Simple Metrics for Evaluating and Conveying Prognostic Model Performance To Users With Varied Backgrounds

Michael E. Sharp

University of Tennessee Nuclear Engineering Department, Knoxville TN
msharp6@utk.edu

ABSTRACT

The need for standardized methods for comparison and evaluation of new models and algorithms has been known for nearly as long as there has been models and algorithms to evaluate. Conveying the results of these comparative algorithms to people not intimately familiar with the methods and systems can also present many challenges as nomenclature and relative representative values may vary from case to case. Many predictive models rely primarily on the minimization of simplistic error calculation techniques such as the Mean Squared Error (MSE) for their performance evaluation. This, however, may not provide the total necessary information when the criticality, or importance of a model's predictions changes over time. Such is the case with prognostic models; predictions early in life can have relatively larger errors with lower impact on the operations of a system than a similar error near the end of life. For example, an error of 10 hours in the prediction of Remaining Useful Life (RUL) when the predicted value is 1000 hours is far less significant than when the predicted value is 25 hours. This temporality of prognostic predictions in relation to the query unit's lifetime means that any evaluation metrics should capture and reflect this evolution of importance.

This work briefly explores some of the existing metrics and algorithms for evaluation of prognostic models, and then offers a series of alternative metrics that provide clear and intuitive measures that fully represent the quality of the model performance on a scale that is independent of the application. This provides a method for relating performance to users and evaluators with a wide range of backgrounds and expertise without the need for specific knowledge of the system in question, helping to aid in collaboration and cross-field use of prognostic methodologies. Four primary evaluation metrics can be used to capture information regarding both timely precision and accuracy for any series or set of prognostic predictions of RUL. These metrics, the Weighted Error Bias, the Weighted Prediction Spread, the Confidence Interval Coverage, and the Confidence Convergence Horizon are all detailed in this work and are designed such that they can easily be combined into a single representative "score" of the overall performance of a prediction set and by extension, the prognostic model that produced it. Designed to be separately informative or used as a group, this set of performance evaluation metrics can be used to quickly compare different prognostic prediction sets not only for the same corresponding query set, but just as simply from differing query data sets by scaling all predictions and metrics to relative values based on the individual query cases.

1. INTRODUCTION

The need for standardization in the area of evaluation for prognostics research has been well documented [Uckun et al 2008]. Work has even been presented on the evaluation of individual features or parameters as to their suitability for use in prognostic modeling [Coble 2010]. Recent effort has been focused on the standardization of prognostic model performance evaluation based on meaningful criteria that can be used to compare the output of prognostic models not only within given application, but across the field of predictive engineering [Saxena 2008]. Unfortunately, despite this large step forward in the evaluation of prognostic models, the existing metrics have yet to see wide spread acceptance and use. This may in part be due to both the seemingly and occasionally complicated nature of evaluating and interpreting these metrics.

This work seeks to incorporate group-based comparison into the offline prognostic algorithm evaluation process, and presents variants on some well-known performance metrics that are built upon a multitude of known cases to which the prognostic model has been applied.

Specifically, four separate updated scalar metrics have been identified that sufficiently characterize and convey meaningful, intuitive information about the output predictions of a prognostic model: Weighted Error Bias (WEB), Weighted Prediction Spread (WPS), Confidence Interval Coverage (CIC), and the Confidence Convergence Horizon (CCH). Each one, detailed below, captures a key aspect and desirable quality of prognostic predictions that can be quickly, easily, and intuitively compared amongst separately developed models to rank and rate output performance. These metrics are built upon the errors and uncertainty associated with each prediction set, rewarding the minimization of both. To calculate both the errors and uncertainty of a prediction set, another descriptive series of values known as the Binned Percent Error is also defined and demonstrated in both use and interpretation in regards to the scalar metrics.

2. BACKGROUND AND MOTIVATION

It has been suggested that depending on the needs of the end user, many different types of effective comparison algorithms could be employed such as cost/benefit analysis [Banks 2007]. However, given that many models may provide comparable results at similar costs, what are robust and useable methods for effectively ranking and expressing their relative effectiveness? Or more generally, what is the best way to convey results of a comparative analysis to someone that is not necessarily well versed in the science of prognostics or to a large audience with varied backgrounds and expertise? A standardized set of evaluation metrics that is both simple to calculate and intuitive to understand is possibly the best answer to this question. Many metrics for determining model error and even prognostic error have been proposed in the past. Building upon these metrics to update the evaluation of prognostic prediction set metrics, the addition of standardization in the formats and values reported can promote the use of the more intuitively descriptive metrics for a more wide spread understanding and standardization of the field of prognostics beyond it's traditional set of core users. Simply, and accurately conveying the capabilities of any prognostic algorithm is key to gaining acceptance and application in real world, and industry scenarios.

2.1. Standard Model Evaluation Metrics

The most basic of metrics are often overlooked in regards to their usefulness for evaluation prognostic predictions. It is true that in many ways these simple error metrics are inadequate to completely and appropriately characterize the type of information pertinent to prognostic performance. However, when conveying information to potential users of a prognostic model or scientists and engineers from other fields, it is often convenient to at first convey information in a manner both simple and familiar to them.

Many of these type metrics exist, but the Mean Absolute Error (MAE) is a fairly easy metric to compute and in many ways, the most intuitive to understand. Unfortunately, this metric could also be argued to be the least informative about the overall performance of the model compared to those presented in later sections of this paper. Defined in Equation 1, MAE is the average absolute difference between the model prediction P_i and the true Remaining Useful Life (RUL_i) at all times t and for all historic query cases i.

Equation 1

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t}^{T_i}\left(\left|\hat{P}_i(t) - RUL_i(t)\right|\right)$$

In other words, the MAE can be thought of as the average error in prediction for each unit, i, that has run to failure and for each time T, that a prediction is made. The primary attraction of this metric is that it quantifies the average expected value any estimate will be from the true value in real units directly comparable with the system lifetime. Similarly, one could also calculate the standard deviation of the prediction error for a measure of the spread of these errors.

These metrics are useful for comparing separate models built upon similar data, or data from systems with comparable lifetime scales, but give no clear indication of prediction performance without some context to the data. Additionally, these standard formula metrics are inflexible to individual requirements about the specifications of the predictions and can be largely susceptible to outliers.

Although MAE has existed in some implicit form for as long as there have been errors to calculate, the direct usefulness of this intuitive error metric to the evaluation of prognostic predictions performance is severely limited. Conceptually, this error metric provides clear and meaningful indications of the expected error of the total lifetime of a system. Unfortunately as far as prognostics is concerned predictions near the end of a unit or system's lifetime are much more critical than those near the beginning of life. The remaining metrics introduced and described in this work help to overcome and fill in the gaps left by MAE and similar standard metrics.

2.2. Traditional Hierarchical Based Metrics

Saxena et al proposed a hierarchical system of prognostic evaluation, which includes four primary metrics that rely on one another to provide meaning [Saxena et al 2009]. The hierarchy includes in order: the Prognostic Horizon, the Alpha – Lambda Performance, the Relative Accuracy, and Convergence. Briefly, these metrics describe in order, the first instance a prediction becomes within acceptable bounds, if predictions remain in the acceptable bounds at a

given time, one minus the percent difference of the prediction to the true value, and finally how quickly the predictions arrive at the final answer.

This system is a good step in establishing a coherent and consistent method for evaluating the performance of a prognostic model. However, these metrics rely on several case specific terms and concepts that do not always lend well to conveying performance to individuals not previously well versed in their application.

As presented, the hierarchy is largely self-reliant. The groups of metrics must be evaluated together, the interpretation of the results from one test have effect on others. For example, a model could have an early lifetime Prognostic Horizon, but this is only truly useful if the model also passes all the Alpha-Lambda tests at every subsequent time instance. Similarly, a rapid convergence should be coupled with a good Relative Accuracy. Understanding the passing of these tests and their significance requires an understanding of the relationships between the metrics that may not be instantly intuitive to experts of other fields.

Also, some of these metrics are not presented in a scalar value manner, making it difficult to assign an overall quantitative value of prediction quality. When presenting the results of a model analysis to prospective users, often a simple and intuitive scalar value comparison is much cleaner and easier to understand than a series of mixed visually and numerically descriptive values. In other words, for certain audiences the hierarchy may unintentionally obscure model evaluation when trying to compare separate models.

In the papers presented by Saxena, the metrics are used to evaluate the prognostic estimations of a single failed unit query case. These metrics each take into consideration only a single query case and only report aspects of that case. The obvious extension of this is to create an average of these individual query based metrics over a large set of query cases. However, this may not always translate well, particularly given the interdependency of the interpretation of the metrics as described above. Metrics built to collectively analyze a group predictions across several failed query cases can help to avoid such skewing of the reported values.

Group based metrics can also help to better estimate a level of uncertainty associated with each predictive model under evaluation. Saxena et al suggest methods for incorporating singular case uncertainties into their metric evaluations, but do not suggest a simple, effective way to propagate these uncertainties [Saxena 2010]. Other recent works have also focused on the evaluation of uncertainty in regards to prognostics. Many interesting ideas concerning both the quantification and evaluation of uncertainty and uncertainty-based metrics have been presented [Orchard et al. 2008], [Leao et al. 2010]. Metrics presented in later sections seek to provide intuitive estimations on model uncertainty based on the set of estimations themselves.

Additional considerations about the interpretation and presentation of some of the metrics should also be mentioned. Prognostic Horizon was originally designed to report the first instance in life where predictions fall within a certain bound, regardless of if the predictions later leave that bound. In later work, Saxena suggests corrections to this by allowing the user to instead quantify the last time it falls in bound without going back out [Saxena 2009]. This practice makes much more sense and should become standard, but again because Prognostic Horizon is calculated over a series of individual cases, there is no standard way to define the value for a set of prediction cases. Should the average value be reported, or would a minimum or maximum be more representative?

Similarly, the convergence metric has the potential to give the same quantification of convergence for vastly different evolutions in the predictions, potentially misleading any blind interpretation of the metric. For example, a prediction set that contains a large outlier early in life (which may be considered trivial) followed quickly by consistent near correct values can show the exact same convergence as one without an early outlier the never quite settles on a consistent prediction value, depending on the application these could be effectively very different results. This work seeks to build upon the initial successes of these metrics, by creating and presenting metrics and methods that are more easily interpreted on a common scale without need for intense understanding of the methods behind them.

2.3. Additional Quantitative Evaluations

Other works have also tried to build upon or propose other standardized metrics. Some of these works, taking a cue from the fields of meteorology and climatology have adapted the concepts of “value” and “skill into the prognostic predictions evaluation vernacular [Tang et al 2011]. Skill, simply put, is the percent improvement of any singular evaluation metric of one prediction set versus some reference prediction set. This can be convenient as a concept for comparing two different models, but provides no additional information not obtained for the original metric itself.

Conversely the concept of “value” is a quantitative metric that can directly be used to evaluate a prediction set. Value is the total set of some user-defined error versus consequence values for a particular application. This allows a user to capture important aspects of low probability but high cost errors that may be of particular interest, such as extreme early life failures. This is very useful for high level decision making and internal review processes; however it

lacks the intuitive ability to be conveyed without some form of reference context for the associated costs, whether they be in safety related hours, repair/downtime costs, or some other arbitrary unit. The value of a system is a wonderful internally created metric and can be used to great effect when properly applied and calculated in the industrial setting. Unfortunately, "value" does not translate well across different systems and industries. A standardized prognostic evaluation system should be expected to be instantly interpretable by people of many different backgrounds. The work presented below proposes solutions to this and other problems inherent in the standard set of prognostic evaluation metrics.

3. PROPOSED UPDATED PROGNOSTIC PREDICTION METRICS

To promote the wide spread usage of a set of standardized evaluation metrics for prognostic predictions, this work presents set of prognostic prediction evaluation metrics that are designed to be both intuitive and informative to users and reviewers with various backgrounds and levels of expertise. These metrics are also designed to be evaluated on and capture pertinent aspects of entire sets of prognostic predictions over many query cases. Each metric captures key aspects of accuracy, precision, and timeliness. For any prediction, there is both an expected error and an associated uncertainty, these metrics help to report the evolution of these values with special regards to the importance of the relative lifetime of the failed system or equipment, also referred to as query units.

3.1. Weighted Error Bias

The Weighted Error Bias (WEB) is the first of the lifetime percentage based metrics. WEB, as defined in Equation 2, is a measure indicating the effective bias in all predictions as a percentage of total unit lifetime.

Equation 2

$$WEB = \frac{100}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} w_i(t) * \frac{\left(\hat{P}_i(t) - RUL_i(t)\right)}{TotalUnitLifeTime_i}$$

where $w_i(t)$ is the importance weighting for unit i at time t. In this equation, negative values indicate that the prognostic predictions tend to be lower than the true RUL where positive means the opposite. Additional metrics, such as the Weighted Prediction Spread (WPS) presented below, can be combined with the WEB help to capture the average absolute deviation or uncertainty of a prediction set.

From this equation, it becomes evident that WEB is very similar to MAE except in two important respects. First, it is tallied and reported as a percentage of the total lifetime of the individual failed query unit, i. This allows for the intuitive inspection of the performance of a series of predictions without the need for some contextual setting. A model whose predictions yield a 10%WEB would be expected to be better than one with a 25%WEB regardless of the systems, equipment, or time scales involved. This also has the added benefit of implicitly scaling the errors such that similar deviations from the true Remaining Useful Life (RUL) values for short-lived components would be weighed heavier than those in longer-lived units, even within the same historic data set. This is intuitively important, as an error of 20 time cycles is less important if the unit in question survives 300 cycles as opposed to if it only survives 100 cycles.

The second difference is in the explicit importance weighting, w_i, of the different errors based on their time in the lifecycle of the historic unit. This importance weighting can easily be tailored to the specific needs or desires of the end user, but in most cases an emphasis on the end of lifetime is the most meaningful towards prognostic predictions. A 10% error near the beginning of unit life when there is 85% of life remaining gives plenty of time to act an take corrective actions, where a 10% error with only 5% of life remaining could result in an unexpected failure if the unit were expected to life through those remaining cycles. A weighting function that accurately reflects this end of life importance is the Gaussian Kernel Function with a mean value set to the lifetime of the unit and a standard deviation, or bandwidth, set to 50% of that lifetime. Although this metric is built with weightings in mind, a weighting function equal to 1/T for all t can easily turn this metric into a simple average of percent difference between the true and estimated values. For this and all weighted metric, comparisons between algorithms using these metrics would only be meaningful if standard weighting functions are used. Additional work and investigation into what the most appropriate standardized weighting function could prove beneficial. However, regardless of the weighting function, the standardized scaling of the metrics can help it be more relatable to generic audiences.

The optimal value for the WEB is zero, indicating that the average prediction value is centered on the true RUL. Positive and negative values simply express the direction of the bias, otherwise this metric can be presented as a representation of averaged percent error, a concept that is widely utilized and accessible to many academic and industry backgrounds. The weighting function can be tailored to any specific need or application, but the fundamental metric remains an easily interpreted percentage of system lifetime.

3.2. Percent Error Value Binning

The final three prognostic prediction performance metrics rely on estimating or inferring the uncertainty of prognostic predictions throughout the total lifetime of a query unit. In order to do this effectively, the 95%

confidence interval (or some similar level of confidence interval) needs to be calculated at various points throughout the unit lifetime. One of the more straight forward methods for doing this is to create a set of bins evenly divided between 0 and 100% of system lifetime, and placing each calculated percent error in the bin corresponding to the true percent of unit life corresponding to that error. In other words, first calculate the percent error for a given historic prediction, $P_i(t)$, such that the percent error is the difference between the predicted RUL and the actual RUL divided by the query unit ,i's, total lifetime.

Equation 3

$$\%Er = 100 * \frac{\hat{P}_i(t) - RUL_i(t)}{TotalLifeTime_i}$$

Next note the corresponding percentage of actual lifetime (POL), defined by the current time, t, divided by the current unit's total lifetime. Finally place the calculated percent error into the POL bin whose edges, B, are defined as: $B_{LOWER} < POL_i(t) < B_{UPPER}$

Repeat for all historic predictions over all query cases, placing them in to the same series of corresponding bins. Converting the numbers into percentages allows for the direct comparison and inclusion of these similarly located values with proper importance weightings applied as based on their lifetime.

Once this series of regular serial bins is populated, a 95% confidence interval around the mean value can be calculated from the 2.5% and 97.5% percentiles of the error set for each bin. Much like the weightings presented with the metrics presented in this paper, these percentages can be altered to suit the specific application requirements. Additionally, the expected value for each individual bin can be calculated, creating an expected error bias that maps throughout the lifetime of a unit as a more detailed representation of the WEB if such is required. This binning is primarily an intermediary form for the metrics presented in this paper, but as will be shown later, it can also be used to create clear visualizations of the evolution of predictions and how they relate to the true values of RUL. Visualizations such as these can be a great aid in communicating a prediction algorithm's performance to an audience not intimately familiar with the algorithm or system in question.

3.3. Weighted Prediction Spread

Uncertainty estimations, though not always straightforward, are a crucial part of evaluating any prediction value. Thus it follows that the quality of any prediction model should also be defined by its' associated uncertainty. Additionally, much like the model prediction error and bias, not all points during the lifetime of the query system should necessarily be treated with equal importance. The predictions of Remaining Useful Life (RUL) made by a model are typically more important near the end of the system's life than they are at the beginning of life, as near the beginning of life there is comparatively much more time to react and compensate, or mitigate any impending faults or failure inferred from the prognostic model.

The spread of the model predictions at various points in life are an important factor in the total considerations of the uncertainty of a series of predictions. The prediction spread for each binned point of system life, is calculated as the difference between the upper and lower bounds of the corresponding 95% confidence intervals from the binned error values discussed previously. Using the same importance weighting function as the Weighted Error Bias (WEB), the Weighted Prediction Spread (WPS) can be defined as by Equation 4.

Equation 4

$$WPS = 100 * \frac{\sum_{bi=1}^{\#Bins} W_{bi} * CI_{bi}}{\sum_{bi=1}^{\#Bins} W_{bi}}$$

In this equation, the weighting function is based on the center value for each reference bin, such that each bin importance weighting, W_{bi}, is defined by the Gaussian kernel in Equation 5.

Equation 5

$$W_{bi} = \exp\left(-\left(\frac{Bin_{bi} - 100\%}{50\%}\right)^2\right)$$

Notice that the typical normalization factor associated with Gaussian kernels is rendered unnecessary due to the inherent normalization factor included in the definition of WUS. Although a kernel bandwidth of 50% is shown, other bandwidths or even a uniform weighting function can easily be substituted to accommodate specific needs. All the factors and values associated in the metrics based on the binned interval error values are listed and manipulated as percentages allowing for quick intuitive evaluation of the effective important uncertainty of any given prediction set.

With this metric, a 0% WPS alone would seem to indicate absolute certainty in all predicted values, but this may be misleading. In fact, all this would indicate is that all predictions made are exactly the same based exclusively on the percent RUL of the system in question. This is why uncertainty is comprised of both a spread and a bias. The WPS metric can be used in conjunction with the WEB to infer the level of model uncertainty according to the equation: $Uncer \approx \sqrt{WPS + WEB^2}$. This modification of the traditional equation for analytic uncertainty allows for more flexibility in defining what an appropriate value of the spread should be.

Another useful criteria to think of is if the predictions do in fact have enough spread to cover the true RUL (i.e. WPS>=WEB). A more explicit and useful metric evaluating this coverage is the Confidence Interval Coverage (CIC) should also be calculated, and is discussed in the next section.

3.4. Confidence Interval Coverage

Another important indication of the quality of a prediction set generated by any model is whether or not the confidence interval of the prediction spread covers the true Remaining Useful Life (RUL). This effectively incorporates information relating to both the error bias and the error variance at given points in life. This metric is simply defined by the total percentage of binned error sets whose 95% confidence interval contains the true RUL. This is more rigorously defined in Equation 6.

Equation 6

$$CIC = 100 * \frac{\sum_{bi=1} \%RUL_{bi} \in B_{bi}}{\#Bins}$$

This equation is interpreted as the sum number of true percent RUL values that are contained within their corresponding error bin set, and divided by the total number of bins and multiplied by 100 to convert to a percentage. This additional metric verifies the total accuracy of the prediction set. An optimal coverage of 100% shows that the true value of any prediction is contained within the prediction spread or approximate confidence interval of the prognostic model's predictions. This when coupled with the previously detailed metrics gives a solid expectation of the accuracy and expected effective error over the total of system life predictions. The final vital element not conveyed by these metrics is the explicit end of life accuracy and precision. The Confidence Convergence Horizon fills this void.

3.5. Confidence Convergence Horizon

This final standalone metric captures and quantifies the end of life quality of both the precision and accuracy of a prediction set. A 10% Confidence Convergence Horizon (CCH), or simply the Convergence Horizon (CH), identifies the percentage of system Remaining Useful Life (RUL) beyond which, all prediction confidence intervals are both less than 10% of the total system life and contain the true RUL. In other words, the CCH identifies a RUL prediction value that once reached, it and all remaining predictions of RUL can trusted to be no more than 10% from the true RUL 95% of the time (assuming a 95% confidence interval was selected as described above). Obviously a CCH of 100% would be optimal, showing that all predictions within the query set are within less than 10% of the true values. Much like the other metrics, the percentage of this metric can be adjusted to suit the specific needs and requirements of any particular application.

Although this seems to be a rather stringent criterion to meet, it nonetheless, is very important. This horizon is a quick and intuitive identifier of the region of most confidence for a particular prediction set. Unfortunately, like any single descriptive metric, the CCH has the potential to be misleading if it is not considered along with the other metrics defined in this section. As an example, consider a model which predicts the RUL of a system within 10% during most of the system life, but due to an artifact of the data, exhibits an 11% bias at the very end of life. This model would produce a CCH of 0% as there is no point in time which you can trust all following predictions to be less than 10%. This does not however mean that the model produces unusable or even inaccurate results.

Information from each of the listed metrics contains and expresses vital information required to develop a full understanding of the models performance, but when relating to potential users of an algorithm it is often convenient to assign a single quantitative value of "goodness" to a particular model and prediction set. Described in the following section is a method for developing such a single metric.

3.6. Total Score Metric

There has been proposed a sort of hierarchical ranking of some of the previously developed metrics [Saxena 2009]. To some degree, this work is able to eliminate the explicit need for this hierarchical system and in its place supplies a single aggregate scoring metric to rank the overall performance of a particular prognostic model's output predictions. Of the metrics detailed in this paper, four in particular can be merged to give a singular quantitative value of "goodness" for a prognostic model prediction set. These metrics, Weighted Error Bias (WEB), Weighted Prediction Spread (WPS), Confidence Interval Coverage (CIC), and the Confidence Convergence Horizon (CCH) each detail a particular yet vital aspect of the total historic prediction produced by a given model. With this in mind, and given that each of these metrics have been constructed to be listed in similar units of percent Remaining Useful Life (%RUL), a simple composite of these metrics can yield a meaningful, accessible, and direct measure of the quality of a model prediction set. Equation 7 can easily be applied for quick quantitative comparison of multiple models' prediction sets.

Equation 7

$$TotalScore = \vec{N} * \begin{bmatrix} 100 - |WEB| \\ 100 - WPS \\ CIC \\ CCH \end{bmatrix}$$

Note that in this equation, both the absolute value of the WEB and the WPS are subtracted from 100 to reflect that the minimums of these values are the desired quantities. The original WEB metric can be negative to indicate direction of bias, but when combining into an overall score, it is the absolute value that is of more interest. N is any normalized vector weighting the importance of the four metrics. For both simplicity, and intuitive interpretation of the resulting number, a simple average of the four modified metrics can be taken, (mathematically this results from a vector of [.25 .25 .25 .25]). This combined metric can easily be used to present the performance of any predictive model out of a perfect score of 100%. Much like other standardized academic testing, this ideal score is ranked based on ideal performance. For nearly all real systems, 100% accurate predictions 100% of the time is essentially impossible, but this still can help to provide an intuitive ranking system familiar to a wide audience. Some of the model metrics contain similar information, this is not useless redundancy, but instead reflects the increased importance of these aspects when the metrics are combined. For example, if a set of model predictions exhibit 0% CIC, that prediction set would also by definition exhibit a 0% CCH. Coverage of the correct RUL within a confidence interval is one of the most important criteria any prognostic model should meet, so with the standard weighting set, the best total score the model could produce would be less than 50%, reflecting that the model has never produced correct answer.

4. Prediction Metric Example Cases Studies

To help further clarify and explain the prediction metrics, consider a standard pump and motor system with a mean failure time of about 275 operating hours with two common modes of failure with different mean failure times. Three separate simulated models were built to predict the Remaining Useful Life (RUL) of these motors. The first is based strictly on statistical conditional time based probability of failure. The second two are built to simulate more effects based modeling types. In order to compare the three models, each one uses a set of 100 predictions about similar sets of query cases and has the metrics detailed above applied to those prediction sets.

Shown in Figure 1, the Model 1 prediction set for all 100 cases completely overlay one another. This is expected and due to the fact that this model's output is based exclusively on the current lifetime of the queried system.

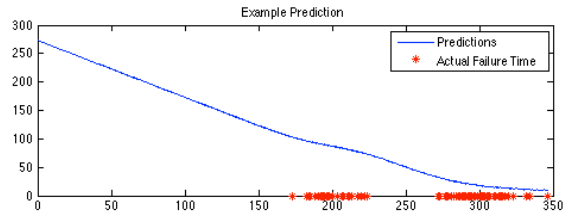

Figure 1 - Model 1 RUL Predictions

Despite the fact that each of the predictions for each individual case are all exactly the same, they represent varying percentages based on the true queried system's lifetime. This is accounted for in the calculations of the performance metrics shown in Figure 2.

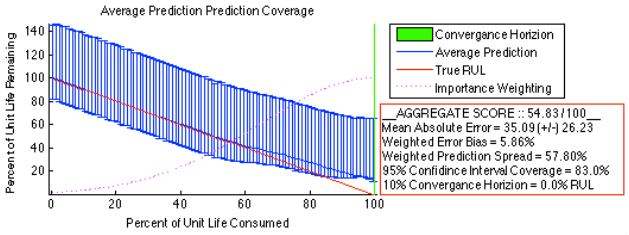

Figure 2 - Model 1 Prediction Performance Metrics

The most intuitive and easily understood metric on this figure is the Mean Absolute Error (MAE), listed as 35.1 hours with an associated standard deviation of 26.2 hours. Considering that the average lifetime is 275 hours, these numbers present values which would easily allow for the rescheduling of duty cycles to accommodate maintenance or similar mitigating actions before the units would be expected to fail. The MAE gives a good basic understanding of how much error to expect out of the model, and is good for comparing models that are run against the same data set. However, the three example models presented here are run with differing query data sets. The sets are taken from similar sets of pump systems, but the individual units and their true total lifetimes are different. Although MAE could be used to compare these models and prediction sets because time units and expected average lifetimes are the same, the percentage-based metrics are more appropriate and generally informative.

The most prominent prediction evaluation tool in this figure is the binned error average estimate and their associated 95% confidence intervals represented by the blue error bars. This contains the most total and useful information about the prediction set. These bind also are used to represent the other metrics as they evolve through time. The solid blue line is the bias at a given time; the error-bars represent the spread; the total number of bars which contain the red (true) RUL represent the confidence coverage. Finally, the Convergence Horizon will be represented as a green box in the following figures, but is not present in this on due to CH being equal to 0.

From this chart it is obvious that early in life the model predicts the correct percentage of RUL on average, but also has high uncertainty, meaning it may in fact never

predict the exact true RUL for a particular unit. This inference is confirmed by examination of the end of life binned error as the average model prediction value departs from the true RUL line at around 62% of life consumed (38% RUL) and loses even the 95% prediction interval coverage at around 85% of life consumed (15% RUL). Because of the fact that this is a strictly time based model, this helps to confirm that the model is unable precisely predict individual systems' RUL, instead only calculating the average RUL over all historic systems. Although this chart of binned error is useful and contains a wealth of information, it does require some degree of examination and analysis in order to compare different model sets. The other percentage based prediction metrics provide that analysis.

The effective bias for this model, as calculated by the Weighted Error Bias (WEB) from Equation 2 is 5.86%. Again, this can be seen in the binned error analysis as the average estimation line begins to deviate from the true RUL line particularly near the end of life. For this system, that means that there is an effective average bias of about 16 hours on average, but this does not mean that the expected error is 16 hours. This value, as well as the Weighted Prediction Spread (WPS), is considered an effective value because of their applied weighting function shown in the figure as a magenta dotted line, which allows them to be more effective at ranking the predictions. If for some reason, the more literal average values are needed, the same equations and metrics can be applied with a simple adjustment of the weighting function. This prediction set's WPS is listed as 58.07% of life, reflecting the fact that there is a considerable amount of uncertainty associated with the predictions.

The final two metrics listed are the Confidence Interval Coverage (CIC) and the Convergence Horizon (CH). Reported at 83% and 0% respectively, these indicate that although the model uncertainty covers the true RUL 83% of the time, it never continuously falls within 10 of that true value towards the end of the unit's life.

All these metrics can be combined according to Equation 7 in order to give this model's prediction set a total ranking of 54.83% out of a possible total score of 100%. This should not be read as an indication that the model's total accuracy is around 50% or that only 50% of the model's estimations are trust worthy. Instead this metric shows a quantitative evaluation of the model's performance for this prediction set. It is a quick and relatable evaluation of the model's "goodness" which can easily be used to compare against other models, or other prediction sets. For example, if Model 1 is compared to Model 2 shown in Figure 3, one can quickly see that Model 2 has a total performance score of 75.02%, much better than Model 1's 54.83%.

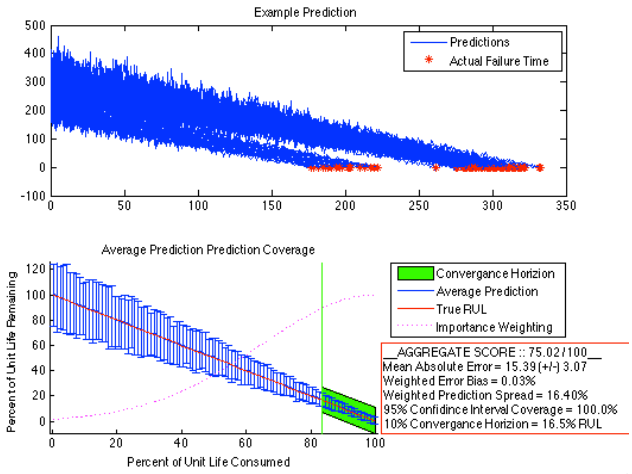

Figure 3 - Model 2 Predictions and Metrics Evaluation

Looking at the individual metrics, it becomes clear why this model is ranked better. First, it has 100% CIC with a 16.5% CH meaning that not only is the model more accurate overall, but it also shows that the accuracy improves near end of life. Next the effective prediction spread is 16.4% of life, much lower than Model 1's WPS. Finally Model 2 has virtually 0% effective bias, meaning that all the predictions are centered on the true RUL.

Clearly, these metrics give a quick, effective, and qualitative method for comparing two different models, and if that were the only end goal the analysis could stop there. However, if there is opportunity to change and improve the models which created the prediction sets, then the scalar metrics alone may not give the complete picture. Consider the prediction set developed by Model 3 in Figure 4.

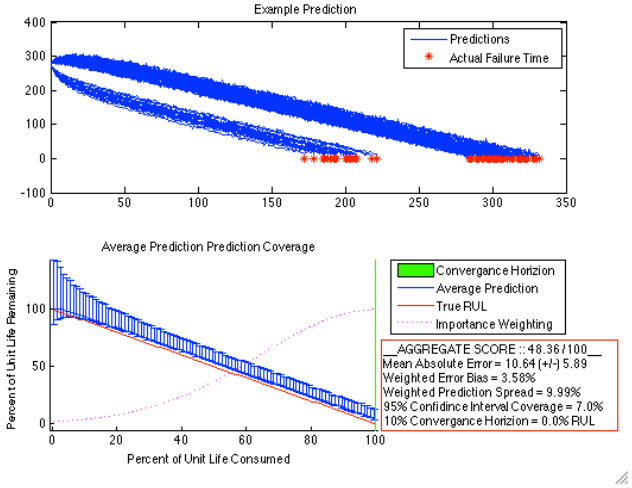

Figure 4 - Model 3 Predictions and Metrics Evaluation

Model 3 has a total performance score of 48.36%, indicating its' performance is worse than either of the two previous models. In fact, the only metrics which it out performs both of the other models are MAE and the WPS. Unfortunately, these alone would not necessarily merit further investigations into the development of this model. However,

when the total binned prediction value map is investigated, it becomes instantly clear that by removing a small bias in this model, these predictions would be expected to out perform both of the previous models. This same conclusion could be inferred from the scalar metrics, but a graphical examination of the binned values map is both more expedient and informative.

5. SUMMARY AND CONCLUSION

The scalar metrics presented in this work help to provide clear and concise evaluations of the performance of prognostic models in a manner easily accessible and largely intuitive to audiences with various backgrounds and expertise. In order to demonstrate and visualize the underlying meanings of each of the metrics, three separate sets of predictions made from three separate simulated prognostic models were compared. From the results listed in Table 1 it is clear that Model 2 is the best performing model by a large margin.

Table 1 – Summary of Model Comparison Results

	Total Score	MAE	WEB	WPS	CIC	CCH
M1	54.83%	35.09 Hrs	5.06%	57.80%	83.0%	0%
M2	**75.02%**	15.39 Hrs	0.03%	16.40%	100%	16.5%
M3	46.36%	10.64 Hrs	3.58%	9.99%	7.0%	0%

Further, Model 3 shows great potential for improvement via a simple bias removal as can be inferred from the low Weighted Prediction Spread (WPS) coupled with the results of the binned prediction value map. A quick summary of each metric is listed below in Table 2.

Table 2 - Metrics Summary

Metric Name	Quality Aspect Reflected	Units
Mean Absolute Error (MAE)	Precision *Average distance from true value*	Real Time Units
Weighted Error Bias (WEB)	Timely Precision *Scaled expected distance from true value*	Percent of Unit Life *Weighted by Lifetime Importance*
Weighted Prediction Spread (WPS)	Timely Accuracy *Scaled uncertainty estimate associated with each prediction*	Percent of Unit Life *Weighted by Lifetime Importance*
Confidence Interval Coverage (CIC)	Accuracy *How often the estimated uncertainty contains the true value*	Percent of Unit Life
Confidence Convergence Horizon (CCH)	Timely Accuracy & Precision *What part of life can all remaining estimates be trusted to within 10%*	Percent of Unit Remaining Useful Life
Binned Prediction Value Map	Timely Accuracy & Precision *Detailed visualization of the evolution of the prognostic predictions. Used to calculate other metrics*	Percent of Unit Life

These novel metrics build upon natural aspects of the prediction data itself to create meaningful and intuitive representations of performance. The goal of this work is to learn from previously introduced metrics and create a set of generic metrics that can be widely used and understood in both academic and industrial settings. All of the metrics detailed in this work can be easily calculated and widely applied and interpreted across many cases allowing for un-obscured, evaluation of predictions from a wide variety of algorithms and methodologies. The balance between case specific adaptability and overall standardization is an area of continual interest and research. This work seeks to provide a set of metrics that provide a level of both in a manner that is accessible and relatable to a wide audience to help promote investigation and collaboration on prognostic projects across many fields.

REFERENCES

Banks, J., J.Merenich. "Cost Benefit Analysis for Asset Health Management Technology". Reliability and Maintainablity Symposium (RAMS), Orlando, Florida. 2007

Coble, Jamie, "Merging Data Sources to Predict Remaining Useful Life – An Automated Method to Identify Prognostic Parameters," Doctorial Dissertation, University of Tennessee, Knoxville TN. 2010

Leao, B.P., J.P.P.Gomes, R.K.H.Galvaro, and T.Yoneyama. "How to Tell the Good from The Bad in Failure Prognostics". IEEE Aerospace Conference Proceedings. 2010

Orchard, M., G.Kacprzynski, K.Goebel, B.Saha, and G.Vachtservanos. "Advances in Uncertainty Representation and Management for Particle Filtering Applied to Prognostics". International Conference on Prognostics and Health Management, 2008.

Saxena, Abhinav, Jose Celaya, E. Balaban, B. Saha, S. Saha, and K. Goebel, "Metrics for evaluating

performance of prognostic techniques". International Conference on Prognostics and Health Management (PHM08), Denver CO, pp. 1-17, 2008

Saxena, Abhinav, Jose Celaya, Bhaskar Saha, Sankalita Saha, and Kai Goebel. "On Applying the Prognostic Performance Metrics." Annual Conference of the Prognostics and Health Management Society (2009)

Saxena, Abhinav, Jose Celaya, Bhaskar Saha, Sankalita Saha, and Kai Goebel. "Metrics for Offline Evaluation of Prognostic Performance". International Journal of Prognostics and Health Management. ISSN 2153-2648, 2010 001. April 2010.

Tang, Liang, Marcos E.Orchard, Kai Gobel, George Vachtevanos, "Novel Metrics for the Verification and Validation of Prognostic Algorithms". Aerospace Conference 2011 IEEE, Big Sky, MT. 5 -12 March 2011.

Uckun,S., K.Goebel, and P.J.F.Lucus. "Standardizing Research Methods for Prognostics. International Conference on Prognostics and Health Management (PHM08). Denver CO. 2008