Assessing the Impact of Metrics on the Choice of Prognostic Methodologies

Bieber, Marie; Verhagen, Wim J.C.; Santos, Bruno F.

DOI
10.2514/1.J063365

Publication date
2024

Document Version
Final published version

Published in
AIAA Journal

AIAA JOURNAL
Vol. 62, No. 2, February 2024

Assessing the Impact of Metrics on the Choice of Prognostic Methodologies

Marie Bieber*
Delft University of Technology, 2629 HS Delft, The Netherlands
Wim J. C. Verhagen†
RMIT University, Carlton, Victoria 3053, Australia
and
Bruno F. Santos‡
Delft University of Technology, 2629 HS Delft, The Netherlands

https://doi.org/10.2514/1.J063365

Over the past years, advanced prognostic models and approaches have been developed. Most existing approaches are tailored to one specific system and cannot adaptively be used on different systems. This can lead to years of research and expertise being put into implementing prognostic models without the capacity to predict system failures, either because of a lack of data or data quality or because failure behavior cannot be captured by data-driven models. In addition, prognostic models are often evaluated using metrics only related to the correctness of predictions, preventing meaningful evaluation of operational performance. This paper makes use of a framework that can automatically choose prognostic settings based on specific system data. It simultaneously optimizes the choice of methodologies using metrics that capture multiple aspects of prediction quality. We apply this framework to both a simulated data set and a real aircraft data set to characterize the impact of metrics on the choice of prognostic methodologies. The results show that the choice of optimization metric greatly impacts the output of the generic prognostic framework and the overall performance. In addition, a definition for data suitability is provided and assessed on the aircraft system data sets.

I. Introduction

WITHIN the framework of condition-based maintenance (CBM), prognostics enable assessment of equipment health and prediction of the remaining useful life (RUL) [1]. Using prognostics in such a context requires properly assessing the quality of predictions [2,3]. An effort to standardize prognostic metrics has been made by Saxena et al. in [4,5]. The metrics commonly used in prognostics are highlighted, and several ways to classify them are presented as ways to interpret and use the metrics. A comprehensive overview of existing metrics to evaluate prognostic performance is given by Ochella and Shafiee [6]. A single metric, such as the mean-squared error (MSE), can arguably not characterize the quality of RUL predictions sufficiently for a thorough assessment within a CBM framework [7]. Instead, the design of prognostic metrics has to be linked to the application and decision-making process [8,9]. In addition, as highlighted in Fig. 1, metrics are needed to define requirements and thoroughly evaluate prognostic performance [10].

Goebel et al. [11] state that a meaningful prediction has three attributes, namely, correctness, timeliness, and confidence (see Sec. II.A). 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 a single metric, which is often one linked to the correctness of the method [10]. Still, previous works on including more advanced metrics or defining more advanced metrics have been done in the literature. For example, Amigó et al. [12] introduce a measurement to combine several metrics and indicate how robust the measured differences are to changes in the relative weights of the individual metrics. Baptista et al. [13] show that prognostic metrics correlate with a Shapley Additive Explanations (SHAP) model's explanation. A performance metric to assess performance, effectiveness, and efficiency of health monitoring models of complex engineering systems is suggested by Lewis and Groth [14].

In addition to a suitable prognostic assessment technique, the question remains of how to translate this toward a prognostic assessment. Such an assessment is and must be application dependent. This study focuses on applying prognostics within a CBM framework for aircraft maintenance. A number of publications have been made on the topic of integrating prognostic models in aircraft maintenance planning [15,16]. A framework for aircraft maintenance design with reliability and cost-efficiency objectives has been provided in [17]. To use prognostic models as input for maintenance planning, those models need to be developed, which is time-consuming and requires expertise. However, what would be desirable was if, instead of spending months on developing prognostic models, there was a way to assess system data toward their suitability for prognostics relatively quickly. One of the main guiding works in the literature on this topic is perhaps the work by Coble and Wesley Hines [18], in which prognostic parameters are retrieved from the system data to do a prognostic assessment before applying actual prognostic methodologies. A method to evaluate data quality before the modeling by clustering the data into different system conditions is suggested in Ref. [19]. Omri et al. [20] propose a set of data quality requirements, especially for health assessment and fault detection. They propose a "detectability" metric to assess the suitability of data for fault detection. Atamuradov et al. [21] present a hybrid feature evaluation with a combined metric. A framework for RUL prediction, including a physics-informed failure mode recognition model that can be applied to different systems with different failure modes, is presented first in Ref. [22] and extended by Xiong et al. [23].

Two challenges arise from the above-presented literature: One, tuning prognostic algorithms without understanding which metrics are needed to assess the algorithm is difficult. Similarly, it is tricky to understand the full impact of choosing prognostic metrics without considering the prognostic algorithm. Two, while the presented data suitability methodologies are demonstrated in several case studies, they are lacking the link toward prognostic algorithms. Furthermore,

Presented as Paper 2022-3966 at the AIAA Aviation 2022 Forum, Chicago, IL, June 27–July 1, 2022; received 11 July 2023; revision received 13 September 2023; accepted for publication 16 November 2023; published online 8 January 2024.

*Ph.D. Candidate, Air Transport and Operations, Faculty of Aerospace Engineering.
†Senior Lecturer, Aerospace Engineering and Aviation.
‡Associate Professor, Air Transport and Operations, Faculty of Aerospace Engineering.

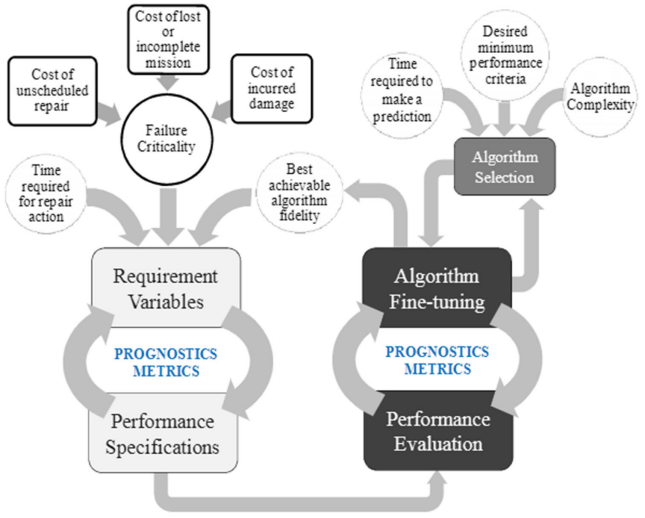


Fig. 1 Prognostic metrics are needed to define requirements and evaluate performance [10].

often statistical methodologies and predefined metrics are used to assess the data quality. This is problematic for several reasons: First, data suitability for prognostics can only truly be assessed when attempting to train a model capable of predicting the system's RUL. Second, artificial intelligence (AI)-based methodologies are in some cases able to detect failures even though the underlying data degradation is not visible or statistically traceable; i.e., statistical methods might not really give us insight into the data suitability for prognostics [24]. Third, in order to go beyond a statistic-based data assessment, prognostic performance metrics should be translated toward a data suitability assessment.

To address the challenges listed above, we, therefore, in this paper, investigate the impact of metrics on the choice of prognostic methodologies. On top of that, we explore how the performance of prognostic methodologies can be translated to an assessment of the suitability of the system for prognostics. Therefore, the following novel contributions to state-of-the-art are made:

1) An integrated framework is presented that selects the optimal prognostic settings, where optimality is assessed in terms of three selected prognostic metrics, representing correctness, confidence, and timeliness of predictions.

2) A study on the impact of different metrics on the choice of prognostic methodologies is conducted.

3) The resulting outcome is used to define the term "system data suitability" for prognostics such that it not only includes the data characteristics but also takes into account the data suitability in a CBM framework.

4) An example of the data suitability assessment for aircraft data sets of different quality is given.

The remainder of the paper is structured as follows: Section II explains the generic prognostic framework (GPF) used in this study and how it can be used to assess the prognostic suitability of system data. To investigate the impact of prognostic metrics and validate the presented data suitability assessment, in Sec. III, two case studies are conducted: one on a simulated turbofan dataset and one on a real aircraft system. In Sec. IV, the results of the two case studies are observed and interpreted to understand two points: first, the impact of metrics on the choice of prognostic methodologies, and, second, how the output of the GPF can be used as an assessment for system's suitability for prognostics. Furthermore, the limitations and directions for further research are highlighted. Section V concludes the paper and highlights the main findings.

II. Methodology

To select the optimal set of prognostic methodologies, a GPF as presented in [25] is used, which contains three steps of prognostics and according to representative techniques. 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 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, further consequences, 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 into the GPF give us insight into the quality of predictions and thereby help to choose appropriate prognostic methods.

The GPF consists of three phases (colored blocks in Fig. 2). In phase one, which is highlighted in green, a genetic algorithm (GA) is applied to find the optimal prognostic settings. This is done using multi-objective optimization based on three different metrics, which are explained in more detail in Sec. II.A. In phase two, highlighted in red and further explained in Sec. II.B, a prognostic model is trained, which then has the capability to output RUL estimates. In the final step of the framework, phase three, highlighted in blue, a data suitability assessment is performed. Based on the resulting accuracies in terms of the selected prognostic metrics, thresholds are defined to determine if the system data are suitable for prognostics. A detailed explanation is given in Sec. II.C.

A. Generic Prognostic Framework

The framework used in this work is a modified version of the GPF presented in [25,26]. It differs from the original framework mainly in optimizing three different prognostic metrics simultaneously. Therefore, we only give a short overview of the elements and functionalities of the GPF and refer the reader to the previous work for more details about the GPF. The GPF consists of three blocks corresponding to three selected steps in prognostics: data rebalancing, feature engineering, and the prognostic algorithm itself, as displayed in Fig. 3. Each of the three blocks contains several representative methodologies for each of the selected steps in prognostics. Imbalanced data occur when one class of data (e.g., faulty behavior) is under-represented when compared to the other class(es) (e.g., healthy behavior). Data rebalancing methods make use of the concepts of undersampling and oversampling: the former consists in removing majority examples, while the latter replicates the minority examples [27]. Three data rebalancing methodologies as introduced in [28] are included: random oversampling (RO), introduction of Gaussian noise (GN), and weighted relevance-based combination strategy (WERCS).

The feature engineering methodologies in the framework are principal component analysis (PCA), correlation-based feature, and importance-based feature selection representing, respectively, feature extraction, filter-based feature selection, and embedded

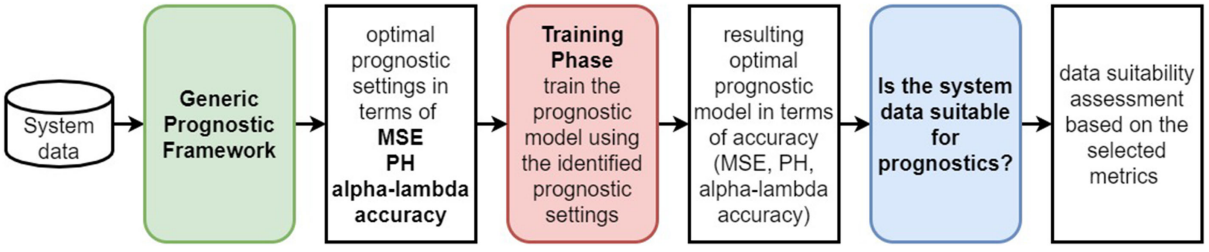


Fig. 2 The generic prognostic framework flow.

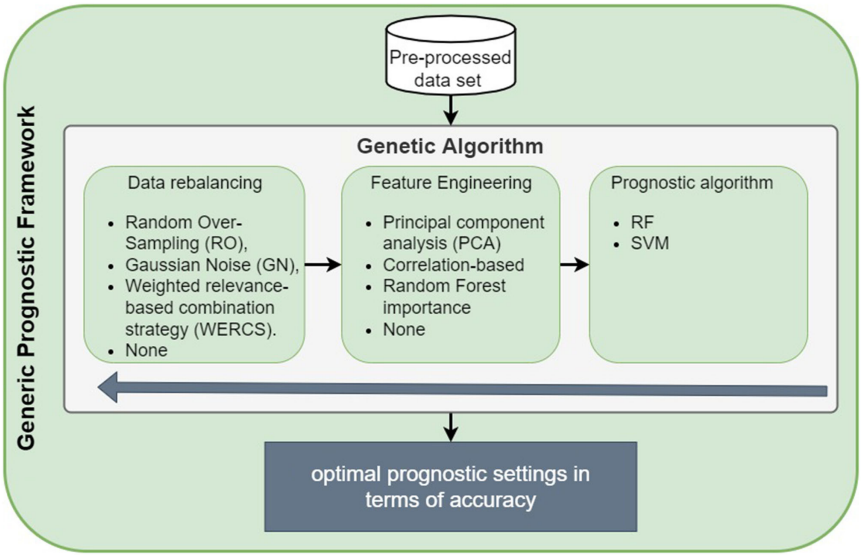


Fig. 3 The elements of the generic prognostic framework.

feature selection techniques. In order to get a first prognostic assessment through the framework, the prognostic algorithms included are a random forest (RF) regression and a support vector regression (SVM). The two selected algorithms are well-established and offer potential advantages in terms of interpretability and explainability [29]. However, they will generally not offer performance on the level of bespoke, advanced models developed for specific applications.

The GPF selects optimal sets of methodologies for each of the three steps in the prognostic framework. Here, "optimal" refers to the best in terms of the MSE, prognostic horizon (PH), and $\alpha - \lambda$ score. In other words, we treat the problem of finding the prognostic settings as a multi-objective optimization problem: The objective function is to simultaneously minimize the MSE, maximize the PH, and maximize the $\alpha - \lambda$ score of the prognostic algorithm together with data rebalancing and feature engineering techniques on the preprocessed data set. To solve the optimization problem, we use a GA. These algorithms are based on the concepts of natural selection and genetics [30]. Due to their flexibility, GAs can solve global optimization problems and optimize several criteria at the same time, like in our case, the simultaneous selection of data rebalancing, feature engineering, and prognostic algorithm techniques [31].

Basically, there are two approaches to multi-objective optimization: The first is to create a single optimization objective by combining the individual objective functions. The second is to move all but one objective to the constraint set [32]. This approach results in a set of solutions, each of which satisfies the objectives at an acceptable level without being dominated by another solution. Due to the fact that GAs are a population-based approach, they are well-suited for multi-objective optimization problems. Sets of solutions are returned in every generation; therefore, multiple solutions can easily be returned [32]. In fact, a majority of the multi-objective optimization problems in current literature are solved using evolutionary approaches [33]. Several multi-objective approaches for GAs have been suggested in the literature, and a comprehensive overview can be found in [32]. We use the Nondominated Sorting Genetic Algorithm II (NSGA-II, introduced in [34]). It ranks candidate solutions with the fast nondominated sorting method and uses a crowding distance as a diversity mechanism. The algorithm is well-tested, has been used in many applications, and is efficient, which makes it a good candidate for this study.

The NSGA-II, in our case, takes the system data as input and outputs the set of Pareto dominant solutions. A solution is Pareto dominant if there does not exist any other feasible solution that dominates it [35]. In this case, a solution is a combination of a data rebalancing technique, a feature engineering methodology, and a prognostics algorithm. If the algorithm identifies that applying no rebalancing or feature engineering technique results in better prognostic outputs, the GPF returns "None" for the according to the block. The three different metrics integrated into the framework are the MSE, PH, and the $\alpha - \lambda$ metric. The metrics account for the three attributes of meaningful predictions, i.e., correctness (MSE), timeliness (PH), and confidence ($\alpha - \lambda$ metric) [4,10,36].

The MSE at time t is given as

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

where RUL_i is the true RUL value and $\hat{\mathrm{RUL}}i$ the predicted RUL value at time step i.

The PH is defined as

$$\mathrm{PH}(t, \alpha) = \mathrm{RUL}(t_{i_\alpha}) \tag{2}$$

with $\mathrm{RUL}(t_{i_\alpha})$ the true RUL at time t_{i_α} and $i_\alpha := \min\{k \in p | \forall j \geq k : \alpha_j^- \leq \hat{\mathrm{RUL}}(t_j) \leq \alpha_j^+\}$, where p is the set of all time indices where predictions are made, $\hat{\mathrm{RUL}}(t_j)$ is the prediction at time index $j \in p$, and the bounds are defined as $\alpha_j^- := \mathrm{RUL}(t_j) - \alpha$ and $\alpha_j^+ := \mathrm{RUL}(t_j) + \alpha$.

The PH is the smallest RUL in which the predicted RUL is still within the specified α bounds. The best score for the PH is obtained when the predicted RUL always falls within the specified accuracy zone, while the worst score is obtained when the predicted RUL is never within the accuracy zone. The PH indicates whether the predicted estimates are within the specified limits, especially toward the end of life (EoL), so that predictions can be considered trustworthy during a specified time span before the system's EoL is reached. It becomes clear that the longer the PH, the more time available to act based on a prediction. It, therefore, gives an indication of the timeliness of an algorithm, in the sense that during a time span before the system's EoL, the predictions can be used to plan according to actions. Note that the above definition of the PH is a slightly modified version of the PH defined in [36]. Instead of being defined as the first time at which the RUL falls within the specified error bounds, we define it as the first instance of time from which on all RUL predictions fall within the specified error bounds. This modification was made because for safety-critical components and systems predictions toward the EoL are much more crucial than predictions during earlier component or system life. Such predictions have to be reliable, especially when used to, e.g., schedule maintenance actions. In the case studies presented in Sec. III, we set $\alpha = 40$ flight cycles, which is the time needed to schedule maintenance for an aircraft in case it is needed.

And finally, the $\alpha - \lambda$ metric is, as in [37], defined as

$$\alpha - \lambda := \begin{cases} 1, & \text{if } \{1-\alpha)\lambda^* \leq \lambda_p \leq (1+\alpha)\lambda^* \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

with $\lambda^* = \text{RUL}(t_\lambda)$ the ground truth, $\lambda_p = \text{RUL}(t_{\lambda_p})$ the prediction, and α an arbitrary chosen accuracy. The two input parameters for the metrics are α, which determines the required level of confidence for the predictions, and λ, which represents a fraction of time between the point when the algorithm starts predicting (t_P) and the actual failure or EoL.

The $\alpha - \lambda$ metric, therefore, measures the prediction quality by determining whether the prediction falls within specified limits at particular times, which—as mentioned above—are presented as a percentage of the total ailing life of the system. To be more precise, the question it seeks to answer is whether the prediction accuracy of the RUL model is within $\alpha \cdot 100\%$ of the actual RUL at a specified time instance t_λ (depending on λ). The output is binary (true or false), stating if the desired condition [Eq. (3)] is met at the specific time instance. It is more stringent than the PH because it requires the predictions to stay within a cone of accuracy, i.e., bounds that shrink as time passes.

The $\alpha - \lambda$ metric can be evaluated and averaged over the whole trajectory with N time steps (i.e., for the entire interval $[t_P, EoL]$), arriving at $\overline{\alpha - \lambda}$, which lies between 0 and 1. It, therefore, returns the confidence that the predictions fall into the α bounds over the entire period of time. This is why it is a good candidate to represent prediction confidence.

B. Training Phase

The output of the GA is the "best individual," i.e., the set of methodologies and hyperparameter settings that lead to the best performance on the dataset. Note that, in order to save computational power and arrive at a solution more quickly, the GPF only takes a reduced dataset as an input for the optimization [25]. In this step, the prognostic model is trained on the full dataset using the optimal settings returned by the GPF. Therefore, the output of this step is a trained prognostic model, which takes as input system data and outputs the RUL.

C. Determining if a System Is Suitable for Prognostics

Once the GPF has identified a set of optimal prognostic models in terms of MSE, PH, and $\alpha - \lambda$ score and outputs the according scores, phase three (indicated in blue in Fig. 2) starts. Based on the output models, this phase aims to identify whether the system data are suitable for prognostics. Of course, the question of whether it makes sense to apply prognostic approaches for given data highly depends on the user, the application, and the underlying requirements. As highlighted in Sec. I we aim to assess data suitability in a prognostic context. This means that we go beyond a simple statistical assessment and instead translate prognostic metrics of basic prognostic machine-learning models trained on the underlying system data into a data suitability assessment. The definition of "system data suitability for prognostics" depends on user inputs, which can be adapted accordingly. The user needs to set bounds for each of the criteria measured: 1) in terms of correctness, MSE_{max}, the upper MSE limit; 2) in terms of timeliness, $\text{PH}_{\text{min}}(a)$, the minimum number of time steps before failure at which the failure needs to be known to take according to actions, which is based on a, the maximum value (measured in time steps) that the prediction is allowed to deviate from the true value; 3) and in terms of confidence, $\overline{(\alpha - \lambda)_{\text{min}}}$, where $0 < \overline{(\alpha - \lambda)_{\text{min}}} < 1$, the minimum ratio of predictions within the α bounds.

System data are defined to be suitable for prognostics if

$$\text{MSE}(t = \text{end of life}) \leq \text{MSE}_{\text{max}} \tag{4}$$

$$\wedge\ \text{PH}(t_j) \geq \text{PH}_{\text{min}}(a) \quad \forall\ j \in p \text{ and specified } a \tag{5}$$

$$\wedge\ \overline{\alpha - \lambda} \geq \overline{(\alpha - \lambda)_{\text{min}}} \tag{6}$$

Only when all three of the conditions are met for a given prognostic model does that model satisfies the data suitability criteria. This can be applied to each model in the set of optimal models returned by the GPF. If a single prognostic model is found that fulfills the above requirements [Eqs. (4–6)], then the system data are assumed to be suitable for prognostics.

III. Results

There are two main aims of the conducted study: First, we want to understand the impact of prognostic metrics on the methodology selection in the different steps of the prognostic framework. Second, an example evaluation is performed for different input system data to understand if the systems are suitable for prognostics. For this purpose, two case studies were conducted: The first case study in Sec. III.A is conducted on a simulated turbofan dataset commonly used in literature and known to be suitable for prognostics. The second case study in Sec. III.B uses a real-world aircraft dataset.

A. Case Study: Simulated Turbofan Dataset

The Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) data set contains simulated run-to-failure data for turbofan engines [38]. Using this tool, four data sets were created [39]. The data sets differ mainly in the number of fault modes and operational conditions simulated in the experiments. An overview is given in Table 1. For our purpose, we use two of the four datasets: First, dataset FD001 is considered the simplest one as it only contains one fault mode and operating condition. The second dataset FD002 is considered to be more complex due to the different operating conditions. 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 captures the operations of each unit until it fails. In the test set, the time series ends at some point before the failure, and the objective is to estimate the RUL, or, in other words, the number of remaining operational cycles before failure. There are 21 sensor measurements, and each row in the data set contains the measurements corresponding to operations during a one-time cycle for a certain unit.

The framework is applied to both datasets and in the following, the according results are presented. We compare the resulting prognostic models to baseline models, namely, using only RF and SVM, respectively, without any data rebalancing or feature engineering. In all cases, we run the GA for 20 generations with a population of 30 individuals.

1. *Results on Dataset FD001*

First, we present the output of the GA, i.e., the Pareto front for dataset FD001. Table 2 contains the set of individuals in the Pareto front, with their respective choices of methodologies for the data rebalancing, feature engineering and prognostic algorithm. In addition,

Table 1 Characteristics of the four turbofan engine data sets [40]

Data set	No. of fault modes	No. of conditions	No. of train units	No. of test units	Relative no. of train units, %	Relative no. of 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

Table 2 Resulting prognostic settings and metrics when running the MOGA GPF with 30 individuals for FD001

Rebalancing	Feature engineering	Prognostic algorithm	MSE	PH	$\overline{\alpha-\lambda}$
None	None	RF	1647.10	144.34	0.524206
GN	PCA	SVM	1774.21	129.47	0.536729
None	PCA	SVM	1759.08	132.25	0.536652
RO	None	SVM	1757.86	132.68	0.529704
WERCS	None	SVM	1755.32	130.92	0.531689

the according metrics (MSE, PH, and alpha-lambda score) for the trained prognostic models are given.

The results in Table 2 show that most of the Pareto optimal solutions use SVM as a prognostic algorithm. Note that the SVM-based solutions outperform RF-based solutions when using feature engineering or rebalancing techniques together with SVM. At the same time, the RF performs well without using any data rebalancing or feature engineering methodologies. The term "outperforms" here refers to in the sense of a lower $\overline{\alpha-\lambda}$ score in terms of best MSE and PH; using only RF proves to be the optimal technique for FD001. Furthermore, the $\overline{\alpha-\lambda}$ scores are all very close to each other. Finally, it can be seen that increasing the performance in terms of MSE also, in most cases, increases the performance in terms of PH, while it usually results in lower $\overline{\alpha-\lambda}$ scores.

Figure 4 shows a two-dimensional representation of both individuals in the Pareto front and dominated individuals. The following are observed:

1) In Fig. 4a, it can be seen that a good score in terms of MSE can be reached without decreasing the performance in terms of alpha-lambda score too much (only from around 0.54 to 0.42).

2) Figure 4b shows that this comes at the cost of reducing the PH to almost 0.

3) Therefore, in Table 2, only individuals with an MSE of around 1750 are in the Pareto front.

4) In this case, when using only the MSE as an optimization metric it would result in models with a poor score in terms of timeliness.

Figure 5 shows, for six randomly selected trajectories in the test set, the true RUL and the predicted values for the individuals in the Pareto front and the baseline models (using purely RF and SVM). For the selected trajectories of dataset FD001, the resulting prognostic models seemingly all perform very well, as do the baseline models. This is especially true for trajectories 5 (Fig. 5a), 24 (Fig. 5c), 46 (Fig. 5e), and 92 (Fig. 5f). Only in Fig. 5a, for trajectory 5, it can clearly be seen that the GPF-based models outperform the baseline algorithms, especially toward the EoL.

2. Results on Dataset FD002

Similarly, the results for the runs on data set FD002 are shown in Table 3, which contains both the choice of methodologies and the scores in terms of the selected metrics. What can clearly be seen is that the Pareto front contains more individuals than the one for FD001 (10 respectively, 5 individuals). Again, similarly to dataset FD001, in most cases, SVM results in better solutions than RF. While adding a resampling or feature engineering step can improve the predictions in single metrics, the Pareto front contains both the baseline scenarios, using purely RF and SVM.

Figure 6 shows a two-dimensional representation of both individuals in the Pareto front and dominated individuals. In the figure, it can clearly be seen that increasing the performance in terms of MSE (decreasing the MSE) results in a better lambda-alpha score (Fig. 6a), but a lower PH (Fig. 6b), which can also be observed in Table 3.

Figure 7 shows six randomly selected trajectories in the test set, the true RUL, and the predicted values for the individuals in the Pareto front and the baseline models. Here, as opposed to in FD001, the quality of results varies much more between the different selected trajectories. For trajectories 5 (Fig. 7a), 46 (Fig. 7e), and 92 (Fig. 7f), the models predict the RUL quite accurately, especially toward the EoL. This is not true for trajectories 18 (Fig. 7b) and 28 (Fig. 7d), for which the prognostic models are not able to accurately predict RUL. Note that the baseline algorithms (only RF and only SVM) are contained in the Pareto front. Therefore, it is no surprise that their predictions' quality is relatively high compared to the other chosen settings of Pareto front individuals.

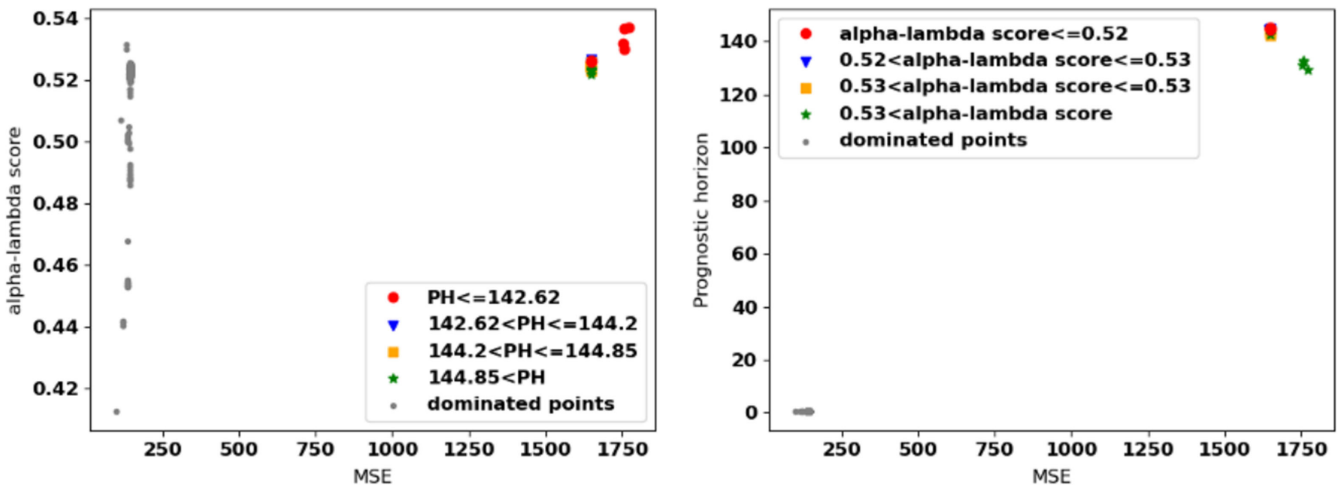
B. Case Study: Aircraft System Data

In the second case study, the GPF is applied to an aircraft pump package installed close to the landing gear. The pump package consists of two redundant pumps: pump 1 and pump 2. The assumption is made that pump 1 and pump 2 failures are independent. Failures happen on the two power boards, presumably due to short circuits.

On each of the pumps, sensors have been installed that measure the following properties: the motor current, the motor speed, the motor temperature, the reservoir fluid level, and the junction temperature (of the liquid).

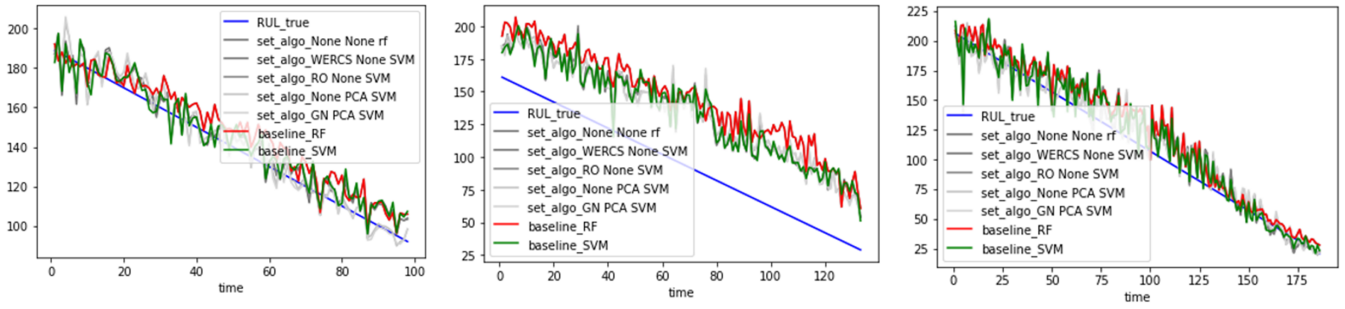
Next to those sensors, the static air temperature (sat) and the calibrated airspeed (cas) reported on the aircraft level are used as input. The sensor measurements are made every second. The per-second data are aggregated per flight phase by mean, maximum, and minimum to remove noise from the raw sensor data. A flight consists of 12 flight phases, from taxi-out until taxi-in. The aggregated data set contains around 35,000 flight phases in total. Of those data, 10% are maintained in the test set, and the rest forms the train set.

The results for the runs on the Pump data set are presented in Table 4. It contains both the choices of methodologies for the three selected



a) A 2D representation of the MSE vs alpha-lambda score on FD001 **b) A 2D representation of the MSE vs PH on FD001**

Fig. 4 A 2D representation of scores for the individuals when running the MOGA GPF with 30 individuals for FD001.



a) True and predicted RULs for trajectory #5 b) True and predicted RULs for trajectory #18 c) True and predicted RULs for trajectory #24

d) True and predicted RULs for trajectory #28 e) True and predicted RULs for trajectory #46 f) True and predicted RULs for trajectory #92

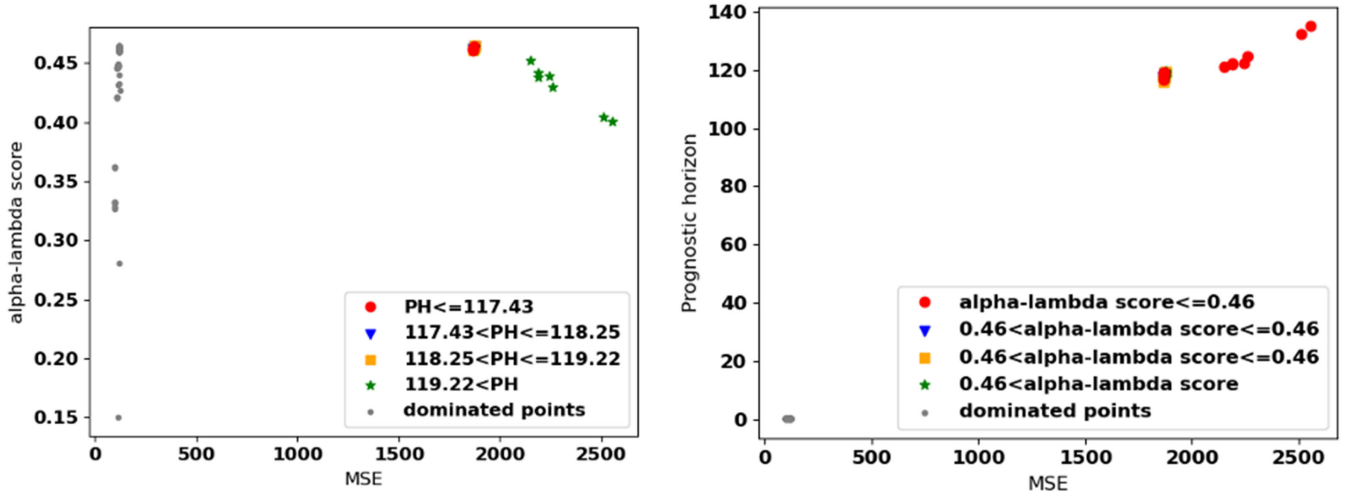
Fig. 5 Predictions of best-found settings vs the two baseline scenarios on example trajectories on FD001.

Table 3 Prognostic settings and metrics when running the MOGA GPF with 30 individuals for FD002

Rebalancing	Feature engineering	Prognostic algorithm	MSE	PH	$\overline{\alpha - \lambda}$
None	None	RF	1873.40	117.11	0.463387
RO	None	RF	1865.68	116.50	0.460887
WERCS	None	RF	1872.08	118.64	0.462034
GN	Correlation	SVM	2241.46	122.10	0.439204
None	Importance	SVM	2262.72	124.53	0.430084
None	None	SVM	2152.96	120.97	0.452355
RO	Importance	SVM	2557.72	134.95	0.400665
RO	None	SVM	2188.97	122.18	0.438469
WERCS	Importance	SVM	2510.06	132.32	0.404166
WERCS	None	SVM	2189.37	122.00	0.442205

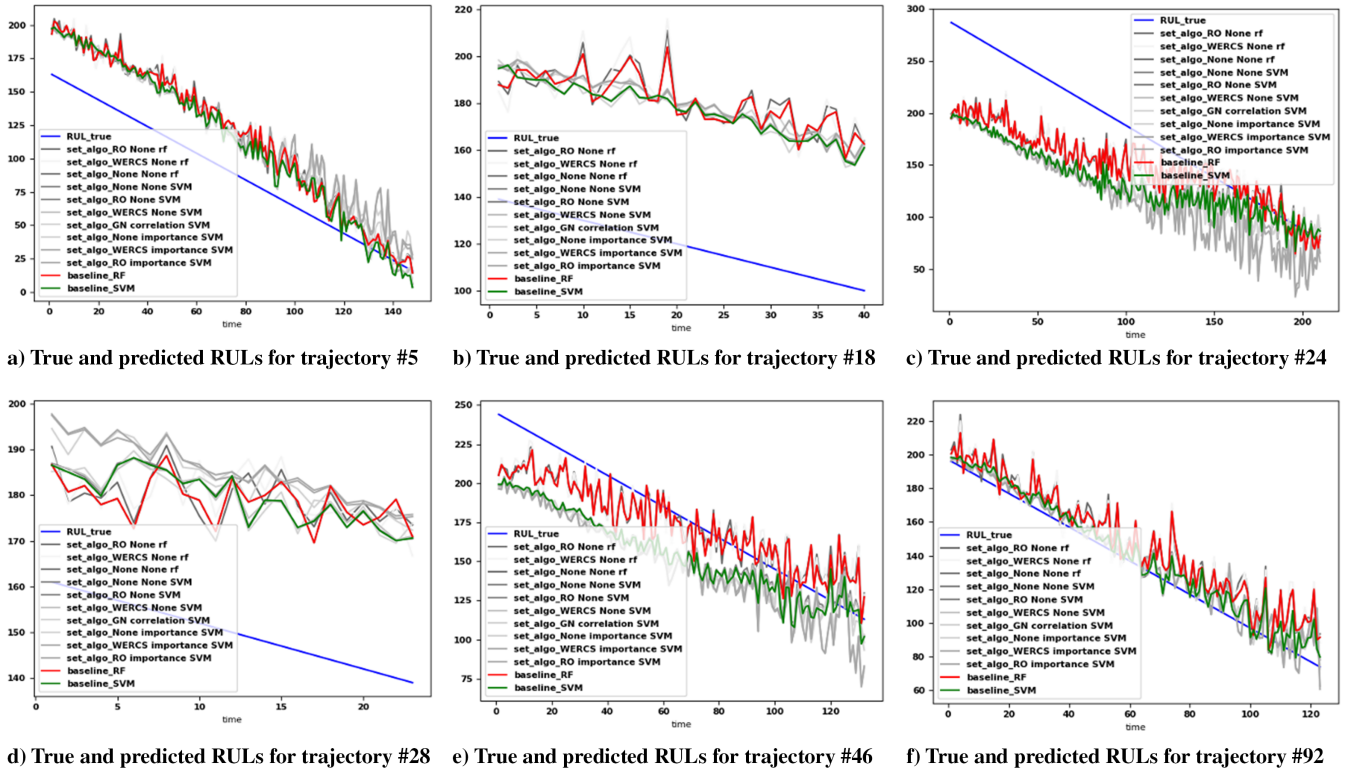
steps in prognostics and the according scores. Again, in this case, the Pareto front contains more individuals than the one of FD001. With its 11 individuals, the size is comparable to that of dataset FD002. As opposed to the simulated aircraft turbofan dataset, in this case, using RF results in better solutions than SVMs. In fact, no SVM solution is contained in the Pareto front. For the RF, almost every combination of rebalancing, feature engineering, and the prognostic algorithm is contained, resulting in very similar scores in terms of MSE. However, differences can be seen in terms of the other metrics. The PH ranges from 10.48 when using WERCS, correlation-based feature selection, and RF to 223.75 when using no rebalancing, importance-based feature selection, and RF. The $\overline{\alpha - \lambda}$ ranges from 0.0523 when using the previous settings to 0.2104 when using WERCS, correlation-based feature selection, and RF.

Figure 8 shows a two-dimensional representation of individuals in the Pareto front and their according scores in relation to each other. In Fig. 8a, it can be seen that optimizing toward a low MSE simultaneously results in a lower PH but increases the $\overline{\alpha - \lambda}$ score. Figure 8b shows the link between the $\overline{\alpha - \lambda}$ score and PH in a clearer way: Increasing the PH at the same time decreases the $\overline{\alpha - \lambda}$ score. This can also be observed in Table 4: The highest scoring solution in terms of PH is also the lowest scoring in terms of $\overline{\alpha - \lambda}$ score.



a) A 2D representation of the MSE vs alpha-lambda score on FD002 b) A 2D representation of the MSE vs PH on FD002

Fig. 6 A 2D representation of scores for the individuals when running the MOGA GPF with 30 individuals for FD002.



a) True and predicted RULs for trajectory #5 b) True and predicted RULs for trajectory #18 c) True and predicted RULs for trajectory #24

d) True and predicted RULs for trajectory #28 e) True and predicted RULs for trajectory #46 f) True and predicted RULs for trajectory #92

Fig. 7 Predictions of best-found settings vs the two baseline scenarios on example trajectories on data set FD002.

Table 4 Prognostic settings and metrics when running the MOGA GPF with 30 individuals for the aircraft pump dataset

Rebalancing	Feature engineering	Prognostic algorithm	MSE	PH	$\overline{\alpha - \lambda}$
GN	None	RF	4.64E+07	29.94	0.1417
GN	Correlation	RF	4.59E+07	11.25	0.1606
None	None	RF	4.63E+07	16.64	0.1507
None	PCA	RF	6.24E+07	191.61	0.0556
None	Correlation	RF	4.56E+07	10.23	0.1725
None	Importance	RF	5.46E+07	223.75	0.0523
RO	None	RF	4.85E+07	32.00	0.1985
WERCS	None	RF	4.82E+07	22.17	0.2147
WERCS	PCA	RF	6.52E+07	58.33	0.1253
WERCS	Correlation	RF	4.66E+07	10.48	0.2104
WERCS	Importance	RF	6.20E+07	57.88	0.0893

IV. Discussion

The aim of the conducted case studies was to explore the two main research questions introduced in Sec. I:

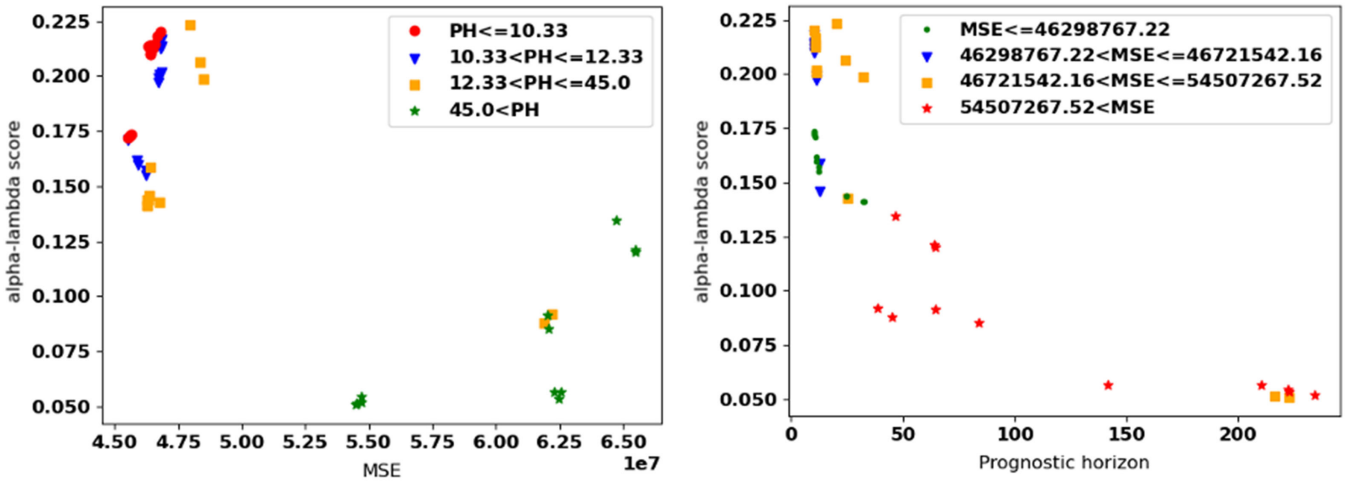
1) What impact do metrics have on the choice of methodologies?

2) How can the performance of prognostic methodologies be translated to an assessment of the system's suitability for prognostics?

Section IV.A analyzes the impact of metrics on the choice of prognostic methodologies. In Sec. IV.B, the system data of the two conducted case studies are analyzed using the definition of data suitability given in Sec. II.C. Finally, in Sec. IV.C addresses the limitations of the presented study, and directions for further research are given.

A. Impact of Metrics on the Choice of Prognostic Methodologies

The results of applying the GPF to the simulated turbofan datasets FD001 and FD002 are presented in Sec. III.A. When trying to



a) A 2D representation of the MSE vs alpha-lambda score on the aircraft pump dataset b) A 2D representation of the PH vs alpha-lamba score of the pareto front individuals in the aircraft pump dataset

Fig. 8 A 2D representation of scores for the individuals when running the MOGA GPF for the Pump data set.

understand the impact of the metrics on the choice of prognostic methodologies, it is of interest to take a closer look at both Tables 2 and 3, listing the chosen methodologies in the Pareto front, and Figs. 4 and 6, showing the links between the different metrics. It can be seen that using a different optimization metric can have an impact as big as a different choice of the prognostic algorithm used. For example, for FD001 in Table 2 we see that optimizing toward confidence results in using SVM for the prognostic model, while optimizing toward correctness results in using RF for this purpose. In FD002 the dynamics are a bit different. Still, the underlying outcome is the same: When optimizing toward confidence, the GPF chooses RF as the optimal prognostic algorithm while optimizing toward timeliness results in the GPF choosing SVM. Those dynamics are visualized in Figs. 9 and 10. This can also be observed in the aircraft Pump data case study: In Table 4, we see that instead of in the choice of prognostic algorithm, the impact metrics have on the selection of techniques is reflected in the rebalancing and feature engineering settings. An example of this effect is given in the example in Sec. III.B for the selection the GPF makes to reach the highest PH or highest $\overline{\alpha - \lambda}$ score.

Therefore, increasing the performance in terms of a single metric comes at the cost of decreasing the performance in terms of another metric; i.e., the metrics do have an influence on the chosen prognostic settings. With the term "prognostic settings," we refer to the combination of data rebalancing, feature engineering, and prognostic algorithm that is used to arrive at a prognostic model. This means that, when making choices for prognostic methodologies, it is important to consider which metric to use for evaluation. To summarize, the following main points are raised:

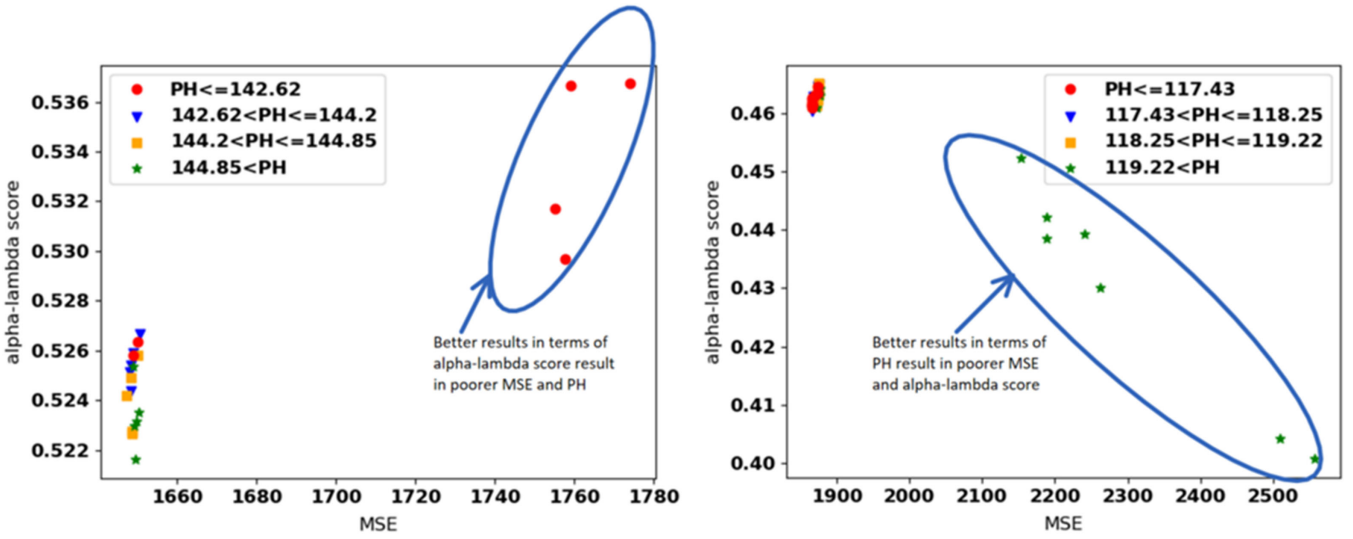
1) The choice of the optimal metric depends on the underlying data set and objective of prognostics, e.g., in what context they are used.

2) A single metric is often not enough to make fully informed choices regarding which prognostic methodology to use.

3) Optimizing toward different prediction attributes, i.e., correctness, timeliness, or confidence, results in different prognostic models and is often a tradeoff.

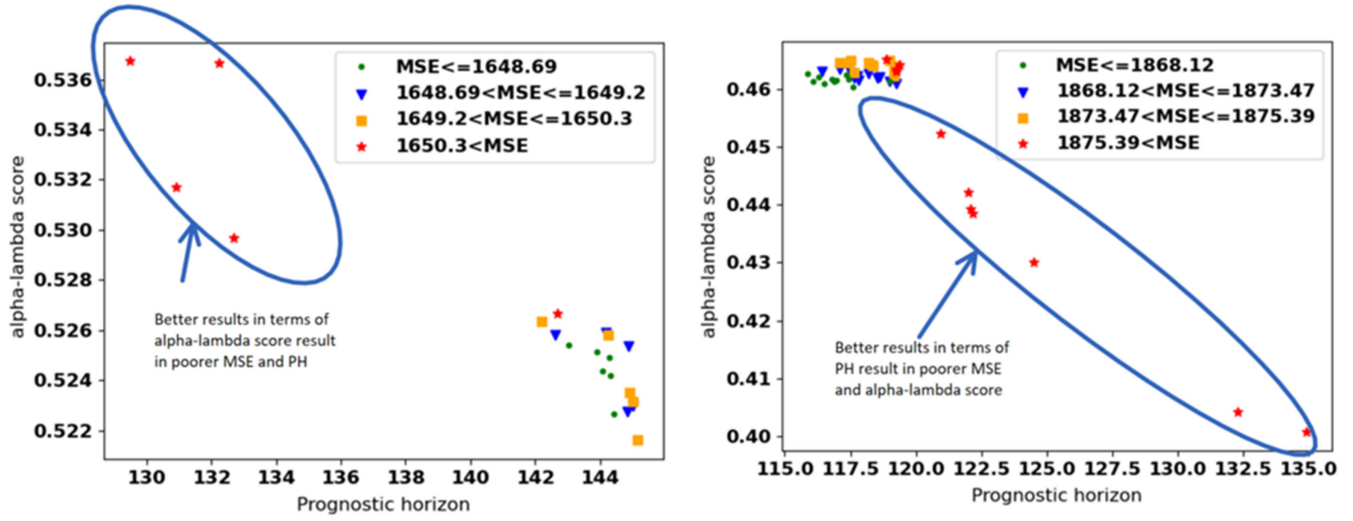
B. Evaluation of the Systems Suitability for Prognostics

In this section, we answer the question of how to assess data suitability for prognostics using the GPF. This is achieved by applying in Sec. II.C the introduced definition and methodology in both the case studies to assess the according systems data suitability. Be aware that, in the following, we provide a suggestion of how to set the boundaries, which is tailored to this case study. We put the focus on



a) A 2D representation of the pareto front individuals of FD001 b) A 2D representation of the pareto front individuals of FD002

Fig. 9 Comparison of the alpha-lambda score vs MSE of the Pareto points for datasets FD001 and FD002.



a) A 2D representation of the pareto front individuals of FD001 b) A 2D representation of the pareto front individuals of FD002

Fig. 10 Comparison of the alpha-lambda score vs PH of the Pareto points for datasets FD001 and FD002.

aircraft and look at it from the perspective of an Maintenance, Repair and Overhaul/airline/aircraft maintenance provider. Such a stakeholder uses the output of the prognostic models to plan and schedule maintenance tasks. Furthermore, we assume that the airline operates short-haul flights mainly with an average aircraft usage of 4 flight cycles (FC) per day. As mentioned in Sec. II.A, the assumption is made that a failure needs to be known at least 40 FC in advance (for the case studies on simulated turbofan engine data, we assume that FCs correspond to time cycles) in order to schedule maintenance. Based on this assumption, we set the following bounds for the criteria specified in Sec. II.C:

1) The upper MSE limit: $\text{MSE}_{\max} = 2000$ FC.

2) The minimum number of time steps before failure at which the failure needs to be known to take according to actions, $\text{PH}_{\min}(a) = 40$ FC, which corresponds to 10 days of operation, with $a = 40$ FC.

3) It is assumed that for this case it is sufficient that 45% of the predictions lie within the α bounds. Therefore, the minimum ratio of predictions within the α bounds, $\overline{(\alpha - \lambda)}_{\min} = 0.45$.

Based on the definition introduced in Sec. II.C, we observe that for all solutions contained in the Pareto front for dataset FD001, the three conditions [Eqs. (8) and (9)] hold; i.e.,

$$\text{MSE}(t = \text{end of life}) \leq 2000 \tag{7}$$

$$\wedge\ \text{PH}(t_j) \geq 40 \quad \forall\ j \in p \quad \text{and} \quad a = 40 \tag{8}$$

$$\wedge\ \overline{\alpha - \lambda} \geq 0.45 \tag{9}$$

And since all the conditions are fulfilled, according to the definition given in Sec. II.C, dataset FD001 proves to be suitable for prognostics. Figure 5 underlines this visually as it can be seen that the predicted value of almost all the models is close to the true RUL. For dataset FD002, Table 3 shows that there are three individuals in the Pareto front, satisfying all three above criteria [Eqs. (8) and (9)]. Those three individuals are the ones using RF as a prognostic algorithm, together with no rebalancing and feature engineering, together with random oversampling as a rebalancing methodology, and together with WERCS as a rebalancing technique respectively. The definition of data suitability in Sec. II.C states that only a single solution in the Pareto front is required to fulfill the conditions in order for the system data to be suitable for prognostics. As a result, also dataset FD002 turns out to be suitable for prognostics. Visually an indication of this can be seen in Fig. 7.

For the aircraft pump dataset, however, both the MSE and the $\overline{\alpha - \lambda}$ score are too high, respectively, too low for all solutions in the Pareto front. Therefore, the pump dataset is not suitable for prognostics according to the here presented definition.

C. Limitations and Further Research

The presented definition of data suitability is only dependent on prognostic metrics, meaning that the assessment of suitability is based merely on the quality of the prognostic model. Of course, as a stakeholder, one could be interested in metrics not just linked to the prognostic model itself. Therefore, a possible direction for further research would be to extend the data suitability assessment toward a more thorough assessment based on stakeholder needs. This could, e.g., be to include a calculation of costs associated with wrong predictions. Depending on how "wrong" the predictions are (in terms of selected prognostic metrics), this can then further be reflected in setting the thresholds for the data suitability assessment presented in Sec. II.C.

In addition, the GPF only includes a limited set of methodologies and steps integrated into prognostics. Of course, those were carefully chosen to represent the most important groups of methodologies and be relatively simple, while still powerful. The framework could be extended to include more advanced methodologies, such as deep learning techniques or even diagnostic approaches. The metrics used in the multi-objective optimization are all evaluated over the entire training data, i.e., component life span. A next step could be to instead of averaging over the entire time use moving averages to evaluate model performance. Another possibility is to include methods recognizing and separating health stages (e.g., healthy or degrading) and evaluating model performance during each of those.

The PH used in the data suitability assessment depends on the parameter a, which we treat as a user constraint in this study and set to 40 FC, representing the time needed to schedule and plan maintenance. In a further study, a range of values for a could be tested to see the effect on the prognostic model assessment. Such a sensitivity analysis could be conducted taking scheduling approaches into account, i.e., assessing a range of parameters and their effect on prognostic performance not only in terms of prognostic algorithms but also in terms of, e.g., costs for rescheduling maintenance. Such an analysis would produce a more thorough assessment of the according values, model qualities, and implications for subsequent CBM use.

Finally, the user is required to specify boundaries for each metric. This can be a challenging task. A way to overcome this could be to implement, as mentioned above, a more thorough assessment, e.g., in terms of costs. Having said that, the approach presented here still goes beyond what has been done in literature so far, adding a novelty here. So far, as highlighted in Sec. I, most studies regarding data suitability focused merely on the system data and their structure and statistical properties. However, when using machine learning approaches, it can be the case that even without trends being visible in the system data, the models can detect or even predict anomalies [41]. The approach presented here does not only provide an integrated way of assessing data suitability by taking into account prognostic machine learning algorithms. It also integrates metrics to capture the three aspects of prognostics namely correctness, timeliness, and confidence and thereby enables a more thorough assessment of the model quality.

V. Conclusions

The objective of the presented study is twofold: The first aim is to investigate the impact metrics have on prognostics. The second aim is to provide the means for a data suitability assessment for 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 GPF that chooses the optimal settings for the three steps of data rebalancing, feature engineering, and prognostic algorithm. A multi-objective optimization is conducted to reflect a selection of metrics, which account for all the aspects of prediction evaluation, including correctness (MSE), timeliness (PH), and confidence ($\overline{\alpha - \lambda}$ score). The results show the following: First, the choice of optimization metric has a big impact on the output of the GPF. This means that depending on the objective and motivation of using prognostics, a suitable metric should be carefully chosen. It can also make sense to use a combination of metrics to reflect multiple prediction evaluation aspects. Especially the PH can play an important role for airlines that want to schedule maintenance time and are dependent on predictions arriving early enough to schedule a corrective action. Therefore this should be taken into consideration when developing and evaluating prognostic methodologies. Second, the framework presented can be used together with a definition we provided to assess a system's suitability for prognostic based on the system data. All in all, this study both highlights the importance of choosing proper prognostic metrics and their impact on the prognostic outputs and gives directions for practitioners as to whether or not it makes sense to invest time and money in the development of prognostic systems based on the available system data.

Acknowledgments

This research is supported by the European Union's Horizon 2020 program under the Real-time Condition-based Maintenance for Adaptive Aircraft Maintenance Planning (ReMAP) project (Grant No. 769288). We are grateful for all the support and input given by the airline technicians and engineers.

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.
https://doi.org/10.1007/s40747-016-0019-3

[2] Brunton, S. L., Kutz, J. N., Manohar, K., Aravkin, A. Y., Morgansen, K., Klemisch, J., Goebel, N., Buttrick, J., Poskin, J., Blom-Schieber, A. W., Hogan, T., and McDonald, D., "Data-Driven Aerospace Engineering: Reframing the Industry with Machine Learning," *AIAA Journal*, Vol. 59, No. 8, 2021, pp. 2820–2847.
https://doi.org/10.2514/1.J060131

[3] 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, Feb. 2022, Paper 108119.
https://doi.org/10.1016/j.ress.2021.108119

[4] 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*, Inst. of Electrical and Electronics Engineers, New York, 2008, pp. 1–17.
https://doi.org/10.1109/PHM.2008.4711436

[5] 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, 2021, pp. 4–23, https://papers.phmsociety.org/index.php/ijphm/article/view/1336.
https://doi.org/10.36001/ijphm.2010.v1i1.1336

[6] Ochella, S., and Shafiee, M., "Performance Metrics for Artificial Intelligence (AI) Algorithms Adopted in Prognostics and Health Management (PHM) of Mechanical Systems," *Journal of Physics: Conference Series*, Vol. 1828, No. 1, 2021, Paper 012005.
https://doi.org/10.1088/1742-6596/1828/1/012005

[7] 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*, The Prognostic and Health Management Soc., 2014, pp. 1–12.
https://doi.org/10.36001/phme.2014.v2i1.1511

[8] 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*, The Prognostic and Health Management Soc., 2014, pp. 533–545.
https://doi.org/10.36001/phmconf.2014.v6i1.2520

[9] Bi, S., Prabhu, S., Cogan, S., and Atamturktur, S., "Uncertainty Quantification Metrics with Varying Statistical Information in Model Calibration and Validation," *AIAA Journal*, Vol. 55, No. 10, 2017, pp. 3570–3583.
https://doi.org/10.2514/1.J055733

[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*, The Prognostic and Health Management Soc., 2009, pp. 1–16, http://papers.phmsociety.org/index.php/phmconf/article/view/1621.

[11] Goebel, K., Celaya, J., Sankararaman, S., Roychoudhury, I., Daigle, M., and Saxena, A., *Prognostics: The Science of Making Predictions*, Createspace Independent Pub., Scotts Valley, CA, 2017.

[12] Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F., "Combining Evaluation Metrics via the Unanimous Improvement Ratio and Its Application to Clustering Tasks," *Journal of Artificial Intelligence Research*, Vol. 42, Dec. 2011, pp. 689–718.
https://doi.org/10.1613/jair.3401

[13] Baptista, M. L., Goebel, K., and Henriques, E. M., "Relation Between Prognostics Predictor Evaluation Metrics and Local Interpretability SHAP Values," *Artificial Intelligence*, Vol. 306, May 2022, Paper 103667.
https://doi.org/10.1016/j.artint.2022.103667

[14] Lewis, A. D., and Groth, K. M., "Metrics for Evaluating the Performance of Complex Engineering System Health Monitoring Models," *Reliability Engineering and System Safety*, Vol. 223, March 2022, Paper 108473.
https://doi.org/10.1016/j.ress.2022.108473

[15] de Pater, I., and Mitici, M., "Predictive Maintenance for Multi-Component Systems of Repairables with Remaining-Useful-Life Prognostics and a Limited Stock of Spare Components," *Reliability Engineering and System Safety*, Vol. 214, Oct. 2021, Paper 107761.
https://doi.org/10.1016/j.ress.2021.107761

[16] Pater, I. D., Reijns, A., and Mitici, M., "Alarm-Based Predictive Maintenance Scheduling for Aircraft Engines with Imperfect Remaining Useful Life Prognostics," *Reliability Engineering and System Safety*, Vol. 221, Jan. 2022, Paper 108341.
https://doi.org/10.1016/j.ress.2022.108341

[17] Lee, J., and Mitici, M., "Multi-Objective Design of Aircraft Maintenance Using Gaussian Process Learning and Adaptive Sampling," *Reliability Engineering and System Safety*, Vol. 218, Feb. 2022, Paper 108123.
https://doi.org/10.1016/j.ress.2021.108123

[18] Coble, J., and Wesley Hines, J., "Identifying Optimal Prognostic Parameters from Data: A Genetic Algorithms Approach," *Annual Conference of the Prognostics and Health Management Society, PHM 2009*, The Prognostic and Health Management Soc., 2009, pp. 1–11, http://www.papers.phmsociety.org/index.php/phmconf/article/view/1404.

[19] Chen, Y., Zhu, F., and Lee, J., "Data Quality Evaluation and Improvement for Prognostic Modeling Using Visual Assessment Based Data Partitioning Method," *Computers in Industry*, Vol. 64, No. 3, 2013, pp. 214–225.
https://doi.org/10.1016/j.compind.2012.10.005

[20] Omri, N., Al Masry, Z., Mairot, N., Giampiccolo, S., and Zerhouni, N., "Towards an Adapted PHM Approach: Data Quality Requirements Methodology for Fault Detection Applications," *Computers in Industry*, Vol. 127, May 2021, Paper 103414.
https://doi.org/10.1016/j.compind.2021.103414

[21] Atamuradov, V., Medjaher, K., Camci, F., Zerhouni, N., Dersin, P., and Lamoureux, B., "Machine Health Indicator Construction Framework for Failure Diagnostics and Prognostics," *Journal of Signal Processing Systems*, Vol. 92, No. 6, 2020, pp. 591–609.
https://doi.org/10.1007/s11265-019-01491-4

[22] Jiao, R., Peng, K., Dong, J., and Zhang, C., "Fault Monitoring and Remaining Useful Life Prediction Framework for Multiple Fault Modes in Prognostics," *Reliability Engineering and System Safety*, Vol. 203, Dec. 2020, Paper 107028.
https://doi.org/10.1016/j.ress.2020.107028

[23] Xiong, J., Zhou, J., Ma, Y., Zhang, F., and Lin, C., "Adaptive Deep Learning-Based Remaining Useful Life Prediction Framework for Systems with Multiple Failure Patterns," *Reliability Engineering and System Safety*, Vol. 235, March 2023, Paper 109244.
https://doi.org/10.1016/j.ress.2023.109244

[24] Braglia, M., Carmignani, G., Frosolini, M., and Zammori, F., "Data Classification and MTBF Prediction with a Multivariate Analysis Approach," *Reliability Engineering and System Safety*, Vol. 97, No. 1, 2012, pp. 27–35.
https://doi.org/10.1016/j.ress.2011.09.010

[25] Bieber, M., and Verhagen, W. J., "A Generic Framework for Prognostics of Complex Systems," *Aerospace*, Vol. 9, No. 12, 2022, pp. 1–27.
https://doi.org/10.3390/aerospace9120839

[26] 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*, The Prognostic and Health Management Soc., 2021, pp. 60–70.
https://doi.org/10.36001/phme.2021.v6i1.2868

[27] Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., and Santos, J., "Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]," *IEEE Computational Intelligence Magazine*, Vol. 13, No. 4, 2018, pp. 59–76.
https://doi.org/10.1109/MCI.2018.2866730

[28] Branco, P., Torgo, L., and Ribeiro, R. P., "Pre-Processing Approaches for Imbalanced Distributions in Regression," *Neurocomputing*, Vol. 343, May 2019, pp. 76–99.
https://doi.org/10.1016/j.neucom.2018.11.100

[29] 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

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

[31] Stanovov, V., Brester, C., Kolehmainen, M., and Semenkina, O., "Why Don't You Use Evolutionary Algorithms in Big Data?" *IOP Conference Series: Materials Science and Engineering*, Vol. 173, No. 1, 2017, Paper 012020.
https://doi.org/10.1088/1757-899X/173/1/012020

[32] Konak, A., Coit, D. W., and Smith, A. E., "Multi-Objective Optimization Using Genetic Algorithms: A Tutorial," *Reliability Engineering and System Safety*, Vol. 91, No. 9, 2006, pp. 992–1007.
https://doi.org/10.1016/j.ress.2005.11.018

[33] Jones, D. F., Mirrazavi, S. K., and Tamiz, M., "Multi-Objective Meta-Heuristics: An Overview of the Current State-of-the-Art," *European Journal of Operational Research*, Vol. 137, No. 1, 2002, pp. 1–9.

[34] Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T., "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, Vol. 6, No. 2, 2002, pp. 182–197.
https://doi.org/10.1109/4235.996017

[35] Hua, Y., Liu, Q., Hao, K., and Jin, Y., "A Survey of Evolutionary Algorithms for Multi-Objective Optimization Problems with Irregular Pareto Fronts," *IEEE/CAA Journal of Automatica Sinica*, Vol. 8, No. 2,

2021, pp. 303–318.
https://doi.org/10.1109/JAS.2021.1003817

[36] Saxena, A., Celaya, J., Saha, B., Saha, S., and Goebel, K., "Evaluating Algorithm Performance Metrics Tailored for Prognostics," *IEEE Aerospace Conference Proceedings*, Inst. of Electrical and Electronics Engineers, New York, 2009, pp. 1–13.
https://doi.org/10.1109/AERO.2009.4839666

[37] Biggio, L., Wieland, A., Chao, M. A., Kastanis, I., and Fink, O., "Uncertainty-Aware Remaining Useful Life Predictor," arXiv preprint arXiv:2104.03613, 2021, pp. 1–14, http://arxiv.org/abs/2104.03613.

[38] Frederick, D. K., DeCastro, J. A., and Litt, J. S., "User's Guide for the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS)," No. E-16205, National Aeronautics and Space Administration (NASA) Glenn Research Center, Cleveland, OH, 2007.

[39] 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*, Inst. of Electrical and Electronics Engineers, New York, 2008, pp. 1–9, http://ieeexplore.ieee.org/document/4711414/.
https://doi.org/10.1109/PHM.2008.4711414

[40] 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.

[41] Liu, R., Yang, B., Zio, E., and Chen, X., "Artificial Intelligence for Fault Diagnosis of Rotating Machinery: A Review," *Mechanical Systems and Signal Processing*, Vol. 108, Aug. 2018, pp. 33–47.
https://doi.org/10.1016/j.ymssp.2018.02.016

F. Yuan
Associate Editor