi^l(i\mid j) - z_\#(i)\right)^2 \sum\left(z_*(i) - z_\#(i)\right)^2}}$, where, $z_*(i)$ is a prediction variable (e.g. $f^l_{*n}(i)$ or $h^l_*(i)$), and $z_\#(i)$ is the corresponding history data value.	Measures correspondence or phase difference between prediction and observations, subtracting out the historical mean at each point. The anomaly correlation is frequently used to verify output from numerical weather prediction (NWP) models. 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.	[-1 1] Perfect score = 1	[34]
Symmetric mean absolute percentage error (sMAPE)	$sMAPE(i) = \frac{1}{L}\sum_{l=1}^{L}\left\lvert\frac{100\Delta^l(i)}{(r^l_*(i) + r^l(i))/2}\right\rvert$	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.	[0,∞) Perfect score = 0	[28, 30]
*Mean squared error (MSE)	$MSE(i) = \frac{1}{L}\sum_{l=1}^{L}\Delta^l(i)^2$	Averages the squared prediction error for multiple UUTs at the same prediction horizon. A derivative of MSE is Root Mean Squared Error (RMSE).	[0,∞) Perfect score = 0	[28]
Mean absolute error (MAE)	$MAE(i) = \frac{1}{L}\sum_{l=1}^{L}\left\lvert\Delta^l(i)\right\rvert$	Averages the absolute prediction error for multiple UUTs at the same prediction horizon. Using median instead of mean gives median absolute error (MdAE).	[0,∞) Perfect score = 0	[28]
Root mean squared percentage error (RMSPE)	$RMSPE(i) = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left\lvert\frac{100\Delta^l(i)}{r^l_*(i)}\right\rvert^2}$	Square root of the average of percentage error of the prediction from multiple UUTs. A similar metric is Root median squared percentage error (RMdSPE).	[0,∞) Perfect score = 0	[28]

[*] these metrics can also be classified into precision based category

Precision Based Metrics

Sample standard deviation (S)	$S(i) = \sqrt{\frac{\sum_{l=1}^{n}\left(\Delta^l(i) - M\right)^2}{n-1}}$ where M is the sample mean of the error	Sample standard deviation measures the dispersion/spread of the error with respect to 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.	[0,∞) Perfect score = 0	[1, 45]
Mean absolute deviation from the sample median (MAD)	$AD(i) = \frac{1}{n}\sum_{l=1}^{n}\left\lvert\Delta^l(i) - M\right\rvert$ where $M = \underset{l}{median}\ (\Delta^l(i))$ and median is the $\frac{n+1}{2}$th order statistic	This is a resistant estimator of the dispersion/spread of the prediction error. It is intended to be used where there is a small number of UUTs and when the error plots do not resemble those of a normal distribution.	[0,∞) Perfect score = 0	[45]
Median absolute deviation from the sample median (MdAD)	$MAD(i) = \underset{l}{median}\left(\left\lvert\Delta^l(i) - M\right\rvert\right)$ where $M = \underset{l}{median}\ (\Delta^l(i))$ and median is the $\frac{n+1}{2}$th order statistic	This is a resistant estimator of the dispersion/spread of the prediction error. It is intended to be used where there is a small number of UUTs and when the error plots do not resemble those of a normal distribution.	[0,∞) Perfect score = 0	[45]

Robustness Based Metrics

Reliability diagram, Brier Score	 Reliability Diagram The Brier Score computed as $BS = \frac{1}{K}\sum_{k=1}^{K}(p_k - o_k)^2$ is a measure of the deviation from the diagonal.	The reliability diagram plots the observed frequency against the predicted probability of a random event. In the context of prognostics, an event may be the RUL of a system lying within a given time interval, or a health feature crossing an alarm threshold within a predetermined time. The prediction of the value of RUL is not considered an event. In other words, the problem of prognostics is transformed into the classification domain. The occurrence of the event is predicted multiple times and the range of probabilities is divided into K bins like 0-5%, 5-15%, 15-25%, etc. Let us say that n_k times out of a total of N, the predicted probability falls in the probability bin k centered around p_k and out of those n_k times, the event occurs m_k times, then the corresponding observed relative frequency o_k is calculated as m_k/n_k. Reliability is indicated by the proximity of the plotted curve to the diagonal. The deviation from the diagonal gives the conditional bias. If the curve lies below the line, this indicates over-forecasting (probabilities too high); points above the line indicate under-forecasting (probabilities too low).	[0 1] Perfect score = 0	[34]
Receiver Operating Characteristic (ROC)	 The area under the ROC curve can be used as a score.	ROC gives a comprehensive overview of the tradeoff between false positives and false negatives, as defined in section VIII. The ideal curve would have zero false positives and zero false negatives. Such a curve cannot realistically be achieved for real-world problems. In addition, 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").	[0 1] Perfect score = 1	[34, 36]
Sensitivity	$S(i) = \frac{1}{L}\sum_{l=1}^{L}\left\{\frac{\Delta M^{l}(i)}{\Delta_{input}}\right\}$	Measures how sensitive a prognostic algorithm is to input changes or external disturbances. Can be assessed against any performance metric of interest. ΔM is the distance measure between two successive outputs for metric M's value and Δ_{input} is a distance measure between two successive inputs, e.g. failure threshold, noise level, available sensor set, sampling rate, etc.	[0, ∞) Perfect score = 0	[1]

C. *Computational Metrics*

Most of the publications in the area of prognostic algorithms for aerospace make no mention of computational performance. Some metrics like *complexity* [46] and *specificity* [47] have been mentioned that touch upon the computational aspects of prognostics. Many authors have been able to avoid the question of computational performance so far because they have not yet deployed their systems. We feel that assessing the computational performance of prognostic algorithms is very important, especially for applications that intend to monitor real-time data to make safety-critical decisions, such as deciding when it is necessary to shut down an engine or to land an aircraft to perform critical maintenance. In this section, we suggest several metrics that could be used to measure computational performance of prognostic algorithms, all of which are already widely used to measure the computational performance of other types of algorithms.

In theoretical computer science, the worst case computational complexity of algorithms is usually described using "Big O" notation [48]. This notation describes the amount of time needed for the algorithm to run, as a function of the size of the input, and does so asymptotically, ignoring constant factors. For example, if the time performance of an algorithm is O(n^2), then the time needed to run the algorithm increases quadratically with the size of the input. Big O notation allows the comparison of different algorithms to be independent from the particular software implementations and from the hardware on which the algorithms are run.

To measure the combined performance of an algorithm, its software implementation, and the hardware on which it is run, one can measure either central processing unit (CPU) time or elapsed time. CPU time measures the amount of time that the CPU spends executing the software, and does not include the time that the CPU spends running other software (in a time-shared system), or the time that the CPU spends waiting for input or output (I/O). The advantage of measuring CPU time instead of elapsed time is that it is more repeatable. Elapsed time (also known as "wall-clock time") simply measures the amount of time that it takes for an algorithm to run, including I/O time. It is not appropriate to use elapsed time as a metric on a time-shared (multi-user) system, since in that situation the activities of other users can affect the elapsed time. CPU time and elapsed time are both appropriate for applications in which the prognostic algorithm is run in "batch mode" on recorded data. They can answer the question of whether the software will run fast enough to produce results within a reasonable amount of time.

For applications in which the data is processed in real-time, the more relevant question is whether or not the software can keep up with the real-time data stream. A metric that can be used to answer this question is how many samples per time unit the software (running on a particular hardware configuration) can handle. For example, an application may require the software to be able to process real-time sensor data at 100 samples per second (100 Hz). These requirements are further stretched by the dimension of data points when each sample consists of multiple channel data. For prognostics, depending on the length of prediction horizon, data processing capabilities may be of greater significance from the design and implementation point of view.

Besides time, the other major consideration in computational performance is memory space. Often if makes sense to separately measure the amount of main memory [such as dynamic random access memory (DRAM)] used, and the amount of storage (such as disk space or flash memory used). In both cases, one can either report the asymptotic space complexity using Big O notation, or the number of bytes used by a particular implementation. Space usage is particularly important in embedded applications, such as algorithms run on the flight computer of an aircraft or spacecraft, since these on-board computers usually have very limited space available.

D. *Cost-benefit Metrics*

The cost-benefit metrics, which appear in Table 3, are intended to measure the benefit provided by prognostics. They are all influenced by the accuracy with which RUL is predicted. For example, in MTBF/MTBUR ratio, the denominator, mean time between unit replacements, can be increased if RUL can be more accurately predicted.

Life cycle cost is the sum of acquisition cost and operations cost. Adding prognostic capability to an existing system causes an increase in acquisition cost, due the cost of additional sensors, additional computing hardware, and the cost of developing the prognostic system. Operations cost will decline if RUL is predicted accurately, resulting in fewer components replaced before they need to be replaced, and potentially fewer costly failures. Of course, operations cost can also increase because of the cost of maintaining and operating the prognostic system. So the change in total life-cycle cost caused by the addition of prognostic capability serves as a measure of the net savings gained by adding prognostics to the system.

Return on investment (ROI) goes beyond life-cycle cost by also considering the time value of money. The reduction in life-cycle cost only tells us whether an investment in prognostics resulted in a net savings, without comparing the size of the savings with the size of the investment. ROI tells us the rate of return on the investment in prognostics, which enables the investment in prognostics to be compared with other competing investments.

Table 3. List of metrics based on economic aspect of prognosis

Metric Name	Definition	Description	Selected References
Cost/Benefit			
MTBF/MTBUR ratio	MTBF/MTBUR (mean time between failure / mean time between unit replacement)	This metric measures the ratio between how long a component lasts and how long it is used before replacing it. Prognostics should enable the reduction of this ratio by allowing components to be used longer, until they are closer to failure, which would save money.	[16]
Life-cycle cost	acquisition cost + operations cost	As a metric, compare the life cycle cost of the system (which includes the cost of building it or acquiring it and the cost of operating it) with and without prognostics. Total Value is the change in life cycle cost. ROI will be positive if adding prognostics reduces life cycle cost.	[15]
Return on Investment (ROI)	gain/investment	An investment in prognostics is expected to save money on maintenance and possibly prevention of downtime or lost hardware over the life of the system. The *gain* is the amount of money saved as a result of using prognostics (that is, the reduction in life-cycle cost), and the *investment* is the cost of developing, installing, and maintaining the prognostic system. The ROI (which is usually annualized) can be seen as the interest rate that a bond would have to pay to provide the same financial return. An investment should only be made if its ROI is at least as high as those of other potential investments with similar risk.	[8-10]
Technical Value	$TV = P_f(D \cdot \alpha + I \cdot \beta) - (1 - P_f)(P_D \cdot \phi + P_i \cdot \theta)$ P_f: Probability of a failure mode D: overall detection confidence metric α: savings realized by fault detection in advance I: overall isolation confidence metric β: savings realized by isolating a fault in advance P_D: false positive detection metric φ: cost of false positive detection P_I: false positive isolation metric θ: cost of false positive isolation	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	[1, 11]
Total Value	$V_{TOTAL} = \sum_{i=1}^{FM} TV_i - A - O - (1 - P_c) \cdot \delta$ A: acquisition and implementation costs O: lifecycle operational and maintenance cost P_c: computer resource requirements δ: cost of computer systems	Describes the value of a PHM technology in a particular application as the summation of the benefits it provides over all the failure modes that it can diagnose or prognose less the implementation cost, operation and maintenance cost, and consequential cost of incorrect assessments	[1, 11]

VIII. Discussion

A survey of a wide variety of domains reveals that some metrics are common to most applications whereas some are very domain specific. In this section we discuss few issues that may be important to keep in mind before using these metrics to evaluate prognostic performance.

A. Use of Statistics in Prognostics

From a theoretical point of view, RUL of a system can be considered a random variable and the prediction of an algorithm can be seen as estimating the RUL ground truth. The study of estimators is generally concerned with the estimation of location (central tendency), spread (dispersion) and shape of the distribution and a performance assessment is made between the estimated moments and predicted quantities. In prognostics, to arrive at the correct (representative of the population) RUL estimates either a large number of experiments should be run or an analytical method should be devised. However, in the absence of both, as is often the case, a comparison is conveniently made between the ground truth data and the predicted quantities. One should be careful while interpreting such results as a good performance in a particular experiment does not guarantee the overall success of the algorithm when applied to similar systems. Therefore, a prediction algorithm should be tailored to minimize the statistical bias and spread instead. Prediction accuracy metrics aim to quantify any bias from sample data and a smaller spread is desired, which is minimized through typical methods for standard error minimization.

Some metrics make assumptions about the probability distribution of the error and use the standard estimators like mean and variance in such cases. One should also keep in mind that only in cases where prediction error resembles a Gaussian distribution metrics like sample mean and sample standard deviation may be applied directly. In other cases where samples size is small or outliers presence is suspected, more resistant metrics like median, MAD, MdAD etc. must be employed to make meaningful inferences.

B. What Can be Borrowed from Diagnostics

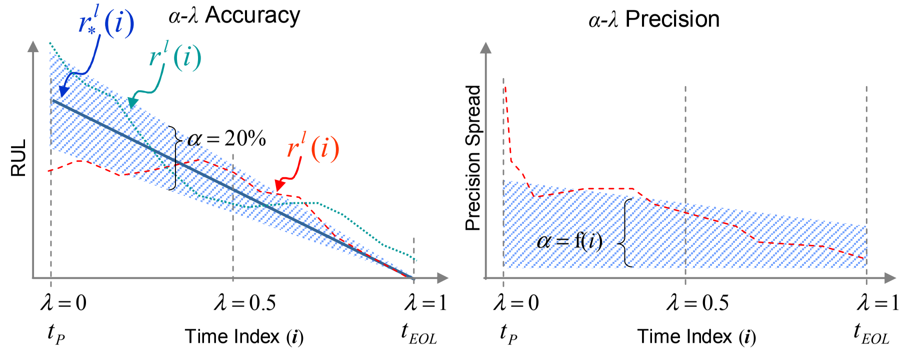
Metrics, previously used for diagnostics, can be modified to be used in prognostics by applying the appropriate modifications. Modifications of some metrics like false positives, false negatives, and the receiver operating characteristics (ROC) curve have been described in Table 2. we illustrate in this section how the false positive and false negative concepts from diagnostics can be transformed for prognostics.

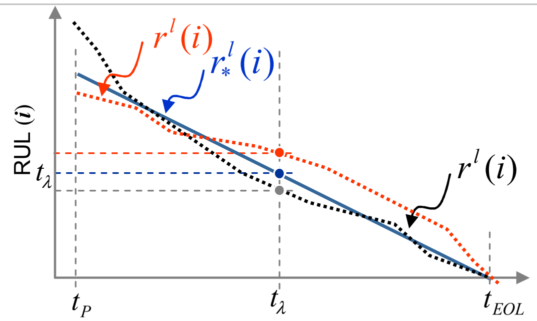
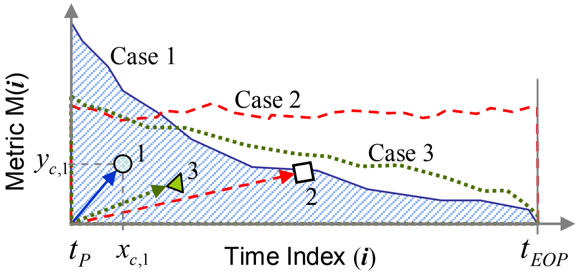
One needs to keep in mind that in prognostics, the utility of the error is most often not symmetric with respect to zero [44]. For instance, if the prediction is too early, the resulting early alarm forces more lead-time than needed to verify the potential failure, monitor the various process variables, and perform a corrective action. On the other hand, if the failure is predicted too late, it means that this error reduces the time available to assess the situation and take a corrective action. The situation deteriorates completely when the failure occurs before an end-of-life prediction is made. Therefore, given the same error size, it is in most situations preferable to have a positive bias (early prediction), rather than a negative one (late prediction). Therefore, one needs to define a limit on how early a prediction can be and still being useful by defining two different boundaries for the maximum acceptable late prediction and the maximum acceptable early one. Any prediction outside of the boundaries can be considered either a false positive (FP) or a false negative (FN). In particular, the focus should be on two instances of the error, $\Delta(c)$, the prediction error at the time t_c when a critical zone (e.g. within the next mission or within next two hours) is reached, and $\Delta(EOL)$, the prediction error at the time when the failure occurs. Other metrics like ROC curve can be then defined based on these modified descriptions. Please refer to Table 2 for brief descriptions on ROC curve and additional details on the false positive and false negative metrics.

C. Suggestions for New Metrics

In addition to various metrics used in different domain, a list of new prognostics specific metrics and their possible formulations have been included in Table 4. Some of these metrics like the α-λ metric are completely new, while the concepts behind some others like convergence have been introduced in literature [43] but have never been as rigorously defined as here. However, we feel there are several other issues that must be considered and may give rise to more metrics as technology matures, some of these issues to consider include (but are not limited to): *Dataset Equivalency* (to compare two different datasets), *Experiment Equivalency* (to compute if two experiments are comparable), *Anticipation Confidence* (to express confidence in future prediction based on a high confidence about the knowledge of future conditions and the accuracy of the past predictions), *Repeatability* (if an algorithm results in repeatable predictions), etc. It is difficult, for obvious reasons, to suggest metrics that will work in all cases - they will necessarily need to be tailored to specific fields or even experiment groups - but we hope to at least initiate a discussion on the subject. For instance, a simple *dataset equivalency* metric can take into account the difference in composition of the feature vectors, the length of the time series, the sampling frequency, and the standard deviation of feature values. An *experiment equivalency* metric could use the condition vectors and compare how the external conditions, such as ambient temperature or pressure, differed throughout the experiments under consideration. The above metrics could, for example, be used to supplement or normalize the accuracy-based metric results for the different runs of the same algorithm.

Table 4. New performance metrics suggested for prognostics in CBM/PHM domain

Metric Name	Definition	Description	Range
Proposed New Metrics for Prognostics			
Prognostic Horizon	$H(i)=EOP - i$	This metric is mentioned in the "Electronics Prognostics R&D Needs Definition" presentation [49], but not explicitly defined. We suggest the following definition: Prognostic Horizon is the difference between the current time index *i* and *EOP* utilizing data accumulated up to the time index *i*, provided the prediction meets desired specifications.	[0, ∞)
Reduced Feature Set Robustness	$FSS(i,f')=\left\|\frac{M(i,f)-M(i,f')}{M(i,f)}\right\|$, where $f'\subset f$ - a subset of the original feature set *f* *and M* is a performance metric of interest	Calculates the effect of an arbitrarily reduced feature set on *M*. This metric does not make a distinction between the essential features of a feature set and the more ancillary ones (as pertinent to the algorithm under consideration). It is simply meant to provide a common way to perform quantitative assessment of the consequences of feature (or features) removal. For instance, loss of signal from an actuator vibration sensor may make the accuracy of RUL estimates unacceptable, while loss of the ambient pressure sensor could still allow useful predictions to be made. In certain fields, such as military aerospace, the tolerance of an algorithm to sensor loss is likely to be an important consideration.	[0, ∞) Perfect score = 0
Prediction Spread	$PS=\sigma\left(M^{l}(i),\ldots,M^{l}(EOP)\right)$, Where σ is any precision measure of choice	This quantifies the variance of prediction over time for any UUT *l*. It can be computed over any accuracy or precision based metric *M*.	[0,- ∞) Perfect score = 0
α-λ Performance	$[1-\alpha]\cdot r_*(t)\le r^{l}(t)\le[1+\alpha]\cdot r_*(t)$ where α accuracy modifier λ: window modifier $t=P+\lambda(EOL-P)$	 Prediction accuracy at specific time instances; e.g., demand accuracy of prediction to be within α*100% after fault detection some defined relative distance λ to actual failure. For example, 20% accuracy (i.e., α=0.2) halfway to failure after fault detection (i.e., λ =0.5).	True/False Perfect value is True

Relative Accuracy	$RA = 1 - \frac{\lvert r_*(t_\lambda) - r^l(t_\lambda) \rvert}{r_*(t_\lambda)}$ where $t_\lambda = P + \lambda(EOL - P)$	 Relative prediction accuracy at specific time instances	[0,1] Perfect score = 1
Cumulative Relative Accuracy	$CRA = \frac{1}{EOL - P + 1} \sum_{i=P}^{EOL} RA$	Normalized sum of relative prediction accuracies at specific time instances.	[0,1] Perfect score = 1
Sampling Rate Robustness	$SRS(\omega_{reference}, \omega) = \frac{1}{L} \sum_{l=1}^{L} \left\{ \frac{\min(M(l, \omega_{reference}), M(l, \omega))}{\max(M(l, \omega_{reference}), M(l, \omega))} \right\}$ M is a performance metric of interest L is the length of the reference data set	Estimates the effect on M from a change in the data set sampling frequency. The estimate is done using a reference frequency that can, for example, be the recommended design frequency for the particular algorithm	[0,1] Perfect score = 1
Data Frame Size Required	$DFS = n;\ n \in (1, \infty)$	Indicates how many consecutive sets of feature values – data frame size - are required to be known at any given time for the algorithm to function within the constraints set on nominal performance. Such constraints can be defined using other metrics, such as minimum accuracy, maximum false positive rate, or others.	[1,∞) Perfect score = 1
Convergence	Let (x_c, y_c) be the center of mass of the area under the curve $M(i)$. then, the convergence C_M can be represented by the Euclidean distance between the center of mass and $(t_p, 0)$, where $C_M = \sqrt{(x_c - t_P)^2 + y_c^2}$, $\mathrm{x_c} = \frac{\frac{1}{2}\sum_{i=P}^{EOP}(t_{i+1}^2 - t_i^2)M(i)}{\sum_{i=P}^{EOP}(t_{i+1} - t_i)M(i)}$, $\mathrm{y_c} = \frac{\frac{1}{2}\sum_{i=P}^{EOP}(t_{i+1} - t_i)M(i)^2}{\sum_{i=P}^{EOP}(t_{i+1} - t_i)M(i)}$ and $M(i)$ is a non-negative prediction error accuracy of precision metric.	Convergence is defined to quantify the manner in which any metric like accuracy or precision improves with time to reach its perfect score. As illustrated below, three cases converge at different rates. It can be shown that the distance between the origin and the centroid of the area under the curve for a metric quantifies convergence. Lower the distance faster the convergence. Convergence is a useful metric since we expect a prognostics algorithm to converge to true value as more information accumulates over time. Further, a faster convergence is desired to achieve a high confidence keeping the prediction horizon as large as possible. 	(0, ∞) Perfect score = ε→0^+
Horizon/Precision Ratio	$HPR_\lambda = \frac{P(i)}{H(i)}$	This metric calculates the ratio of the precision over the horizon. It quantifies the spread as function of distance to EOL.	

D. Some Issues to Keep in Mind

Assessing Multiple Applications- Prognostic applications lack sufficiently large datasets (only a few UUTs) in most cases, which makes it difficult to arrive at statistically significant conclusions. One should be very careful in combining performance results obtained from different applications while assessing an overall performance metric for an algorithm. References from the forecasting domain show that MSE based metrics do not work for combining results from different applications/processes [50]. These metrics are often scale dependent and are not reliable for summarizing results from different applications [51, 52].

Predictions in Prognostics: A prediction in prognostics is always done on a non stationary time series that is often heteroscedastic too. The underlying process constantly changes since the fault leading to a failure physically alters the system as time goes by. This implies that the traditional time series prediction methods do not apply directly and the same is true for the performance metrics. In the forecasting community in general, the selection of performance metric has often been a matter of personal choice and usually with little/no justification [53]. The use of a particular performance metric must be based on several factors like reliability, validity, sensitivity to small changes in errors, resistance to outliers, and how it relates to the health management that the prognostic information will trigger. Our survey indicates that there is no single metric that will capture all the complexities of an algorithm and that the selection of the forecasting method and the evaluation metric is always situation dependent as earlier also pointed out in [51, 54].

Metrics not Based on Error: Although the metrics based on prediction error largely dominate the performance assessment, it is necessary to transform the prediction error into a loss function associated with the decision making process (e.g. design specifications, mission planning, economic justification, etc.) so that prognostics information may be utilized effectively [53, 54]. As enumerated in earlier section, other classes of metrics like cost-benefit and computational complexity, should be given equal importance.

IX. Conclusions & Future Work

In this paper we have provided a concise review of several domains and collected a variety of commonly used metrics to evaluate prediction performance. A list of concepts specific to CBM/PHM requirements has been compiled and these concepts have been molded into a notational framework to facilitate unambiguous descriptions. Several possible categorizations of these metrics have been provided to enhance the understanding of commonalities and differences between varied usages of similar methods. Towards the end some new metrics have been suggested that specifically cater to PHM requirements. Although effort has been made to cover most requirements, further refinements in concepts and definitions are expected as prognostics matures. The intent of this endeavor has been to initiate an open discussion within the research community to standardize the performance evaluation of prognostic systems. We encourage researchers and practitioners in systems prognostics to use a standard set of metrics, such as the ones presented here, to facilitate comparison of alternative approaches and to measure the value provided by prognostics.

References

[1] G. Vachtsevanos, F. L. Lewis, M. Roemer, A. Hess, and B. Wu, *Intelligent Fault Diagnosis and Prognosis for Engineering Systems*, 1st ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2006.

[2] I. Jolliffe, T. and D. Stephenson, B., *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 1st ed. West Sussex, UK: John Wiley & Sons Ltd., 2003.

[3] G. W. Brier and R. A. Allen, "Verification of Weather Forecasts," in *Compendium of Meteorology*, T. F. Malone, Ed. Boston: American Meteorological Society, pp. 841-848.

[4] M. Schwabacher, "A Survey of Data Driven Prognostics," in *AIAA Infotech@Aerospace Conference*, 2005.

[5] M. Schwabacher and K. Goebel, "A Survey of Artificial Intelligence for Prognostics," in *AAAI Fall Symposium*, Arlington, VA, 2007.

[6] ISO, "Condition Monitoring and Diagnostics of Machines - Prognostics part 1: General Guidelines," in *ISO13381-1:2004(E)*. vol. ISO/IEC Directives Part 2, I. O. f. S. (ISO), Ed.: ISO, 2004, p. 14.

[7] B. L. Ferrell, "JSF Prognostics and Health Management," in *IEEE Aerospace Conference*, 1999, p. 471.

[8] D. L. Goodman, S. Wood, and A. Turner, "Return-on-Investment (ROI) for Electronic Prognostics in Mil/Aero Systems," in *IEEE Autotestcon* Orlando, FL, 2005.

[9] S. Vohnout, "Electronic Prognostics System Implementation on Power Actuator," in *IEEE Aerospace Conference* Big Sky, Montana, 2008.

[10] S. Wood and D. Goodman, "Return-on-Investment (ROI) for Electronic Prognostics in High Reliability Telecom Applications," in *Annual International Telecommunications Energy Conference*, 2006.

[11] J. E. Dzakowic and G. S. Valentine, "Advanced Techniques For The Verification And Validation Of Prognostics & Health Management Capabilities," in *Machinery Failure Prevention Technologies (MFPT 60)*, Virginia Beach, VA, 2007.

[12] D. W. Brown, P. W. Kalgren, C. S. Byington, and R. F. Orsagh, "Electronic Prognostics - A Case Study Using Global Positioning System (GPS)," in *IEEE Autotestcon* Orlando, FL, 2005.

[13] J. Gu, D. Barker, and M. Pecht, "Prognostics Implementation of Electronics Under Vibration Loading," *Microelectronics Reliability,* vol. 47, 2007.

[14] J. P. Hofmeister, P. L. T. Walter, D. Goodman, E. G. Ortiz, M. G. P. Adams, and T. A. Tracy, "Ball Grid Array (BGA) Solder Joint Intermittency: Detection: SJ BIST," in *IEEE Aerospace Conference* Big Sky, Montana, 2008.

[15] C. Wilkinson, D. Humphrey, B. Vermeire, and J. Houston, "Prognostic and Health Management for Avionics," in *IEEE Aerospace Conference*, 2004.

[16] C. Teal and B. Larsen, "Technology Update II: Wire Systems Diagnostics & Prognostics," in *Digital Avionics Systems Conference*, 2003.

[17] N. Vichare and M. Pecht, "Enabling Electronic Prognostics Using Thermal Data," in *12th International Workshop on Thermal investigations of ICs*, 2006.

[18] N. Vichare, P. Rodgers, V. Eveloy, and M. Pecht, "Environment and Usage Monitoring of Electronic Products for Health Assessment and Product Design," *Quality Technology & Quantitative Management,* vol. 4, pp. 235-250, 2007.

[19] J. W. Simons and D. A. Shockey, "Prognostics Modeling of Solder Joints in Electronic Components," in *IEEE Aerospace Conference*, 2006.

[20] M. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford, UK: Oxford University Press, 2003.

[21] F. E. Harrell Jr., K. L. Lee, and D. B. Mark, "Tutorial in Biostatistics: Multivariate Prognostics Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors," *Statistics in Medicine,* vol. 15, pp. 361-387, 1996.

[22] I. Zelic, N. Lavrac, P. Najdenov, and Z. Rener-Prime, "Working Notes, Workshop on Prognostic Models in Medicine: Artificial Intelligence and Decision Analytic Approaches," in *Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making, AIMDM'99*, Aalborg, Denmark, 1999.

[23] A. Abu-Hanna and P. J. F. Lucas, "Prognostics Models in Medicine: AI and Statistical Approaches," *Methods of Information in Medicine,* vol. 40, pp. 1-5, 2001.

[24] E. L. Kaplan and P. Meier, "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association,* vol. 53, pp. 457-481, 1958.

[25] L. J. Bond, S. R. Doctor, D. B. Jarrell, and J. W. D. Bond, "Improved Economics of Nuclear Plant Life Management," in *Second International Symposium on Nuclear Power Plant Life Management* Shanghai, China: IAEA, 2007, p. 26.

[26] D. B. Jarrell, "Completing the Last Step in O&M Cost reduction," Pacific Northwest National Laboratory, p. 8.

[27] R. Carbone and J. Armstrong, "Note. Evaluation of Extrapolative Forecasting Methods: Results of a Survey of Academicians and Practitioners," *Journal of Forecasting,* vol. 1, pp. 215-217, 1982.

[28] R. J. Hyndman and A. B. Koehler, "Another Look at Measures of Forecast Accuracy," *International Journal of Forecasting,* vol. 22, pp. 679-688, 2006.

[29] S. Makridakis, A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler, "The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition," *Journal of Forecasting,* vol. 1, pp. 111-153, 1982.

[30] S. Makridakis and M. Hibon, "The M3-Competition: Results, Conclusions, and Implications," *International Journal of Forecasting,* vol. 16, pp. 451-476, 2000.

[31] F. X. Diebold and R. Mariano, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics,* vol. 13, pp. 253 – 265, 1995.

[32] H. Theil, *Applied Economic Forecasting*. Chicago, IL: Rand McNally and Company, 1966.

[33] W. Marquering and M. Verbeek, "A Multivariate Nonparametric Test for Return and Volatility Timing," *Finance Research Letters,* vol. 4, pp. 250 – 260, 2004.

[34] B. Ebert, et.al., "Forecast Verification - Issues, Methods and FAQ." Last accessed. 2007, url: http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html.

[35] D. W. Wichern, Flores, B. E., "Evaluating forecasts: a look at aggregate bias and accuracy measures," *Journal of Forecasting,* vol. 24, pp. 433-451, 2005.

[36] T. N. Palmer, A. Alessandri, U. Andersen, P. Cantelaube, M. Davey, P. De le cluse, M. De qu, E. Diez, F. J. Doblas-Reyes, H. Feddersen, R. Graham, S. Gualdi, J. F. Gueremy, R. Hagedorn, M. Hoshen, N. Keenlyside, M. Latif, A. Lazar, E. Maisonnave, V. Marletto, A. P. Morse, B. Orfila, P. Rogel, J. M. Terres, and M. C. Thomson, "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 2004.

[37] G. Giebel, R. Brownsword, and G. Kariniotakis, "The State-of-The-Art in Short-Term Prediction of Wind Power - A Literature Review," Riso National Laboratory, Rosklide, Denmark 2003.

[38] H. Madsen, G. Kariniotakis, H. A. Nielsen, T. S. Nielsen, and P. Pinson, "A Protocol for Standardizing the Performance Evaluation of Short-Term Wind Power Prediction Models," Technical University of Denmark, IMM, Lyngby, Denmark, Deliverable ENK5-CT-2002-00665, 2004.

[39] O. Gusikhin, N. Rychtyckyj, and D. Filev, "Intelligent Systems in the Automotive Industry: Applications and Trends.," *Knowledge and Information Systems,* vol. 12, pp. 147-168, 2007.

[40] F. L. Greitzer and R. A. Pawlowski, "Embedded Prognostics Health Monitoring," in *Proceedings of the International Instrumentation Symposium Embedded Health Monitoring Workshop*, United States, 2002, p. Size: vp.

[41] L. Jianhui, M. Namburu, K. Pattipati, A. L. Q. Liu Qiao, M. A. K. M. Kawamoto, and S. A. C. S. Chigusa, "Model-Based Prognostic Techniques [Maintenance Applications]," in *AUTOTESTCON 2003. IEEE Systems Readiness Technology Conference. Proceedings*, 2003, pp. 330-340.

[42] D. Djurdjanovic, J. Lee, and J. Ni, "Watchdog Agent--an Infotronics-Based Prognostics Approach for Product Performance Degradation Assessment and Prediction," *Advanced Engineering Informatics,* vol. 17, pp. 109-125, 2003.

[43] G. Vachtsevanos and P. Wang, "Fault Prognosis Using Dynamic Wavelet Neural Networks," in *AUTOTESTCON Proceedings, 2001. IEEE Systems Readiness Technology Conference*, 2001, pp. 857-870.

[44] K. Goebel and P. Bonissone, "Prognostic Information Fusion for Constant Load Systems," in *Proceedings of the 7th Annual Conference on Information Fusion*. vol. 2, 2005, pp. 1247 – 1255.

[45] D. C. Hoaglin, F. Mosteller, and J. W. Tukey, "Understanding Robust and Exploratory Data Analysis," in *Probability and Mathematical Statistics*: John Wiley & Sons, 1983.

[46] C. S. Byington, Roemer, M.J., Kalgren, P.W., "Verification and Validation of Diagnostic/Prognostic Algorithms," in *Machinery Failure Prevention Technology Conference (MFPT 59)* Virginia Beach, VA, 2005.

[47] G. Vachtsevanos, "Performance Metrics for Fault Prognosis of Complex Systems," in *AUTOTESTCON 2003. IEEE Systems Readiness Technology Conference.*, 2003, pp. 341- 345.

[48] A. V. Aho, J. D. Ullman, and J. E. Hopcroft, *Data Structures and Algorithms*: Addison Wesley, 1983.

[49] "Electronics Prognostics R&D Needs Definition," in *Electronics Prognostics Workshop II* Miami, FL: Electronics Prognostics Technology Task Group, Defence Technical Information Center (DTIC) 2006.

[50] P. A. Thompson, "A Statistician in Search of a Population," *International Journal of Forecasting,* vol. 8, pp. 103-104, 1992.

[51] J. S. Armstrong and F. Collopy, "Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons," *International Journal of Forecasting,* vol. 8, pp. 69-80, 1992.

[52] C. Chatfield, "A Commentary on Error Measures," *International Journal of Forecasting,* vol. 8, pp. 100-102, 1992.

[53] D. A. Ahlburg, "Error Measures and the Choice of a Forecast Method," *International Journal of Forecasting,* vol. 8, pp. 99-100, 1992.

[54] R. L. Winkler and A. H. Murphy, "On Seeking a Best Performance Measure or a Best Forecasting Method," *International Journal of Forecasting,* vol. 8, pp. 104-107, 1992.