Aircraft component health analysis for predictive maintenance:

Using a dilated convolutional autoencoder and KL divergence

Pim de Ruijter

Delft University of Technology

TU Delft

Thesis, TU Delft, Faculty of 3ME, MSc Robotics
Delft, The Netherlands, compiled August 28, 2023.

Aircraft component health analysis for predictive maintenance: using a dilated convolutional autoencoder and KL divergence

Pim de Ruijter

I Abstract

The detection of anomalous behaviour is fundamental to component health analysis techniques. However, detecting anomalies is a difficult and time consuming task if their form, location, and frequency are unknown. This research introduces an innovative unsupervised predictive maintenance pipeline that requires minimal domain knowledge and time to create competitive and insightful health monitoring models. First, a Dilated Convolutional Autoencoder learns to recreate healthy sensor data. Then, a Kullback-Leibler (KL) divergence based health analysis transforms discrepancies between the reconstruction and the sensor data into a single performance metric per sensor per flight. A novel evaluation method based on the KL divergence metric allows for quantitative evaluation and hyperparameter tuning of the autoencoder. Results provide new insights and show competitive performance on analysing the fuel level measuring system. Additionally, in a generalisability study on the braking system of a different aircraft type the proposed method outperforms the currently employed health monitoring model in precision and F1 score. The main advantages of the proposed method are; the ability to rapidly create unbiased health indicators on a sensor level, the capability to generalise to other components, and a framework to quantitatively evaluate the model's performance when no truth labels are available.

II Introduction

In an industry characterised by global competition, in which each competitor has access to the exact same aircraft, it is vital to outperform the competition in efficiency on the ground and in aircraft utilisation. In other words, the fraction of time an aircraft is in the air. Predictive maintenance offers a pathway to increase aircraft utilisation by replacing or repairing components prior to failure based on data derived insights. Successful implementation of predictive maintenance can prevent unscheduled maintenance and increase efficiency during scheduled maintenance. This research focuses on creating an unsupervised health monitoring system. The method is composed specifically for the sensors tasked with measuring the fuel quantity in a wide-body aircraft. However, the applicability to other components or systems is paramount to maximising the added value of the proposed method.

II-A Fuel level measuring system

The fuel probes are ultrasonic sensors inside the tanks of the aircraft that are tasked with measuring the fuel level. A graphical representation is given in Figure 10. Two types of fuel probes work in conjunction to gather the required information. Namely targeted and regular fuel probes. The fuel probes consist of a vertical perforated tube such that the kerosene level in the tube is equal to the level in the tank. On the bottom a piezoelectric unit is attached, which can create and read ultrasonic pulses. Throughout the tanks 12 targeted fuel probes are spread out evenly, which measure the speed of sound in the kerosene. The speed of sound needs to be measured because it depends on the density, which in turn depends on the temperature, which varies vastly during operation. In the targeted fuel probes, target rings are located at set distances. When the ultrasonic pulse is created, by the piezoelectric unit at the bottom of the tube, it propagates through the kerosene and bounces back off the target rings. The time spent covering the known distance translates to the speed of sound. The 64 regular probes, as well as the 12 targeted fuel probes, record the time required for the ultrasonic wave to reflect off the fuel level. The time and speed of sound is sent to a central processing unit at a frequency of 1Hz, which subsequently calculates the distance. The central processing unit employs a rolling average over each individual sensor's measurements to remove noise. These processed distance measurements are stored on the aircraft and uploaded to servers once landed. In Appendix B, a graphical representation of the fuel tanks, the fuel probes, and the fuel probe positions are provided.

II-B Data characteristics and anomalies

For a variety of reasons, the fuel probes produce anomalous data. The anomalies range from noise to sudden fuel level drops to flat lining. In Appendix C, a variety of observed errors are displayed in addition to examples of healthy data. For some of the observed anomalies, the cause is known. A constant reading of 0.2 dm indicates a degraded insulation pad on the bottom of the sensor, allowing for the ultrasonic pulse to be reflected off the bottom of the tank. Targeted fuel probes may incorrectly interpret the return signal from one of its rings as the signal from the fuel level. This causes the data for the targeted probes to be far more jittery. For other anomalies, there is only speculation on the cause. However, it is suspected that in certain conditions, the vibration from the engine may interfere with the readings of some sensors.

The data coming from the aircraft is unlabelled. An onboard detection system does log events where the fuel level drops suddenly. Such drops are usually the result of faulty sensors. However, the fuel level measuring system, like any commercial aircraft system, is built with redundancy in mind. Therefore, a fuel drop is the result of multiple failures at once. This

makes it a rare occurrence and the system only indicates that some of the fuel probes in a particular tank are faulty without specifying exactly which. Alternatively, manually creating truth labels is impractical and time-consuming due to the anomalies' high dimensionality and non-discrete nature. These labels would be subjective and biased toward the types of anomalies the annotator can discern. Concisely, the main challenges are the absence of labels, a data imbalance between normal and anomalous data, and an unclear decision boundary between healthy and unhealthy sensors.

II-C Research questions

- How can we compute meaningful health indicators for components in commercial aircraft based on unlabelled sensor data?
- How can we optimise and quantitatively evaluate the unsupervised anomaly detection method?
- How does the proposed method compare to currently employed methods?
- How well does the proposed solution generalise to a different component?

II-D Design objectives

The following design objectives have been determined in cooperation with a major European airline.

- Maximise the discriminative ability of the anomaly detection method; The method should recreate the normal sensor behaviour as accurately as possible. This allows for precise detection of anomalies which can be translated into reliable health metrics. Accurate reconstructions make the proposed method a more capable and valuable tool for justifying predictive maintenance actions.
- Maximise the regularisation capability of the proposed method; a method that is easily adapted to a different component can either be used as a predictive maintenance tool on its own or provide rapid insights into new data for creating traditional predictive maintenance models more efficiently.
- Minimise the required human effort in data exploration and feature engineering; this allows specialists to focus on obtaining high-level insights instead of spending time on searching for sparse clues in high-dimensional data.
- Minimise the computational cost; from an environmental and financial standpoint, it is undesirable to have a solution that is excessively energy-demanding.
- Maximise the interpretability of the model; good interpretability allows the users to translate numerical results from the proposed method into in-depth insights by evaluating the intermediate steps or considerations made by the model.

II-E Contributions

The contributions of this research paper collectively address the challenges of predictive maintenance practices in the commercial aviation industry by using deep-learning techniques. A novel approach to component health analysis is introduced for tracking the health of aircraft components, supported by an evaluation framework and a demonstration of real-world generalisability. The key contributions made in this paper are:

- Innovative Predictive Maintenance Pipeline: A novel combination of a dilated convolutional autoencoder with a KL divergence-based health analysis is presented. The proposed method is based on unlabelled sensor data and outputs a health score per sensor per flight. The performance of the proposed pipeline is verified on two separate components.
- Quantitative Evaluation Framework: A quantitative framework is provided for evaluating and tuning the autoencoder and health analysis combination using real flight data interlaced with artificial anomalies. The framework offers an objective metric to assess the performance when no truth labels are available.
- Practical Impact and Real-World Generalisability: The practical relevance of the proposed method is demonstrated through comparisons with the currently employed methods. First, for anomaly detection of the fuel probes. Second, in a generalisability study a different aircraft's brake system was evaluated. The proposed method provided additional insights in both cases and outperformed the current model employed for the brakes. No definitive performance metrics could be attributed to the fuel probes due to the absence of ground truth.

III Related Work

Anomaly detection has extensively been studied across various domains. Fields such as medical diagnostics [1] [2], financial fraud detection [3], and cyber security [4] have seen a large number of publications. It is outside the scope of this paper to provide a complete overview of all available methods. However, surveys on the topic can be found in [4] and [5]. Anomaly detection in ultrasonic fuel probes has not been studied publicly before. Fortunately, it is not the physical mechanism behind the data that is decisive for the functioning of a machine learning method. Instead, it is the data and its characteristics that are most important [6]. This allows for research on different topics, based on multivariate time-series data, to be relevant for this research.

III-A Autoencoders

Autoencoders consist of two distinct parts that are executed sequentially. First, an encoder is tasked with compressing the input space into a smaller latent vector. Second, is the decoder, which tries to reconstruct the input data from the limited information in the latent vector. Theoretically, in the process the information on any anomalies is lost. Autoencoders were first used for anomaly detection by Japkowicz et al. (1995) [7]. Since then, it has been a tried and true method with new variations being developed over the years, such as convolutional (2011) [8], variational (2013) [9], and recurrent like the Long-Short Term Memory autoencoder (LSTM-AE) (2015) [10]. A dilated convolutional autoencoder has first been applied to time-series data by Yu et al. (2017) [11], with later variations including skip connections (2021) [1].

III-B Alternative methods

Time Series Forecasting (TSF): TSF [12] works by iteratively predicting the measurements for $t + 1$. The deviation of the predicted value and the actual value is the error metric. A wide range of neural networks can be used in this method, but often a type of Recurrent Neural Network (RNN) [13] is used. RNNs provide flexibility to handle time series of varying lengths. However, they are computationally more demanding compared to Convolutional Autoencoders. In previous research on comparable data types CNN have proven to perform superior to RNN-based and LSTM-based TSF [14]. Fully-connected TSF is also among the available options but it is more sensitive to changes in the length of the time series. Since the number of trainable parameters is directly related to the input size. This makes the method less suited for generalisation to other components.

Generative Adversarial Networks (GAN): GANs [15] work by making two networks (a generative network and a discriminative network) compete against each other. After training the discriminator can be used to identify anomalous flights. Unfortunately, GANs are notoriously challenging to train [16], potentially resulting in slow convergence or mode collapse. Additionally, GANs provide an anomaly score for each input sample. This score is less interpretable than the reconstructions created by an autoencoder which can be compared visually to verify the goodness of fit.

IV Method

The proposed method is based on an autoencoder tasked with recreating the input without any anomalies such that the anomalies can be determined by comparing the input and output. A health analysis then translates the observed anomalies into a health score. Figure 1 depicts a high-level overview of the proposed method. When the model is deployed, its input is a single flight and the output a health score per sensor for that specific flight. Table V presents all variables and hyperparameters, accompanied by their corresponding method of selection or reasoning.

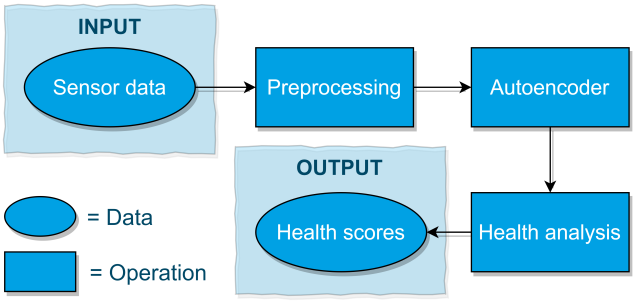


Fig. 1: High-level overview of the proposed method.

IV-A Preprocessing

Compression: Flights on this aircraft type can last up to 14 hours. This, in combination with a sampling rate of 1Hz, creates time series of up to 50.000 data points. Fortunately, the anomalies in the fuel probes often present themselves for prolonged periods. Therefore, a reduction in the number of data points can be made. To make the data quantity more manageable sub-sampling is performed by selecting every 60th data point. The onboard flight computer performs a rolling average on the sensor data before storage. Therefore, the sub-sampling is not greatly influenced by noise.

Additionally, the autoencoder requires the input size of the time series to be constant. However, flight times are variable. Therefore, the sub-sampled signal is linearly interpolated to a fixed length of 300. A side effect of the compression to a fixed length is the loss of inter-flight temporal information. In other words, the rate of fuel being used in flight A compared to flight B. However, this information is not indicative of the functioning of the fuel probes, instead it is determined by the burn rate of the engines.

Filtering erroneous data: Some of the data from the fuel probes has been corrupted in the aircraft or during export. That causes some flights to contain constant zero values, while in other flights the measurements have doubled. Through a set of filters, a large portion of the flights containing these data corruptions have been removed.

Normalisation: The fuel probe data has different ranges depending on the height of the sensor. The autoencoder will be trained with a loss function that responds stronger to larger errors. Having features with different ranges will, therefore, negatively affect the capability of the autoencoder to create a solution that is equally sensitive to anomalies in each sensor. To solve this issue, feature scaling is applied. Various preprocessing methods are widely adopted, such as feature scaling (min-max scaling) [17] [18] and Z-score normalisation [19] [20]. In this research, a variation on feature scaling is used. One of the drawbacks of feature scaling is that the presence of anomalies can greatly impact the minimum and maximum values. This, in turn, impacts the scaling performed on every data point, meaning that the normal operating range will not be compressed between zero and one but to a smaller subset. Trimming is used to combat this issue. We trim the top and bottom 2.5% from the data when determining the minimum and maximum values. This solution works excellently on the fuel probe data because the measurements stay at their respective maximum and minimum for prolonged periods.

IV-B Autoencoder architecture

The goal of the autoencoder is to create a reconstruction of the input without recreating the anomalies that were present in the input data. If that condition is met, the reconstruction of the input data can be compared with the input data to detect differences and thus, anomalies. Example figures of reconstructions compared to original data can be found in Appendix G. An autoencoder functions on the principle of restricting the flow of information. The limited available bandwidth at the centre, also known as the latent space, forces the autoencoder to distil the most essential information that is required to reconstruct the input. If the size of the latent space is set correctly, there is

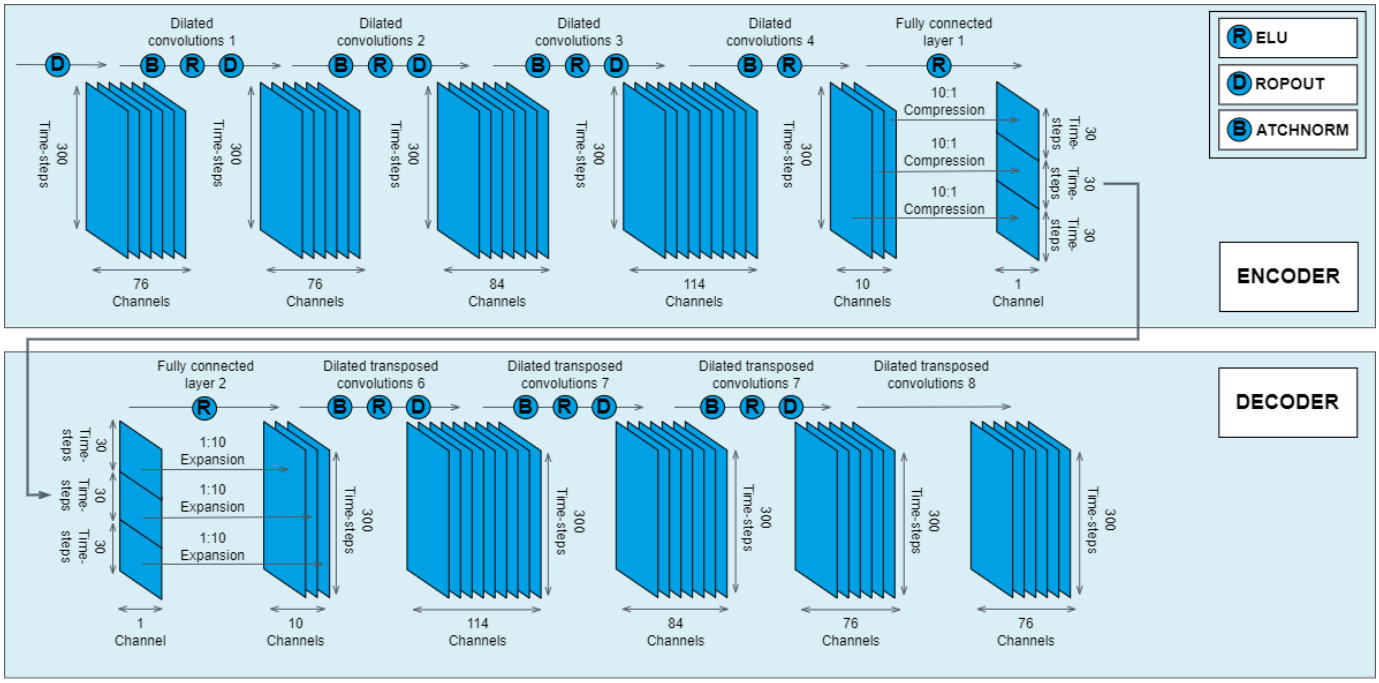


Fig. 2: The internal structure of the dilated convolutional autoencoder.

no capacity to pass on information of rare occurrences. A vital restriction flows from this reasoning. The data used for training must be predominantly healthy, otherwise the autoencoder will learn to encode and recreate the anomalous behaviour.

Dilated Convolutional AutoEncoder (DCAE): The autoencoder employed in this research is a dilated convolutional autoencoder. Convolutional networks are most frequently applied to image-related applications. However, the implementation for time series is analogous. The primary difference is that the kernel is one-dimensional (length) for a time series and two-dimensional for an image (height and width). In images, each colour is assigned a channel. In the case of the fuel probes, each input channel represents a sensor. A filter is the combination of one kernel per channel glued together and moving in its entirety along the time axis. The output dimensions of the convolutional layer will be the time series' length by the number of filters. This means that the output node can harness information from all the sensors and thus, their interrelations. However, due to the way the filter moves along the time axis, each node on the next layer only "sees" a limited part of the time series. When multiple layers are used sequentially, the receptive field increases. It is, however, undesirable to have very deep neural networks due to gradient vanishing/exploding [21] [22], increased computational complexity, and an increased risk of overfitting [23]. In order to create a larger receptive field, with a limited number of layers, dilated convolutions are used. Figure 3 shows a visualisation of the receptive field. Equation 1 gives the formula for calculating the size of the receptive field.

$$r_{i+1} = r_i + (k-1) \cdot d \tag{1}$$

Where: r_i is the receptive field of the ith layer, k is the kernel size, and d is the dilation rate.

In addition to limiting the number of layers, the use of batch normalisation [24] and ReLu [25] also reduce the effect of gradient vanishing/exploding [22]. The issue of overfitting is addressed by incorporating dropout [26] and limiting the information flow through the autoencoder's bottleneck [23].

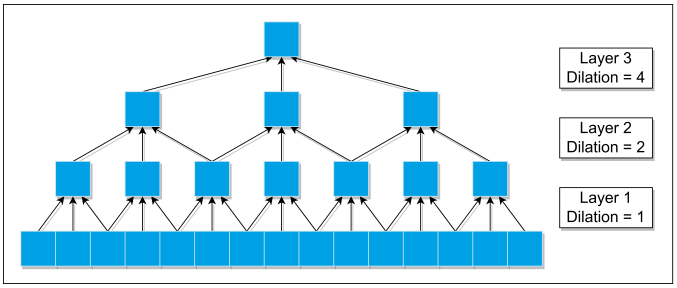


Fig. 3: Visualisation of the receptive field when convolutions with exponentially growing dilations are used. The kernel size is 3 in this illustration.

The structure of the proposed DCAE is illustrated in Figure 2. It consists of four dilated convolutional layers, which reduce the bandwidth by a factor 76:10. Followed by a fully connected (FC) layer which applies a 10:1 reduction. On the decoder side the same steps are performed in reverse. Using a FC layer as the last layer before the latent space ensures that the receptive field covers the entire input space. However, first decreasing the number of channels through dilated convolutions reduces the number of trainable parameters in the FC layer by a factor of eight, as shown in Appendix F. Cohen et al. [27] showed that stacking convolutional layers has a greater impact on the expressive power than adding the equivalent number of nodes to a single layer. The result is that the DCAE is much faster and memory efficient to train while maintaining the same expressive power, especially when utilising a GPU [28].

Loss function: The Mean Square Error (MSE) has been employed as the loss function for the autoencoder and is defined in equation 2. MSE tends to punish larger errors more strongly than the Mean Absolute Error (MAE). This is an undesirable characteristic if the training data contains large anomalies. It would push the DCAE towards learning to represent those anomalies. Unexpectedly, the DCAE underperformed when the MAE was used as is shown in the ablation study V-E. This outcome may be attributed to the fact that the MSE has a more gradual gradient towards the minima allowing for a more optimal solution to be found. From this analysis it can be concluded that the DCAE does not learn to recreate the anomalies more when using the MSE as the loss function as opposed to MAE in the specific case of the fuel probes.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

Where: n is the length of the time series, y_i is the measurement at time i, and $\hat{y}_i$ is the predicted value at time i.

IV-C Health analysis

As described in section IV-B, the autoencoder outputs a reconstruction of the sensor data. The difference between the reconstruction and the original sensor data is the predicted error. These predicted errors can be used to induce the component's performance and thus, health. Various methods have been considered to quantify the health. A straightforward approach is through the use of a threshold. However, some anomalies may be small but long-lasting. Such long-lasting anomalies can be equally indicative of failure as larger short-lived anomalies. When tasked with detecting smaller errors, a threshold-based approach would face difficulty when presented with non-perfect reconstructions created by the autoencoder. Alternatively, a learning-based method could be employed to detect anomalous patterns in the reconstruction errors. However, this would either require labelled data or another unsupervised approach. Stacking another autoencoder, or other unsupervised learning method, onto the existing autoencoder further decreases the interpretability of the pipeline. Instead, a statistics-based solution is proposed.

Histograms: First, the reconstruction errors per time-step, for one sensor and one flight, are sorted, discretised, and turned into a histogram. This histogram can be seen as a fingerprint. Examples of reconstruction error histograms of flights of various levels of anomalousness are illustrated in Appendix H. Applying the aforementioned method to a large number of flights and combining the histograms results in the multi-flight (reference) histogram. The assumption is made that the vast majority of sensor data is healthy. Therefore, the multi-flight histogram can be considered as the fingerprint of normal operation. Examples of multi-flight histograms for various sensors can be found in Figures 19b, 19c, and 19d.

Kullback-Leibler (KL) divergence: The KL divergence [29] is used to quantify the dissimilarity between the single and multi-flight histograms. The single and multi-flight histograms are first turned into probability density functions (PDF). The KL divergence calculates the relative entropy between the distributions. A large KL divergence indicates large differences in the distributions of the reconstruction error. Therefore, a large KL divergence is indicative of abnormal sensor data. The bin width of the histograms can be adjusted to affect the sensitivity to small errors. Fewer bins result in decreased sensitivity to small errors as a larger range of small errors are grouped in the first, and largest, bin. To fine-tune the performance of this metric, a weighting vector is applied. Specifically, a linear increasing vector $\mathbf{w}$ as defined in Equation 4 is used to penalise larger errors more heavily. Vector $\mathbf{w}$ can be adjusted to compensate for imperfections of the DCAE reconstruction. This is because it can decrease the influence of small errors that occur when the reconstruction is imperfect. The formula for the weighted KL divergence is given in equation 3. The main advantage of using the KL divergence instead of simply the mean of the MSEs is that the KL divergence-based method automatically adjusts the intensity of the response based on how erratic the average flight is.

$$KL = \sum_{i=1}^{N_b} p_i \cdot w_i \cdot \log\left(\frac{p_i}{q_i}\right) \tag{3}$$

$$\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{N_b} \end{pmatrix}, \; w_i = 1 + k \cdot \frac{i}{N_b}, \; i \in \{0, 1, \dots N_b\} \tag{4}$$

Where:

- KL is the weighted KL divergence.
- p is the probability density function of reconstruction errors for the evaluated flight.
- q is the reference probability density function of reconstruction errors created by averaging the PDFs of a large number of flights.
- N_b is the number of bins used in histograms of p and q.
- k is a hyperparameter that defines the slope of the linearly increasing vector $\mathbf{w}$.

Reconstruction error metric: For the selection of the reconstruction error metric, three methods have been considered, namely, MSE, MAE, and Dynamic Time Warping (DTW) [30]. Each metric offers distinct advantages and drawbacks. However, ultimately MSE has been selected. This is due to it being orders of magnitude faster than DTW. Additionally, when compared to MAE, MSE has the advantage of penalising large errors more heavily while being lenient towards small errors. This is because the bins used for creating the histograms are spaced linearly while MSE quadrates the errors. Therefore, making a larger range of small errors fall into the same bin. Conversely, larger errors become more spread out.

Meanwhile, DTW does present a notable advantage over MSE and MAE. DTW sidesteps the difficulty experienced by the other methods when presented with steep gradients. To be precise, when a small timing error occurs between the reconstruction and the measurements, and the gradient is steep,

both MSE and MAE register disproportionately large errors. DTW instead finds the optimal alignment path between two time series. From this alignment path, a distance metric can be calculated for each point. Small errors in timing will, therefore, only result in small reconstruction errors. In Figure 30, the effect of using DTW compared to MSE for the reconstruction error histograms is illustrated.

IV-D Evaluating the autoencoder

The autoencoder can be evaluated by visually comparing the original measurements and the reconstructions of individual flights. Because a well-functioning autoencoder accurately recreates the healthy sensor data points but does not recreate any anomalies. However, evaluating the autoencoder based on quantitative results takes human bias out of the equation. Additionally, some types of anomalies are difficult to identify manually, especially when the appearances of anomalies are unknown. Instead, to quantitatively evaluate the performance artificial pseudo labels are used.

Introducing Artificial Anomalies: The first step is identifying the types of anomalies that occur. After which, artificial anomalies that are similar to the real anomalies can be inserted programmatically. By definition, the labels of the inserted anomalies are known. However, the dataset is only partially labelled, as the synthetic anomalies are inserted in addition to preexisting real anomalies. In Appendix D, data sets with synthetic anomalies are visualised. Two distinct types of anomalies are inserted in separate fuel probe data sets:

1) Spikes and prolonged constant errors (Spikes): The original sensor data is augmented by adding or subtracting a variable number of spikes of random height and length. A fraction of the spikes' duration is extended by a large factor to form prolonged constant errors. E.g. Figures 13b and 14.
2) Inter-sensor fuel level discrepancy (Time-shift): The original sensor data is shifted forward or backwards in time to simulate a sensor having a constant deviation from the actual fuel level. E.g. Figure 13a.

Evaluation: In section IV-C, it was determined that supervised learning methods were unfit due to a lack of labels. However, even when artificial pseudo labels are available, a supervised learning solution would be challenging to train and tune effectively. This is because only the artificial anomalies are labelled. Therefore, only true positives and false negatives can be determined. Instead, optimisation through maximising the increase of weighted KL divergence as a result of inserted anomalies is proposed. The increase of the weighted KL divergence is notated as ΔKL and defined as the difference between the KL after inserting artificial anomalies and the KL prior to inserting artificial anomalies. This method does not rely on false positives and true negatives. The objective is to maximise the weighted KL divergence increase for sensors where synthetic anomalies are inserted. However, the weighted KL divergence needs to remain unchanged for unaltered sensors. This objective is formularised by creating a weighting variable v_i for every sensor in Equation 6. Variable v_i is multiplied by the change in weighted KL divergence per sensor in Equation 5. v_i is crafted such that the sum of all v_i is zero. Consequently, the importance of maximising the ΔKL for altered sensors is set equal to the importance of maintaining $\Delta KL = 0$ for unaltered sensors.

$$\Delta KL_c = \sum_{i=1}^{N_s} \left(KL_i^{anom} - KL_i^{ref} \right) \cdot v_i \tag{5}$$

$$\text{where} \quad v_i = \begin{cases} 1 - F_a & \text{if } S_i \in S_{SA} \\ -F_a & \text{if } S_i \notin S_{SA} \end{cases} \tag{6}$$

Where:

- ΔKL_c is the combined increase in KL divergence of all sensors as a result of inserting artificial anomalies and applying weighting variable v_i.
- N_s is the number of sensors.
- KL_i^{ref} is the weighted KL divergence (KL) of sensor i prior to adding synthetic anomalies.
- KL_i^{anom} is the weighted KL divergence (KL) of sensor i after synthetic anomalies are added.
- F_a is the average fraction of sensors that have had synthetic anomalies added.
- S_i is the i^{th} sensor.
- S_{SA} is the set of sensors that have had synthetic anomalies added.

The ΔKL_c is calculated per flight. However, evaluating the performance of the DCAE on a single flight will produce noisy results. Therefore, the average ΔKL_c is taken over a large number of flights, denoted as $\overline{\Delta KL}_c$. The $\overline{\Delta KL}_c$ can then be used as a metric to evaluate the performance of the DCAE.

V Experiments

V-A Hyperparameter Optimisation

Set-up: The goal of hyperparameter optimisation is to obtain the hyperparameters for which the discriminative ability of the autoencoder and health analysis combination is maximised. The performance indicator being optimised is the $\overline{\Delta KL}_c$. Keeping the health analysis method constant allows us to induce the relative performance of the DCAE variation. Due to computational constraints, not all hyperparameters are evaluated. The following two hyperparameters are deemed most influential, as these determine the size of the latent space.

- H_1: Number of filters in 4th and 5th convolutional layer.
- H_2: Reduction factor realised by fully connected layers.

Figure 4 illustrates a single cycle in the hyperparameter tuning process. This process is repeated for each new set of hyperparameters. The hyperparameter pairs are selected by means of a grid search [31].

Results: In Appendix J the hyperparameter optimisation results are visualised as loss landscapes of the hyperparameter space. In table I, the best-performing hyperparameter combinations are provided for both artificial anomaly types. Sudden drops (spikes) are most indicative of fuel probe failure.

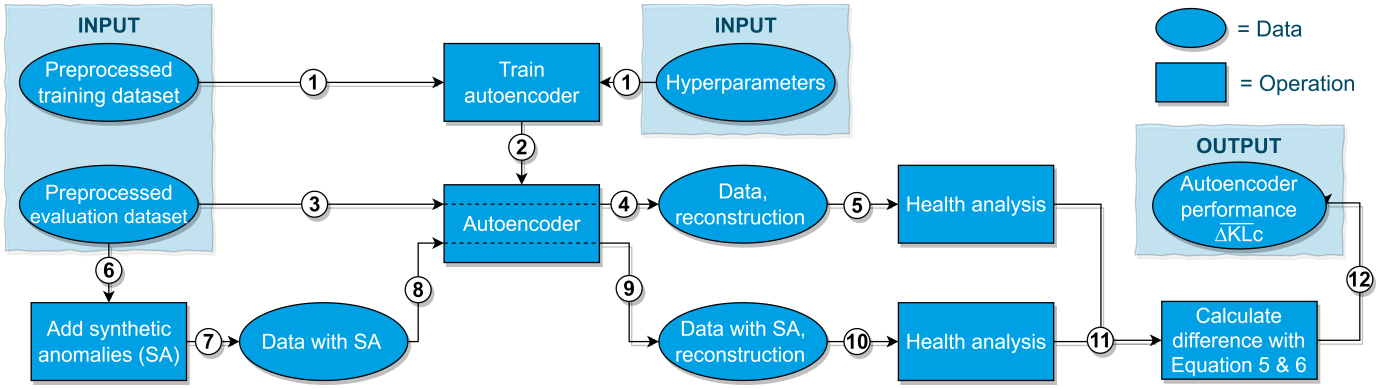


Fig. 4: A flowchart showing the optimisation process for the autoencoder.

Therefore, extra emphasis is given by considering small and large spikes separately. The best overall hyperparameters are determined by ranking hyperparameter pairs based on the performance per anomaly type and summing the three ranks. The hyperparameter combination with the lowest sum of ranks is $H_1 = 10$ and $H_2 = 0.1$. Thus, for these hyperparameters the best overall performance is observed.

TABLE I: Best performing DCAE hyperparameter combinations for different types of artificial anomalies.

Rank	Time-shift		Small spikes		Large spikes	
	H_1	H_2	H_1	H_2	H_1	H_2
1st	10	0.1	10	0.1	17	0.1
2nd	10	0.3	25	0.5	10	0.1
3rd	31	0.5	22	0.9	13	0.5
4th	17	0.5	7	0.7	19	0.3
5th	28	0.7	7	0.9	28	0.5

Interpretation: The results are in line with expectations. The size of the latent space, as a result of these hyperparameters, is only 1.3% of the input size. Such high rates of compression indicate that the underlying logic exhibited by the fuel probes is efficiently captured by the autoencoder. It is important to note that the training process of the autoencoder is stochastic. The goodness of fit, therefore, varies not only as an effect of changing the hyperparameters but also partially by randomness. This too, can be observed in the figures in Appendix J, as local minima and maxima can be found right next to each other. This stochastic behaviour is also the reason why a grid search was used. Bayesian [32], Genetic [33] or other informed optimisation techniques often obtain better results with fewer iterations [34]. However, multiple runs of a Bayesian optimiser generated highly varying optimal solutions. Consequently, this provided minimal insights into the loss landscape of the hyperparameter space. Lastly, it can be concluded that the performance of the DCAE is not solely related to the size of the latent space. Setting $H_1 = 1$ and $H_2 = 0.9$ results in a considerably less capable autoencoder compared to setting $H_1 = 10$ and $H_2 = 0.1$, even though the size of the latent space is near identical. This observation can be explained by considering that setting $H_1 = 1$ means that a single filter has to compress 114 channels down to 1 with very limited freedom on how to rearrange the data. A FC layer tasked with the same compression has more trainable parameters and is not limited by a kernel, therefore, it has more freedom to encode the data. These findings provide additional insight into the complex interdependence of hyperparameters, system architecture and performance.

V-B Quantitative Evaluation

Set-up: No ground truth labels are available for quantifying the performance. Therefore, the quantitative evaluation of the proposed method is performed through sensitivity studies instead. The objective of the sensitivity studies is to provide insight into the proposed method's response to anomalies. In the sensitivity studies, the weight vector $\mathbf{w}$ from equation 4 is set to a constant value of 1. This is done to improve the interpretability of the results. We are interested in the natural response of the DCAE and health analysis combination, not one that is altered by a variable weight vector.

Results: Figure 5 shows the responses of the proposed method as a result of inserting artificial spikes and prolonged constant errors for three different fuel probes. The response is in terms of $\Delta\overline{MSE}$ and $\overline{\Delta KL}$.

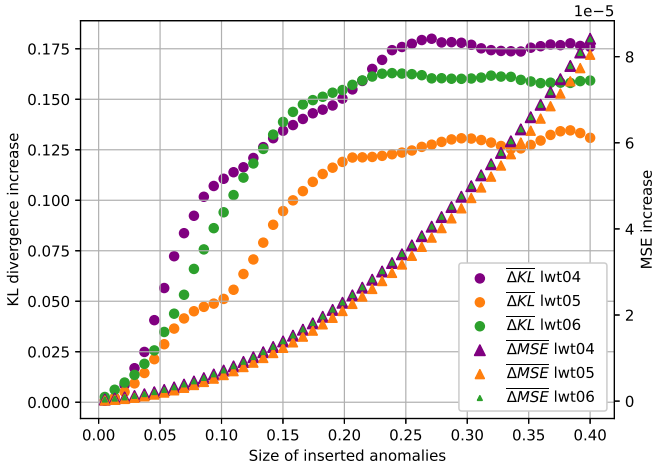


Fig. 5: Sensitivity study on 3 separate fuel probes (lwt04, lwt05, lwt06) with artificial anomalies of type: spikes and prolonged constant errors. Weighting vector $\mathbf{w}$ is set to 1.

Figure 6 shows the combined response to spike anomalies of all fuel probes. The response is shown in terms of the sum of

$\Delta\overline{MSE}$ over all sensors and $\Delta\overline{KL}_c$. Additionally, in Figure 16b the response as a result of inserting inter-sensor fuel level discrepancies (time-shifts) is illustrated.

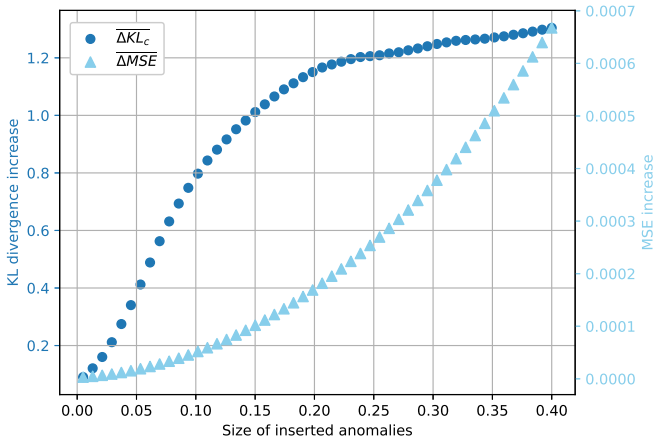


Fig. 6: Sensitivity study on all sensors with artificial anomalies of type: spikes and prolonged constant errors. Weighting vector $\mathbf{w}$ is set to 1.

Interpretation: Figure 5 illustrates the desirable effect of using a metric based on KL divergence. The $\overline{\Delta KL}$ response is dependent on the level of anomalousness of the average flight for a specific sensor. In all three sensors the same artificial anomalies are added. However, lwt05 is a targeted fuel probe and has the highest levels of real anomalies per average flight of the three depicted sensors. The $\overline{\Delta KL}$ response is therefore the weakest for lwt05. In Figure 19 the $\overline{\Delta KL}$ responses of lwt04, lwt05, and lwt06 are also shown in combination with the corresponding reconstruction error histograms.

From Figure 6 it can be concluded that inserted spike anomalies of all sizes result in an increase of the $\overline{\Delta KL}_c$. The KL divergence increase starts to level off when the errors pass the 0.1 mark. The primary reason for this is that the KL divergence is based on the likelihood of a specific size anomaly to occur. From the histogram of all sensors combined in Figure 16c it can be seen that errors of size 0.4 are only around 10 times as rare as errors of size 0.1. In contrast, errors of size 0.1 are 10.000 times as rare as errors of size 0.

Even though the KL divergence increase is directly related to the distribution of the reconstruction errors, the response seen in Figure 6 and histogram in Figure 16c do not match exactly. This is partly due to synthetic anomalies that occur on top of preexisting anomalies, effectively creating an anomaly of a rarer size which disproportionately increases the KL divergence for such a small synthetic anomaly. Additionally, when calculating the $\overline{\Delta KL}_c$, each sensor is evaluated and scored separately with the probability density function of that specific sensor and weighted according to Equation 6.

V-C Qualitative evaluation

Set-up: Sensitivity studies were used to evaluate the behaviour of the proposed method in a quantitative manner in response to well defined anomalies. However, it is also essential to understand how the DCAE responds to the wide variety of situations present in the real (non-artificial) data. In what cases does the proposed method behave unexpectedly?

Results: In Appendix L, weighted KL divergences are plotted per flight per sensor over multiple flights. From this overview, individual flights have been selected and evaluated on goodness of fit and to identify potential shortcomings of the proposed method. The general observation is that in the vast majority of flights the DCAE accurately reconstructs the sensor measurements regardless of the level of anomalousness. However, two types of undesirable behaviour have been identified:

1) Flights that are of an extreme length prove to be difficult for the DCAE to recreate. Examples of such a flight can be observed in Figures 24c and 26c.
2) Anomalies that affect a large number of closely related fuel probes negatively influence the quality of the recreation. In Figures 29a and 29b a case is observed where from 12/2021 to 12/2022 a large portion of sensors had constant zero readings, leading to an incorrect constant zero reconstruction, and subsequent incorrect zero reconstruction error.

Interpretation: Convolutional networks are often thought of as translation invariant. This is due to the kernel that passes over the entire time series, or image, which detects patterns regardless of the location. However, Kauderer et al. [35] and Biscione et al. [36] show the limitations and highlight the importance of data augmentation in order to train the model on the entire input space. The undesirable behaviour stated above in point one, concurs with these findings. The interrelations between the sensors are the same regardless of flight length. Only the events in the data, such as the moment where the measured fuel level starts to drop or when it hits zero, are occurring at a later time. In other words the data is translated along the time axis, but the reconstruction becomes poor.

The finding in point two, is one that highlights the importance of data validation. It is debatable whether a large scale data corruption is an error that can be accredited to the proposed method or whether this is besides its designed objective. It is a failure not of the fuel probes but of a completely different system. The proposed method does register anomalies at the edge cases. Namely, the fuel probes that neighbour the corrupted fuel probes. These edge cases output high KL divergences for the affected flights.

Overlaying the output of the proposed method per flight (exemplar Figure 25a), with the replacement dates of fuel probes, would provide a valuable insight into the effect of fuel probe replacements and the model's functioning. Additionally, it would allow for the precision and recall to be calculated. Unfortunately, due to the employed replacement policy, a large number of sensors is replaced rather than just the problematic fuel probes. Therefore, it is impossible to determine if a replaced probe, for which the proposed method did not output a high weighted KL divergence, is a false negative or true negative. Additionally, faulty fuel probes can remain undetected for prolonged periods due to the redundancy in the system.

False positives in the proposed method could, therefore, also be true positives. For these reasons, no evaluation is made based on replacements as these cannot reliably be regarded as ground truth labels of the fuel probe data.

V-D Comparison to baseline

Set-up: The baseline method, which is currently employed at a major European airline, is based on a set of rules drafted from expert knowledge. In order to compare the behaviour of both approaches quantitatively, a sensitivity study as introduced in Section V-B is applied. The output of the baseline method is binary, either healthy or unhealthy. By evaluating the baseline over a large number of flights the average response is determined. The flights onto which synthetic anomalies are added, already contain real anomalies. Therefore, the baseline method marks a percentage of flights as unhealthy prior to any synthetic anomalies being added. These flights are removed from the sensitivity study of the baseline.

In addition to the quantitative sensitivity studies, the predictions of the baseline method are qualitatively compared to the results of the proposed method to identify differences and (dis)advantages in Appendix L.

Results: The response to artificial anomalies of the baseline and the proposed method for a single sensor is illustrated in Figure 7. Responses of other sensors are shown in Appendix K.

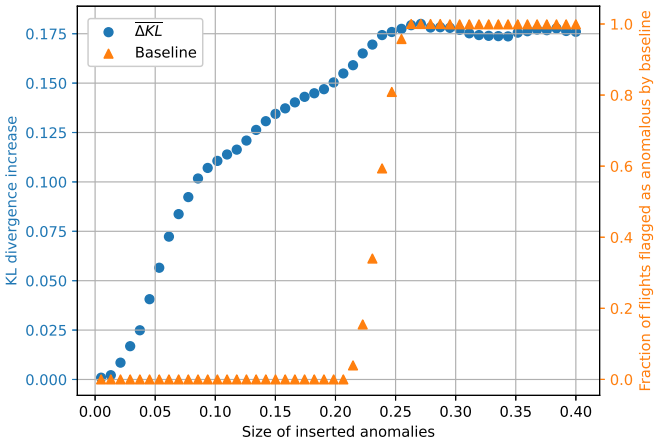


Fig. 7: Responses of the baseline and proposed method to injected spike anomalies of incremental sizes. Weighting vector $\mathbf{w}$ is set to 1.

Interpretation: The baseline uses a threshold for the size of anomalies in order to determine outliers. That is why the baseline method has an acute rise in detections once the synthetic anomalies reach this size. It is remarkable that the threshold used in the baseline seems to coincide with the saturation of the KL divergence increase of the proposed method. This implies that the size range of anomalies considered equally unlikely by the proposed method coincides with the size range of anomalies selected by an industry expert as anomalous in the baseline method prior to this research.

The binary approach of the baseline method limits the quality of the insights that can be obtained. In contrast, the proposed method provides a continues output allowing for the level of anomalousness to be determined and trends to be observed. In Figure 25a a clear trend can be observed where the level of anomalousness is lower during the summer months and higher in the winter. This is in line with experts' suspicions on the negative effect that kerosene anti-freeze additives have on the fuel probe's performance. This trend could not be detected with the baseline method.

V-E Ablation Study

Set-up: An ablation study is performed to verify the necessity, and quantify the influence, of various components in the DCAE architecture. Sensitivity studies are performed with variations in the model architecture on two types of synthetic anomalies. The performance is evaluated relative to the proposed method according to Formula 7.

$$P_a = \frac{1}{N_a} \sum_{j=1}^{N_a} \frac{\overline{\Delta KL}_{c,ablation}(a_j)}{\overline{\Delta KL}_{c,proposed}(a_j)} \tag{7}$$

Where:

- P_a is the averaged performance of the ablated model relative to the proposed model.
- a_j is the artificial anomaly size (e.g. spike height or amount of time-shift).
- N_a is the number of artificial anomaly sizes evaluated in the sensitivity study.
- $\overline{\Delta KL}_{c,ablation}(a_j)$ is the combined increase in KL divergence of all sensors, averaged over multiple flights, as a result of inserting artificial anomalies of size a_j into the ablated model.
- $\overline{\Delta KL}_{c,proposed}(a_j)$ is the combined increase in KL divergence of all sensors, averaged over multiple flights, as a result of inserting artificial anomalies of size a_j into the proposed model.

Results: In table II, the averaged relative results over a range of anomaly sizes are displayed numerically. In Figure 31 the results of the sensitivity studies are displayed graphically.

TABLE II: Ablation study based on KL divergence increase in sensitivity studies, graphically illustrated in Figure 31.

	Relative avg. perf. (P_a) Artificial Spikes	**Relative avg. perf. (P_a) Artificial Time-shift**
Proposed model	1.000	1.000
No dropout	1.029	0.917
No batchnorm	0.726	0.754
MAE loss func.	0.980	0.923
-2 Conv. layers	0.923	0.892
+2 Conv. layers	0.969	0.922

Interpretation: The results show that the undertaken ablations all result in a decrease in performance with the sole exception of ablating dropout on spike anomalies. A likely explanation is that the DCAE requires little information on other

sensors to recreate sensor data without the spikes. However, dropout forces the reliance on other sensors as during training some channels are nullified. However, for detecting time-shift anomalies, information from other sensors is essential, this results in a strong decrease in performance when dropout is ablated. Including dropout makes the architecture better suited to handle more complicated systems where capturing sensor interrelations is paramount to identifying anomalies.

V-F Generalisability Study

Set-Up: One of the main intended advantages of the proposed method is its generalisability. In order to validate this assertion, the proposed method has been applied to a different component in a different aircraft type. Specifically, the braking system, which consists of 32 actuators evenly spread across 8 tires. For each actuator the current and displacement is logged at a frequency of 1Hz, resulting in 64 channels. The braking system was chosen because it, like the fuel probes, is a high dimensional system with sensor interdependencies.

In this generalisability study, only the preprocessing step has been altered. Specifically, no interpolation was used and the output time-series' lengths are shorter. This is because the brakes only function for a short window of time during landing. Contrary to the fuel probes, which gradually change over the course of the entire flight. All other steps and hyperparameters remain unchanged.

A vital difference between the brakes and the fuel probes, is that the brakes are changed on an individual basis and the exact defects are logged. This allows for performance metrics to be constructed based on the findings obtained during the performed maintenance. The results of the maintenance reports can be used as ground truth labels. The output of the model, which is in terms of weighted KL divergence, is thresholded in order to calculate the precision, recall, and subsequent F1 score. The values used for the thresholds are selected to maximise precision. Non-optimal precision would lead to additional costs for unnecessary maintenance and might lead to a decreased end-user trust in the model's predictions.

Results: The test set spans a 10 month period in which 10 known actuator failures have occurred and 21 brake pads were replaced due to wear. Each actuator has two sensors, one measuring displacement, the other current. Examining the weighted KL divergence response for each sensor separately generated the following insights.

- Displacement sensors: High weighted KL divergences are observed when an actuator is about to break or broken (Exemplar Figure 32a).
- Current sensors: High weighted KL divergences are observed when the brake pads are worn and need replacement within a few months (Exemplar Figure 32b).

Table III shows the performance metrics of the proposed method based on thresholding of the weighted KL divergences of the displacement sensors. Additionally, the currently employed method's performance metrics are given. The implementation details of the proposed method's thresholds can be found in Table V. In Figure 32a a true positive example of an actuator failure, based on a displacement sensor, is illustrated.

TABLE III: Performance metrics on detecting actuator failure

	Precision	**Recall**	**F1 score**
Current method	0.5	0.3	0.375
Proposed method	1.0	0.3	0.462

Table IV shows the performance metrics of the proposed method for detecting brake pad wear. These are obtained by setting a threshold for the weighted KL divergences of the current sensors as specified in Table V. Figure 32b shows an example of a true positive for brake pad wear based on a current sensor. No method is currently in use to detect break wear to which the proposed method can be compared.

TABLE IV: Performance metrics on detecting brake pad wear

	Precision	**Recall**	**F1 score**
Proposed method	1.0	0.619	0.765

Interpretation: The proposed DCAE based method outperforms the existing method on precision and F1 score. It is also likely that with optimisation of the hyperparameters an even better result can be obtained.

In the case of the brakes the two failure modes manifest as anomalies in two separate sensor types. However, this does bring to light the possibility where multiple failure modes exist, and anomalies created by a multitude of failure modes manifest in a single sensor. It then becomes practically impossible to discern from the output of the proposed method what failure mode is occurring.

VI Conclusion

In this paper, a novel predictive maintenance pipeline was introduced. The purpose of which is to detect unhealthy components in commercial aircraft based on anomalous behaviour in unlabelled sensor data. It is the first implementation of a dilated convolutional autoencoder combined with a KL divergence-based health analysis. The proposed method has quantitatively proven the ability to accurately reconstruct the healthy behaviour of the fuel probes when the measurement data contains spikes or inter-sensor fuel level discrepancies through a series of sensitivity studies.

Additionally, the hyperparameter tuning solution driven by an increase in combined weighted KL divergences ($\Delta\overline{KL_c}$) allowed for quantitative tuning of the autoencoder without the need for real labelled data. Only a low-level understanding of the occurring types of anomalies was required to create datasets with comparable labelled synthetic anomalies.

It has been shown through the generalisability study that the proposed method can be of added value as a general predictive maintenance tool as it outperformed the currently employed solution for the brakes in both precision and F1 score. Additionally, this result was achieved without any component specific hyperparameter tuning, substantiating the time saving ability of the proposed method.

The limited interpretability of the DCAE's inner workings may complicate implementation in some cases. In such cases, the proposed framework can be effectively used as an exploratory tool to pinpoint sensors and specific flights that contain informative irregular behaviour. In this role, it would aid in accelerating the development of more traditional predictive maintenance models.

VI-A Future work

The following topics have been identified in which future work may provide additional insights or improved performance of the proposed method.

- Rarely will all sensors fail at once. Therefore, if the reconstruction errors on all sensors are high, it is an indication that the DCAE has likely created a poor reconstruction. This information can be used to either fit a different variation of the DCAE on this flight to obtain a better reconstruction, or to nullify the results.
- Extremely long flights were found to be a cause of poor reconstructions. Data augmentation can aid in providing the DCAE with more training data of such long flights.
- The level of noise in the loss function of the hyperparameter tuning could be reduced by training multiple models with the same hyperparameters and averaging the results.
- Additional information, useful for creating accurate reconstructions, may be available from other sensors. These data streams would not require reconstruction as they do not convey information on the evaluated component(s). However, the DCAE could be made asymmetrical, allowing for more input channels than output channels. This would require a modified loss function, but may allow for better reconstructions.
- Systematically verify on what types of components the proposed method is able to perform well. What are the data characteristics that determine if the model generalises well to a new component?

VII Acknowledgements

I would like to express my gratitude to Holger Caesar as my supervisor at the TU Delft, as well as Dennis van den Berg and Martijn Oerlemans as daily supervisors, for providing invaluable guidance and support throughout this thesis.

References

[1] M. Thill, W. Konen, H. Wang, and T. Bäck, "Temporal convolutional autoencoder for unsupervised anomaly detection in time series," *Applied Soft Computing*, vol. 112, p. 107751, 2021.

[2] N. Shvetsova, B. Bakker, I. Fedulova, H. Schulz, and D. V. Dylov, "Anomaly detection in medical imaging with deep perceptual autoencoders," *IEEE Access*, vol. 9, pp. 118 571–118 583, 2021.

[3] M. Ahmed, A. N. Mahmood, and M. R. Islam, "A survey of anomaly detection techniques in financial domain," *Future Generation Computer Systems*, vol. 55, pp. 278–288, 2016.

[4] D. Kwon, H. Kim, J. Kim, S. C. Suh, I. Kim, and K. J. Kim, "A survey of deep learning-based network anomaly detection," *Cluster Computing*, vol. 22, pp. 949–961, 2019.

[5] H. Wang, M. J. Bah, and M. Hammad, "Progress in outlier detection techniques: A survey," *Ieee Access*, vol. 7, pp. 107 964–108 000, 2019.

[6] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264.

[7] N. Japkowicz, C. Myers, M. Gluck *et al.*, "A novelty detection approach to classification," in *IJCAI*, vol. 1. Citeseer, 1995, pp. 518–523.

[8] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21*. Springer, 2011, pp. 52–59.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[10] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International conference on machine learning*. PMLR, 2015, pp. 843–852.

[11] Y. Yu, J. Long, Z. Cai *et al.*, "Network intrusion detection through stacking dilated convolutional autoencoders," *Security and Communication Networks*, vol. 2017, 2017.

[12] S. Schmidl, P. Wenig, and T. Papenbrock, "Anomaly detection in time series: a comprehensive evaluation," *Proceedings of the VLDB Endowment*, vol. 15, no. 9, pp. 1779–1797, 2022.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[14] S. Selvin, R. Vinayakumar, E. Gopalakrishnan, V. K. Menon, and K. Soman, "Stock price prediction using lstm, rnn and cnn-sliding window model," in *2017 international conference on advances in computing, communications and informatics (icacci)*. IEEE, 2017, pp. 1643–1647.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2014.

[16] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[17] A. Jiménez-Cordero and S. Maldonado, "Automatic feature scaling and selection for support vector machine classification with functional data," *Applied Intelligence*, vol. 51, pp. 161–184, 2021.

[18] Y. N. Kunang, S. Nurmaini, D. Stiawan, A. Zarkasi *et al.*, "Automatic features extraction using autoencoder in intrusion detection system," in *2018 International Conference on Electrical Engineering and Computer Science (ICECOS)*. IEEE, 2018, pp. 219–224.

[19] A. Dallali, A. Kachouri, and M. Samet, "Classification of cardiac arrhythmia using wt, hrv, and fuzzy c-means clustering," *Signal Processing: An Int. J.(SPJI)*, vol. 5, no. 3, pp. 101–109, 2011.

[20] N. Fei, Y. Gao, Z. Lu, and T. Xiang, "Z-score normalization, hubness, and few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 142–151.

[21] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber *et al.*, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," 2001.

[22] S. Basodi, C. Ji, H. Zhang, and Y. Pan, "Gradient amplification: An efficient way to train deep neural networks," *Big Data Mining and Analytics*, vol. 3, no. 3, pp. 196–207, 2020.

[23] M. M. Bejani and M. Ghatee, "A systematic review on overfitting control in shallow and deep neural networks," *Artificial Intelligence Review*, pp. 1–48, 2021.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[25] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[26] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[27] N. Cohen, O. Sharir, and A. Shashua, "On the expressive power of deep learning: A tensor analysis," in *Conference on learning theory*. PMLR, 2016, pp. 698–728.

[28] K. Chellapilla, S. Puri, and P. Simard, "High performance convolutional neural networks for document processing," in *Tenth international workshop on frontiers in handwriting recognition*. Suvisoft, 2006.

[29] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[30] R. Bellman and R. Kalaba, "On adaptive control processes," *IRE Transactions on Automatic Control*, vol. 4, no. 2, pp. 1–9, 1959.

[31] P. Liashchynskyi and P. Liashchynskyi, "Grid search, random search, genetic algorithm: a big comparison for nas," *arXiv preprint arXiv:1912.06059*, 2019.
[32] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," *Advances in neural information processing systems*, vol. 25, 2012.
[33] N. M. Aszemi and P. Dominic, "Hyperparameter optimization in convolutional neural network using genetic algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, 2019.
[34] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020.
[35] E. Kauderer-Abrams, "Quantifying translation-invariance in convolutional neural networks," *arXiv preprint arXiv:1801.01450*, 2017.
[36] V. Biscione and J. S. Bowers, "Convolutional neural networks are not invariant to translation, but they can learn to be," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 10 407–10 434, 2021.
[37] D. Choi, C. J. Shallue, Z. Nado, J. Lee, C. J. Maddison, and G. E. Dahl, "On empirical comparisons of optimizers for deep learning," *arXiv preprint arXiv:1910.05446*, 2019.
[38] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.
[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
[40] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.
[41] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.
[42] M. M. McKerns, L. Strand, T. Sullivan, A. Fang, and M. A. Aivazis, "Building a framework for predictive science," *arXiv preprint arXiv:1202.1056*, 2012.
[43] Tosaka. (2010) Jet-liner's main fuel tanks. [Online]. Available: https://commons.wikimedia.org/wiki/File:Jet-liner%27s_main_fuel_tanks(B-777).PNG

Appendix A
Table of variables and implementation details

TABLE V: Variables with selected values and method of selection.

Group	Variable name	Value / Range	Reasoning / Method of determining value
Preprocessing	Down sampling factor	60	One measurement per minute
	Minimum input length cutoff	400	Objective is to maximise to have the largest number of flights available. Threshold set around retaining 90% of flights.
	Fixed length after compression	300	Objective is to minimise to reduce computational expense. Manual inspection showed most anomalies remained present.
	Percentile threshold for clipping in minmax normalisation	Min 2.5%, Max 97.5%	Evaluate that values used for clipping match to the length of thefuel probes.
Autoencoder architecture	Dropout	0.2	Brute force search.
	Number of filters layer 1	Input channels (76)	Brute force search.
	Number of filters layer 2	1.1 · Input channels (84)	Brute force search.
	Number of filters layer 3	1.5 · Input channels (114)	Brute force search.
	Number of filters layer 4	10	Result of hyperparameter optimisation.
	Size of output dense layer	0.1 · Input length · Channels layer 4 (300)	Result of hyperparameter optimisation.
	Learning rate	0.01, 0.001, 0.0001	Verified visually with training and validation loss.
	Epochs	200, 500, 100	Verified visually with training and validation loss, loss no longer decreases.
	Kernel size	5	Brute force search.
	Number of convolutional layers	4	Brute force search.
	Activation function	Relu	Based on choice in [1].
	Optimizer	Adam	Adam generally performs best when hyperparameters are tuned well [37].
	Loss function	MSE	Verified in ablation study.
Artificial anomalies: Spikes	Max fraction of sensors with added anomalies	0.2	Close to what was observed in the real data by manual inspection.
	Anomaly length	1 to 4 time steps	Close to what was observed in the real data by manual inspection.
	Anomaly frequency	3 per flight	The baseline method requires at least 3 spikes to trigger. Additionally, minimize the chance of anomalies overlapping.
	Size for small spikes data set	0.01 to 0.1	Based on the behaviour observed in mostly healthy non-targeted fuel probes.
	Size for large spikes data set	0.1 to 0.4	Based on the behaviour observed in targeted fuel probes.
Artificial anomalies: Time-delay	Max fraction of sensors with added anomalies	0.2	Difficult to observe, set equal to spikes.
	Size of time delay	0.005 to 0.05	Close to what was observed in the real data by manual inspection.
Health analysis	Metric for reconstruction error	MSE	Discussed in Section IV-C.
	Number of bins	300	Set equal to the length of the time-series.
Generalisability study: Preprocessing	Input length	80	Time required for touchdown to stand still.
Generalisability study: Thresholding Actuator failure	Weighted KL divergence	2 or higher	Brute force, optimise towards max. precision.
	Occurrences	3 or more	Brute force, optimise towards max. precision.
	Timeframe	Within 50 flights	Approximate number of flights in one month.
Generalisability study: Thresholding Brake pad wear	Weighted KL divergence	2.8 or higher	Brute force, optimise towards max. precision.
	Occurrences	3 or more	Brute force, optimise towards max. precision.
	Timeframe	Within 50 flights	Approximate number of flights in one month.

TABLE VI: Implementation details

Library name	**Usage**	**Version**	**Reference**
Python	General	3.9.7	Van Rossum et al [38]
PyTorch	Create the DCAE architecture	2.0.0 + cu117	Paszke et al. [39]
Optuna	Visualise the hyperparameter tuning results	3.3.0	Akiba et al. [40]
FastDTW	Calculate the DTW path	0.3.4	Salvador and Chan [41]
Dill	Save and load trained PyTorch models	0.3.7	Mckerns et al. [42]

.

Appendix B
Fuel level measuring system

The fuel in the aircraft's tanks is compartmentalised in the left-wing, centre, and right-wing tanks. First, the fuel in the centre tank is consumed by the engines. The last 1000kg of fuel from the centre tank is not fed to the engines. Instead, it is pumped to the other tanks. The fuel from the wing tanks will be used for the remaining flight time. As the wings provide lift, they tend to curve upwards. This causes the tanks to be curved as well. The fuel level will, therefore, first start to drop in the tips of the wings. The most outwards fuel probes register this decrease in fuel level. As the flight progresses, the fuel level moves closer towards the fuselage. In Figure 11a this behaviour can clearly be observed. The fuel level in the section of the wing where the fuel probe of Figure 11a is located only starts decreasing around timestamp 150.

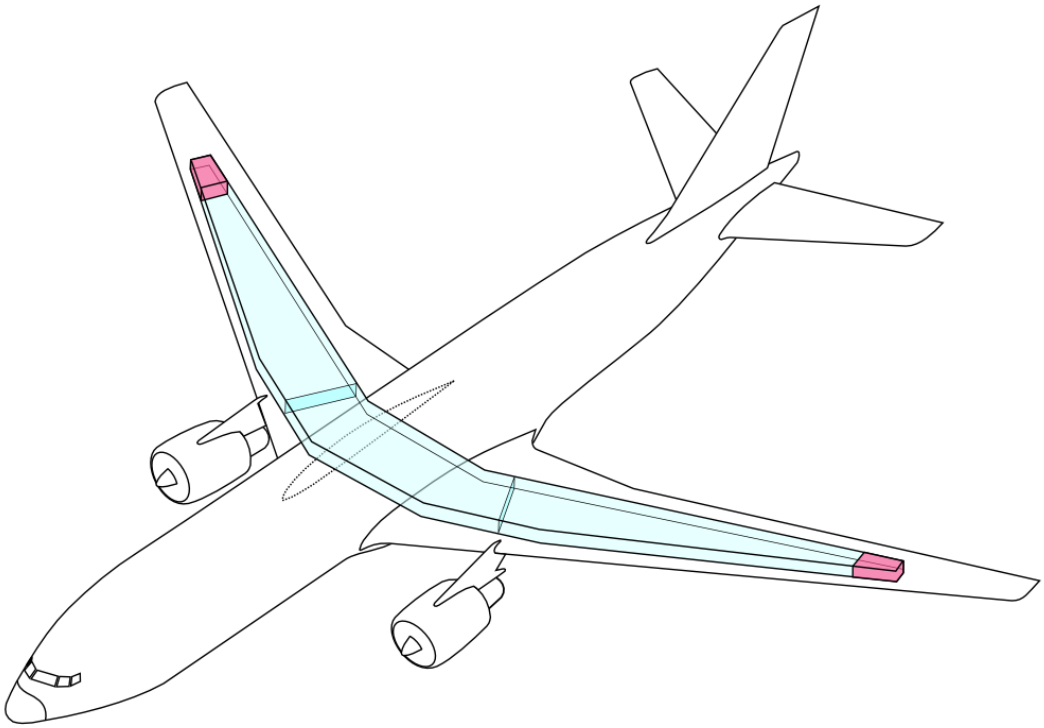


Fig. 8: The position of the fuel tanks in the wide-body commercial aircraft [43].

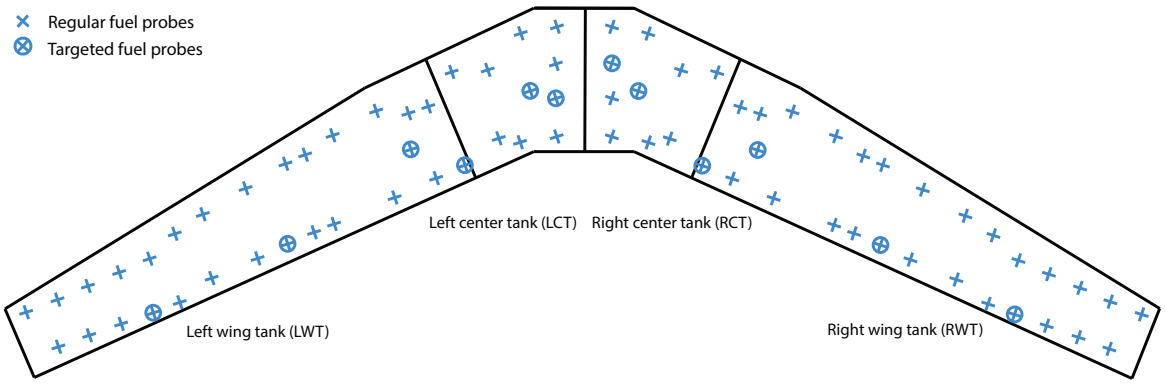


Fig. 9: The type of fuel probes and their locations in the fuel tanks.

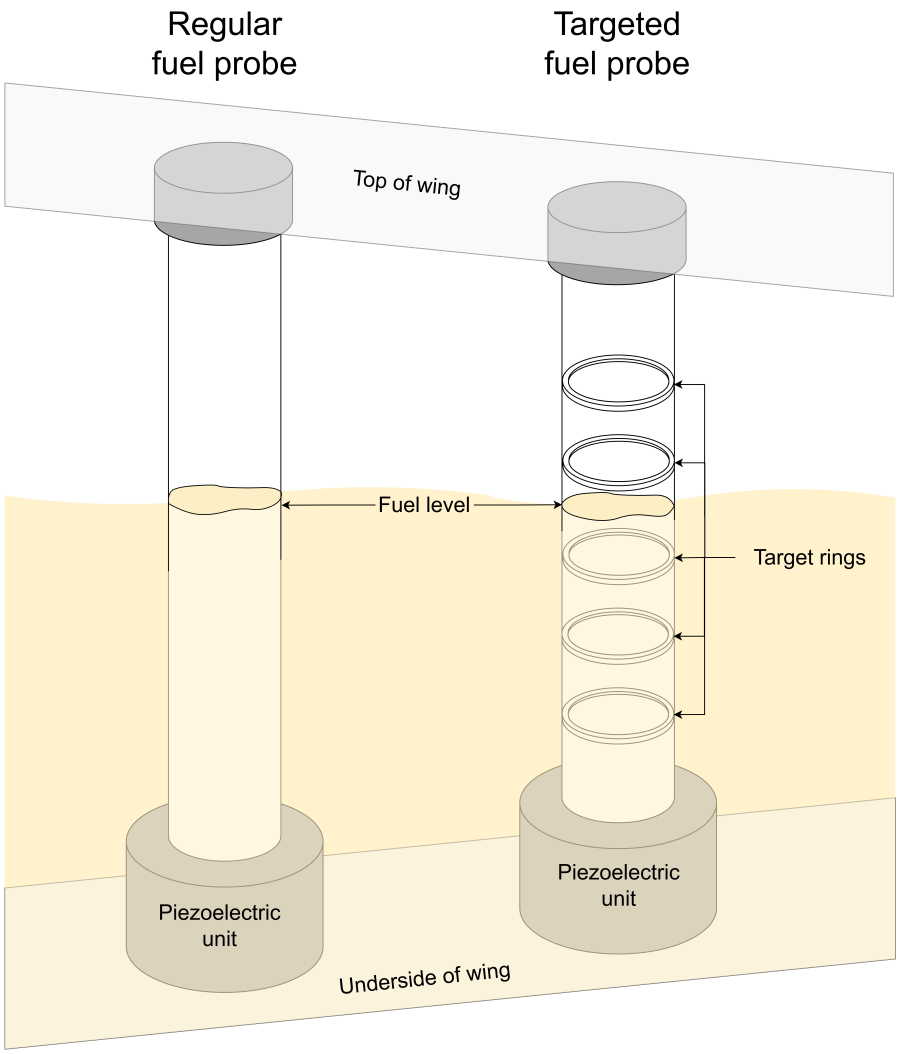
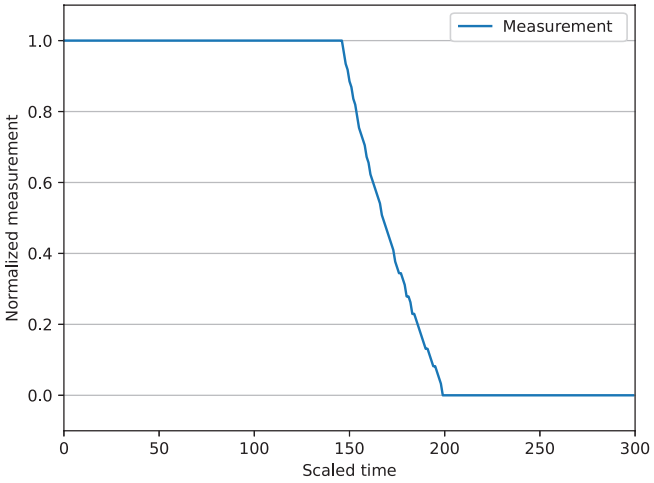


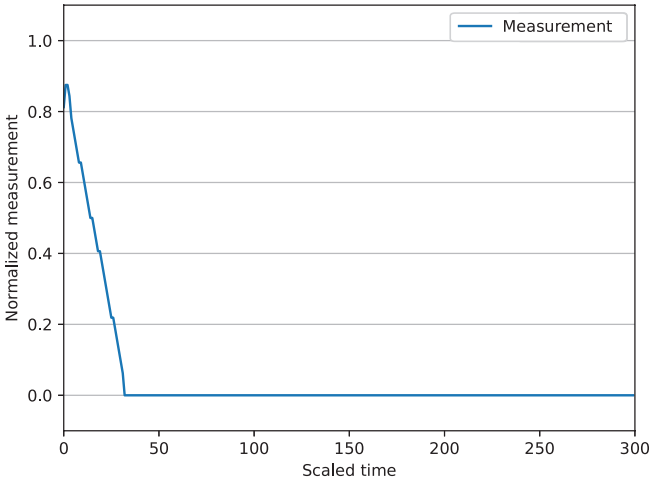
Fig. 10: Graphical representation of regular and targeted fuel probes.

Appendix C
Observed anomalies

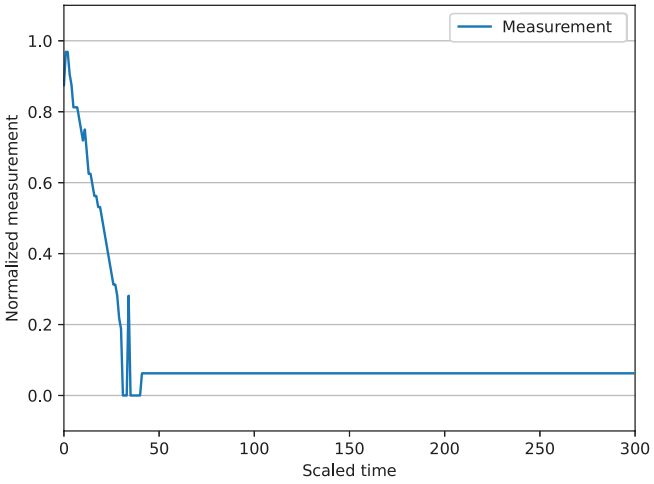
A detailed explanation of the behaviour exhibited by the fuel probes can be found in Appendix B.



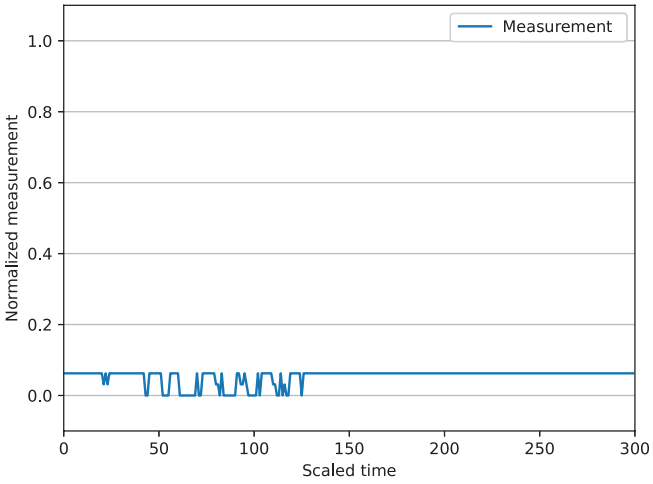
(a) Healthy behaviour of a wing tank fuel probe.



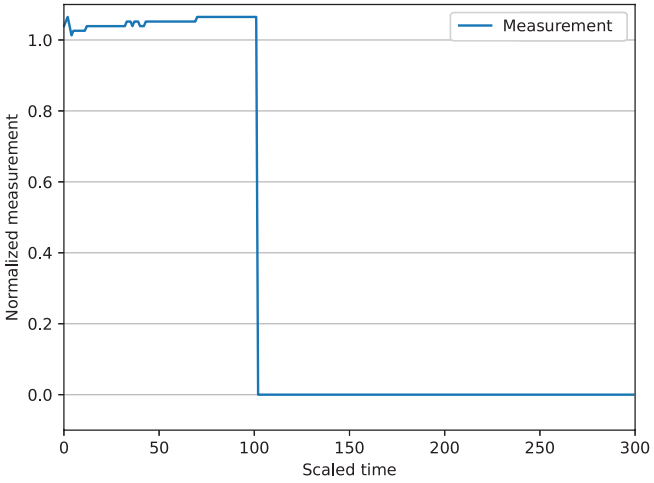
(b) Healthy behaviour of a center tank fuel probe.



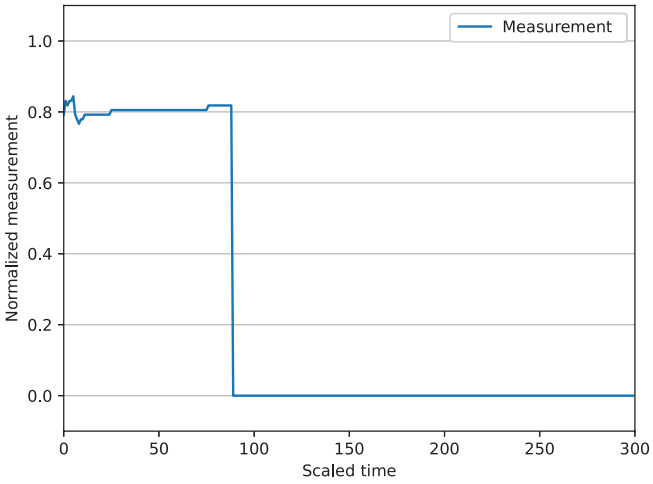
(c) Spiking and erroneously measuring non-zero when empty.



(d) Erroneously measuring non-zero when empty.

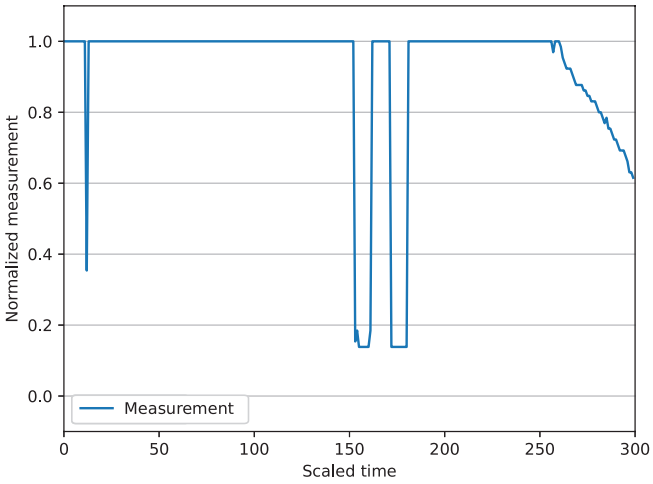


(e) Sudden drop from completely full to empty.

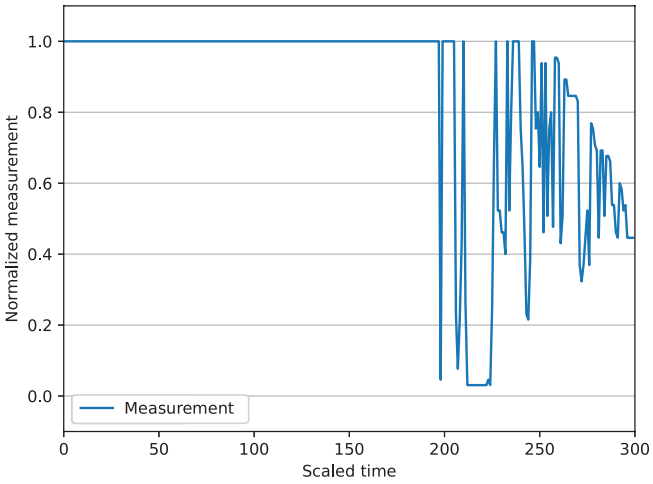


(f) Sudden drop for partially full to empty.

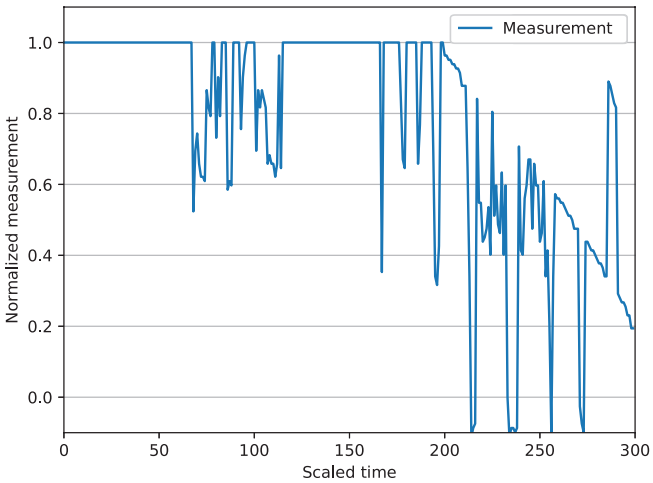
Fig. 11: Observed types of anomalies in the fuel probe data.



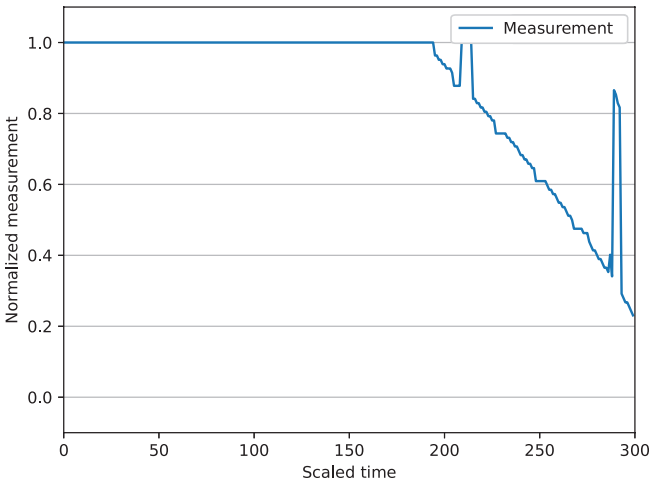
(a) Spikes.



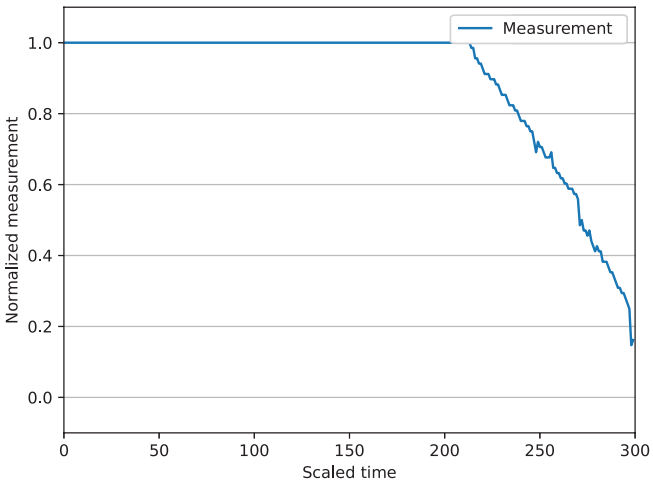
(b) Spikes.



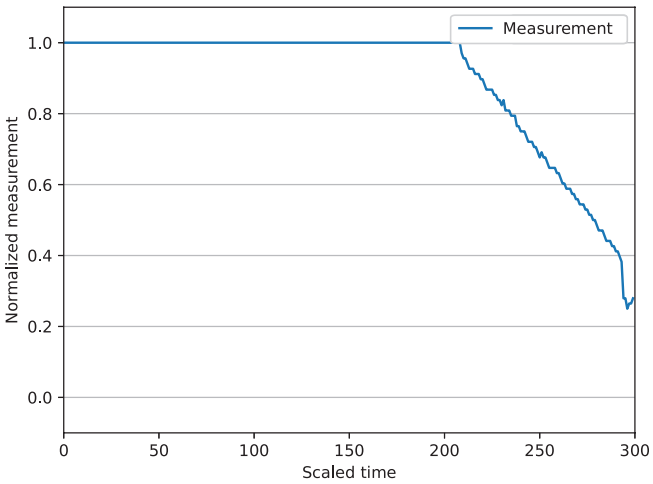
(c) Spikes.



(d) Spikes.



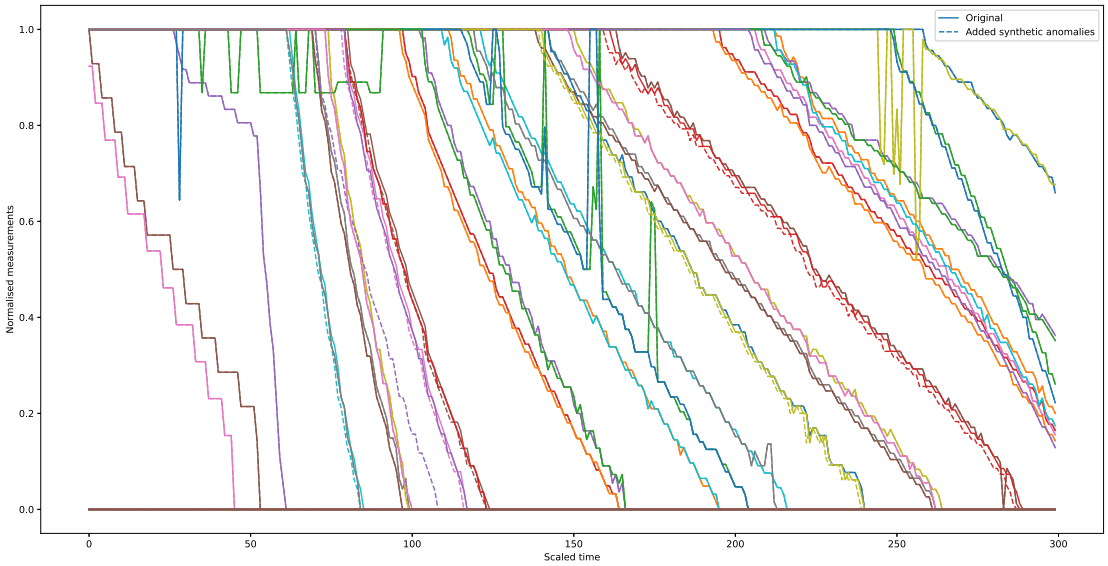
(e) Small imperfections.



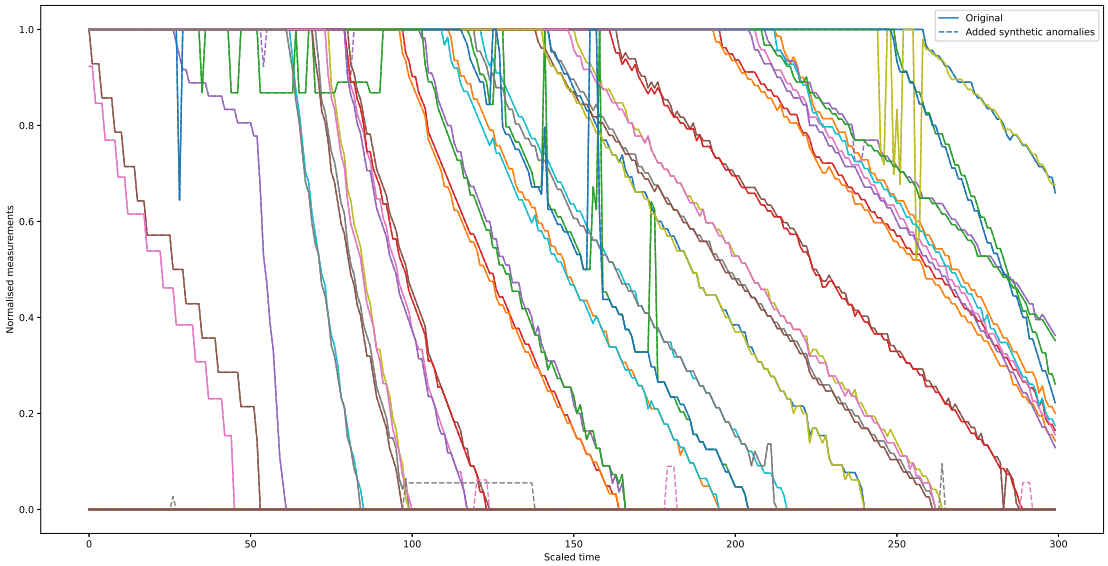
(f) Small imperfections.

Fig. 12: Observed types of anomalies in the fuel probe data.

Appendix D
Synthetic anomalies



(a) Normalised measurements with inter-sensor fuel level discrepancy (time-shift) anomalies.



(b) Normalised measurements with small spikes and prolonged constant errors injected.

Fig. 13: Visualisation of the inserted synthetic anomalies in combination with normalised measurement data.

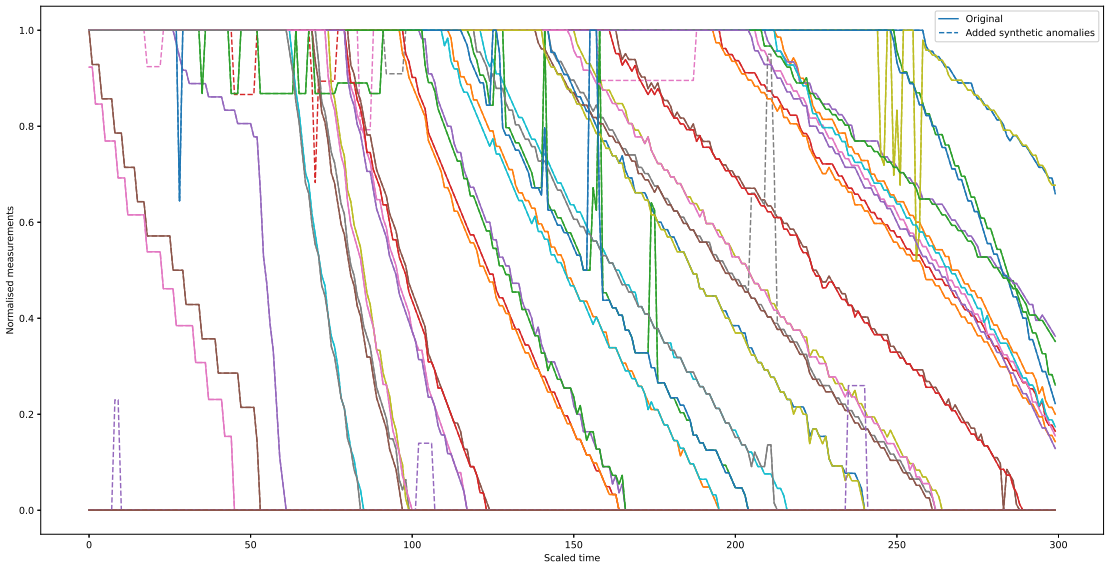
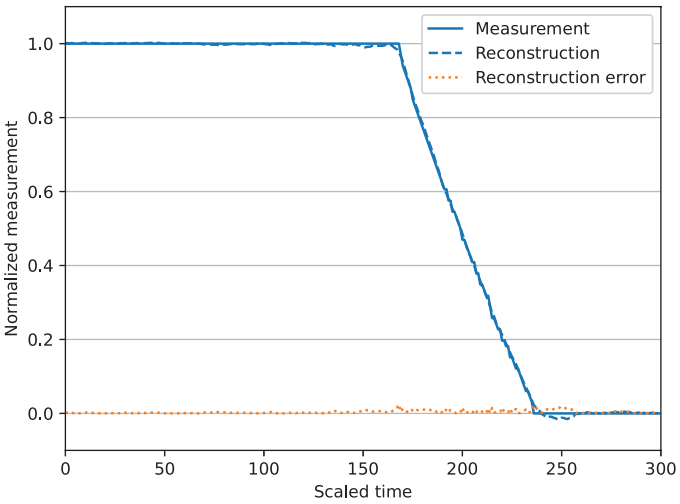


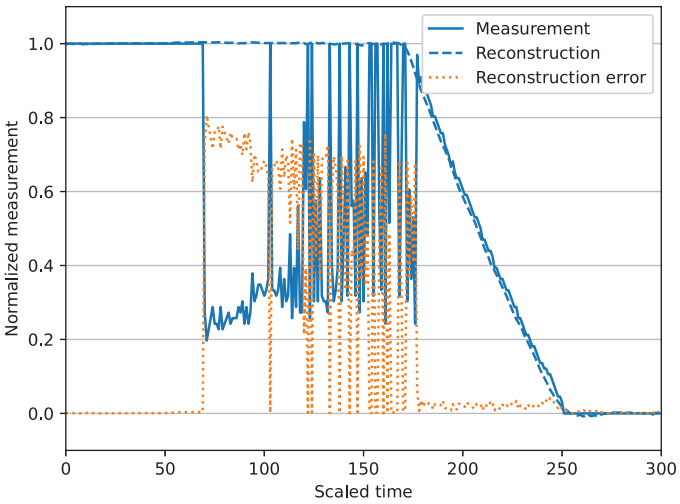
Fig. 14: Normalised measurements with large spikes and prolonged constant errors injected.

.

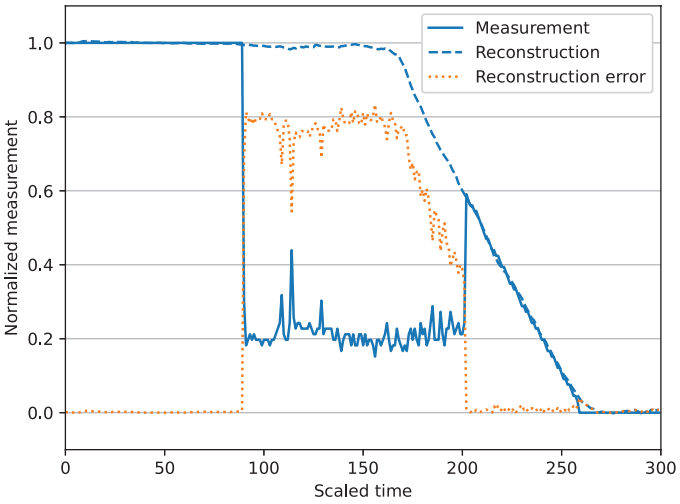
Appendix E
Observed anomalies with reconstructions and reconstruction loss



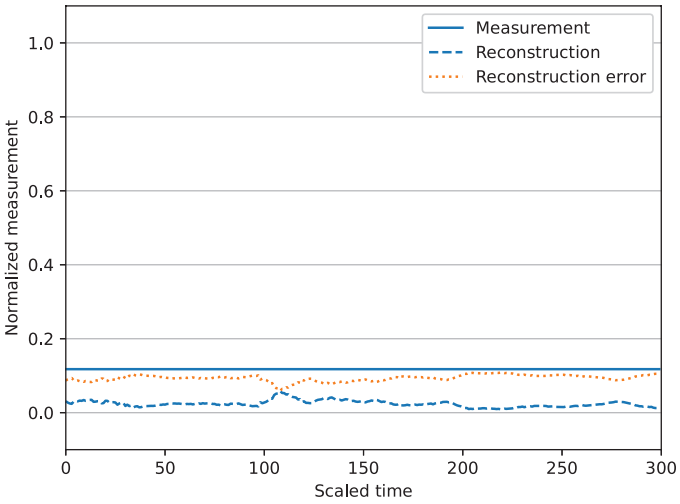
(a) Healthy behaviour from regular fuel probe lwt10.



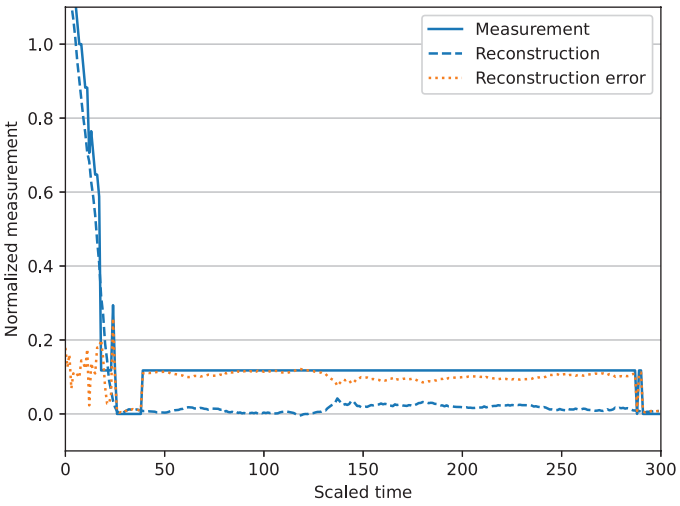
(b) Unhealthy behaviour from regular fuel probe lwt10.



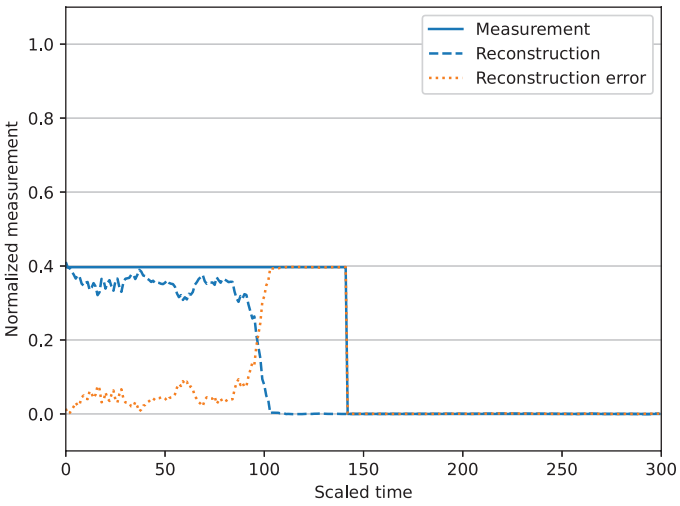
(c) Unhealthy behaviour from regular fuel probe lwt10.



(d) A prolonged constant error in regular fuel probe rct01.



(e) A prolonged constant error and spiking in regular fuel probe rct01.



(f) A sudden drop in regular fuel probe rct03.

Fig. 15: Observed types of anomalies in the fuel probe data

Appendix F
Number of trainable parameters

These are the equations for determining the number of trainable parameters in a fully connected layer:

$$n_{tp} = (C_i \cdot l_i + 1) \cdot (C_o \cdot l_o) \tag{8}$$

Where n_{tp} is the number of trainable parameters, C_i is the number of input channels, C_o the number of output channels, l_i the input length, and l_o the output length.
The number of trainable parameters in a convolutional layer:

$$n_{tp} = (k \cdot C_i + 1) \cdot C_o \tag{9}$$

Where k is the kernel size.

In the case of the fuel probes, which has 76 channels, and a time series of length 300. If we want the output to have the same dimensions, the number of trainable parameters are:

- 521.572.800 in a fully connected layer.
- 28.956 in a convolutional layer with a kernel of size 5.

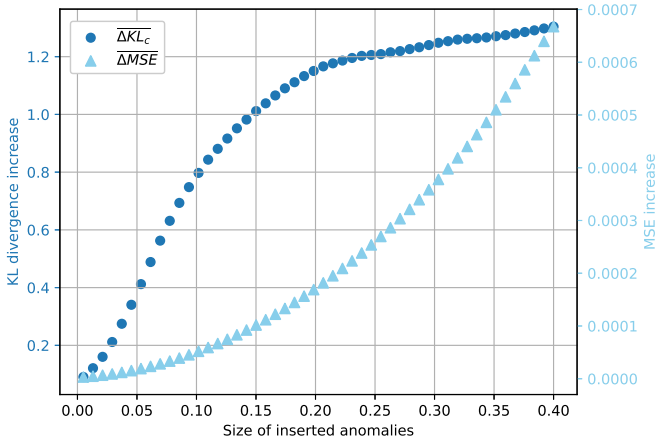
If the number of input channels remains 76, and the input and output time-series' length remain 300, but only a single output channel is required. The number of trainable parameters become:

- 6.862.800 in a fully connected layer.
- 381 in a convolutional layer with a kernel of size 5.

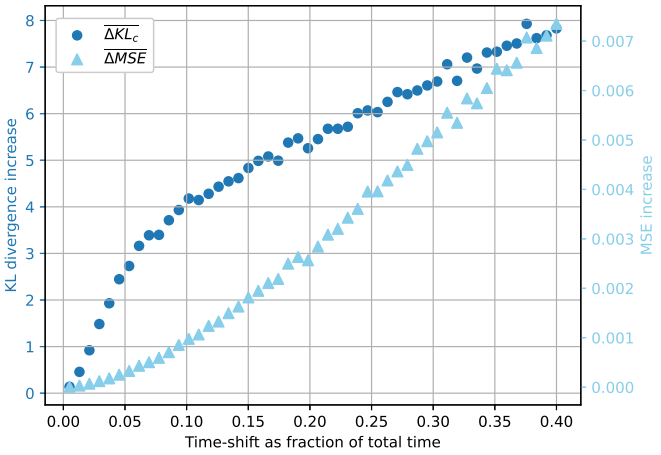
If the number of input channels is first reduced to 10, the input and output time-series' length remain 300, and only a single output channel is required. The number of trainable parameters become:

- 903.000 in a fully connected layer.
- 51 in a convolutional layer with a kernel of size 5.

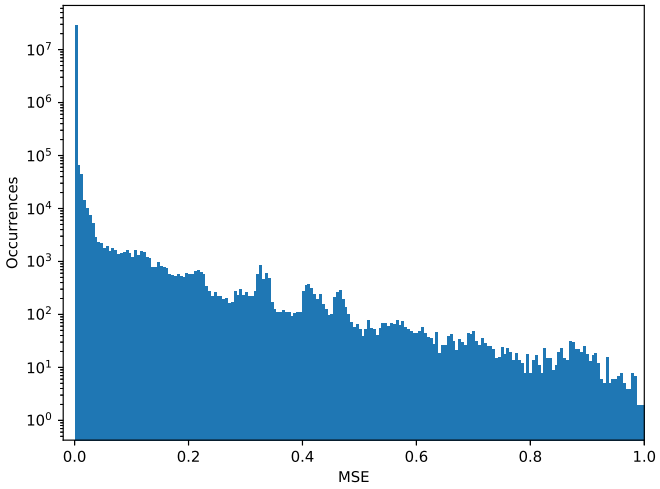
Appendix G
Sensitivity analysis



(a) Increase of $\overline{\Delta KL}_{c,w=1}$ and mean MSE over all sensors resulting from inserting spikes of varying sizes.



(b) Increase of $\overline{\Delta KL}_{c,w=1}$ and mean MSE over all sensors resulting from inserting time-shifts of varying lengths.

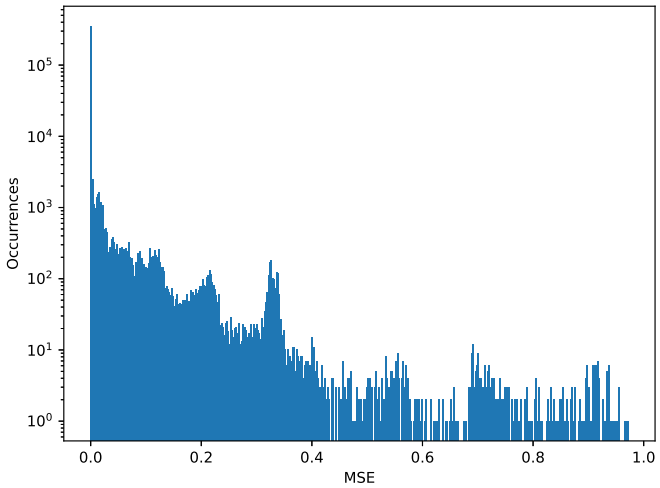


(c) Histogram of reconstruction errors (MSE) of all sensors combined.

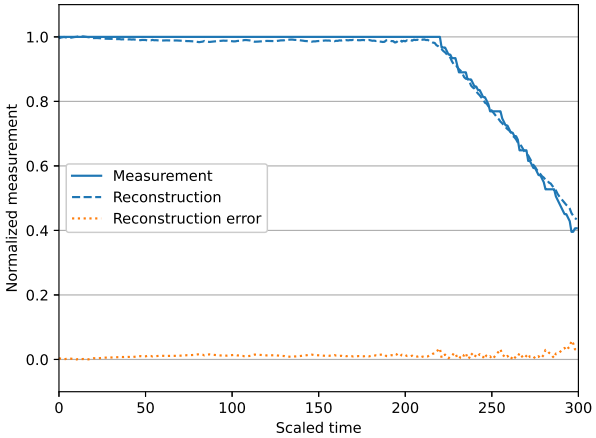
Fig. 16: Sensitivity analysis of all sensors combined.

Appendix H
Flights and corresponding reconstruction error histograms

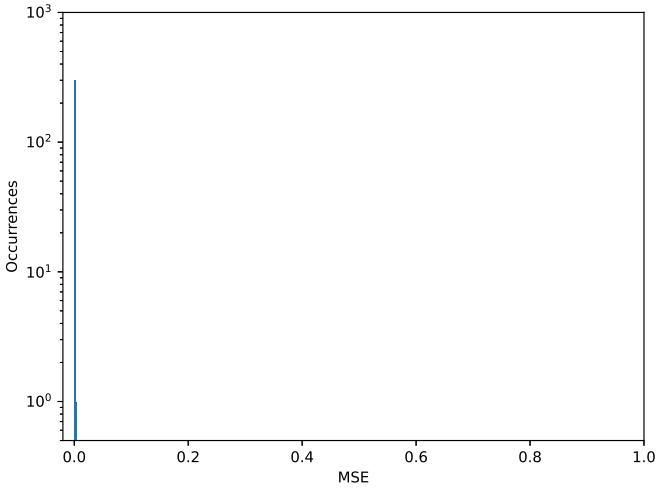
Figure 17a shows the histogram of lwt05 over a large number of flights and various aircraft. This is the baseline used for comparing the single flight histogram to, in order to calculate the KL divergence. The remaining figures are all of the same targeted fuel probe lwt05. The left column of images show; the measurement data, the reconstruction by the DCAE, and reconstruction errors of a single flight. The right column shows the accompanying histograms of these single flights.



(a) Histogram of MSE's of a large number of flights combined

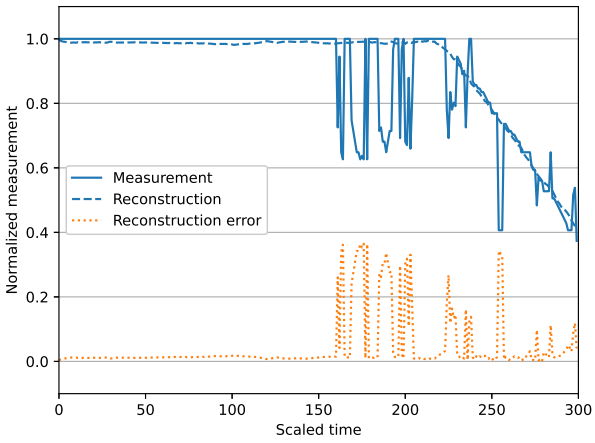


(b) Flight 1: Measurement, reconstruction, and reconstruction error of a single flight for lwt05.

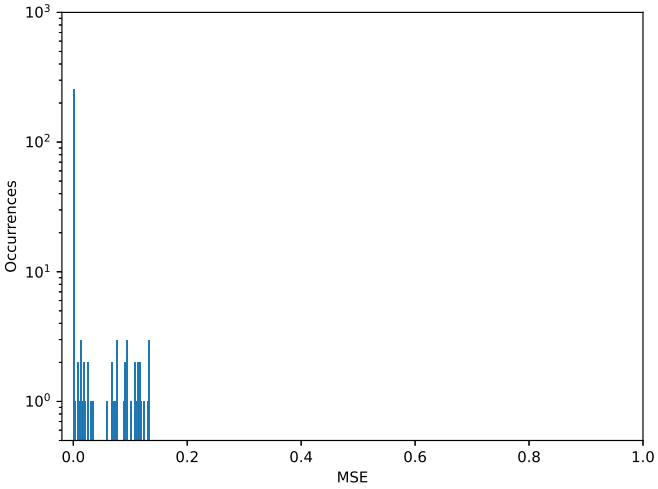


(c) Flight 1: Histogram of MSE's of sensor lwt05. Resultant weighted KL divergence = 0.019.

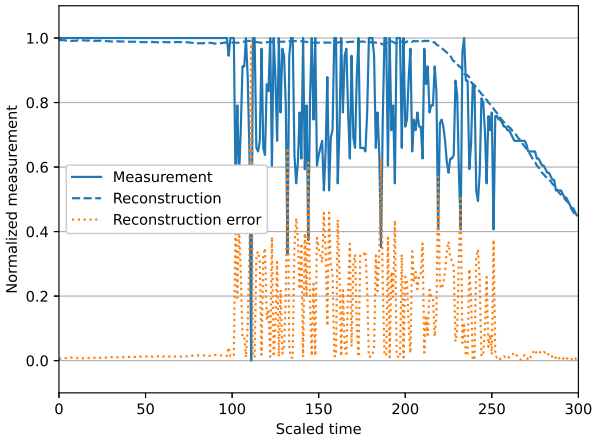
Fig. 17: Observed types of anomalies in the fuel probe data



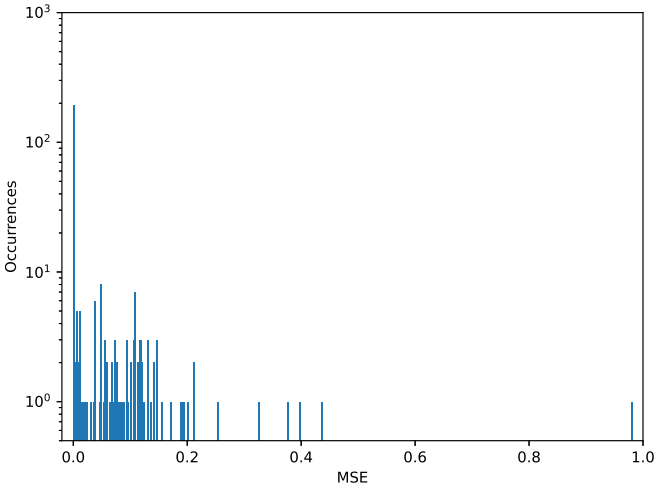
(a) Flight 2: Measurement, reconstruction, and reconstruction error of a single flight for lwt05.



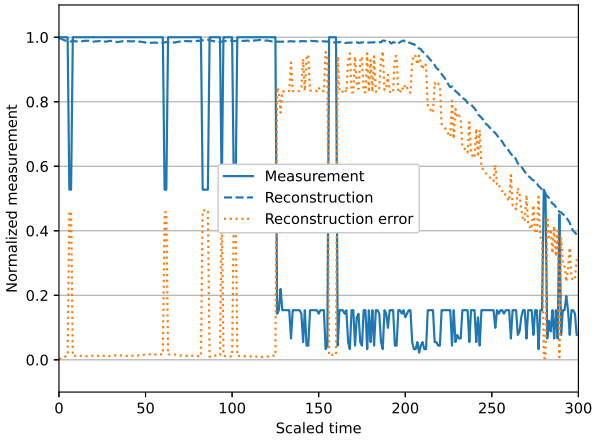
(b) Flight 2: Histogram of MSE's of sensor lwt05. Resultant weighted KL divergence = 0.798.



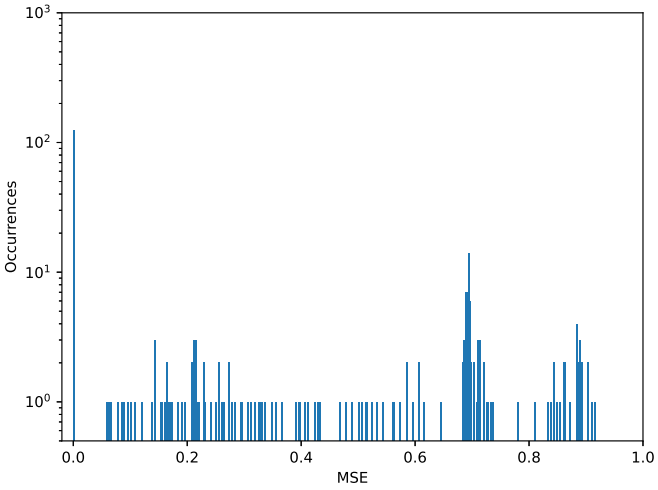
(c) Flight 3: Measurement, reconstruction, and reconstruction error of a single flight for lwt05.



(d) Flight 3: Histogram of MSE's of sensor lwt05. Resultant weighted KL divergence = 3.961.



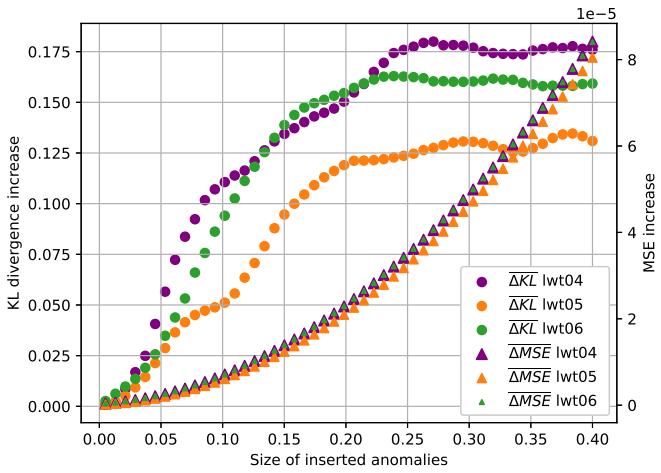
(e) Flight 4: Measurement, reconstruction, and reconstruction error of a single flight for lwt05.



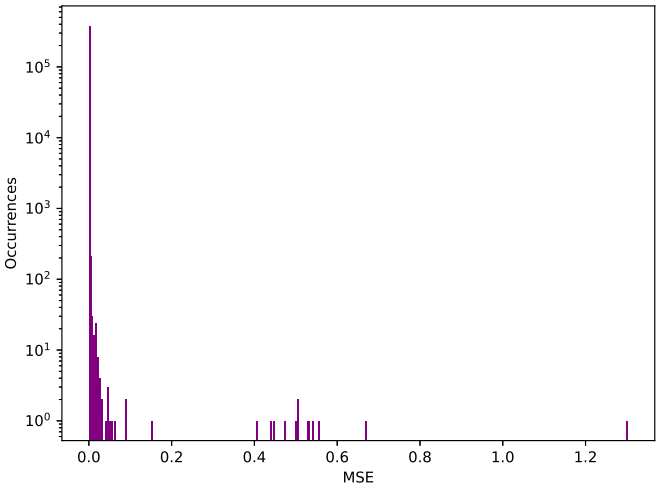
(f) Flight 4: Histogram of MSE's of sensor lwt05. Resultant weighted KL divergence = 54.363.

Fig. 18: Single sensor measurements, reconstruction, and reconstruction errors. In combination with resultant reconstruction error histograms.

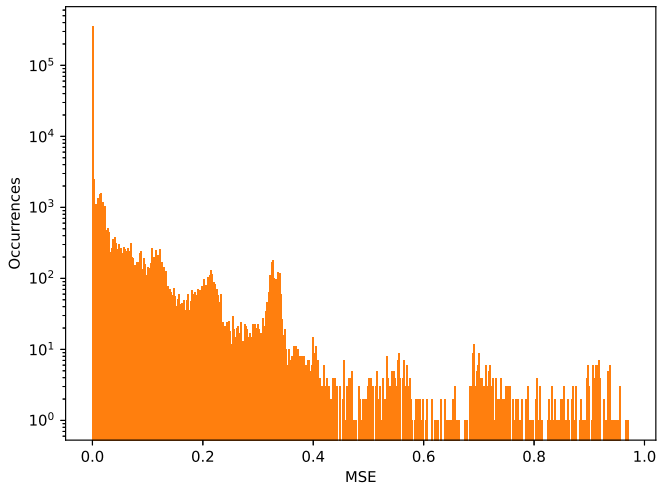
Appendix I
Responses of KL divergence based health analysis



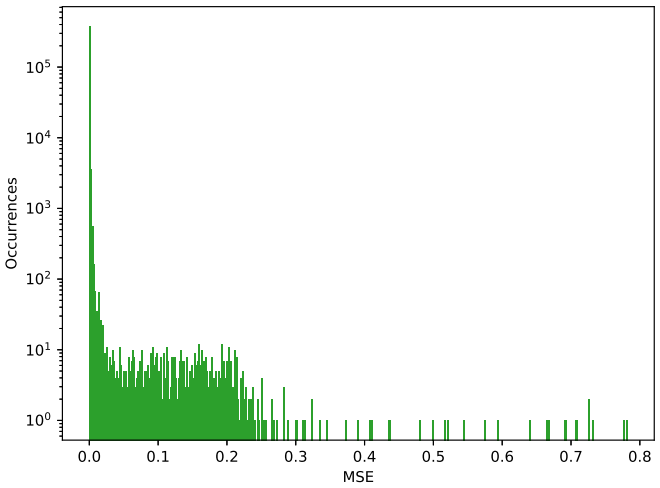
(a) Responses in KL divergence and MSE for 3 sensors.



(b) Histogram of MSE's of sensor lwt04.



(c) Histogram of MSE's of sensor lwt05.



(d) Histogram of MSE's of sensor lwt06.

Fig. 19: Reconstruction error histograms of three fuel probes and the corresponding average $\Delta KL_{w=1}$ responses.

Appendix J
Hyperparameter optimisation

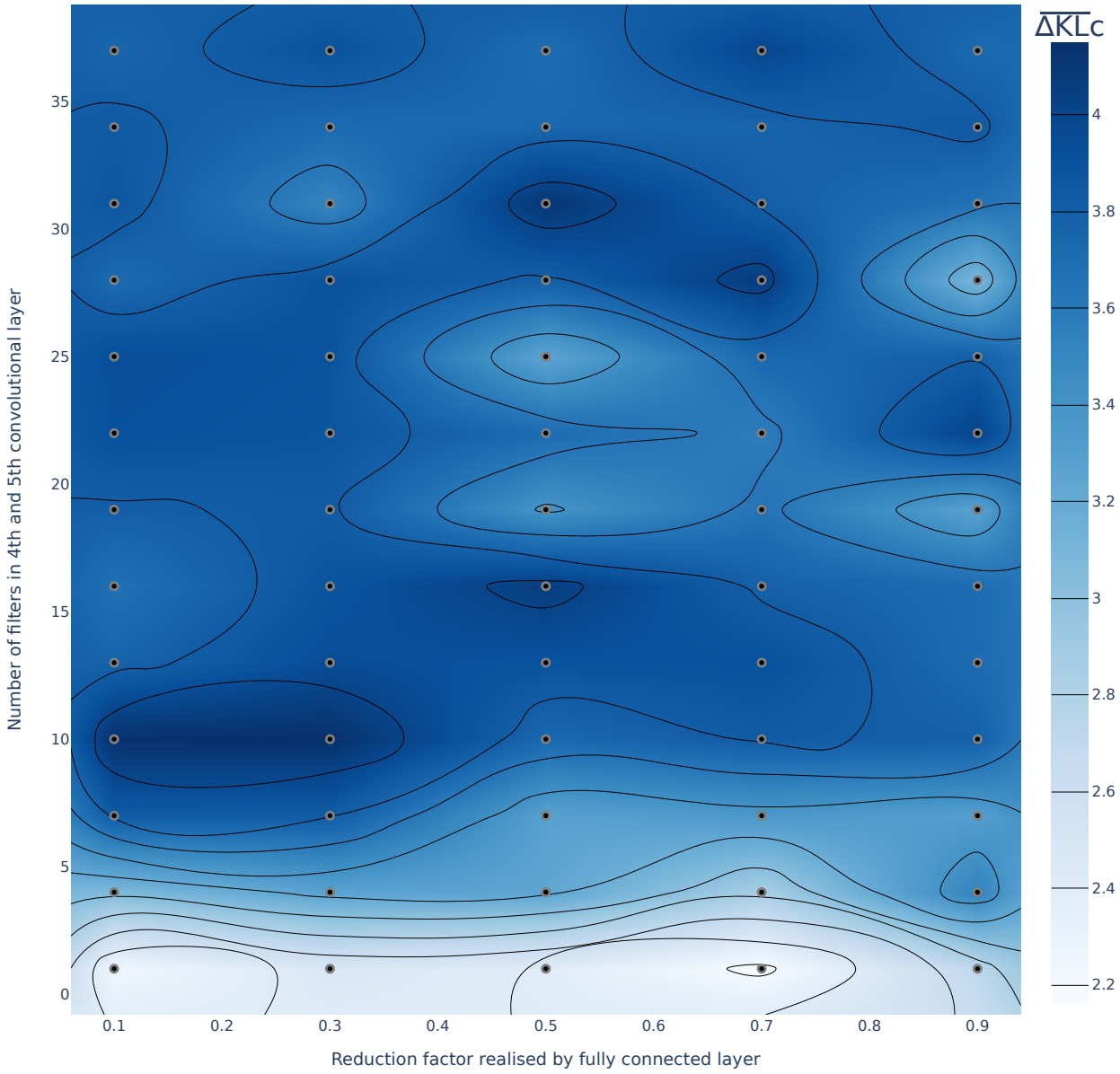


Fig. 20: Results of hyperparameter optimisation on a dataset (Figure 13a) containing inter-sensor fuel level discrepancy (time-shift) anomalies.

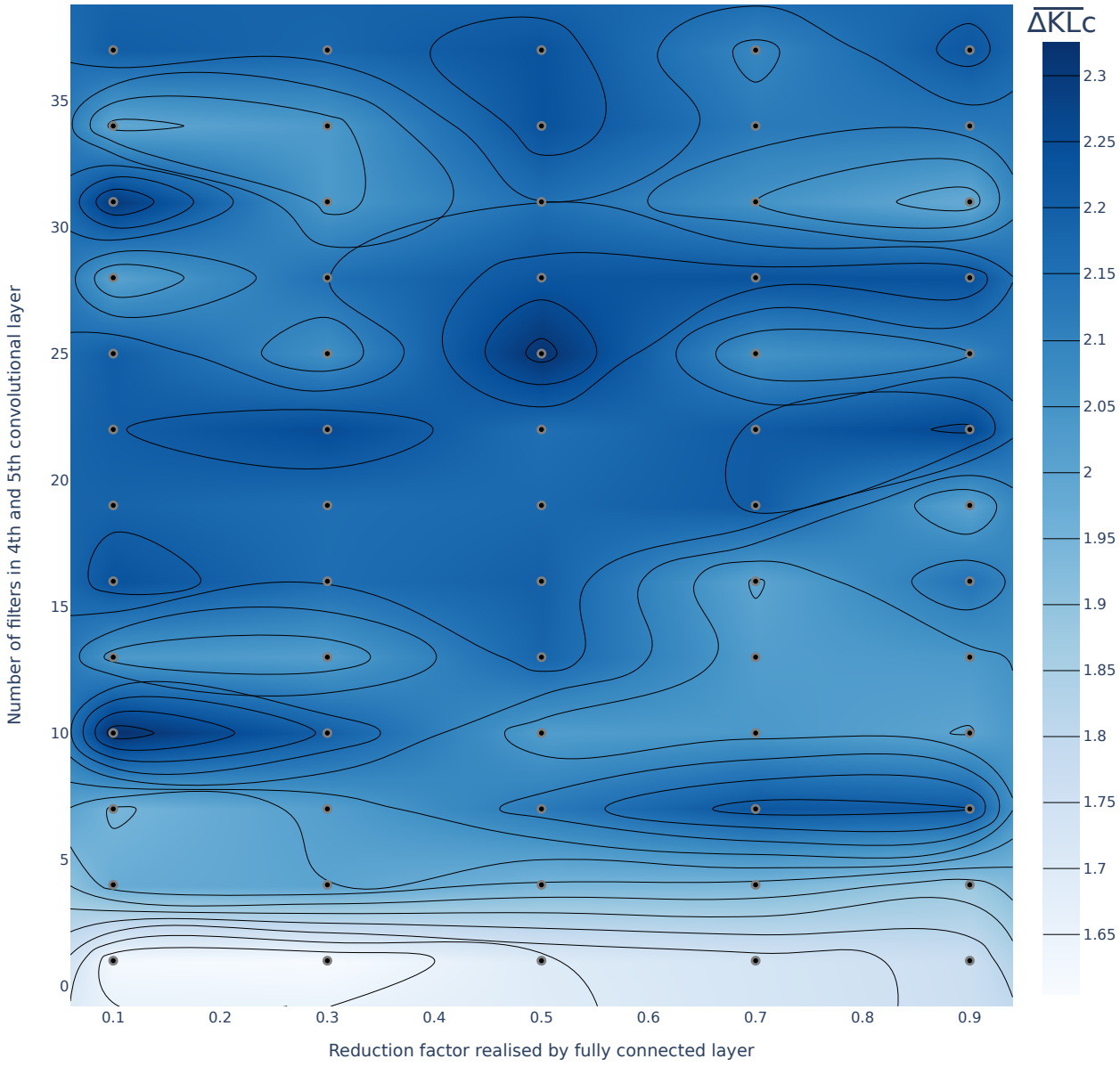


Fig. 21: Results of hyperparameter optimisation on a dataset (Figure 13b) containing small spikes and prolonged constant errors.

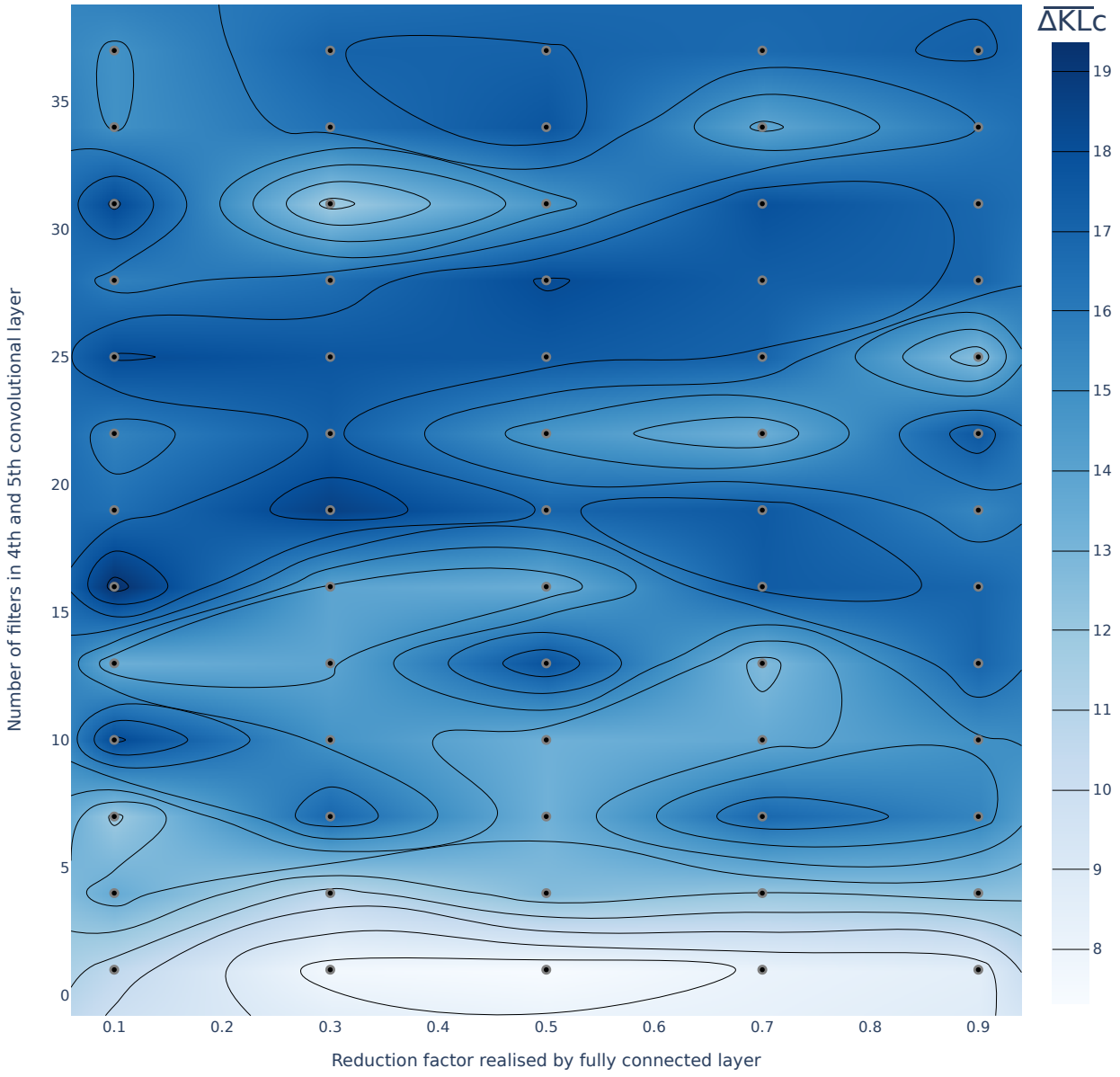
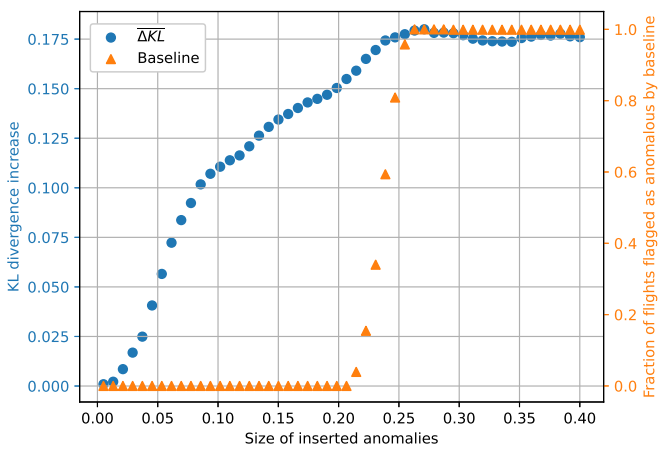
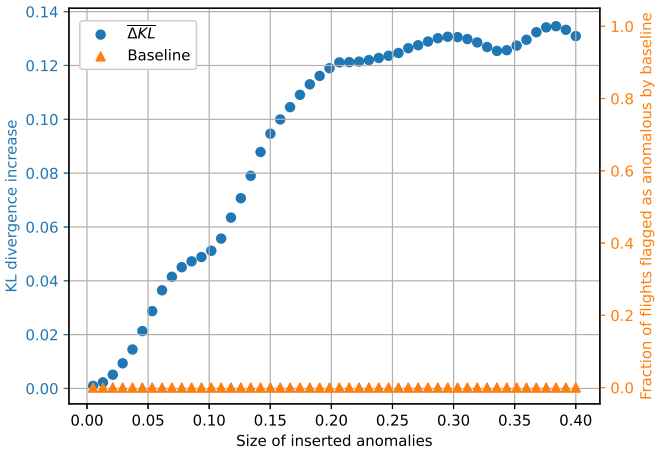


Fig. 22: Results of hyperparameter optimisation on a dataset (Figure 14) containing large spikes and prolonged constant errors.

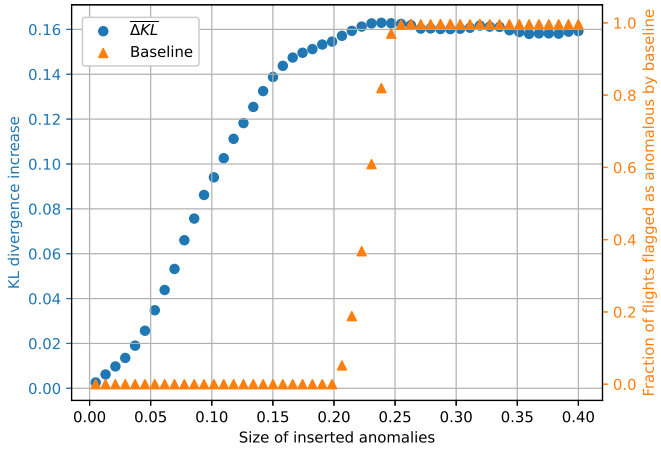
Appendix K
Quantitative comparison to baseline



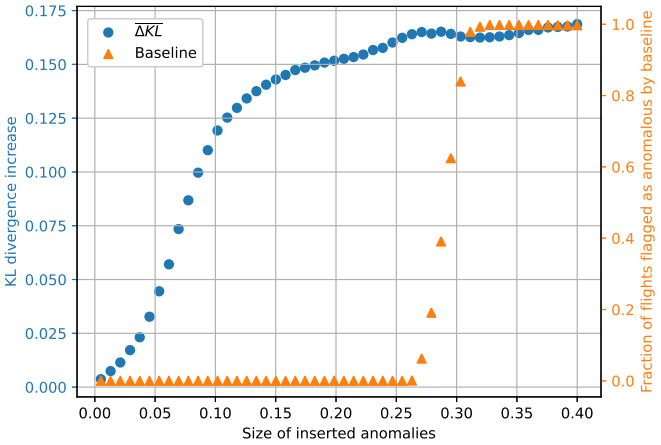
(a) Responses for sensor lwt04, a regular fuel probe.



(b) Responses for sensor lwt05, a targeted fuel probe.



(c) Responses for sensor lwt06, a regular fuel probe.

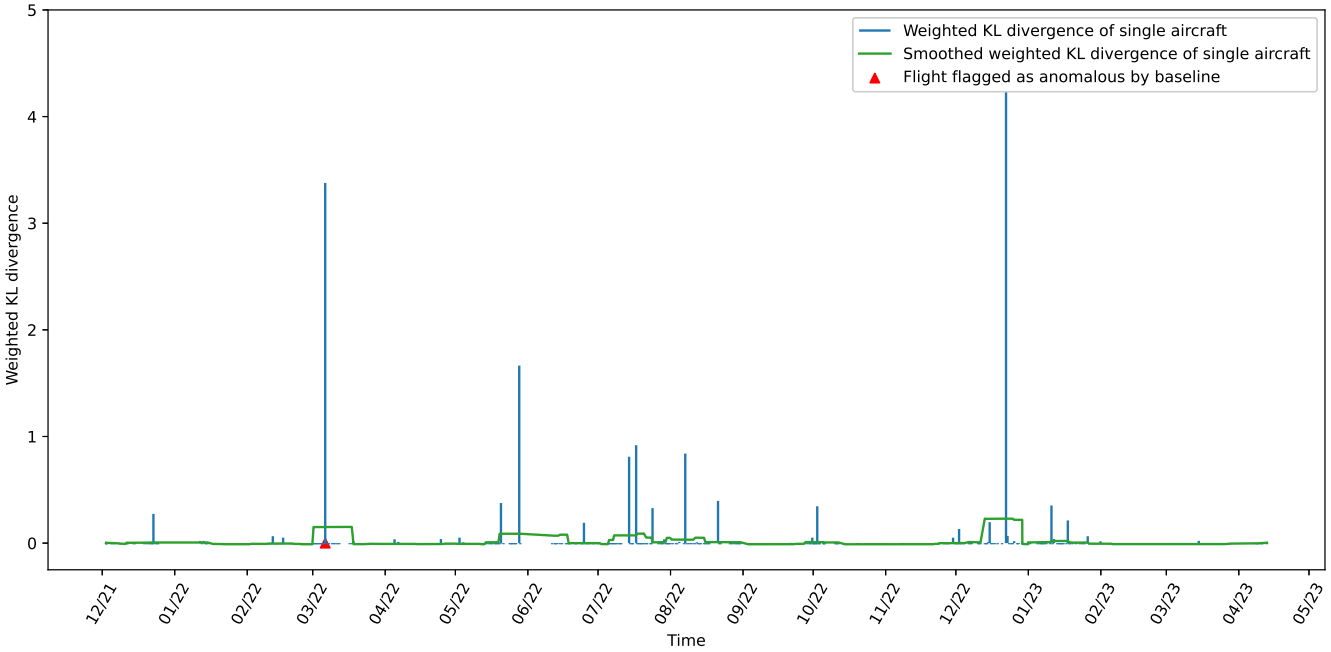


(d) Responses for sensor lwt11, a regular fuel probe.

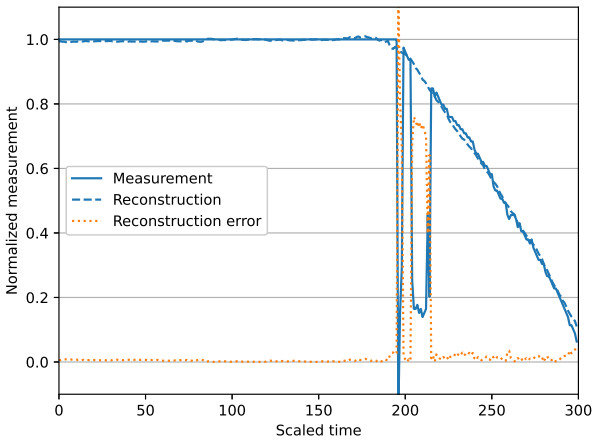
Fig. 23: Responses of the baseline and the proposed method to inserted spike anomalies of varying size.

Appendix L
Qualitative evaluation & Comparison to baseline

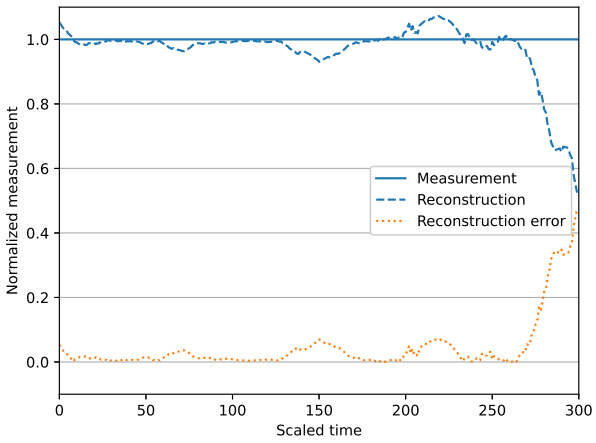
Figures 24a, 25a, 26a, 27a, 28a 28b, 29a, and 29b show the weighted KL divergence results of the proposed method (blue lines), along with the binary results of the baseline method (red triangles), per flight for a single fuel probe. Higher blue lines indicate a higher weighted KL divergence and, therefore, according to the proposed method, a more anomalous flight. Not all figures are of the same aircraft, but all sub-figures within a figure are of the same aircraft.



(a) Fuel probe lwt04: A fuel probe that can be considered healthy. Very few flights are marked as anomalous by the proposed method and only a single one by the baseline method.

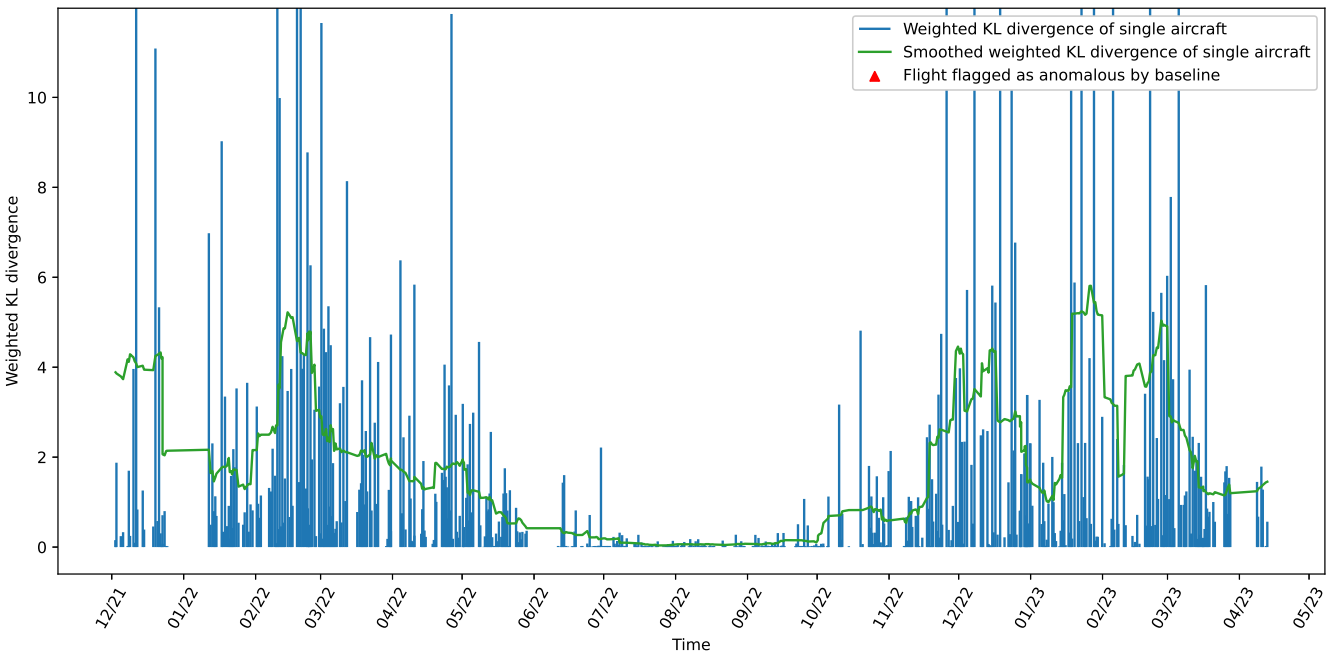


(b) lwt04, 06/03/22: An example of a good reconstruction by the proposed method.

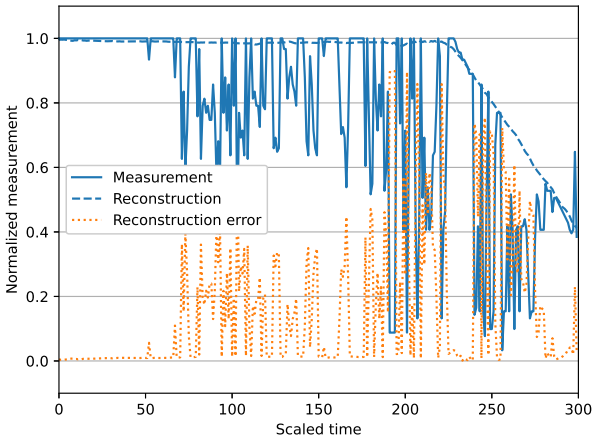


(c) lwt04, 22/12/22: Poor reconstruction due to an extremely long flight time. The reconstructions of the flights on 28/05/22, 14/07/22, 17/07/22 are also poor due to a far above normal flight duration.

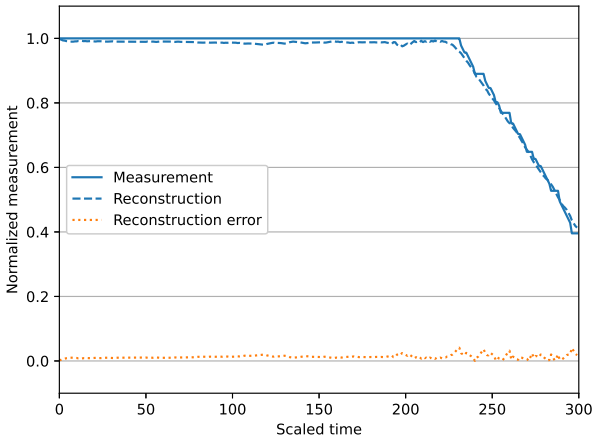
Fig. 24: Weighted KL divergence of proposed method with binary results from currently employed (baseline) method. Additionally, interesting individual flights are singled out.



(a) Targeted fuel probe lwt05: This is an interesting case as there is a clear trend. Between 06/22 and 10/22 the targeted fuel probe worked a lot better than prior and after those dates. Note that the baseline does not flag a single flight as anomalous, this is because the baseline method for targeted fuel probes is forced to work on different rules compared to regular fuel probes due to the erratic behaviour of targeted fuel probes. A suspected culprit are the different additives present in the kerosene depending on the temperature. In the winter the kerosene has to be able to withstand lower temperatures, these additions seem to influence the readings in some cases.

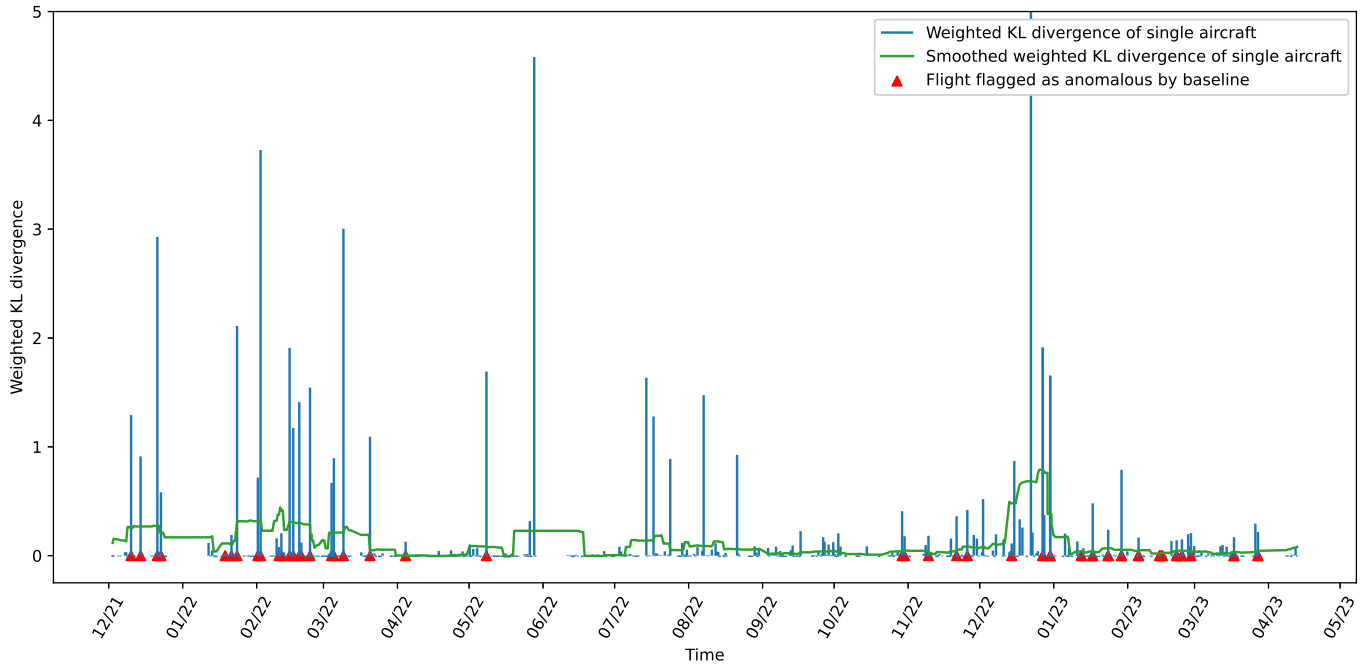


(b) lwt05, 11/02/22: Good reconstruction on a poorly functioning targeted fuel probe.

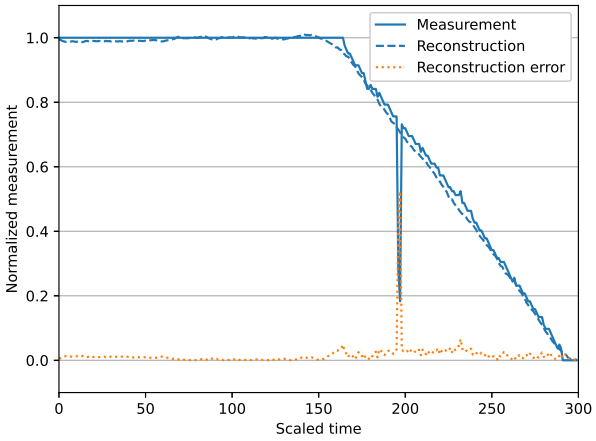


(c) lwt05, 08/08/22: Good reconstruction of a well functioning targeted fuel probe

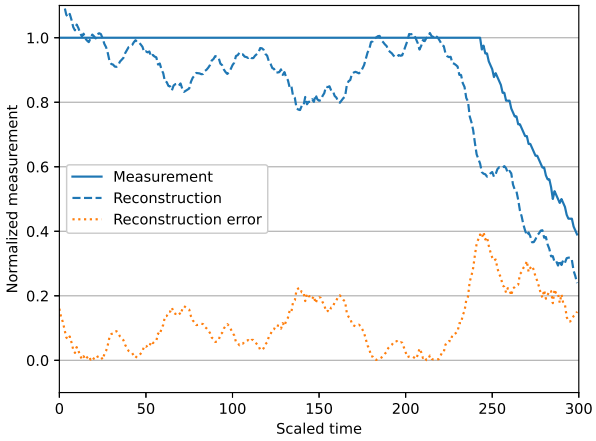
Fig. 25: Weighted KL divergence of proposed method with binary results from currently employed (baseline) method. Additionally, interesting individual flights are singled out.



(a) Regular fuel probe lwt06

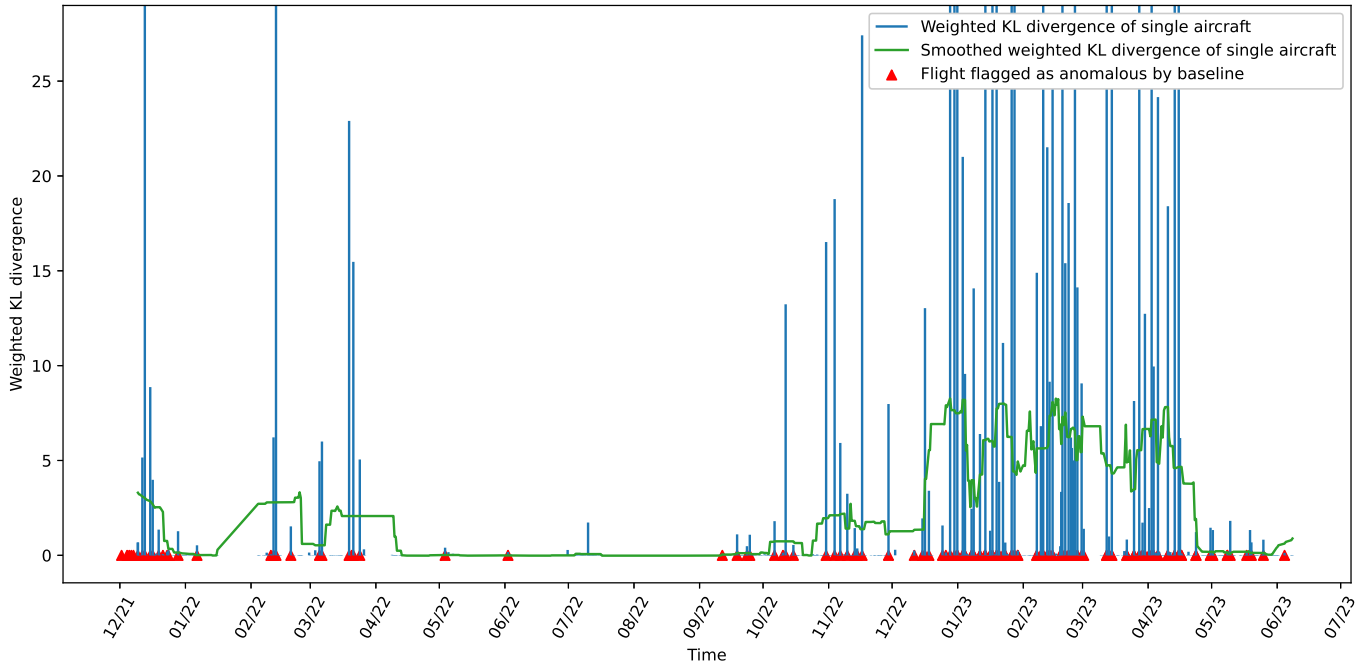


(b) lwt06, 26/05/22: An example of a spike that does not trigger the baseline but does create a small KL divergence peak.

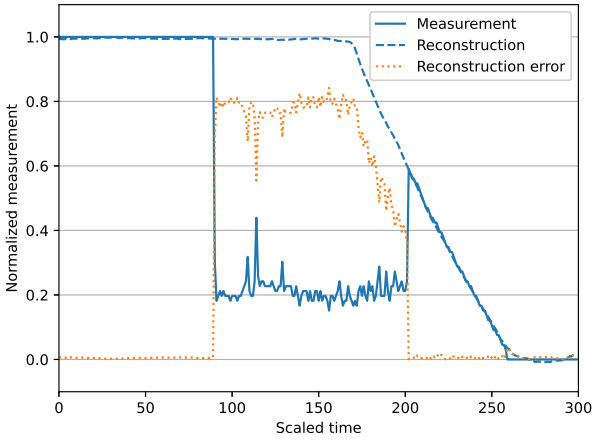


(c) lwt06, 22/12/22: The same extremely long flight as illustrated in figure 24c also causes a poor recreation in lwt06 among many other fuel probes.

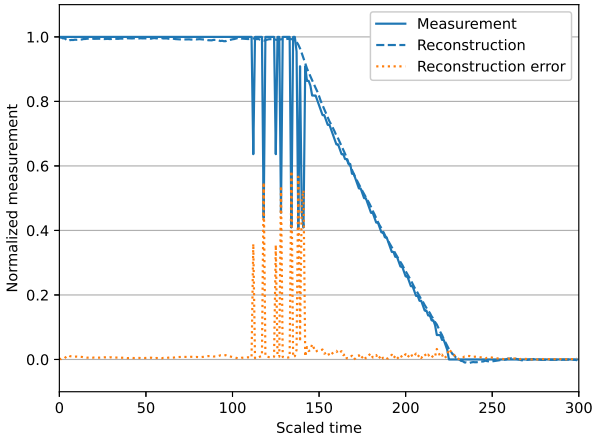
Fig. 26: Weighted KL divergence of proposed method with binary results from currently employed (baseline) method. Additionally, interesting individual flights are singled out.



(a) Regular fuel probe lwt10: The baseline method and the DCAE based method concur on most flights. The DCAE based method does, however, provide additional insights in the degree of anomalousness. It can be observed that in the last 1.5 months the degree of anomalousness has decreased.

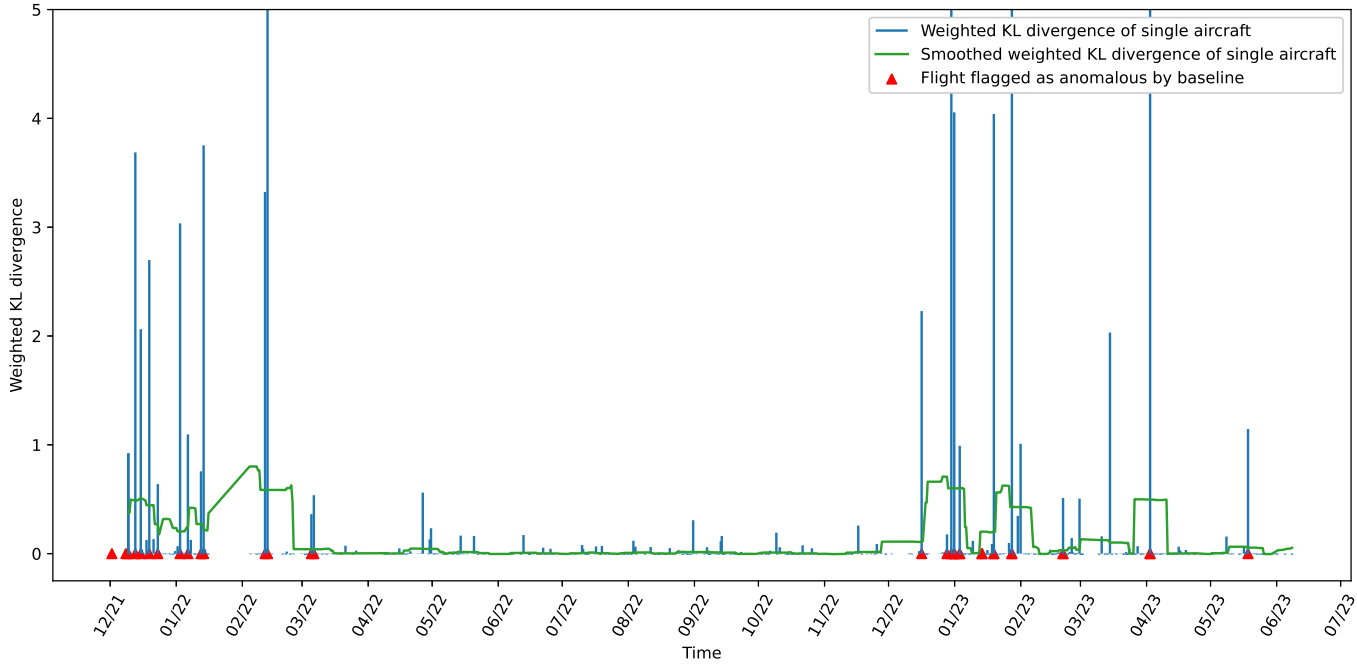


(b) lwt10, 12/02/22: A flight that is flagged by the baseline. The proposed method results in a very high KL divergence.

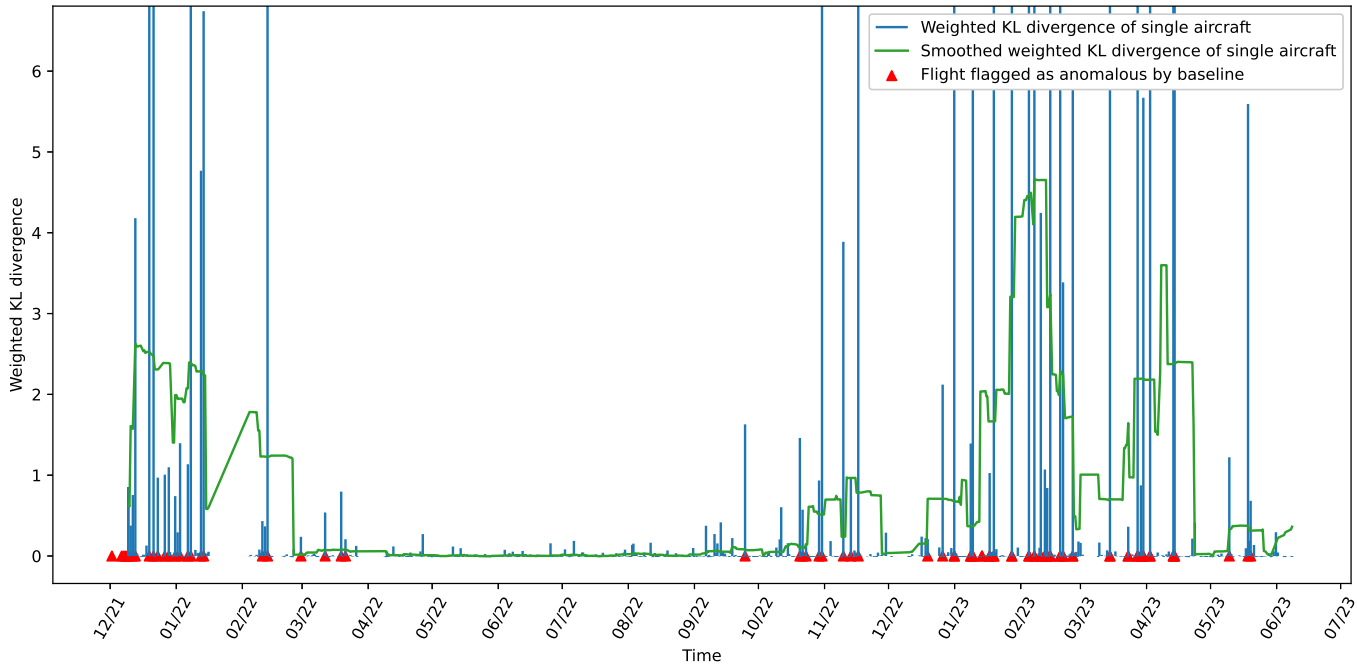


(c) lwt10, 19/02/22: This flight is also flagged by the baseline. However, the proposed method provides additional insights because it gives a lower KL divergence than in Figure **??**.

Fig. 27: Weighted KL divergence of proposed method with binary results from currently employed (baseline) method. Additionally, interesting individual flights are singled out.

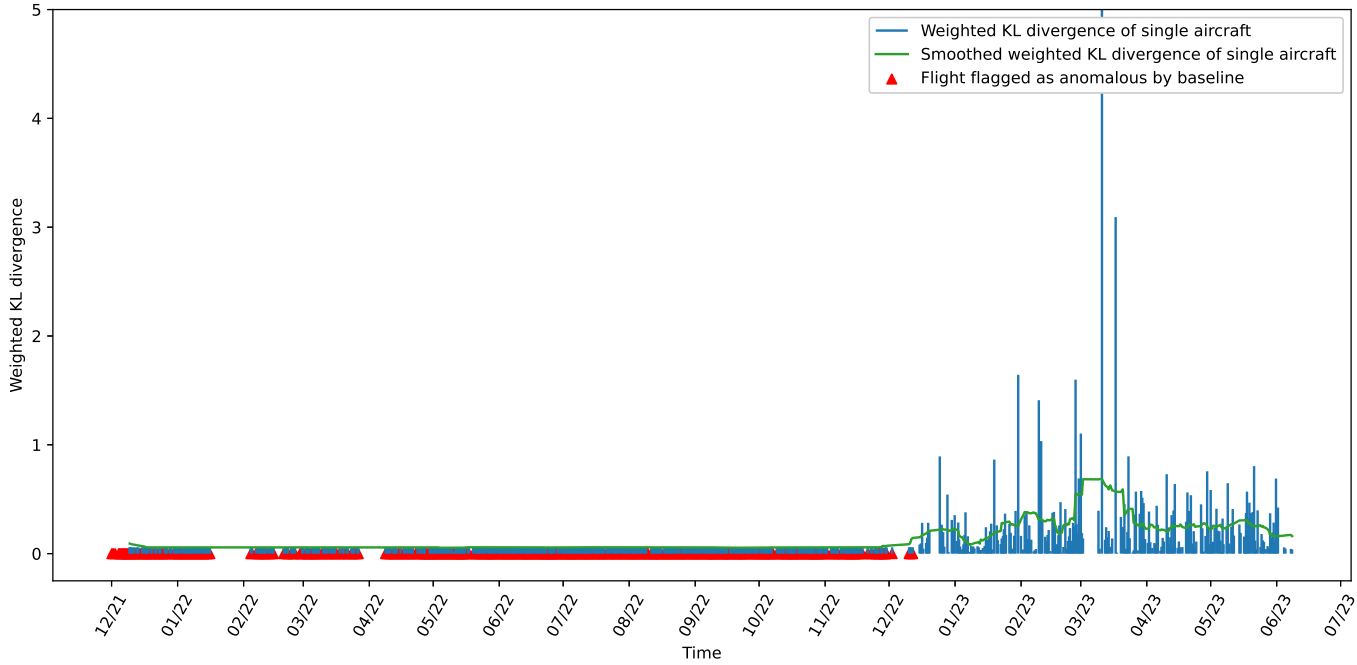


(a) Regular fuel probe lwt15

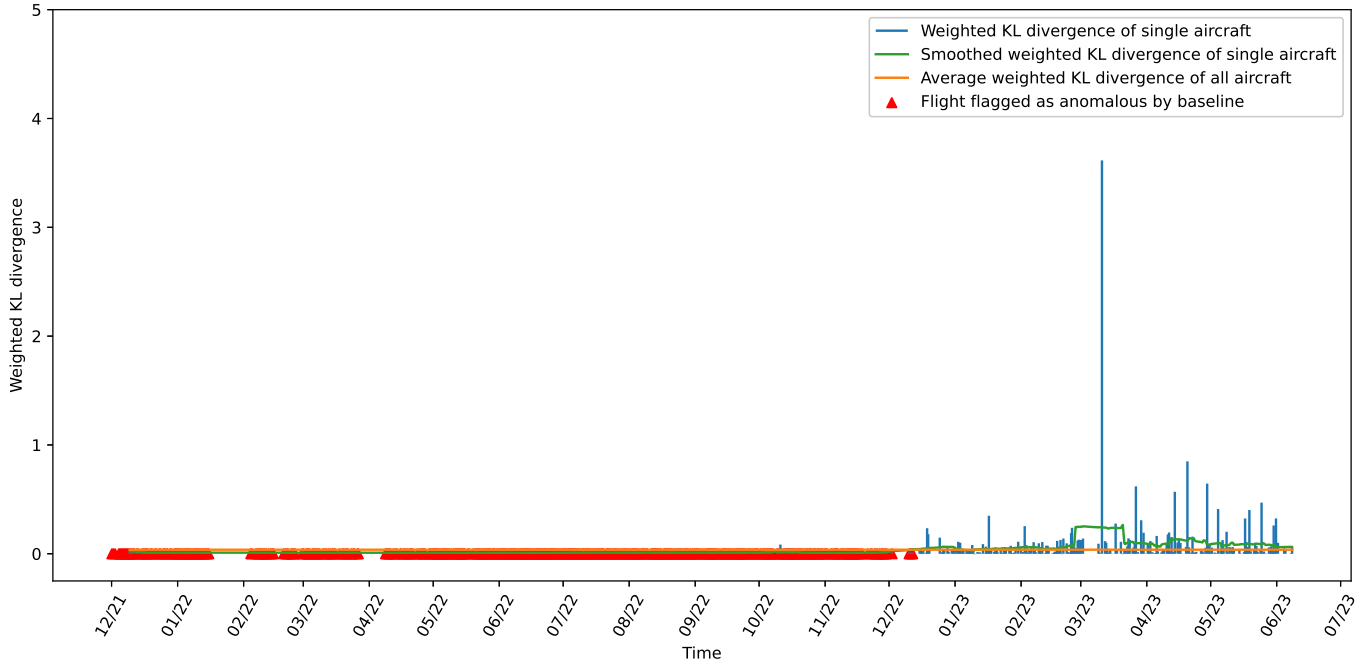


(b) Regular fuel probe lwt18:

Fig. 28: Weighted KL divergence of proposed method with binary results from currently employed (baseline) method.



(a) Fuel probe rwt25: A wide spread malfunction or data corruption caused rwt21 - rwt 28 and lwt 21 - lwt 28 to register constant 0 from 12/21 until 12/22. The baseline method was able to pick up the flatline. The proposed method did not, likely because the proposed method makes a prediction based on the surrounding and opposing probes, as they are most closely related. If all of these register 0 then the reconstruction will be 0 as well. Additionally, the flights from 01/23 onward all show high weighted KL divergence scores, this is due to the fact that rwt25 is very far in the tip of the wing. This part experiences a large amount of turbulence and warping making the data jittery. The far end of the wing also only contains a small amount of fuel so the rate at which the fuel level drops is very high, this can create large reconstruction errors if the reconstruction is just slightly too early or late.

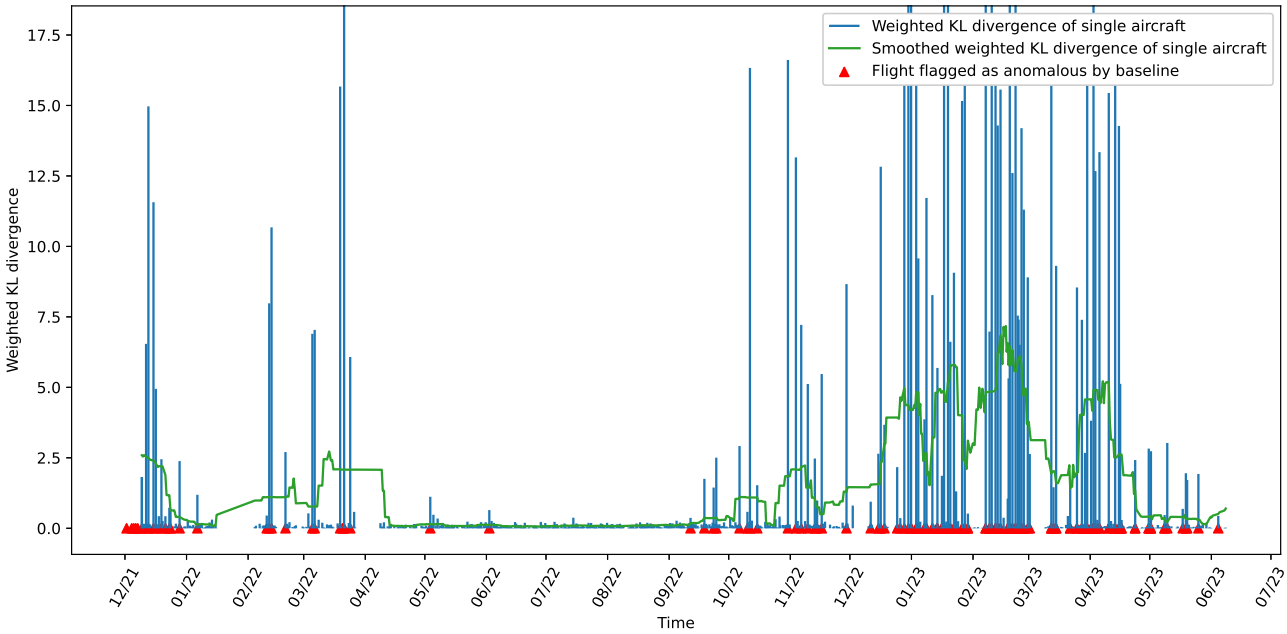


(b) Fuel probe rwt27: The same widespread malfunction as is described for sensor rwt25 is observed. rwt27 is at the very tip of the wing and usually contains no fuel when cruising altitude is reached. Therefore, the large weighted KL divergence observed in rwt25 are not seen in rwt27.

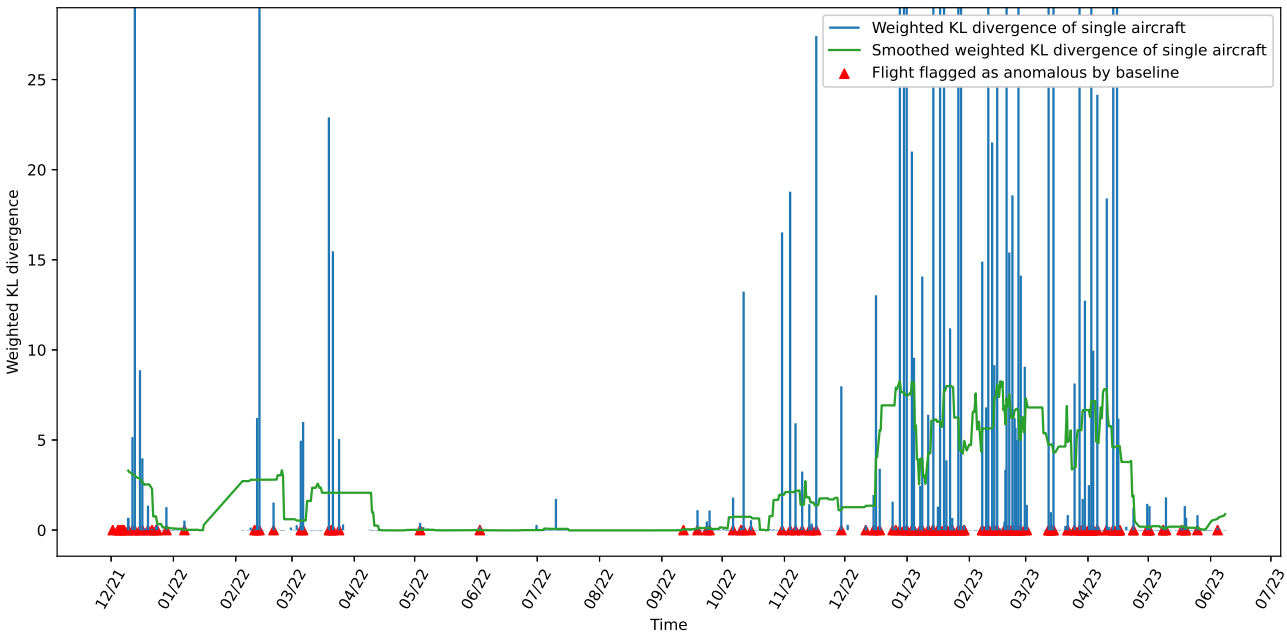
Fig. 29: Weighted KL divergence of proposed method with binary results from currently employed (baseline) method.

Appendix M
Comparison DTW and MSE for health analysis

The Figures 30a and 30b show the effect that choosing a different error metric for the health analysis has. The images show the weighted KL divergence for each flight of a single aircraft and a single fuel probe over an 18 month time span. Higher weighted KL divergence means a higher level of anomalousness according to the proposed method. The red triangles are binary indicators on whether the baseline considers the flight to be anomalous. In general, it can be observed that the proposed method based on DTW matches the baseline method better. The figures are of a single fuel probe (lwt10), however, these findings are consistent over the other sensors. This is likely due to the fact that the baseline method only flags flights if they have a predetermined number of spikes. DTW is more sensitive to spikes and less sensitive to dissimilarities between sensors anomalies as discussed in Section IV-C.



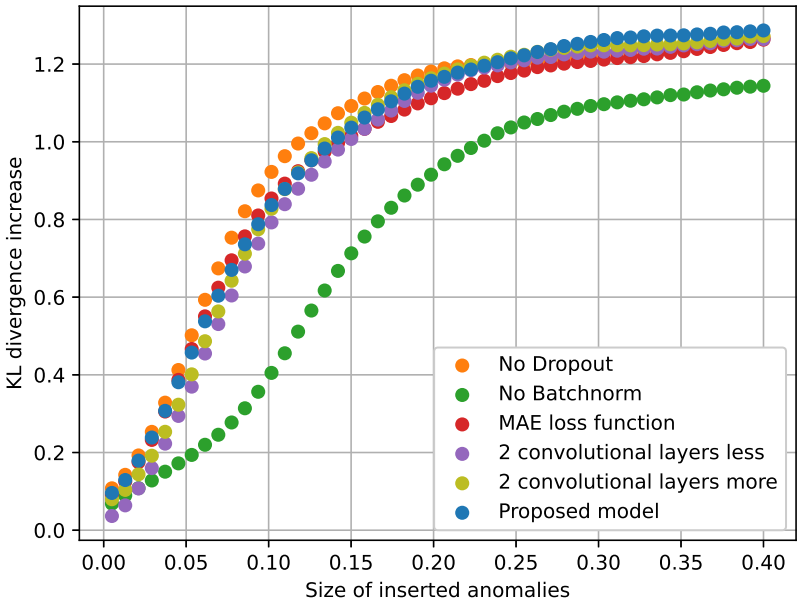
(a) Health analysis based on KL divergence using DTW.



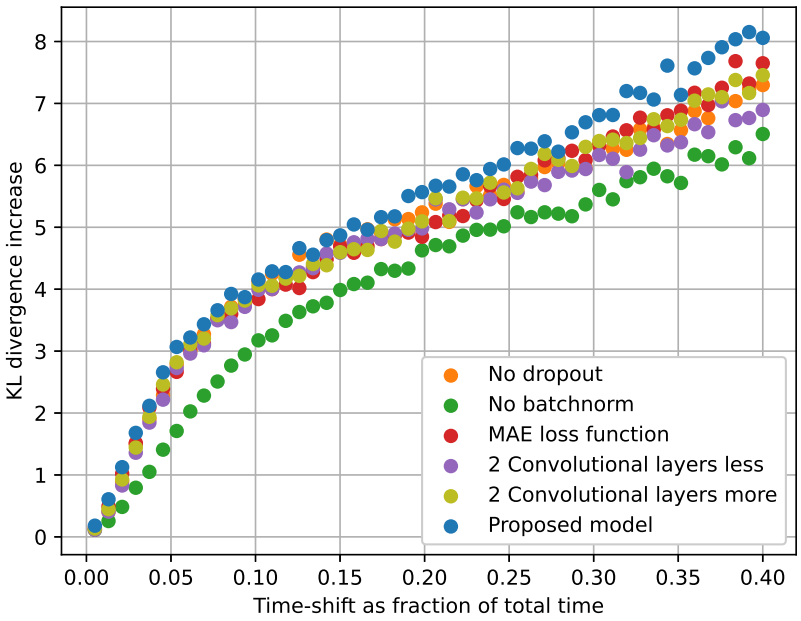
(b) Health analysis based on KL divergence using MSE.

Fig. 30: A comparison of MSE and DTW based errors used for the KL divergence. All other parameters are kept identical, as well as the flights and specific fuel probe (lwt10).

Appendix N
Ablation study sensitivity plots



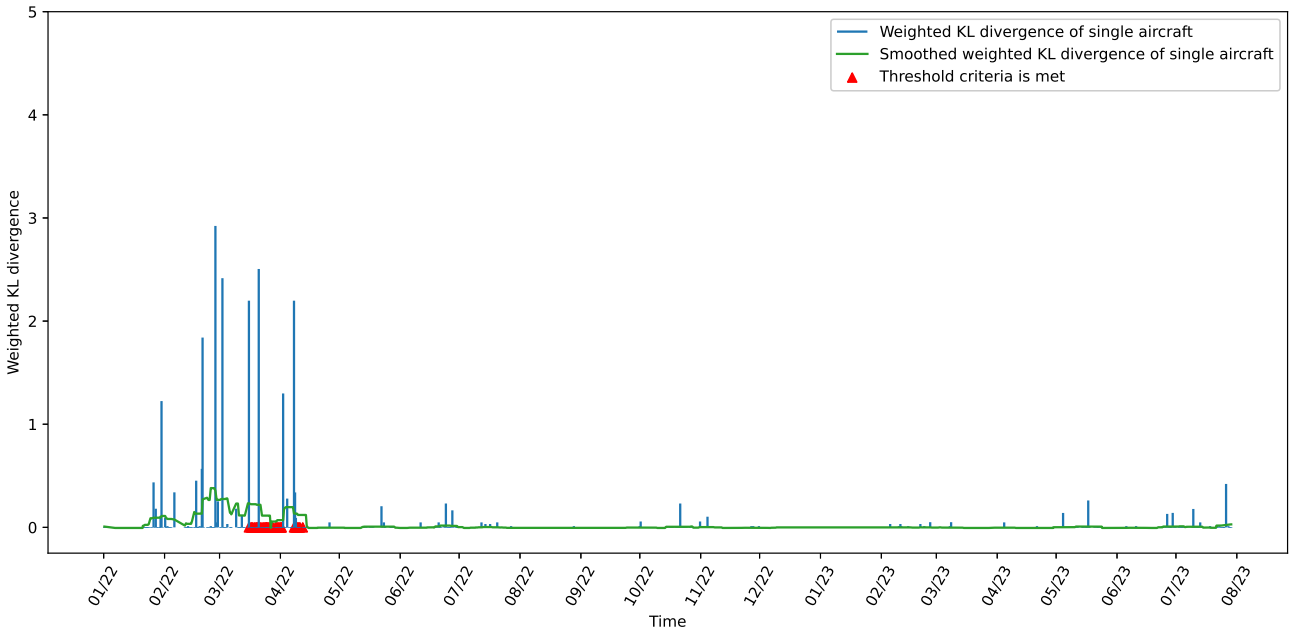
(a) Ablation study based on synthetic spikes and prolonged constant errors.



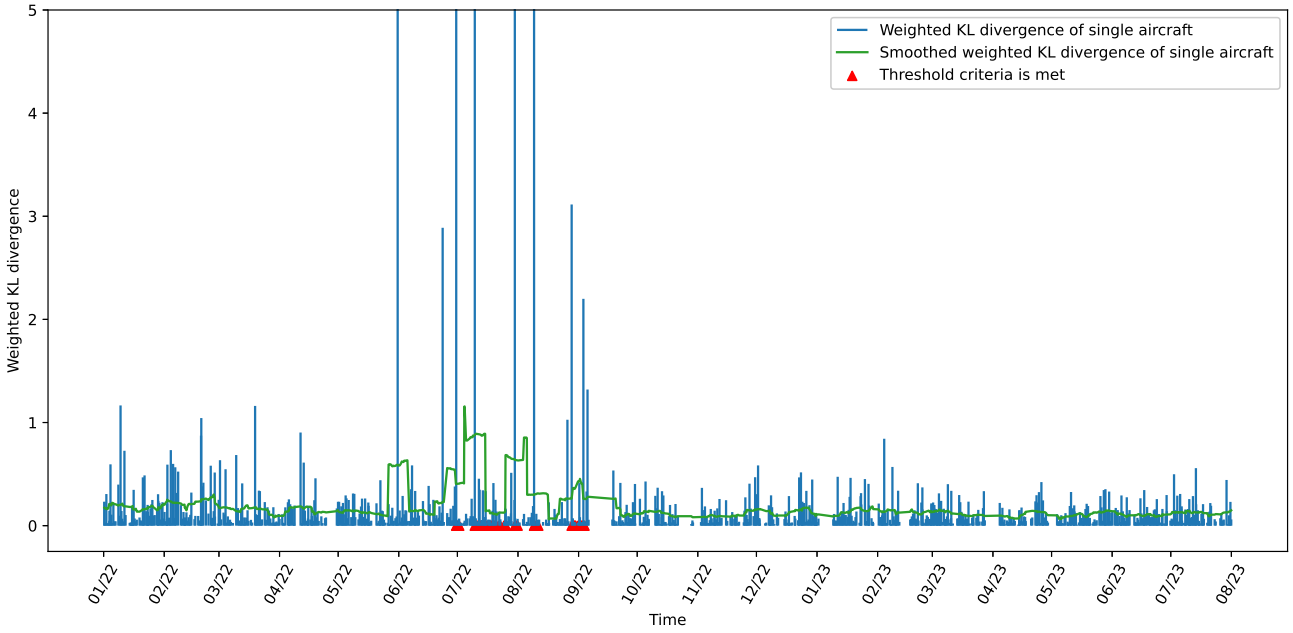
(b) Ablation study based on synthetic inter-sensor fuel level discrepancies.

Fig. 31: Ablation study based on sensitivity studies with various types of synthetic anomalies.

Appendix O
Generalisability study - the brake system



(a) The weighted KL divergence response on the displacement measurements prior to a faulty actuator replacement.



(b) The weighted KL divergence response on the current measurements prior to a worn brake disk replacement.

Fig. 32: Weighted KL divergences on flights from Jan. 2022 to Aug. 2023 for individual sensors of the braking system.