# The Impact of Metrics on the Choice of Prognostic Methodologies

Bieber, M.T.; Verhagen, W.J.C.; Santos, Bruno F.

**DOI**
10.2514/6.2022-3966

**Publication date**
2022

**Document Version**
Final published version

**Published in**
AIAA AVIATION 2022 Forum

AIAA AVIATION Forum
June 27-July 1, 2022, Chicago, IL & Virtual
AIAA AVIATION 2022 Forum

10.2514/6.2022-3966



# The Impact of Metrics on the Choice of Prognostic Methodologies

Marie Bieber*
*Delf University of Technology, Delft, 2629HS, The Netherlands*

Wim J.C. Verhagen†
*RMIT University, Carlton, Victoria, 3053, Australia*

Bruno F. Santos‡
*Delft University of Technology, Delft, 2629HS, The Netherlands*

**Metrics play an important part in the development and application of prognostic methodologies as they provide the capability to characterize and assess the quality of remaining useful life predictions. Although there is a wide range of both, prognostic metrics and prognostic methodologies available, the choice of those often is a demanding and time consuming task. Additionally, they are often treated as two separate problems to solve, while the choice of metrics has an impact on the choice of prognostic methodology and vice versa. In this paper, we therefore present a framework with the capability to automatically choose prognostic settings given specific system data to account for five different prognostic metrics. We then apply this framework to an aircraft data set to characterize the impact of metrics on the choice of prognostic methodologies. The results show that the choice of optimization metric has a big impact on the output of the generic prognostic framework and on the overall prognostic performance.**



## I. Introduction

The development and application of data-driven prognostic approaches has seen a growth in past years. Within the framework of Condition-Based Maintenance (CBM) prognostics enable assessment of equipment health and prediction of the remaining useful life (RUL) [1]. This not only improves system reliability, safety, and availability, but also reduces the life-cycle operational costs of components [2], [3]. However, applying prognostics within a CBM framework for applications such as aircraft maintenance, also brings with it some challenges [4]. One of those challenges is the selection of appropriate prognostic algorithms. This can be time consuming and requires a lot of expert knowledge. Another challenge is the assessment of the performance of the prognostic models, which is crucial considering that the output of the models is used for decision making in a CBM framework. Often those challenges are addressed separately. However, it can be difficult to tune prognostic algorithms without understanding which metrics are needed to assess the algorithm. Similarly, it can be tricky to understand the full impact of choosing prognostic metrics without taking into account the prognostic algorithm.

In many cases the development of a prognostic framework starts with a specific data set to which feature engineering methods and prognostic algorithms are tailored, ultimately resulting in increasingly better remaining useful life estimates. This process not only requires a lot of expertise and technical knowledge, but also in some cases translates to years of research conducted. To address this issue we have developed a framework to support the identification of applicable prognostic methods and the automatic configuration of the method settings, given a specific data set. Prior studies of such frameworks have yielded promising results. An autonomous diagnostics and prognostics framework is suggested by [5] that consisting of several steps, including the data pre-processing, clustering to distinguish operating conditions and the diagnostics and prognostics. Several parameters, including the number of observations for initialisation and optimization of cluster adaption rates, have to be set manually and it can be tricky to tune the algorithm in an optimal way. To account for this, [6] provide a generic prognostic framework that can be instantiated to various applications. However, no specific machine learning algorithms are used in this framework and it is more of a guideline as to how

*PhD Candidate, Faculty of Aerospace Engineering, Delft University of Technology, Delft, 2629HS, The Netherlands.
†Senior Lecturer, Aerospace Engineering and Aviation, RMIT University, Carlton, Victoria, 3053, Australia.
‡Associate Professor, Air Transport and Operations, Faculty of Aerospace Engineering, Delft University of Technology, Delft, 2629HS, The Netherlands.

to build a generic prognostic framework. The authors in [7] suggest a prognostics method based on an ensemble of genetic algorithms that includes most of the steps, from feature engineering until the RUL estimation. One limitation of their proposed method is that the optimization and selection of methodologies is based on a commonly used metric in prognostics, the mean-squared error (MSE), but it is not further evaluated for robustness across multiple error metrics.

After finding a suitable prognostic algorithm for the system data, the next step is to use the prognostic output as an input for the decision making step in CBM. Using prognostics in such a context requires a proper assessment of the quality of predictions. A metric, such as the MSE, can arguably not characterize the quality of RUL predictions sufficiently for this purpose [8]. Instead, the design of prognostic metrics has to be linked to the application and decision making process [9]. In addition, as highlighted in Figure 1 metrics are needed to define requirements and thoroughly evaluate prognostic performance [10].

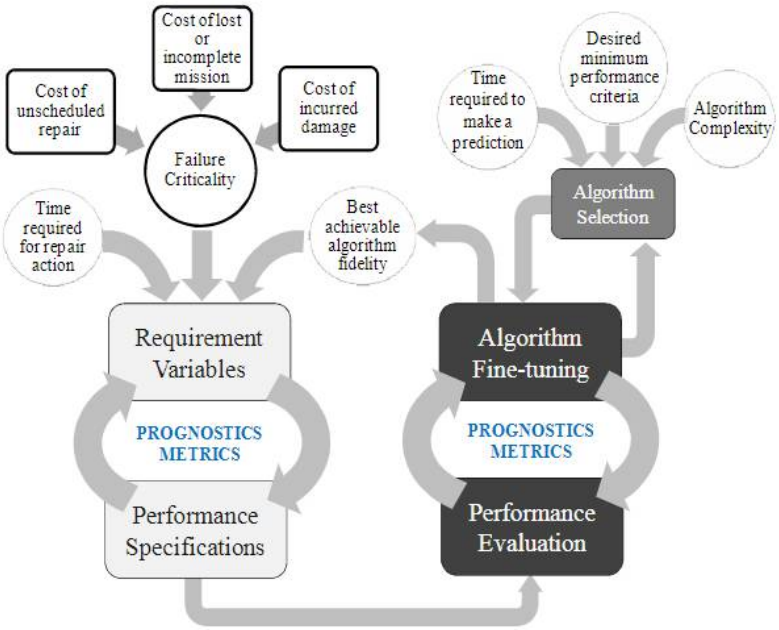


**Fig. 1 Prognostic metrics are needed to define requirements and evaluate performance [10]**

An effort to standardize prognostic metrics has been made by Saxena et al. in [11], [12]. The metrics commonly used in prognostics are highlighted and several ways to classify them are presented as well as ways in which to interpret and use the metrics. Goebel et al. state in [13] that a meaningful prediction has three attributes, namely correctness (measured by accuracy and prediction), timeliness, and confidence. Performance evaluation of prognostic methodologies should enhance all three of those aspects. However, the vast majority of literature published in the field of prognostics uses only one metric, which is often one linked to correctness of the method. To address this gap, we present a generic prognostic framework (GPF) with the capability of automatically choosing prognostic settings by optimizing the prognostic performance in terms of specified prognostic metrics. This framework is applied to an aircraft data set and a sensitivity analysis is conducted to understand the impact of the choice of metrics on the prognostic settings themselves. Or put differently, the question we ask is: How can various prognostic metrics be integrated in a prognostic framework to guide the choice of suitable prognostic methodologies based on a thorough assessment of the quality of the predictions?

## II. Methodology

A generic prognostic framework is suggested, which contains representative techniques for the sequential steps of a data-driven prognostics approach and, given a data set, selects the best techniques to be used for each case. This means that in addition to incorporating different methodologies, the framework includes a selection step in which the best set of techniques is chosen. Note that the essence of the work as presented in this paper lies in assessing and optimizing

the set of prognostic techniques. The way we measure and evaluate the chosen techniques defines the prognostic settings and in further consequence the quality of the predictions. In order to evaluate the prognostic performances, we therefore use different prognostic metrics to account for different aspects of prediction evaluation. Those metrics integrated in the GPF give us insight into the quality of predictions and thereby help to choose appropriate prognostic methods.

The generic prognostic framework consists of two phases (Figure 2). In phase one, which is highlighted in green, a Genetic algorithm is applied to find the optimal prognostic settings. This is done using each of the five selected metrics, which are explained in more detail in Section II.A. In phase two, highlighted in red and further explained in Section II.B, for each of the four possible prognostic settings, a prognostic model is trained, which then has the capability to output RUL estimates.

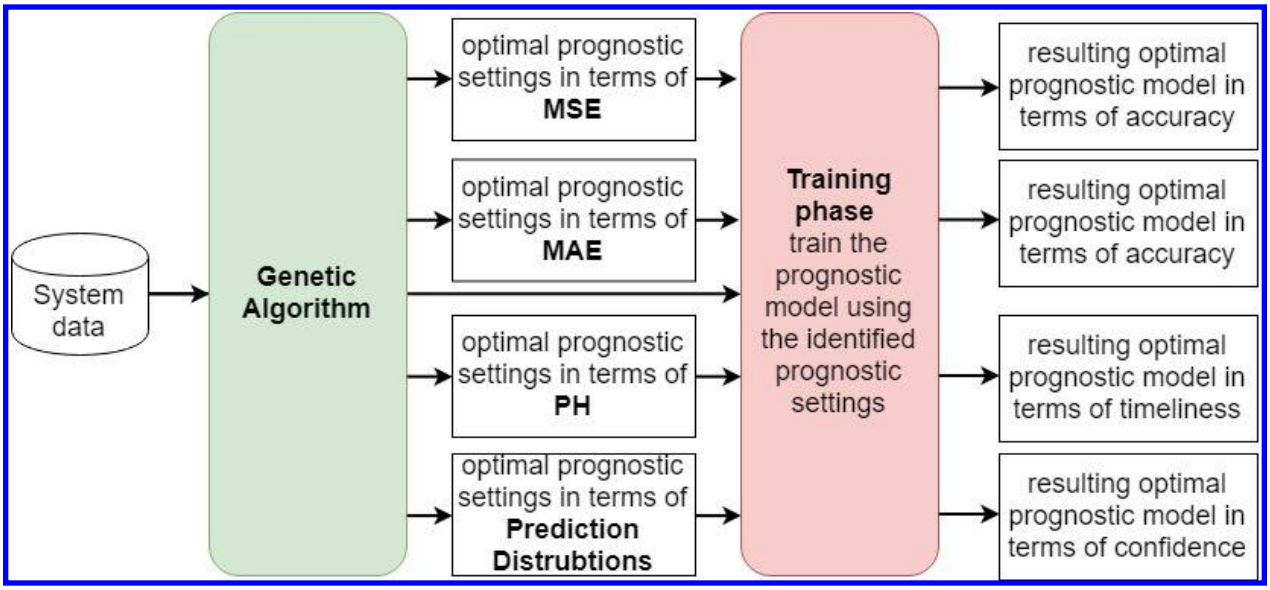


**Fig. 2 The generic prognostic framework flow.**



### A. Generic Prognostic Framework

Prognostics involves several steps, including data pre-processing and feature engineering methodologies. All of those contribute to the quality of predictions. The generic prognostic framework therefore consists of three blocks, as displayed in Figure 3, which are simultaneously optimized using a genetic algorithm. The framework is a modified version of the generic prognostic framework presented in [14]. In the following subsections II.A.1, II.A.2 and II.A.3 we describe the three basic modules of the GPF shown in Figure 3 and then in subsection II.A focus on the optimization metrics used in the framework.

#### *1. Data Rebalancing*

Imbalanced data sets can have an impact on the quality of predictions, especially when system failures are rare as in the case studies presented in this work. Since we are estimating RUL and therefore have to solve a regression problem, it is not so straightforward to implement data re-balancing methods as for classification problems. However, such methods have been explored in literature and in our framework we use the techniques described by [15]. Among those, we use

- Random Over-Sampling (RO),
- Introduction of Gaussian Noise (GN) and
- Weighted relevance-based combination strategy (WERCS).

While we do not go into details about these methods and refer interested readers to [15], we introduce the underlying basic concepts as follows. The main idea behind re-balancing methods for continuous target variables is the construction of bins based on a relevance function. The relevance function maps the values of the target variable into a range of importance, where 1 corresponds to maximal importance and 0 to minimum relevance. With this, the bins classify the data in normal ($BIN_N$) and relevant samples ($BIN_R$). In our setup, we use a sigmoid relevance function as defined in

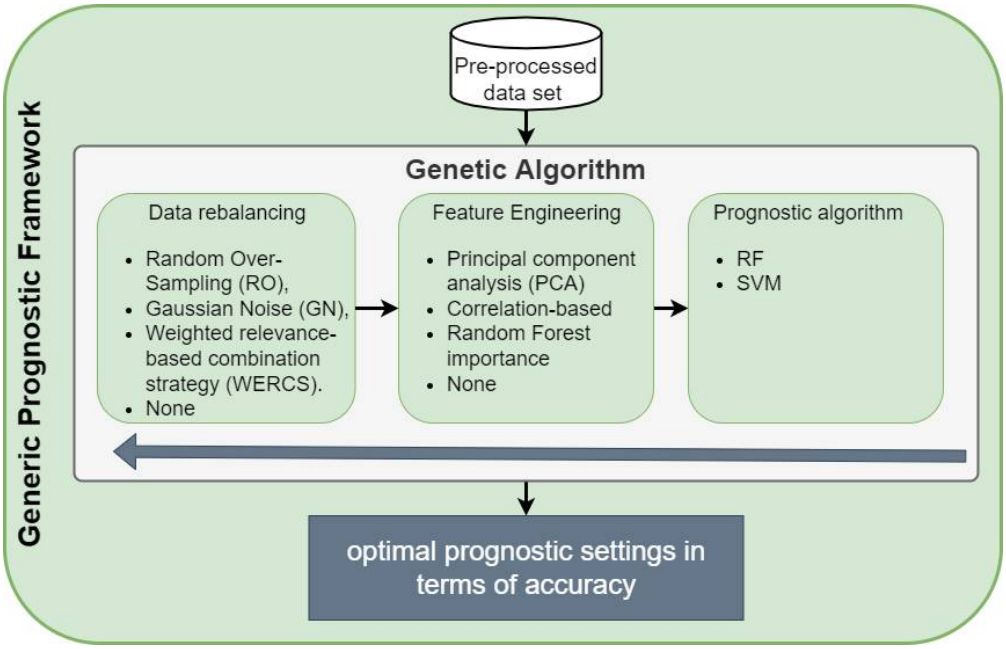


**Fig. 3 The elements of the Generic Prognostic Framework.**



[16] with a relevance threshold, $t_r$ of 0.4. Furthermore, we set all values with a RUL of less then the threshold $cl = 10$ to be of importance.

*2. Feature Engineering*

Feature engineering in general describes the manipulation and transformation of features, i.e. in our case sensor data, before applying a prognostic algorithm to it. The most commonly used feature engineering techniques in prognostics are feature extraction and feature selection.

Feature extraction is performed to obtain useful information from raw signals [17]. Since the scope of this analysis are RUL estimation models for mechanical or electrical systems with run-to-failure data, it is assumed that underlying signals come in the form of time-series data. The simplest way to handle time series data is by calculating characteristic features as descriptive statistics from the data themselves. Of the existing methodologies, we chose to use Principal Component Analysis (PCA). PCA is a widely used technique making use of the singular value decomposition of the data to project it to a lower dimensional space.

Feature selection is identifying features that help finding faults in the monitored systems [18]. According to [19], feature selection techniques can be classified into four categories: Filter approach, in which features are selected without the use of a learning algorithm, wrapper approaches, in which learning algorithms are used to evaluate accuracy produced by the selected features, embedded approaches, where features are selected during training and which are specific to machine learning algorithms and hybrid approaches, which are a combination of filter and wrapper approaches. In the GPF, we include a filter and an embedded approach. The filter approach is a correlation based approach, which chooses the best features based on univariate statistical tests. The embedded approach is based on the random forest importance, i.e. it chooses the most important features identified by a random forest estimator.

*3. Prognostic Algorithms*

In order to get a first prognostic assessment through the framework, the prognostic algorithms included are a Random Forest Regression (RF) and a Support Vector Regression (SVM). The two selected algorithms are well-established and offer potential advantages in terms of interpretability and explainability, which is necessary to understand systems retrospectively and prospectively [20].

*4. Optimization Metrics*

The GPF selects an optimal methodology with the optimal hyper parameter settings for each step in the prognostic framework. Here, 'optimal' refers to the best in terms of a specified metric. In other words, we treat the problem of finding the prognostic settings as an optimization problem: The objective function is to minimize the accuracy (given by the specified metric) of the prognostic algorithm together with data re-balancing and feature engineering techniques on the pre-processed data set. To solve the optimization problem, we use a genetic algorithm (GA). These algorithms are based on the concepts of natural selection and genetics [21]. Due to their flexibility, GAs are able to solve global optimization problems and optimize several criteria at the same time, like in our case the simultaneous selection of data re-balancing, feature engineering and prognostic algorithm techniques [22]. This is what makes them good candidates for our optimization problem. The GA in our case takes as an input the system data and the selected metric and outputs the optimal combination of data re-balancing technique, feature engineering methodology and prognostics algorithm. Note that, in case it identifies that applying no re-balancing or no feature engineering technique results in better prognostic outputs, the GPF returns 'None' for the according block.

To assess the importance of metrics in the context of prognostics, five different metrics are therefore implemented and tested in the framework, namely the mean squared error (MSE), mean absolute error (MAE), prognostic horizon (PH), the alpha-lambda metric and the predicted distributions. The metrics account for the three attributes of meaningful predictions, i.e. correctness (MSE and MAE), timeliness (PH) and confidence (alpha-lambda metric and distribution of predictions) [23].

The MSE at time $t$ it is given as

$$MSE(t) = \frac{1}{t}\sum_{i=1}^{t}(RUL_i - \hat{RUL}_i)^2, \tag{1}$$

where $RUL_i$ is the true RUL value and $\hat{RUL}_i$ the predicted RUL value at timestep $i$.

Similarly, the MAE at time $t$ is given as

$$MAE(t) = \frac{1}{t}\sum_{i=1}^{t}|RUL_i - \hat{RUL}_i|, \tag{2}$$

with $RUL_i$ and $\hat{RUL}_i$ as defined above.

The prognostic horizon (PH) is defined as

$$PH(t, \alpha) = RUL_{true}(t_{i_\alpha}), \tag{3}$$

with $RUL_i$ the true RUL at time index $t_{i_\alpha}$ and $i_\alpha := min\{j \in p | \alpha^- \leq \hat{RUL}(j) \leq \alpha^+\}$, where

- $p$ is the set of all time indices where predictions are made,
- $\hat{RUL}_j$ is the prediction at time index $j$
- and the alpha bounds are defined as $\alpha^- := RUL_i - \alpha$ and $\alpha^+ := RUL_i + \alpha$.

The prognostic horizon is therefore the smallest RUL in which the predicted RUL is still within the specified $\alpha$ bounds and in our case study we set $\alpha = 40$ flight cycles, which is the time needed to schedule maintenance for an aircraft in case it is needed.

And finally in order to define the distribution of predictions, we first need the definition of the $\alpha - \lambda$ metric which is in [24] defined as

$$\alpha - \lambda := \begin{cases} 1, & if(1-\alpha)\lambda^* \leq \lambda_p \leq (1+\alpha)\lambda^* \\ 0, & otherwise, \end{cases} \tag{4}$$

with $\lambda^*$ the ground truth, $\lambda_p$ the prediction and $\alpha$ an arbitrary chosen accuracy. The $\alpha - \lambda$ metric therefore measures if the prediction accuracy of the RUL model is within $\alpha$% error at a specific time instance during the life of the system. It can be evaluated and averaged over the whole trajectory with $N$ time steps.

**Table 1 The hyper parameters and combination of settings explored during the grid search for each of the four prognostic algorithms.**

| Prognostic algorithm | Hyper parameter | Description | Possible settings |
|---|---|---|---|
| rf | n estimators | number of trees | {200, 800, 1400} |
| | max features | maximum number of features to consider when looking for the best split | {'auto', 'sqrt', 'log2'} |
| | min samples leaf | minimum number of samples required to be at a leaf node | {1, 2, 4} |
| SVM | C | learning rate | {0.001, 0.01, 0.1, 10} |
| | gamma | kernel coefficient | {0.001, 0.01, 0.1, 1} |



The authors also use the probabilistic version of the $\alpha - \lambda$ metric, for which we fit a Gaussian distribution $\mathcal{N}(\mu, \sigma)$, to each prediction and calculate the probability of the given prediction being inside the $\alpha$ boundaries. If $\mathbf{F}(x, \mu, \sigma)$ is the according cumulative distribution function, the probabilistic $\alpha - \lambda$ for a single prediction is given as

$$\mathbf{P}_{\alpha-\lambda} = \mathbf{F}((1+\alpha)\lambda^*, \mu, \sigma) - \mathbf{F}((1-\alpha)\lambda^*, \mu, \sigma). \tag{5}$$

The final metric is derived by averaging the probabilistic $\alpha - \lambda$ over the entire trajectory. The authors specify 20% to be a commonly chosen value for $\alpha$, which is therefore also what we use in this study.

### B. Training Phase

Once the Genetic Algorithm outputs are created, the training phase starts. In the training phase the prognostic models are trained using the identified prognostic settings. A grid search is used to find the optimal hyper parameter settings for the identified prognostic technique, which can either be random forest regression (rf) or support vector machine (SVM). The according hyper parameters and their possible settings explored during the grid search are given in Table 1. The so found settings are the ones then used as the settings for the prognostic algorithms which are than trained on the underlying prognostic data set. The output of this step is a trained prognostic model, which takes as an input system data and outputs the remaining useful life (RUL).

## III. Results

The aim of the conducted case study is to understand the impact of prognostic metrics on the methodology selection in the different steps of the prognostic framework. This leads to the further question of how prognostic metrics can guide the choice of suitable prognostic methodologies. The five prognostic metrics introduced in section II are in turn used as optimization metric to find a set of methodologies to train the prognostic model.

For this purpose, the framework is applied to the C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) data set, containing simulated run-to-failure data for turbofan engines [25] [26]. Using this tool, 4 data sets were created. The data sets differ mainly in the number of fault modes and operational conditions simulated in the experiments. An overview is given in Table 2. The relative number of train and test units is the ratio between the number of train/ test units and the total data set size and gives an indication over the bias in the data set towards healthy behaviour. Each engine is considered to be from a fleet of engines of the same type and each time series, also often referred to as trajectory, is from a single unit. The engines are operated until failure, i.e., the time series capture the operations of each unit until it fails. Each row in the data set contains measurements during one time cycle of 21 sensors for a certain unit.

In the following, we present the results of applying the generic prognostic framework on each of the four C-MAPSS data sets with the different optimization metrics. We run the genetic algorithm for 10 generations with population sizes of 30 and 50 individuals to capture the effects of the optimization process. First, we present the output of the genetic

**Table 2 Characteristics of the four turbofan engine data sets [27]**

| Data set | #Fault modes | #Conditions | #Train units | #Test units | relative #Train units | relative #Test units |
|---|---|---|---|---|---|---|
| #1 | 1 | 1 | 100 | 100 | 0.485% | 0.485% |
| #2 | 1 | 6 | 260 | 259 | 0.484% | 0.762% |
| #3 | 2 | 1 | 100 | 100 | 0.405% | 0.603% |
| #4 | 2 | 6 | 249 | 248 | 0.407% | 0.602% |

algorithm, i.e. the choices of methodologies for each of the four data sets. Table 3 contains the choice of methodologies for the data rebalancing, feature engineering and prognostic algorithm when using the MSE, the MAE, the alpha-lambda score, the prognostic horizon or the prediction distributions respectively for runs on data set FD001.

**Table 3 The resulting best prognostic settings found for different optimization metrics when running the GPF for data set FD001.**

| **Optimization Metric** | **Population size of 30/ 50** | | |
|---|---|---|---|
| | **Choice of Method for** | | |
| | **Rebalancing** | **Feature engineering** | **Prognostic Algorithm** |
| **MSE** | RO | None | RF |
| **MAE** | None | None | SVM |
| **alpha-lambda** | RO | correlation | SVM |
| **PH** | RO | correlation | SVM |
| **Prediction Distributions** | RO | correlation | SVM |

Similarly, the results for the runs on data set FD002 result in the choice of methodologies shown in Table 4 and the results for those on data set FD003 in Table 5. In Table 6 the resulting choices of methodologies of applying the generic prognsotic framework to data set FD004 are presented for the different optimization metrics. The according results in terms of metrics are given in Tables 7, 8, 9 and 10.

**Table 4 The resulting best prognostic settings found for different optimization metrics on data set FD002.**

| **Optimization Metric** | **Population size of 30/ 50** | | |
|---|---|---|---|
| | **Choice of Method for** | | |
| | **Rebalancing** | **Feature engineering** | **Prognostic Algorithm** |
| **MSE** | GN | None | RF |
| **MAE** | GN | None | RF |
| **alpha-lambda** | RO | PCA | SVM |
| **PH** | WERCS | PCA | RF |
| **Prediction Distributions** | RO | PCA | SVM |

It can be observed that the choices of methodologies for each of the five optimization metrics are consistent when running the GA with a population size of 30 or 50. This stability of the results over different population sizes is an indication of the stability of the generic prognostic framework regarding the choices of methodologies.

An interesting question to ask now is: How is the choice of optimization metric reflected in the stability of the genetic algorithm? There are two main points that can be observed from the above Tables 3, 4, 5 and 6 to help answer the above

**Table 5 The resulting best prognostic settings found for different optimization metrics when running the GPF for data set FD003.**

| Optimization Metric | Population size of 30 Choice of Method for | | | Population size of 50 Choice of Method for | | |
|---|---|---|---|---|---|---|
| | Rebalancing | Feature engineering | Prognostic Algorithm | Rebalancing | Feature engineering | Prognostic Algorithm |
| **MSE** | None | importance | SVM | GN | importance | SVM |
| **MAE** | None | importance | SVM | GN | importance | SVM |
| **Alpha_lambda** | None | PCA | RF | None | PCA | RF |
| **PH** | None | PCA | RF | None | PCA | RF |
| **Prediction Distributions** | GN | PCA | RF | GN | PCA | RF |

**Table 6 The resulting best prognostic settings found for different optimization metrics when running the GPF for data set FD004.**

| Optimization Metric | Population size of 30 Choice of Method for | | | Population size of 50 Choice of Method for | | |
|---|---|---|---|---|---|---|
| | Rebalancing | Feature engineering | Prognostic Algorithm | Rebalancing | Feature engineering | Prognostic Algorithm |
| **MSE** | None | None | RF | None | None | RF |
| **MAE** | None | None | RF | None | None | RF |
| **alpha-lambda** | RO | PCA | SVM | RO | PCA | SVM |
| **PH** | GN | PCA | SVM | None | PCA | SVM |
| **Prediction Distributions** | RO | PCA | SVM | RO | PCA | SVM |

**Table 7 The resulting metrics found for different optimization metrics when running the Genetic Algorithm for data set FD001.**

| Population size | Optimization Metric | Metrics after training algorithm on full dataset | | | | |
|---|---|---|---|---|---|---|
| | | MSE | MAE | Alpha_lambda | PH | Prediction Distributions |
| **30/50** | **MSE** | 1657.17 | 30.74 | 0.5244 | 141.77 | 0.3423 |
| | **MAE** | 1775.05 | 31.02 | 0.5312 | 130.62 | 0.3388 |
| | **alpha-lambda** | 2734.27 | 39.87 | 0.4126 | 95.64 | 0.2516 |
| | **PH** | 2734.27 | 39.87 | 0.4126 | 95.64 | 0.2516 |
| | **Prediction Distributions** | 2734.27 | 39.87 | 0.4126 | 95.64 | 0.2516 |

question:

First, choosing the $\alpha - \lambda$ metric as optimization metric produces in most cases the same results as optimizing towards the prediction distributions. In fact, they result in the same settings for the methodology choices except for the runs on data set FD003, where the combination of No rebalancing, PCA and RF and GN, PCA and RF result in almost the same score in terms of $\alpha - \lambda$ metric (0.3112 and 0.3173 respectively) and in the same score in terms of prediction distribution (0.1895), as can be seen in Table 9.

Second, often the $\alpha - \lambda$ metric, the prediction distributions and in quite some cases also the PH produce similar outputs

**Table 8 The resulting metrics found for different optimization metrics when running the Genetic Algorithm for data set FD002.**

| Population size | Optimization metric | Metrics after training algorithm on full dataset | | | | |
|---|---|---|---|---|---|---|
| | | MSE | MAE | Alpha_lambda | PH | Prediction Distributions |
| 30/50 | **MSE** | 1877.88 | 33.65 | 0.4633 | 118.75 | 0.3010 |
| | **MAE** | 1877.88 | 33.65 | 0.4633 | 118.75 | 0.3010 |
| | **alpha-lambda** | 18875.29 | 122.11 | 0.0135 | 168.93 | 0.02401 |
| | **PH** | 4281.47 | 50.42 | 0.3316 | 93.79 | 0.1845 |
| | **Prediction Distributions** | 18875.29 | 122.11 | 0.0135 | 168.93 | 0.02401 |

**Table 9 The resulting metrics found for different optimization metrics when running the Genetic Algorithm for data set FD003.**

| Population size of | Optimization Metric | Metrics after training algorithm on full dataset | | | | |
|---|---|---|---|---|---|---|
| | | MSE | MAE | Alpha_lambda | PH | Prediction Distributions |
| 30 | **MSE** | 4284.62 | 45.83 | 0.4632 | 151.28 | 0.2974 |
| | **MAE** | 4284.62 | 45.83 | 0.4632 | 151.28 | 0.2974 |
| | **alpha-lambda** | 7207.84 | 64.13 | 0.3173 | 101.46 | 0.18957 |
| | **PH** | 7207.84 | 64.13 | 0.3173 | 101.46 | 0.18957 |
| | **Prediction Distributions** | 7266.72 | 64.45 | 0.3112 | 103.22 | 0.1895 |
| 50 | **MSE** | 4670.12 | 47.35 | 0.4695 | 138.33 | 0.2961 |
| | **MAE** | 4670.12 | 47.35 | 0.4695 | 138.33 | 0.2961 |
| | **alpha-lambda** | 7207.84 | 64.13 | 0.3173 | 101.46 | 0.18957 |
| | **PH** | 7207.84 | 64.13 | 0.3173 | 101.46 | 0.18957 |
| | **Prediction Distributions** | 7266.72 | 64.45 | 0.3112 | 103.22 | 0.1895 |



of the framework especially in terms of prognostic algorithm and the same is true for MSE and MAE. Of course this has to do with how the metrics are defined and their similarities or differences within the definitions, but still, it reflects back to different outcomes for different prediction evaluation criteria, e.g. optimizing for correctness results in different prognostic settings than optimizing for timeliness. It also shows a stability of the framework even with respect to similar metrics. This brings us to another important observation, more focused on the results presented in Table 7-10, i.e. on the measurable quality of the predictions for the different metric scenarios.

In many cases in order to optimize the prognostic output with respect to one metric, it is done to the price of bringing another metric up quite high. This can be even more clearly seen in Figure 4. Since the definition of the metrics targets different objectives, the choice of optimizing a metric is always a trade off. Those objectives are the prediction attributes highlighted in section II.A.4 and pointed out by [28]. Therefore, this is not only an indication of the impact of the choice of metrics on the selection of optimal prognostic methodologies, but also shows the impact of optimizing towards different prediction attributes on the choice of prognostic methodologies. Figure 4 shows that the effect is especially big in data sets FD002 and FD004. In Table 8, we see indeed that a very low lambda alpha score and Prediction distribution (of 0.0135 and 0.02401), in other words high confidence, can be achieved, but only with a high MSE and MAE (of 18875.29 and 122.11), i.e. lower prediction correctness. Vice versa the same is true when lowering the MSE and MAE (1877.88 and 33.65), i.e. increasing the correctness, which at the same time increases the lambda alpha score and

**Table 10 The resulting metrics found for different optimization metrics when running the Genetic Algorithm for data set FD004.**

| Population size | Optimization Metric | Metrics after training algorithm on full dataset | | | | |
|---|---|---|---|---|---|---|
| | | MSE | MAE | Alpha_lambda | PH | Prediction Distributions |
| 30 | **MSE** | 4559.05 | 50.12 | 0.4076 | 168.32 | 0.2487 |
| | **MAE** | 4559.05 | 50.12 | 0.4076 | 168.32 | 0.2487 |
| | **alpha-lambda** | 28449.29 | 143.02 | 0.02176 | 175.27 | 0.02758 |
| | **PH** | 8999.52 | 69.04 | 0.3041 | 136.75 | 0.1451 |
| | **Prediction Distributions** | 28449.29 | 143.02 | 0.02176 | 175.27 | 0.02758 |
| 30 | **MSE** | 4559.05 | 50.12 | 0.4076 | 168.32 | 0.2487 |
| | **MAE** | 4559.05 | 50.12 | 0.4076 | 168.32 | 0.2487 |
| | **alpha-lambda** | 28449.29 | 143.02 | 0.02176 | 175.27 | 0.02758 |
| | **PH** | 4281.47 | 50.42 | 0.3316 | 93.79 | 0.1845 |
| | **Prediction Distributions** | 28449.29 | 143.02 | 0.02176 | 175.27 | 0.02758 |



prediction distribution (to 0.4633 and 0.3010), i.e. lowers the confidence. We see a similar behaviour for FD004 in Table 10. In some cases, however, the GPF fails to find the optimal solution in terms of the selected optimization metric, e.g. for data set FD001 as shown in Table 7, the combination of RO, No feature engineering and RF produces a in terms of MAE better solution than the by the GPF for this case returned option of using only SVM. This can be due to the fact that the optimal solution was simply not part of all the possible solutions explored by the Genetic Algorithm. The differences, in terms of MAE and MSE, are minor though (MAE of 30.74 and 31.02 respectively), especially when compared against the solutions found when using PH, the $\alpha - \lambda$ metric or prediction distributions as optimization metrics (resulting in an MAE of 39.87).

Another interesting way to evaluate predictions is by plotting the true RUL and the predicted values. In Figures 5, 6, 7 and 8 the resulting plots are shown for six selected trajectories from the test data sets of FD001, FD002, FD003 and FD004 respectively when running the framework with a population size of 50.

For FD001 the predictions closer to the RUL for the selected IDs at least are the once produced with the optimization metric set to MSE or MAE, the predictions produced by optimizing towards the PH and Prediction Distribution, i.e. timeliness and confidence, always perform worse and seem to be more unstable. For this dataset it would perhaps make sense to choose the MSE or MAE as optimization metric when applying the framework. Especially since doing otherwise results in a MSE of almost double the optimized as can be seen in Table 7 and 4.

The opposite is the case on data set FD003, where the Prediction Distribution and PH predictions both for most trajectories outperform the once produced by the MSE and MAE. In data set FD002 and FD004 no such trend is really visible and the overall performance of the predictions is worse. This is to be expected though, as they are the data sets on which it is harder to produce RUL estimates as shown in Table 2. In both, FD002 and FD004, using the prediction distributions to optimize the prognostic settings results in poorly performing prognostic models. For the other metrics the trend is not so clearly visible. While the MSE and MAE produce the same resulting outputs on those datasets, see Table 4 and 6, the prognostic horizon outperforms them on some trajectories (id 5 and 18), while on others it performs worse (id 24 and 92).

Throughout all the four data sets the following findings are made:

- In most cases, a population size of 30 individuals produces the in terms of optimization metric optimal solution when using the generic prognostic framework and usually it is consistent with the solution found for a population

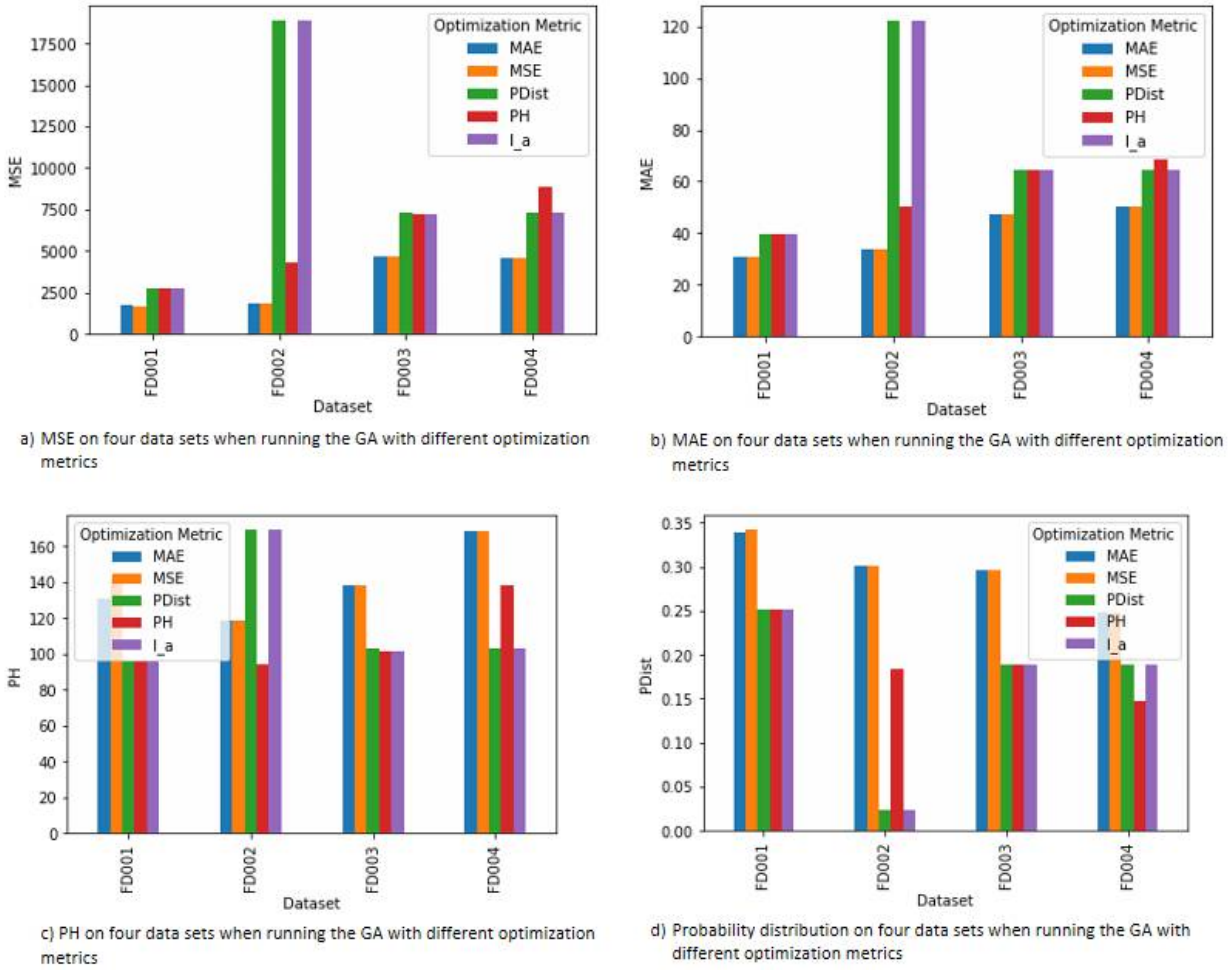


**Fig. 4 The resulting scores of each metric depending on the chosen optimization metric in the Genetic Algorithm when running with a population size of 50.**

size of 50 individuals.

- Within different population sizes, however, different optimization metrics lead to different prognostic settings with patterns according to the evaluation criteria the metric focuses on.
- The choice of the optimal metric is dependent on underlying data set and objective of prognostics, e.g. in what context they are used.
- A single metric does often not suffice in making informed and appropriate choices of prognostic methodologies
- Optimizing towards different prediction attributes, i.e. correctness, timeliness or confidence, results in different choices of prognostic methodologies and is often a trade-off.

## IV. Conclusion

The objective of our study is to understand the impact metrics have on prognostics. To account not only for different prognostic algorithms, but also for other steps involved in prognostics, such as data rebalancing and feature engineering, we use a generic prognostic framework which chooses the optimal settings for the three steps data rebalancing, feature engineering and prognostic algorithm, optimal with respect to a selected metric. The optimization metric is varied to reflect a selection of metrics, which account for all the aspects of prediction evaluation, including correctness (MSE and MAE), timeliness (PH) and confidence ( $\alpha - \lambda$ score and prediction distributions). The results show that the choice of optimization metric has a big impact on the output of the generic prognostic framework. This means that depending on the objective and motivation of using prognostics, a suitable metric should be carefully chosen. It could make sense to

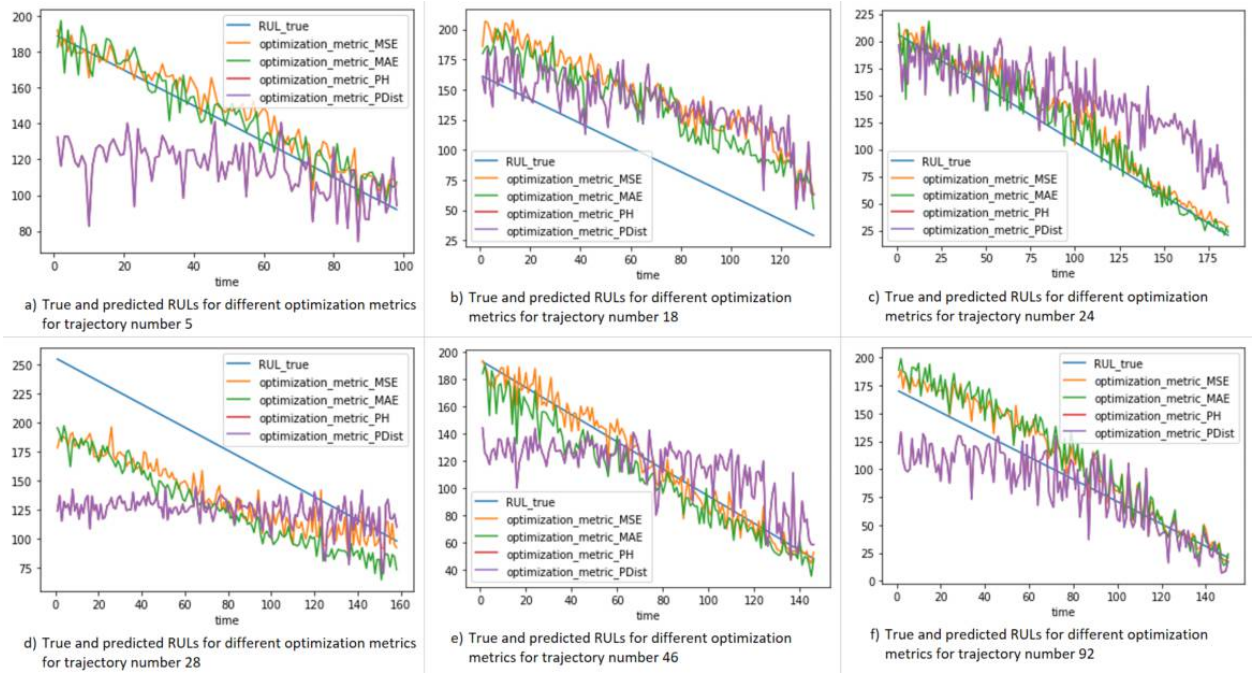


**Fig. 5 Predictions for different settings of optimization metrics on example trajectories for a population size of 50 on data set FD001.**

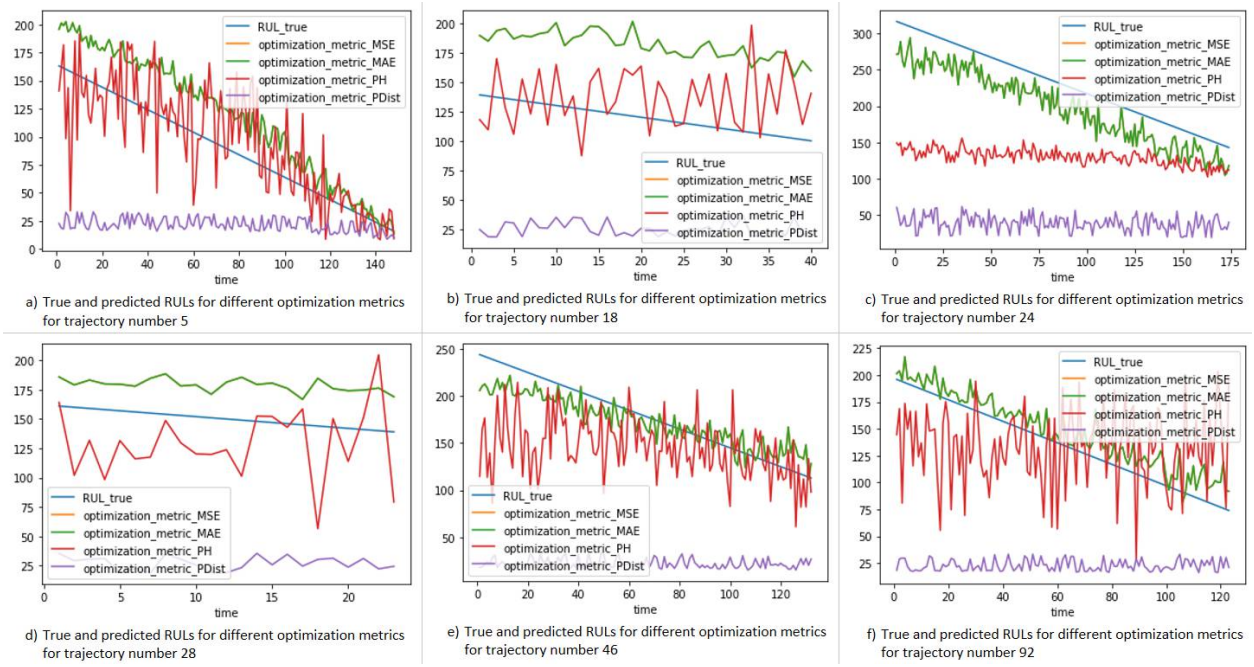


**Fig. 6 Predictions for different settings of optimization metrics on example trajectories for a population size of 50 on data set FD002.**

use a combination of metrics to reflect multiple prediction evaluation aspects. Especially the Prognostic horizon can play an important role for airlines which want to schedule maintenance timely and are dependent on predictions arriving early enough to schedule a corrective action. Therefore this should be taking in consideration when developing and evaluating prognostic methodologies. Further research can be done on combining multiple metrics and producing an

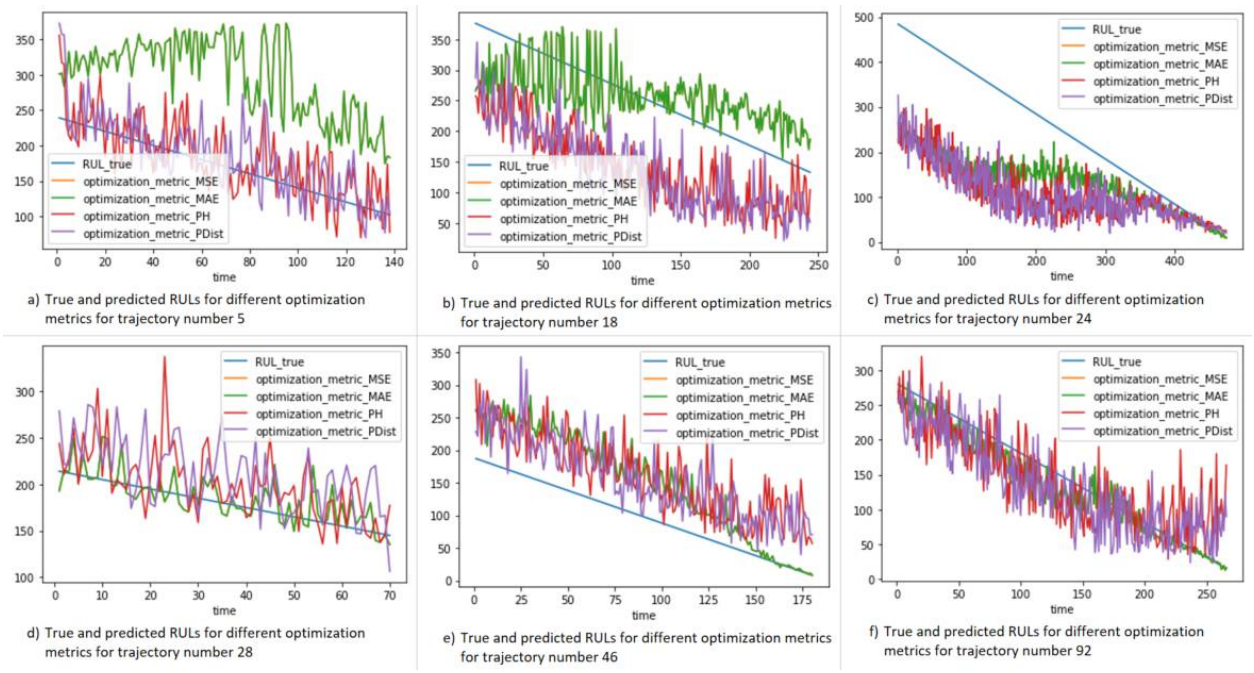


**Fig. 7 Predictions for different settings of optimization metrics on example trajectories for a population size of 50 on data set FD003.**

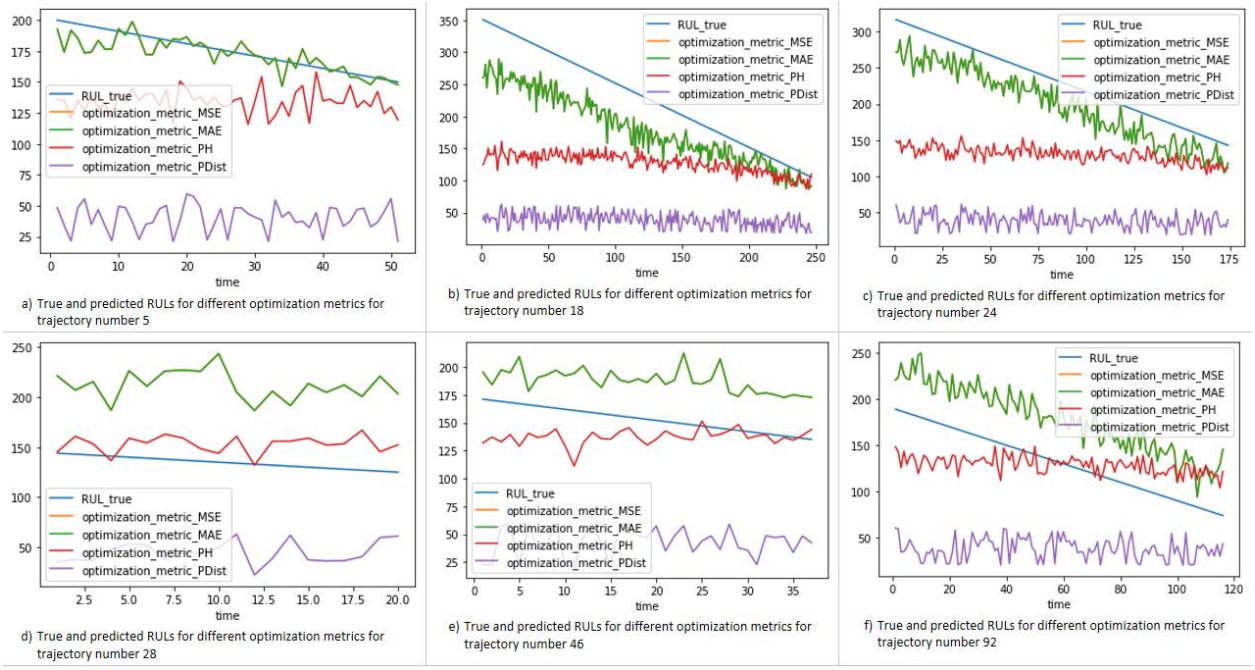


**Fig. 8 Predictions for different settings of optimization metrics on example trajectories for a population size of 50 on data set FD004.**

aggregated metric taking into account uncertainties in predictions to evaluate prognostics. Another idea could be to use multiple metrics simultaneously to arrive at more robust prognostic results. All in all, this study highlights the importance of choosing proper prognostic metrics and their impact on the prognostic outputs.

## Acknowledgment

This research is supported by European Union's Horizon 2020 program under the ReMAP project, grant No 769288. We are grateful for all the support and inputs given by the airline technicians and engineers.

## Contribution Statement

Marie Bieber: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing - original draft; Writing - review & editing.

Wim J.C. Verhagen: Conceptualization; Supervision; Writing - review & editing.

Bruno F. Santos: Conceptualization; Supervision; Writing - review & editing.

## References

[1] Elattar, H. M., Elminir, H. K., and Riad, A. M., "Prognostics: a literature review," *Complex & Intelligent Systems*, Vol. 2, No. 2, 2016, pp. 125–154.

[2] Zhang, J., Wang, P., Yan, R., and Gao, R. X., "Deep Learning for Improved System Remaining Life Prediction," *Procedia CIRP*, Vol. 72, 2018, pp. 1033–1038. https://doi.org/10.1016/j.procir.2018.03.262, URL https://doi.org/10.1016/j.procir.2018.03.262.

[3] Lei, Y., Li, N., Guo, L., Li, N., Yan, T., and Lin, J., "Machinery health prognostics: A systematic review from data acquisition to RUL prediction," *Mechanical Systems and Signal Processing*, Vol. 104, 2018, pp. 799–834.

[4] Zio, E., "Prognostics and Health Management (PHM): Where are we and where do we (need to) go in theory and practice," *Reliability Engineering and System Safety*, Vol. 218, No. PA, 2022, p. 108119. https://doi.org/10.1016/j.ress.2021.108119, URL https://doi.org/10.1016/j.ress.2021.108119.

[5] Baruah, P., Chinnam, R. B., and Filev, D., "An autonomous diagnostics and prognostics framework for condition-based maintenance," *IEEE International Conference on Neural Networks - Conference Proceedings*, 2006, pp. 3428–3435. https://doi.org/10.1109/ijcnn.2006.247346.

[6] Voisin, A., Levrat, E., Cocheteux, P., and Iung, B., "Generic prognosis model for proactive maintenance decision support: Application to pre-industrial e-maintenance test bed," *Journal of Intelligent Manufacturing*, Vol. 21, No. 2, 2010, pp. 177–193. https://doi.org/10.1007/s10845-008-0196-z.

[7] Trinh, H. C., and Kwon, Y. K., "A data-independent genetic algorithm framework for fault-type classification and remaining useful life prediction," *Applied Sciences (Switzerland)*, Vol. 10, No. 1, 2020. https://doi.org/10.3390/app10010368.

[8] Saxena, A., Sankararaman, S., and Goebel, K., "Performance Evaluation for Fleet-based and Unit-based Prognostic Methods," *European Conference of the Prognostics and Health Management Society*, 2014, pp. 1–12.

[9] Sankararaman, S., Saxena, A., and Goebel, K., "Are current prognostic performance evaluation practices sufficient and meaningful?" *PHM 2014 - Proceedings of the Annual Conference of the Prognostics and Health Management Society 2014*, 2014, pp. 533–545.

[10] Saxena, A., Celaya, J., Saha, B., Saha, S., and Goebe, K., "On applying the prognostic performance metrics," *Annual Conference of the Prognostics and Health Management Society, PHM 2009*, 2009, pp. 1–16.

[11] Saxena, A., Goebel, K., Simon, D., and Eklund, N., "Damage propagation modeling for aircraft engine run-to-failure simulation," *2008 International Conference on Prognostics and Health Management, PHM 2008*, 2008. https://doi.org/10.1109/PHM.2008.4711414.

[12] Saxena, A., Celaya, J., Saha, B., Saha, S., and Goebel, K., "Metrics for offline evaluation of prognostic performance," *International Journal of Prognostics and Health Management*, Vol. 1, No. 1, 2010.

[13] Goebel, K., Celaya, J., Sankararaman, S., Roychoudhury, I., Daigle, M., and Saxena, A., *Prognostics: The Science of Making Predictions*, 2017.

[14] Bieber, M., Verhagen, W. J. C., and Santos, B. F., "An Adaptive Framework For Remaining Useful Life Predictions Of Aircraft Systems," *European Conference of the Prognostics and Health Management Society*, 2021, pp. 60–70.

[15] Branco, P., Torgo, L., and Ribeiro, R. P., "Pre-processing approaches for imbalanced distributions in regression," *Neurocomputing*, Vol. 343, No. xxxx, 2019, pp. 76–99. https://doi.org/10.1016/j.neucom.2018.11.100, URL https://doi.org/10.1016/j.neucom.2018.11.100.

[16] Gado, J. E., Beckham, G. T., and Payne, C. M., "Improving Enzyme Optimum Temperature Prediction with Resampling Strategies and Ensemble Learning," *Journal of Chemical Information and Modeling*, Vol. 60, No. 8, 2020, pp. 4098–4107. https://doi.org/10.1021/acs.jcim.0c00489.

[17] Jardine, A. K., Lin, D., and Banjevic, D., "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mechanical Systems and Signal Processing*, Vol. 20, No. 7, 2006, pp. 1483–1510. https://doi.org/10.1016/j.ymssp.2005.09.012.

[18] Kothamasu, R., Huang, S. H., and Verduin, W. H., "System health monitoring and prognostics - A review of current paradigms and practices," *Handbook of Maintenance Management and Engineering*, , No. July 2006, 2009, pp. 337–362. https://doi.org/10.1007/978-1-84882-472-0{_}14.

[19] Hoque, N., Bhattacharyya, D. K., and Kalita, J. K., "MIFS-ND: A mutual information-based feature selection method," *Expert Systems with Applications*, Vol. 41, No. 14, 2014, pp. 6371–6385. https://doi.org/10.1016/j.eswa.2014.04.019, URL http://dx.doi.org/10.1016/j.eswa.2014.04.019.

[20] Ward, F. R., and Habli, I., "An Assurance Case Pattern for the Interpretability of Machine Learning in Safety-Critical Systems," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 12235 LNCS, 2020, pp. 395–407. https://doi.org/10.1007/978-3-030-55583-2{_}30.

[21] Holland, J. H., *Adaptation in Natural and Artificial Systems*, 1992. https://doi.org/10.1016/S0376-7361(07)53015-3.

[22] Stanovov, V., Brester, C., Kolehmainen, M., and Semenkina, O., "Why don't you use Evolutionary Algorithms in Big Data?" *IOP Conf. Ser.: Mater. Sci. Eng.*, Vol. 173, No. 1, 2017. https://doi.org/10.1088/1757-899X/173/1/012020.

[23] Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., and Schwabacher, M., "Metrics for evaluating performance of prognostic techniques," *2008 International Conference on Prognostics and Health Management, PHM 2008*, 2008. https://doi.org/10.1109/PHM.2008.4711436.

[24] Biggio, L., Wieland, A., Chao, M. A., Kastanis, I., and Fink, O., "Uncertainty-aware Remaining Useful Life predictor," , No. Cv, 2021, pp. 1–14. URL http://arxiv.org/abs/2104.03613.

[25] Frederick, D. K., DeCastro, J. A., and Litt, J. S., "User's guide for the commercial modular aero-propulsion system simulation (C-MAPSS)," 2007.

[26] Saxena, A., Goebel, K., Simon, D., and Eklund, N., "Damage Propagation Modeling for Aircraft Engine Prognostics," 2008.

[27] Ramasso, E., and Saxena, A., "Performance benchmarking and analysis of prognostic methods for CMAPSS datasets," *International Journal of Prognostics and Health Management*, Vol. 5, No. 2, 2014, pp. 1–15.

[28] Goebel, K., Saxena, A., Saha, S., Saha, B., and Celaya, J., "Prognostics: The Science of Making Predictions." Createspace Independent Publishing Platform, 2017, Chap. 5. Prognos, pp. 149–171.