Investigating Lab-scaled Offshore Wind Aerodynamic Testing Failure and Developing Solutions for Early Anomaly Detections

Yuksel R. Alkarem[1], Ian Ammerman[2], Kimberly Huguenard[1], Richard W. Kimball[2], Babak Hejrati[2], Amrit Verma[2], Amir R. Nejad[3], Reza Hashemi[4], and Stephan Grilli[4]

[1]Civil and Environmental Engineering Department, University of Maine, 35 Flagstaff Road, Orono, Maine 04469, USA
[2]Mechanical Engineering Department, University of Maine, 35 Flagstaff Road, Orono, Maine 04469, USA
[3]Department of Marine Technology, Norwegian University of Science and Technology (NTNU), Jonsvannsveien 82, 7050 Trondheim, Norway
[4]Ocean Engineering, University of Rhode Island, Sheets Building, 15 Receiving Road, Narragansett, Rhode Island 02882, USA

Correspondence: Yuksel R. Alkarem (yuksel.alkarem@maine.edu)

Abstract. ~~Experimental demonstration of offshore wind components and systems plays an increasingly critical role in advancing the industry. As these systems increase in complexity, the likelihood~~ As offshore wind systems become more complex, the risk of human error or ~~software malfunction increases , leading to costly equipment damage~~equipment malfunction increases during experimental testing. This study ~~examines a laboratory incident which occurred during aerodynamic characterization of~~ investigates a lab-scale incident involving a 1:50 scale 5 MW ~~reference wind turbineand presents an efficient early anomaly detection method. During the experiment, the model generator disengaged causing a catastrophic~~ wind turbine, where a generator failure led to rotor overspeed and ~~subsequent~~ a blade-tower strike. ~~Applying~~ To improve early fault detection, we propose a data-driven ~~approach to predict system's dynamics from measurements , this study investigates the potential to enhance reaction time and prediction quality~~ method based on multivariate long short-term memory (LSTM) models. High-frequency measurements are projected onto principal components, and anomalies are identified using reconstruction error and its time derivative. Two models are trained on different healthy datasets and tested using single- and multi-principal component ~~models. Early system malfunctions are detected with the single-principal component model showing better performance. Sensitivity analyses show gains in reaction time with increasing sample frequency, lending this work in particular to~~ (1PC and MPC) variations. Results show that combining both error and error derivative improves detection accuracy. The 1PC model detects faults faster, has a higher recall rate, and achieves a 43% improvement in anomaly detection accuracy, while the MPC model yields higher precision. This approach provides a simple and effective tool for early anomaly detection in lab-scale ~~systems that operate at high sample rates to reduce future incidents~~experiments, helping to reduce the risk of future failures during the testing of new technologies.

1 Introduction

Model scale laboratory testing is a necessity for early development of grid scale on- and offshore wind energy technologies, and recent industry trends have driven increased demand for such testing (Mehlan and Nejad, 2024; Soares-Ramos et al., 2020). In the case of offshore wind energy projects, operation and maintenance costs can amount to a third of ~~the projects~~ a project's life-cycle cost, often quantified as the levelized cost of energy (LCOE)~~, meaning that~~. Small-scale validation and testing improve the maturity of new technologies ~~for deployment must be improved through small-scale validation and testing~~ (Mehlan and Nejad, 2024; Association, 2009; Leahy et al., 2016; Wang et al., 2022).

To meet this demand, lab-scale turbine systems are designed to match the performance of full-scale offshore commercial wind plants~~to facilitate accurate coupling of wind turbine dynamics with~~, enabling accurate coupling between wind turbine aerodynamics and the hydrodynamic forces on the substructure Fowler et al. (2023); Kim (2014); Cao et al. (2023). ~~As a consequence of the low-Reynolds wind environment at lab-scale, model turbine blades rely on~~ Due to the low Reynolds number at lab scale, thin airfoil sections are used for the model turbine blades, such as the SD7032, to achieve ~~sealed rotor performance, increasing flexibility and reducing strengthof the blades. Furthermore, due to tight mass considerations, particularly for floating models, system redundancy in the case of~~ full scale rotor performance. However, this increases blade flexibility and reduces structural strength. Additionally, because of strict mass constraints-particularly for floating configurations-system redundancy that accommodate equipment malfunctions is ~~not generally designed for Parker (2022). Consequently~~typically not included in the design (Parker, 2022). As a result, lab-scale turbines are highly sensitive ~~pieces of equipment requiring acute care~~ systems that require careful handling by operators to ensure safe and reliable operation throughout a test campaign.

In experimental testing campaigns, ~~and~~ particularly when testing novel control algorithms, the likelihood of fault events ~~is increased and the consequences of a fault or erroneous command~~ increases, and their impacts can be severe. These ~~events could fold as consequences of an operator error, erroneous~~ faults may arise from operator errors, incorrect control commands, or instrumentation malfunctions (Peng et al., 2023). Such ~~errors can lead to costly damage to lab equipment~~ incidents can result in costly equipment damage, violations of laboratory safety standards, and ~~cause significant project delays; therefore~~substantial project delays. Therefore, efforts to develop efficient methods of detecting operational faults are critical to improving the testing process ~~.~~ (Leahy et al., 2016; Lu et al., 2024).

~~Currently, studies of conceptual vibration-based~~ Vibration-based condition monitoring techniques, ~~such as velocity and acceleration measurements, were proposed for rapid and early online fault detection~~often evaluated using the root mean square (RMS) of velocity or acceleration signals, are widely used for drivetrain fault detection, particularly to determine whether signal amplitudes exceed the thresholds defined by standards such as ISO 10816-21 (ISO, 1996). For example, Nejad et al. (2018) demonstrated that angular velocity measurements already available in existing control systems can be repurposed for fault detection, avoiding the need for costly additional instrumentation. Their approach was motivated by the challenge of identifying faults in ~~commercial-scale systems (Nejad et al., 2018); however, in~~ complex systems ~~composed of multiple~~ with many interconnected components, where vibration signals may originate from ~~multiple sources and the data captured by individual sensors may offer limited insight. Consequently~~various internal sources at different frequencies. In such cases, incorporating multiple ~~data channels is essential for constructing a comprehensive representation~~ sensor channels is often necessary to obtain a more complete understanding of system behavior~~, though doing so can come at the expense of greater~~. However, this added complexity can increase computational cost and the risk of misinterpreting ~~otherwise irrelevant signal data. Such risks can be mitigated through dimension reduction strategies during system~~ irrelevant or noisy signal components.

To mitigate these issues, dimensionality reduction techniques, such as principal component analysis (PCA), are often employed during pre-processing to ~~improve algorithm efficiency (Dibaj et al., 2022). Other approaches may utilize adaptive filters, such as~~ retain the most informative features while reducing data redundancy. For instance, Dibaj et al. (2022) applied PCA to multi-point raw vibration data as a means of compressing the dataset prior to classification, thereby improving computational efficiency without sacrificing key diagnostic information. These reduced-dimensional signals were then input to a convolutional neural network (CNN) for automated fault classification and pattern recognition. Similarly, adaptive filtering techniques, including linear and non-linear Kalman filters~~for efficient adaptive fault detection , although these strategies can be challenging to implement for increasingly complex systems (Zhou and Zhu, 2023; Le and Matunaga, 2014; Ammerman et al., 2024) . Online~~ , have been used to enhance fault detection capabilities in dynamic environments, though their implementation can become increasingly complex for large-scale systems (Zhou and Zhu, 2023; Le and Matunaga, 2014; Ammerman et al., 2024) . Overall, data-driven ~~prediction and forecast modelshave also been used in the past to detect changes in the systems' state or any~~ models, when combined with feature extraction or filtering techniques, provide a robust framework for detecting changes in system state and identifying early signs of failure or adverse environmental conditions (Dibaj et al., 2022; Alkarem et al., 2024, 2023).

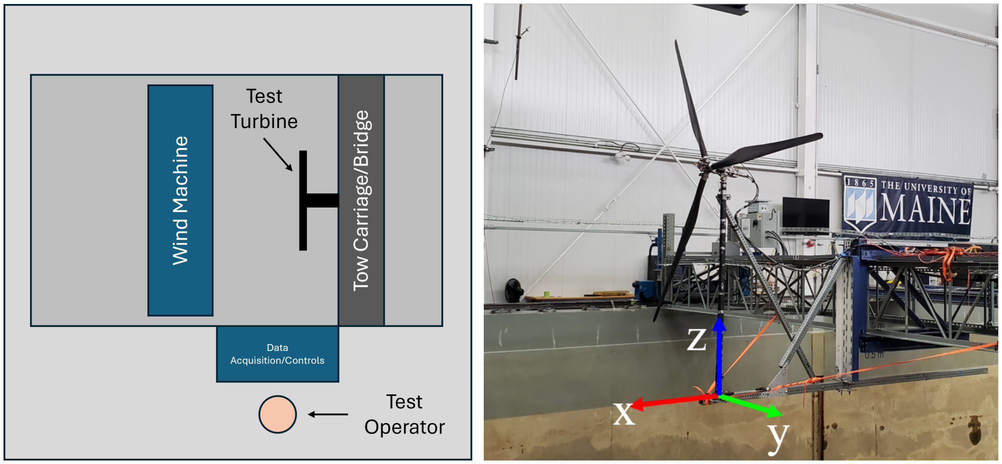
These and similar methods can also be applied to lab-scale models, with the additional caveat that computational efficiency is even more critical. Due to time scaling and typically higher frequencies of motion at lab-scale, fault detection strategies on models must be able to operate quickly and with minimal overhead. To meet this need, pre-trained data-driven approaches offer significant performance benefits over non-linear physics-based models.

The case study in this work comes from a fault incident which occurred during a standard scale model characterization test, wherein the turbine generator disengaged during an experiment, causing the turbine to spin out of control and one of the blades to strike the tower. The resulting damages caused significant delays in the campaign. Using this incident as a real example of the need for online fault detection and mitigation strategies, a data-driven approach was applied to develop an efficient online monitoring system which can detect failures or anomalous behavior before significant system effects are realized, increasing reaction time for operators or enabling automated shutdown procedures to take place.

2 Methodology

2.1 Experimental setup

The experimental data for this study comes from a wind turbine characterization test performed on a scale model, at the Harold Alfond Ocean Engineering ~~Lab~~ Laboratory at the University of Maine's Advanced structures and Composites Center. The layout of the experiment is shown in Figure 1a, illustrating the arrangement of the wind machine and turbine model. During the experiments, the turbine was controlled and monitored by 1 or 2 test operators stationed ~~to~~ on the side of the basin, via a data acquisition system (DAQ) based on the National Instruments cRIO platform. Figure 1b shows the installed experimental turbine before testing began. To properly characterize the turbine's aerodynamic performance, it was installed in a fixed configuration within the wind field. Cross-bracing was installed to keep the turbine tower and mounting surface rigid during the test to target rotor performance only.

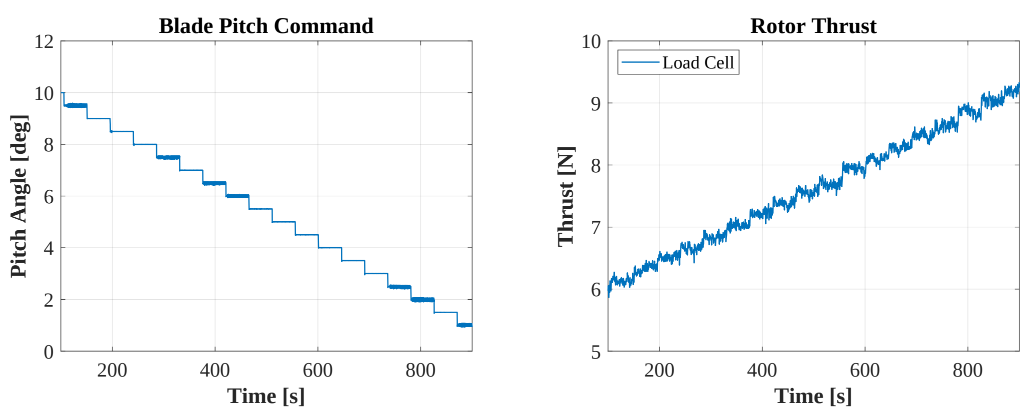


(a) Basin layout for scaled turbine characterization experiment.

(b) Photograph of the experimental test turbine installed in the basin.

Figure 1. Experimental test setup: (a) an overview, and (b) an image of installed turbine (b).

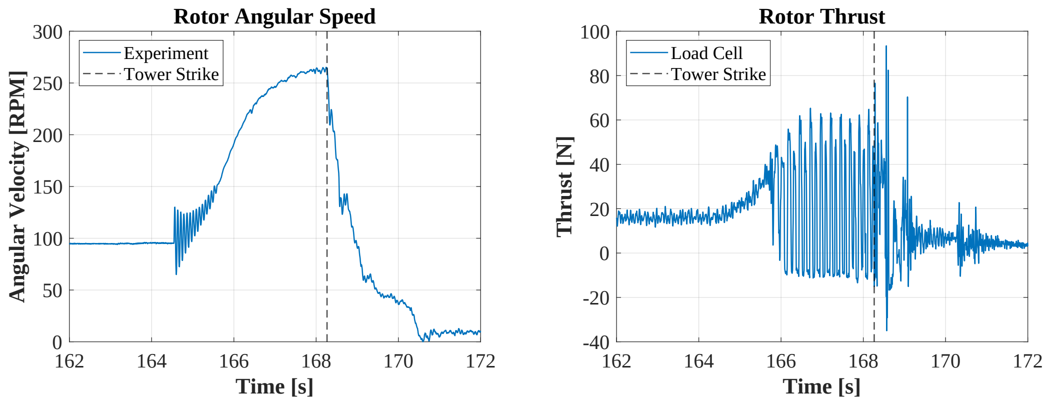
To fully characterize the rotor, experiments were performed at various wind ~~speed-angular velocity~~ speed/RPM pairs. Each experiment used a previously generated setpoint file to cycle through blade pitch setpoints. Figure 2 shows blade pitch (2a) and rotor thrust (2b) from one of the experiments. Results from these tests were then used to form rotor performance surfaces for future experiment design.



(a) Blade pitch schedule~~for sample characterization experiment~~.

(b) Rotor thrust measured~~for sample characterization experiment~~.

Figure 2. (a) Scheduled blade pitch and (b) measured rotor thrust for sample characterization experiment run.



(a) Rotor angular velocity in RPM. (b) Rotor thrust ~~during incident~~force.

Figure 3. Blade strike incident: (a) angular velocity in RPM, and (b) rotor thrust force. The dashed line represents the blade-strike instance.

2.2 Failure incident

During one of the characterization runs, an operator mistakenly triggered an emergency stop on the turbine generator. As a result, the rotor began accelerating unrestricted until a blade strike occurred with the tower~~and the wind generator could be shut down~~. Plots of rotor speed and thrust load during the incident are shown in Figure 3, with a vertical line indicating when the blade struck the tower.

2.3 Predictive model description

Detecting early signs of anomalies in testing campaigns can be beneficial. It can either provide data where operators can act upon with informative decisions and/or it can be automated to abort the test in case certain thresholds are exceeded. However, signaling a possibility of an anomaly requires real-time processes of incoming ~~measurements~~ measurement data, which can be best done using deep machine learning algorithms. Such algorithms indeed make it possible for predictive models to be trained on certain healthy data ~~then, when processing upcoming measurements in the lab during similar testing, to provide predictions or forecast of a certain state of the system~~and provide predictions of the systems' states during similar runs.

Accidents with lab equipment can be costly and labor intensive and can cause delays. ~~to~~ To mitigate such incidents, we propose an early anomaly detection model to improve response times and reduce human error. To this effect, a multi-step, multivariate Long Short-Term Memory (~~MLSTM~~LSTM) model — a type of recurrent neural network (RNN) designated to address the vanishing gradient issue that traditionally prevents models to capture long-term dependencies — was developed and trained on data from a healthy aerodynamic characterization tests with similar wind speeds. When an anomaly occurs, the error between the predicted signal and the measured signal increase which can be used to inform the operator of such an incident. The model parameters were initially estimated intuitively, but these could be further refined for enhanced predictive accuracy.

2.4 Anomaly detection over the span of multiple channels

In complex systems such as offshore wind testing, there are numerous measurements and data channels, which can be used to understand the overall behavior of the system. However, for anomaly detection purposes, it can be overwhelming and computationally expensive to manually and in real-time search the data space for deviation in measured data. The operator might not have sufficient time to abort the test before the anomaly becomes too consequential. Additionally, an anomaly might not be detectable based on any single data channel to comprehend the full state of the system. Therefore, the predictive model must be based on multiple data channels related to the test being conducted while providing the operator with a single, concise, anomaly detection capability based on the most relevant information. To accomplish that, principal component analysis (PCA) was carried out. A PCA creates combinations of variables that explain the largest amount of variance in the data.

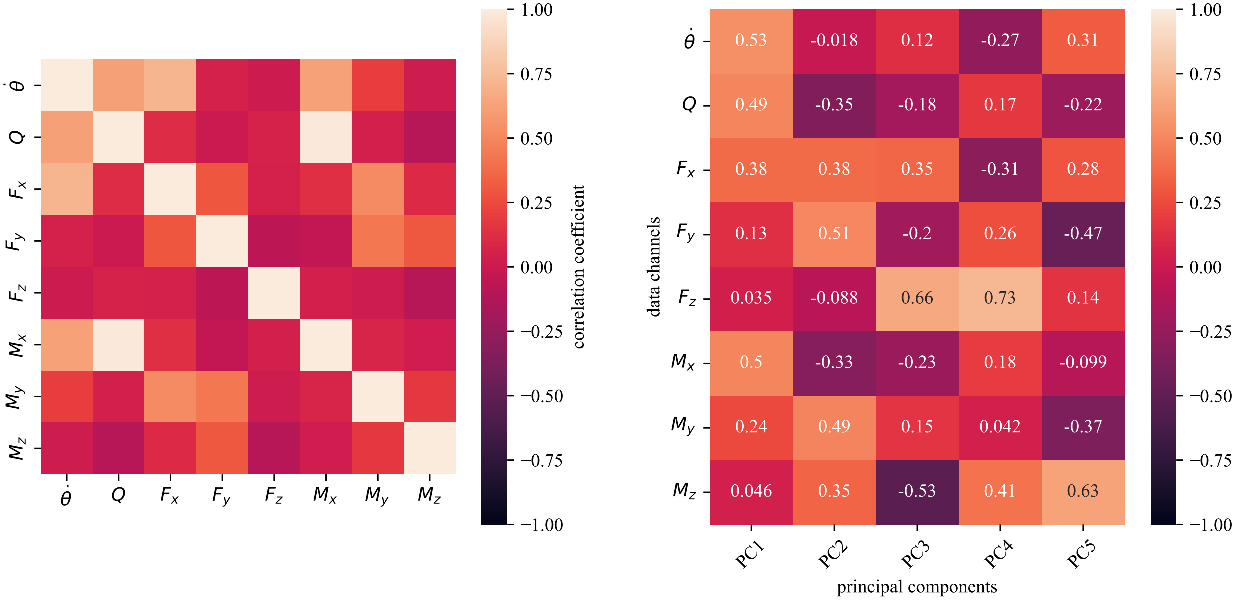
~~Prior to performing this analysis, earlier data reported~~ Before performing the analysis, the raw data recorded by the data acquisition system ~~is cleaned~~ were pre-processed to remove idle measurements ~~or~~ and non-numeric ~~values. Then, the data are prepared for the PCA, by standardizing them, using their mean and standard deviation, which ensures that~~ entries. Data channels collected comprises wind speed, angular velocity of the rotor, azimuth angle, all blades pitch angles, generator torque, rotor torque, forces and moments at the base of the tower. We assume the digital twin only has access to some of these channels (i.e., angular velocity, $\dot{\theta}$, rotor torque, Q, and tower base forces and moments: $F_x, F_y, F_z, M_x, M_y, M_z$) to simulate cases where some measurements can be restricted by turbine manufacturers and validate the model's operability under restrictive data access. Figure 4a illustrates the correlation matrix between the channels of interest.

Data are then standardized to ensure all channels (features) ~~are on the same scale to prevent features with larger ranges dominate. The covariance matrix is then computed for the standarized variables , which is then rotated to become a diagonal matrix, with transformed variable, a.k.a.~~ have a mean value, μ, of 0 and a standard deviation, σ, of 1:

$$x_i = \frac{x_i - \mu_i}{\sigma_i}, i = 1, ..., N, \tag{1}$$

where N is the number of channels included in the model. Following standardization, the covariance matrix of the variables was computed and then diagonalized through eigendecomposition, yielding a set of orthogonal transformed variables—i.e., the principal components (PCs)~~, ranked from those describing the largest to the lowest fraction of the total variance . One can finally decide which components to retain for the purpose of reducing the problem dimensionality, which still explaining the largest possible amount of~~ —ordered by the amount of total variance ~~.~~ they explain. Based on this ranked structure, a subset of components can be selected to reduce the dimensionality of the problem while preserving as much of the original variance as possible. For instance, the first 5 PCs and channel loads/contributions to them is illustrated in Fig. 4b

The PCs ~~, which are expressed as weighted sums of the earlier variables,~~ were then used to train the RNN model(s) that will later be used for prediction. As new data is acquired, it is ~~transofmed~~transformed/projected onto the same PCs that were used in training the models. For the purpose of anomaly detection, the mean absolute error (MAE) is computed between measurements and predictions from the RNN models, and the error derivative is calculated, to estimate rapid fluctuations in the quality of



(a) Correlation matrix between channels of interest. (b) Covariance loadings matrix.

Figure 4. Data pre-processing: (a) correlation matrix between available channels used in the models, and (b) covariance loadings of the first 5 principal components.

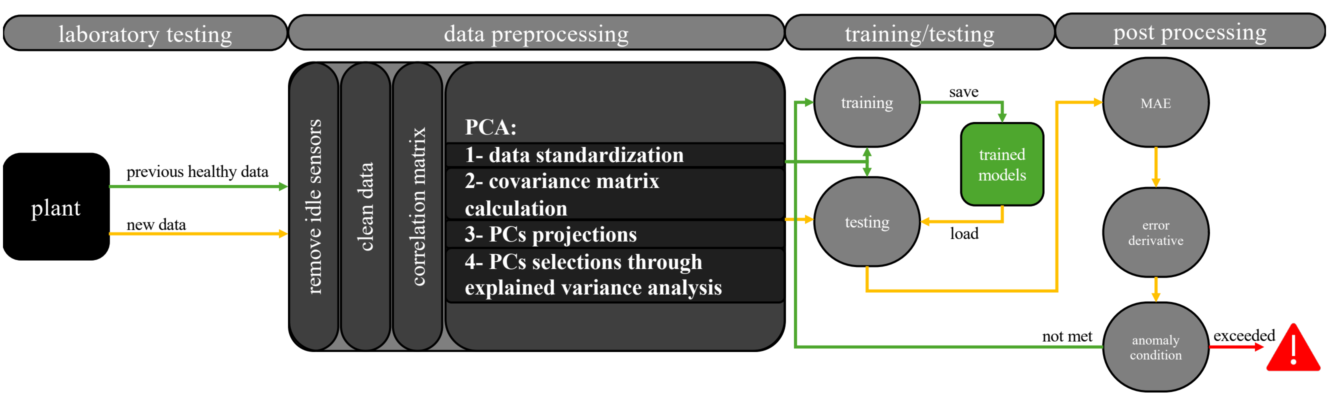


Figure 5. Flowchart describing data stream, data ~~preprocessing~~pre-processing, training the ~~MLSTM~~ LSTM model and using it for anomaly detection.

the predictions. An anomaly alert is reported to the operator when certain anomalous conditions are met. In this research, we investigate conditions when both the error and its derivative were crossing certain thresholds. This procedure is illustrated in Figure 5 and is explained in section 2.6.

2.5 Principal component selection

Two models developed vary in their projected principal component selection. The projected PC results from training data are presented in ~~Figure~~ Fig. 6. The first model compresses the data by retaining only the first PC; it is therefore named '1PC'. The second model selects the group of (M) PCs that cumulatively explain 90% of the total variance, thus only neglecting the remaining 10%; this model is hence called 'MPC'.

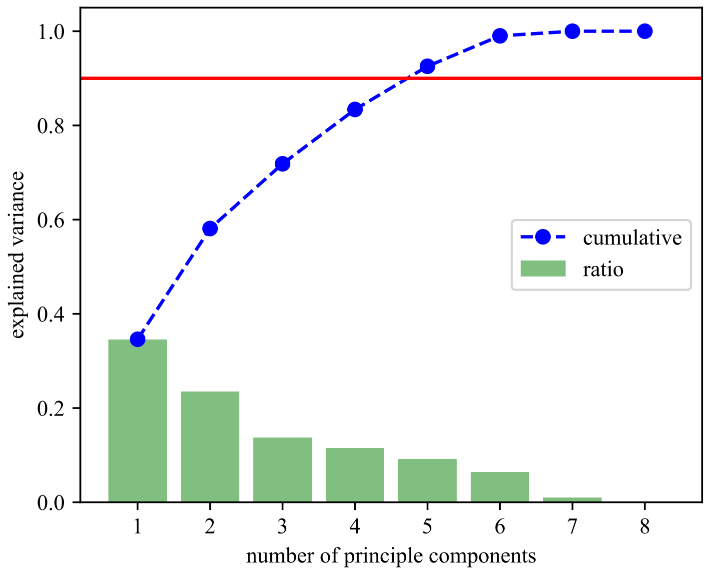
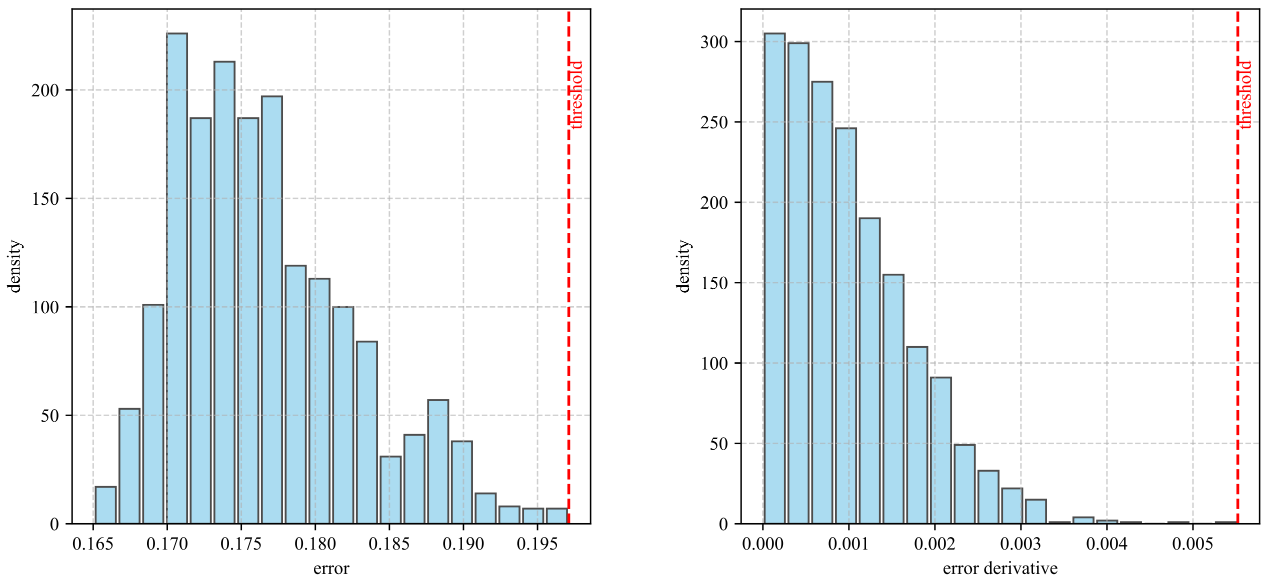


Figure 6. Explained variance ratio and cumulative of all principal components.

2.6 Error and error derivative thresholds selection for single/multiple PCs

The histogram of ~~the derivative of the error~~ error metrics between the trained model and the training data for the 1PC and MPC models are shown in ~~Figure 4.~~ Fig. 7 for the 1PC model and in Fig. 8 for the MPC model. Inspired by the work of Dibaj et al. (Dibaj et al., 2024), the thresholds were selected to be the highest values in the histogram for the training data error. ~~This will later be usedto assess whether or not the predictive model is diverging from the~~ However, in case of multiple principal components being used. the weighted average of maximum errors (and error derivatives) of all principal components was computed. The per-PC thresholds are weighted by the explained variance ratio of the corresponding principal component shown in Fig. 6. These threshold values were used to assess the accuracy of the predictive model against measured data during ~~the~~ testing/anomaly detection stage.



(a) Error histogram in 1PC model.

(b) Error derivative histogram in ~~MPC~~ 1PC model.

Figure 7. 1PC Histograms of the (a) error, and (b) error derivative generated by comparing the model with the training data, and the maximum error/error derivative as the selected threshold.

2.7 Performance metrics

The overall accuracy of the model(s) was measured by a single score that combines precision and recall in its calculation (Miele et al., 2022; Wang et al., 2019). Precision, P, illustrates the proportion of anomalies detected that are true, while recall, R, indicates which proportion of true anomalies are detected. They can be computed as:

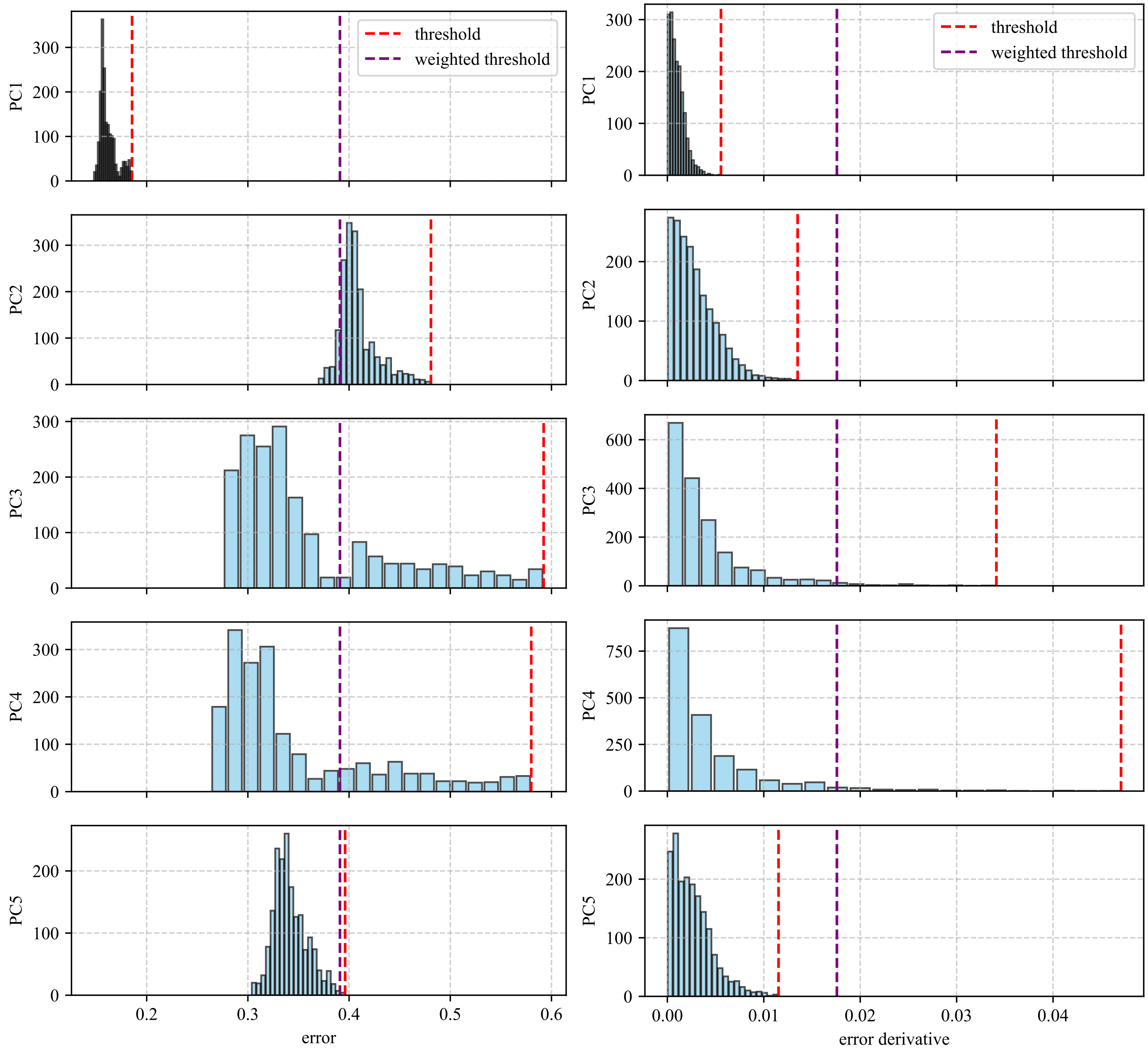
$$P = \frac{T^+}{T^+ + F^+}, \; R = \frac{T^+}{T^+ + F^-} \tag{2}$$

where T^+ represents the count of true positives (the identified anomalies are true), F^+ are the count of false positives (i.e., for which the identified anomalies are not true), and F^- are the false negatives (i.e., the unidentified true anomalies). These contribute to an overall FI score that ranges between 0 and 1 with 1 being a perfectly precise model and is expressed as:

$$FI = 2 \times \frac{P \times R}{P + R}. \tag{3}$$

3 ~~Results~~Problem statement

3.1 ~~Pre-strike anomaly detection~~



(a) Error ~~derivative~~ histogram in ~~1PC~~ MPC model.

(b) Error derivative histogram in MPC model.

Figure 8. MPC Histograms of the (a) error~~for 1PC~~, and (b) error derivative for ~~1PC~~the first 5 PCs, ~~(c) error for MPC, and (d) error derivative for MPC throughout~~ generated by comparing the model with the training data, and the maximum error/error derivative ~~as~~ per PC and the ~~selected~~ weighted average threshold.

Three datasets, $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$, were gathered during the test campaign. While wind speeds were kept constant (variation < 1%), the rotor angular velocity for $\mathcal{D}_2$ dataset was slightly lower by 12% and higher by 51% for $\mathcal{D}_3$, relative to $\mathcal{D}_1$. The angular velocity, and the resulting thrust force variations are illustrated in Fig. 9a and Fig. 9b, respectively. The blade pitch varied the same way for these cases based on the pre-generated setpoints. The actual anomaly and blade strike occurred near the end of $\mathcal{D}_3$, which was truncated to <200 s, while $\mathcal{D}_1$ and $\mathcal{D}_2$ each span 1000 s. In addition, three altered variants of $\mathcal{D}_2$ were generated to introduce a synthetic anomaly for further evaluation of the developed models: $\mathcal{D}_2^{(a1)}$, $\mathcal{D}_2^{(a2)}$, and $\mathcal{D}_2^{(a3)}$. The synthetic anomaly was imposed by modifying the tower base fore-aft bending moment, M_y, through a time-varying amplification factor. This factor was applied starting from an arbitrary onset time (225 s), increased linearly to a maximum value by 250 s, and then reduced back to unity by 275 s. The variants amplify the signal by 0.25%, 0.5%, and 1.00% per Δt for $\mathcal{D}_2^{(a1)}$, $\mathcal{D}_2^{(a2)}$, and $\mathcal{D}_2^{(a3)}$, respectively. Table 1 summarizes the model setup and intervals of datasets utilized during training, validation, anomaly criteria threshold selection, and testing tasks.

~~During one of the high wind speed tests for the aerodynamic characterization of a 1: 50 scaled wind turbine, an erroneous activation of the emergency stop led to the shutdown of the generator. This caused the rotor to spin rapidly increasing thrust forces on the blades, causing them to bend more. Within less than four seconds, one of the blades struck the tower, resulting in severe damage as shown in figure 13~~Channels used in training the models include angular velocity, rotor torque, and tower base forces and moments. For most of the analyses presented in this paper, model $\mathcal{M}_1$ was employed. This model is trained on a previously available healthy dataset, $\mathcal{D}_1$, and serves as the primary reference. The rationale for this approach is based on the practical constraint that datasets containing anomalies rarely have a corresponding healthy segment recorded immediately beforehand. As such, training a model in real-time using only the healthy portion of a dataset that later exhibits an anomaly is typically infeasible. Nonetheless, for comparative purposes, we also evaluate model $\mathcal{M}_3$, which is trained on the healthy portion of dataset $\mathcal{D}_3$, under the hypothetical assumption that similar data had been recorded under identical conditions in advance.

Models $\mathcal{M}_1$ and $\mathcal{M}_3$ were configured with identical training hyperparameters, except for the number of training epochs. Both models utilize a prediction horizon of a single timestep and a look-back to prediction ratio of $n/m = 10$. The network architecture consists of a single hidden layer with 100 neurons, trained using a batch duration of 60 seconds, a learning rate of 0.001, and no dropout regularization. Model $\mathcal{M}_1$ was trained for 60 epochs, whereas model $\mathcal{M}_3$ required an extended training schedule of 1000 epochs. This increase was motivated by the significantly shorter duration of training data available for $\mathcal{M}_3$, which spans only from 70 to 119 seconds due to the presence of an anomaly later in the dataset, as detailed in Table 1.

~~Blade damage after blade-tower collision due to high thrust forces.~~

~~The model used to detect an anomaly is based on a double condition criteria;~~ Three combinations of anomaly detection criteria were investigated. The symbols E and ΔE refer to threshold-exceeding conditions based on the model prediction error and its time derivative, respectively. The detection logic tested includes:

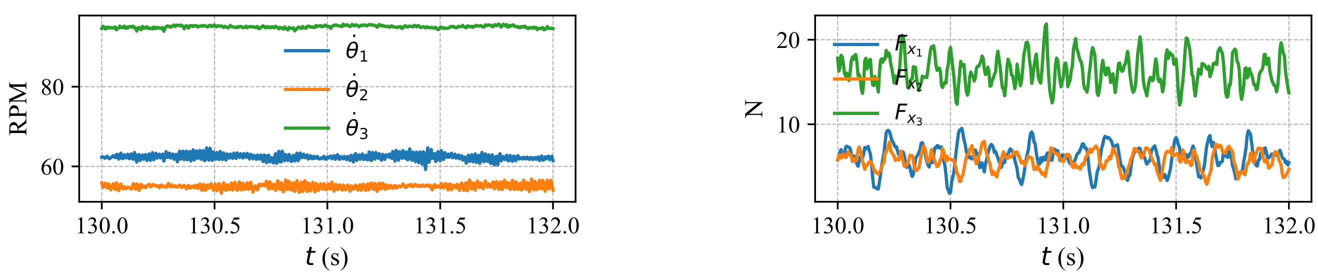
1. ΔE - the derivative of the error must exceed a threshold

2. $\Delta E \vee E$ - either the error or its derivative must exceed its threshold

Table 1. Dataset usage by models $\mathcal{M}_1$ and $\mathcal{M}_3$ for different tasks. Time intervals are in seconds.

Model	**Task**	**Dataset(s)**	**Interval**
$\mathcal{M}_1$	Training	$\mathcal{D}_1$	$[100, 450]$
	Validation	$\mathcal{D}_1$	$(450, 675]$
	Error threshold	$\mathcal{D}_1$	$[100, 1000]$
	Testing	$\mathcal{D}_2, \mathcal{D}_2^{(a1,a2,a3)}, \mathcal{D}_3$	$[100, 1000], [100, 350], (135, 190]$
$\mathcal{M}_3$	Training	$\mathcal{D}_3$	$[70, 119]$
	Validation	$\mathcal{D}_3$	$(119, 135]$
	Error threshold	$\mathcal{D}_3$	$[70, 135]$
	Testing	$\mathcal{D}_3$	$(135, 190]$

3. $\Delta E \wedge E$ - both the error and its derivative must ~~exceed the training error thresholds. This was shown to provide the most accurate model (from a sensitivity study carried out and presented below). The trained model tested against the run where the anomaly took place. The evolution of the error derivative of the first component for~~ simultaneously exceed their respective thresholds



(a) Angular velocity of three experimental dataset.

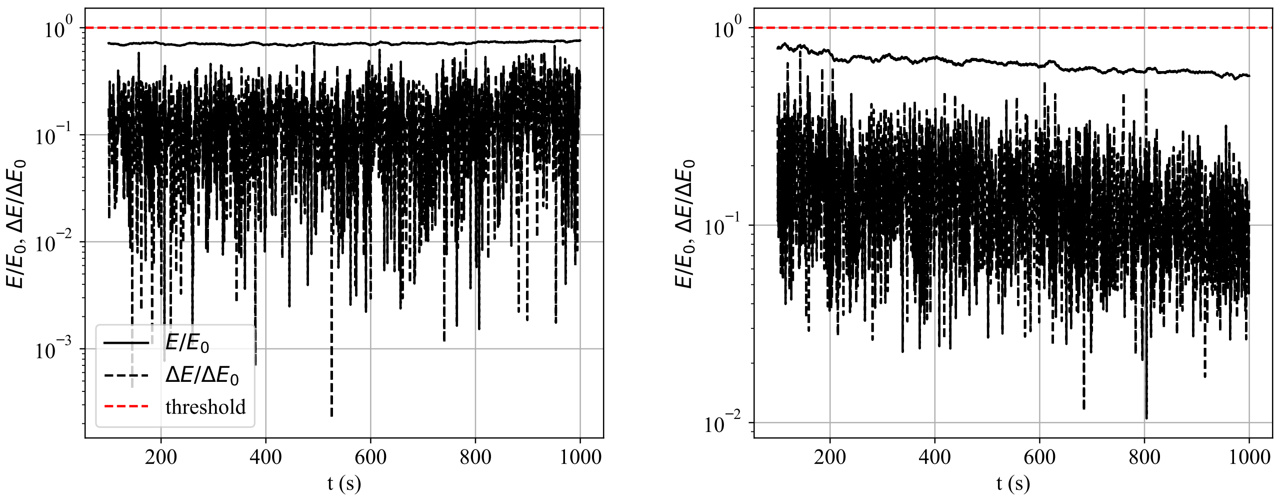
(b) Thrust force measurements of three experimental dataset.

Figure 9. Three experimental datasets and their variations in (a) angular velocities and (b) thrust forces.

4 Results

4.1 Model performance during healthy conditions

The performance of the $\mathcal{M}_1$ model, in terms of normalized error and error derivative to their respective threshold values, when tested against measured data during healthy operations, $\mathcal{D}_2$, are shown in Fig. 10. When using the lead principal component (i.e., 1PC variation of $\mathcal{M}_1$ model), the error values were consistent throughout the test. The MPC variation experienced a slight decline in error values as time progressed. As desired, both model variations exhibited no predicted anomalies based on any of the exceeding threshold criteria discussed.



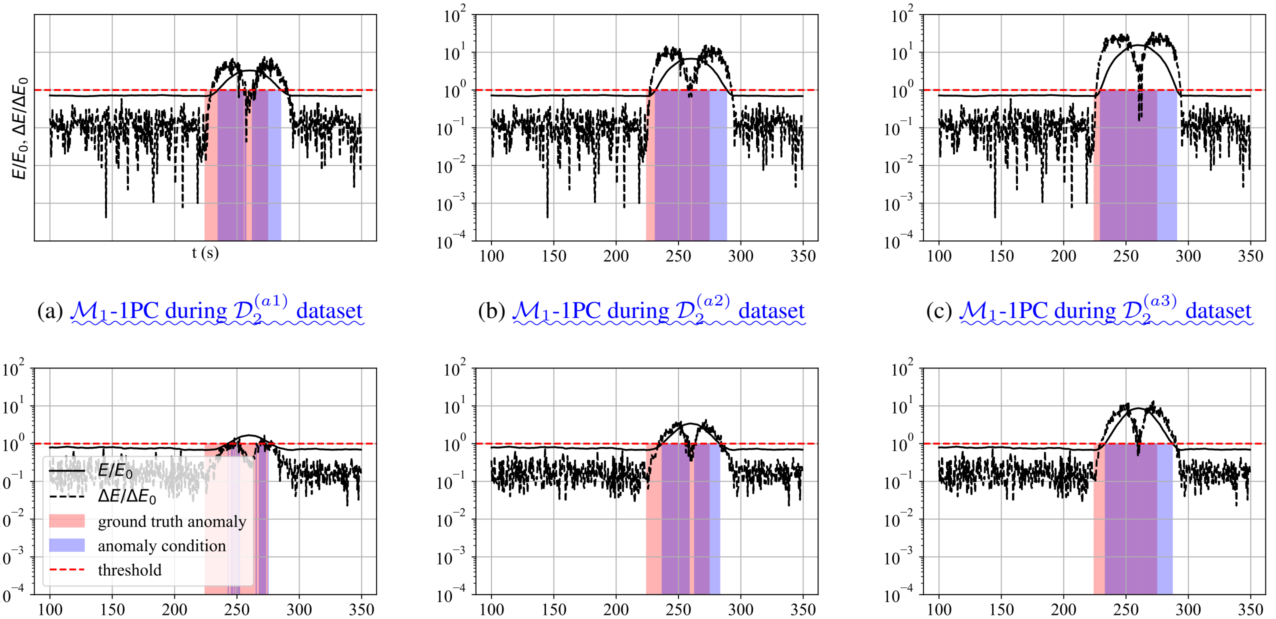
(a) $\mathcal{M}_1$ 1PC model performance

(b) $\mathcal{M}_1$ MPC model performance

Figure 10. Error and error derivative curves between measured and $\mathcal{M}_1$ model during healthy $\mathcal{D}_2$ testing dataset when (a) a single or (b) multiple principal components are used.

4.2 Performance under synthetic anomaly realizations

Model $\mathcal{M}_1$ was tested on the synthetically altered variations of $\mathcal{D}_2$: datasets $\mathcal{D}_2^{(a1)}$, $\mathcal{D}_2^{(a2)}$, and $\mathcal{D}_2^{(a3)}$, using both 1PC and MPC variations. The results are presented in Fig. 11, where the first row (Figs. 11a, b, and c) shows the 1PC model ~~is shown in Figure ?? and the average error derivative of the first five components for MPC model in Figure ??. The vertical black dashed line is the inception of the anomaly and the vertical reddashed line is the earliest detection~~ responses, and the second row (Figs. 11d, e, and f) presents the MPC model responses. Anomaly criterion selected for this analysis is the joint condition ($\Delta E \wedge E$). The 1PC variation demonstrates overall enhanced coverage (highlighted in blue) and reduced detection delay relative to the onset of the ground-truth anomaly (highlighted in red). The reduction in detection delay is also represented in Fig. 12 when comparing 1PC to MPC. Additionally, detection performance generally improved with increasing severity of the synthetically introduced anomaly. This is indicated in Fig. 12 which shows detection delay in seconds between synthetically introduced anomaly and the predicted anomaly by the models~~; note, as it is impossible to detect an actual anomaly before its inception, the red line can only exist on the right side of~~ . The figure also shows a sensitivity analysis of the models to the ~~black line. Before the anomaly, both models have error derivatives below the threshold , indicating an accurate representation of a healthy system~~



(a) $\mathcal{M}_1$-1PC during $\mathcal{D}_2^{(a1)}$ dataset (b) $\mathcal{M}_1$-1PC during $\mathcal{D}_2^{(a2)}$ dataset (c) $\mathcal{M}_1$-1PC during $\mathcal{D}_2^{(a3)}$ dataset

(d) $\mathcal{M}_1$-MPC during $\mathcal{D}_2^{(a1)}$ dataset (e) $\mathcal{M}_1$-MPC during $\mathcal{D}_2^{(a2)}$ dataset (f) $\mathcal{M}_1$-MPC during $\mathcal{D}_2^{(a3)}$ dataset

Figure 11. Anomaly detection based on $\Delta E \wedge E$ criterion during synthetically altered dataset variations of $\mathcal{D}_2$ and $\mathcal{M}_1$ model detection response, with (1) 1PC - $\mathcal{D}_2^{(a1)}$, (b) 1PC - $\mathcal{D}_2^{(a2)}$, (c) 1PC - $\mathcal{D}_2^{(a3)}$, (d) MPC - $\mathcal{D}_2^{(a1)}$, (e) MPC - $\mathcal{D}_2^{(a2)}$, and (f) MPC - $\mathcal{D}_2^{(a3)}$ variations.

~~befxore the anomaly. The reddots are when anomalous conditions were met. In the present case, we see that the~~ timestep at which the data is sampled. Small timesteps (high sampling frequency) can provide reduce anomaly detection delay but at the expense of computational cost.

4.3 Pre-strike anomaly detection

During a high rotor angular velocity test, $\mathcal{D}_3$, an unexpected anomaly caused the rotor to accelerate rapidly. The resulting increase in thrust forces caused significant blade deflection, and within four seconds, one of the blades struck the tower, leading to severe damage, as shown in Figure 13.

Models $\mathcal{M}_1$ and $\mathcal{M}_3$ were evaluated using both 1PC ~~model able to detect the anomaly earlier than the MPC model as the error derivative breaches the threshold earlier. The reason of having a drop in~~ and MPC variations. The normalized error and error derivative, each scaled by their respective threshold values, are presented in Fig. 14. Anomalies are identified based on the joint exceedance of both criteria ($\Delta E \wedge E$). As shown in the figure, the predicted anomaly region (blue) aligns well with the ground-truth anomaly (red), demonstrating the efficacy of the detection method. Additionally, anomaly conditions are detected prior to the blade strike, suggesting that such models could be used as preventive measures against consequential incidents.

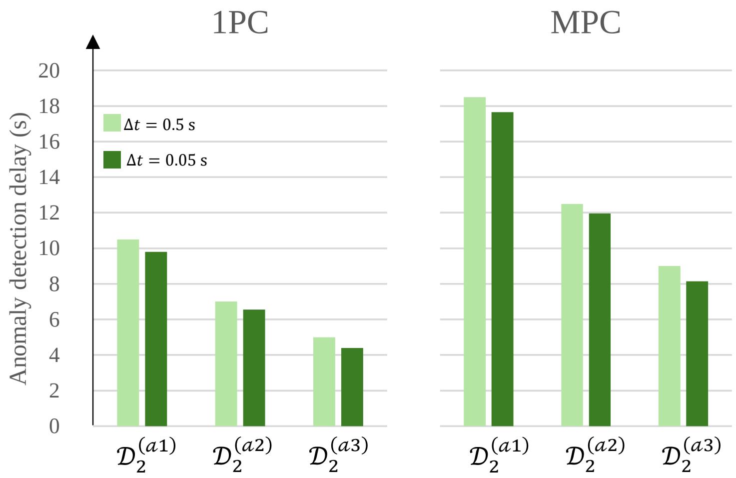


Figure 12. Anomaly detection delay in seconds for $\mathcal{M}_1$ (both 1PC and MPC variations) when tested during various altered $\mathcal{D}_2$ datasets for two timestep realizations.

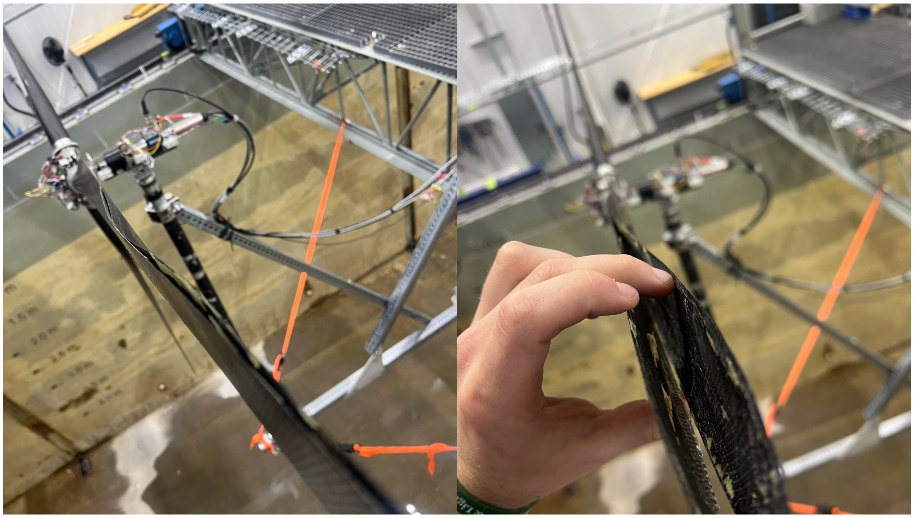


Figure 13. Blade damage after blade-tower collision due to high thrust forces.

For all models, the error derivative remains below the threshold prior to the anomaly, indicating that system behavior was consistent with healthy operation. However, the ~~error derivative after an anomalyhas started is because the error at that time reaches its peak and reverses down as the prediction model works to acquire to the available historical data.~~ 1PC variation of model $\mathcal{M}_1$ shows threshold violations in the error metric E, before the onset of the actual anomaly. This can be attributed to a mismatch in operating conditions: $\mathcal{M}_1$ was trained on dataset $\mathcal{D}_1$, where the turbine operated at significantly lower angular velocity, as illustrated earlier in Fig. 9a. This discrepancy introduces errors when applied to data from $\mathcal{D}_3$, which exhibits 51% higher angular velocity.

Notably, this premature threshold crossing is not observed in the MPC variation of $\mathcal{M}_1$. By incorporating multiple principal components, the MPC approach distributes the reconstruction error across several components, thereby diluting the influence of a mismatch from a single channel. This is further supported by the principal component contribution analysis in Fig. 4b, which shows that angular velocity $\dot{\theta}$ is the dominant contributor to the leading principal component. Consequently, in the 1PC case, discrepancies in angular velocity have a large impact on the error.

Despite exceeding error threshold in $\mathcal{M}_1$-1PC prior to the anomaly, the error derivative ΔE remains within acceptable bounds, ensuring no false positive detection. When model $\mathcal{M}_3$, trained on the healthy segment of dataset $\mathcal{D}_3$, is used instead, the predicted anomaly coincides precisely with the true event. This underscores the importance of matching operating conditions between training and deployment for reliable anomaly detection.

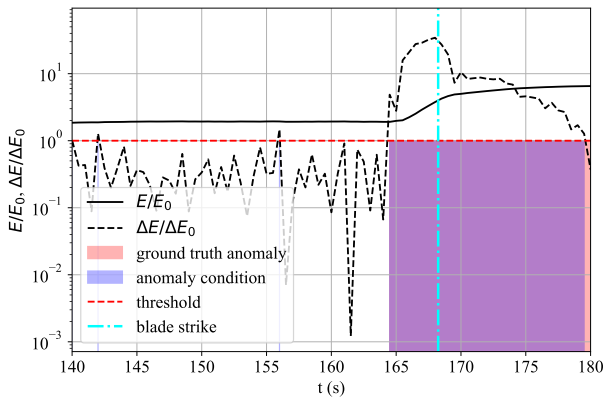
4.4 ~~Variation of anomaly detection criteria combination to the accuracy of the models~~

5 Discussion

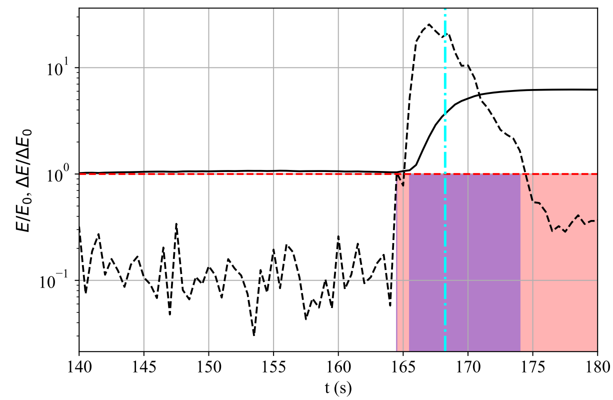
~~Three combinations of anomaly detection criteria and their effects on the FI score were investigated. Table **??** and Table **??** provide a descript of these variations for the 1PC and MPC models, respectively.~~ Table 2 summarizes the anomaly detection performance of model $\mathcal{M}_1$ with its 1PC and MPC variations, evaluated on the healthy dataset $\mathcal{D}_2$, synthetically altered anomaly datasets $\{\mathcal{D}_2^{(a1)}, \mathcal{D}_2^{(a2)}, \mathcal{D}_2^{(a3)}\}$, and the ~~$E$ and ΔE symbols refer to the error and the error derivative threshold exceeding criteria, respectively tested under normal healthy run operations (HR) or anomaly runs (AR). The combinations studied were 1) sole derivative, 2) a combination of absolute error value and its derivative must be met, and 3)~~blade-tower strike dataset $\mathcal{D}_3$. From these results, the following key observations can be made:

- The 1PC variation generally yields higher F1-scores compared to the MPC variation (43% enhancement under $\Delta E \wedge E$ criterion);
- The combined threshold criterion $\Delta E \wedge E$ provides the most consistent and reliable detection performance across datasets;
- While the 1PC model achieves higher recall (R), the MPC model tends to produce higher precision (P).

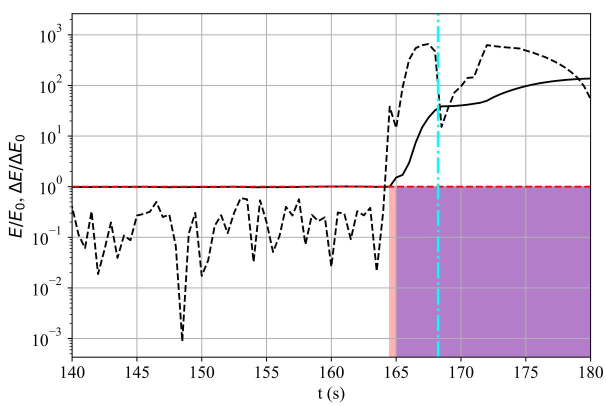
Importantly, both model variations produce no false positive detections under healthy conditions ($\mathcal{D}_2$), regardless of the threshold criterion employed. The 1PC model typically reacts more rapidly to actual anomalies, as it is not constrained by the



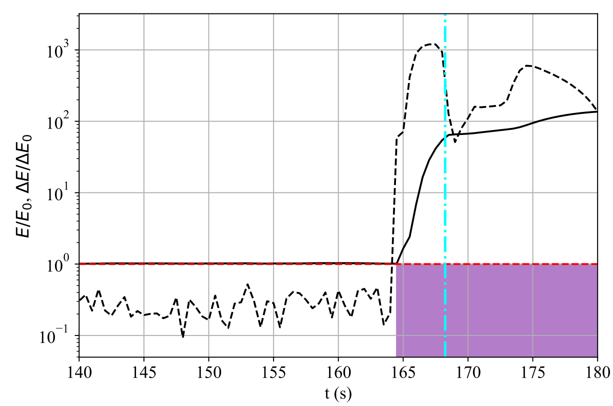
(a) $\mathcal{M}_1$ model - 1PC ~~anomaly detection~~variation



(b) $\mathcal{M}_1$ model - MPC ~~anomaly detection~~variation



(c) ~~Anomaly detection during the run when anomaly occurred and~~ $\mathcal{M}_3$ model ~~detection response using~~ - 1PC ~~(a) and MPC (b) modes.~~variation



(d) $\mathcal{M}_3$ model - MPC variation

Figure 14. Anomaly detection based on $\Delta E \wedge E$ criterion during $\mathcal{D}_3$ anomaly dataset and model detection response of the (1) $\mathcal{M}_1$ - 1PC variation, (b) $\mathcal{M}_1$ - MPC variation, (c) $\mathcal{M}_3$ - 1PC variation, and (d) $\mathcal{M}_3$ - MPC variation.

averaging of reconstruction errors across multiple principal components. This responsiveness contributes to its higher recall scores. However, this same sensitivity can lead to over-detection, which reduces precision. In contrast, the MPC model's error aggregation results in more conservative detection behavior, improving its precision at the expense of some detection latency.

Figure 15 presents the relative percentages of true positives, false negatives, false positives, and true negatives for model $\mathcal{M}_1$ across the same set of testing datasets. In the horizontal bar charts, darker shades correspond to the presence of anomalies in the data—hence their absence in the healthy dataset $\mathcal{D}_2$. The sign of each classification outcome indicates whether the model successfully detected an anomaly (positive) or failed to do so (negative). Color is used to convey prediction quality and context: green denotes correct classifications, while red indicates incorrect ones. This visual encoding effectively communicates both the correctness of model predictions and the operational context in which they occur, thereby emphasizing the model's ability to distinguish between healthy and anomalous system states.

The error derivative appears to be the dominant criterion for accurate anomaly detection. As shown in Fig. 15, the combined threshold criterion $\Delta E \wedge E$ results in fewer incorrect classifications (i.e., reduced red regions), whereas more flexible criteria—where either the error or its derivative ~~must be exceeded for the anomaly to be triggered. The precision and recall as well as the~~ $\cancel{FI}$ ~~score are presented. As the models were trained on healthy data, when being tested under other healthy data (with slight variations in blade pitch angle being tested), the HR, since it does not contain any anomalies, reports zero~~ $\cancel{T^+}$ ~~and~~ $\cancel{F^-}$ ~~values, but a non-zero~~ $\cancel{F^+}$ ~~value for~~ $\cancel{\Delta E}$ ~~and~~ $\cancel{\Delta E | E}$ ~~combinations because some normal samples were misclassified as anomalies . The~~ $\cancel{\Delta E \& E}$ ~~condition, however, does not trigger false anomalies and it also shows the highest~~ $\cancel{FI}$ ~~score. The MPC model shows similar patterns but with lower~~ $\cancel{FI}$ ~~scores.~~ alone exceeds the threshold—lead to increased misclassifications. Notably, the reconstruction error E serves as a useful indicator for identifying deviations due to previously unseen operating conditions. In contrast, the error derivative ΔE is particularly effective in capturing abrupt transitions between the reconstructed and measured signals, making it well-suited for detecting sudden-onset anomalies such as the one present in this paper.

5.1 ~~Sensitivity study of sampling frequency to anomaly detection delay, simulation speed, and accuracy~~

~~The two models were tested FOR the test run that includes the anomaly and the blade-tower strike. The anomaly to strike duration about 3.7 seconds, and is shown as a red dashed line in Figure **??**. Any delay less than that duration is considered a successful anomaly detection and the earlier the better. As is apparent in the results, the 1PC model the superior option as it provides: 1) a shorter delay in anomaly detection, 2) a faster computation due to the fact that only a single PC needs to be predicted, compared to 5 PCs in the counterpart MPC model, as seen in Figure **??**, and 3) a higher~~ $\cancel{FI}$ ~~score across the sampling frequencies tested as illustrated in Figure **??**.~~

~~anomaly detection delay simulation to real-time ratio~~

~~FI scores Sensitivity study of sampling frequency on (a) anomaly detection delay, (b) simulation nto real-time ratio, and (3) FI score for both 1PC and MPC models.~~

6 ~~Discussion~~

Table 2. ~~1PC quality of anomaly prediction quantification~~Anomaly detection performance for models tested on datasets $\mathcal{D}_2$, $\mathcal{D}_2^{(a1)}$, $\mathcal{D}_2^{(a2)}$, $\mathcal{D}_2^{(a3)}$, and $\mathcal{D}_3$, under different detection criteria: ΔE, $\Delta E \vee E$, and $\Delta E \wedge E$.

Criterion	~~T^+~~ Dataset ~~Precision~~	~~F^+~~ 1PC		~~F^-~~ MPC		
		~~Recall~~ T^+	~~F1 Score~~ F^-	F^+	T^-	P
ΔE	~~AR~~ $\mathcal{D}_2$	~~33~~ 0	~~32~~ 0	~~4~~ 0	~~0.452~~ 1800	~~0.892~~ N/A
	~~AR~~ $\mathcal{D}_2^{(a1)}$	75	26	30	369	0.74
	$\mathcal{D}_2^{(a2)}$	92	9	36	363	0.91
	$\mathcal{D}_2^{(a3)}$	95	6	38	361	0.94
	~~0.762~~ $\mathcal{D}_3$	49	2	4	125	0.96
~~MPC quality of anomaly prediction quantification.~~ $\Delta E \vee E$	~~HR~~ $\mathcal{D}_2$	0	0	0	1800	N/A
	~~HR~~ $\mathcal{D}_2^{(a1)}$	~~0~~ 90	11	~~0~~ 30	369	0.89
	~~T^+~~ $\mathcal{D}_2^{(a2)}$	~~F^+~~ 95	~~F^-~~ 6	~~Precision~~ 36	~~Recall~~ 363	~~F1 Score~~ 0.94
	~~AR~~ $\mathcal{D}_2^{(a3)}$	~~25~~ 97	~~10~~ 4	~~12~~ 38	361	0.96
	~~HR~~ $\mathcal{D}_3$	51	0	~~14~~ 129	0	1.00
$\Delta E \wedge E$	~~AR~~ $\mathcal{D}_2$	~~19~~ 0	0	~~18~~ 0	1800	N/A
	~~AR~~ $\mathcal{D}_2^{(a1)}$	~~35~~ 65	~~10~~ 36	~~2~~ 21	378	0.64
	~~HR~~ $\mathcal{D}_2^{(a2)}$	~~0~~ 87	~~16~~ 17	28	371	0.83
	$\mathcal{D}_2^{(a3)}$	89	12	32	367	0.88
	$\mathcal{D}_3$	49	2	4	125	0.96

~~Results shows that the error and error derivative criteria ($\Delta E \& E$) the better option in terms of anomaly response time and accuracy of the response. Applying PCA on the data before training the model(s) helps reducing the problem size, which can be very helpful when there are many data channels that make it extremely laborious to do intensive search to detect an anomaly. When only considering the first principal component (1PC), the anomaly detection quality surpasses that of a multiple components model such as with the MPC model by 12% under the $\Delta E \& E$ anomaly criteria. Additionally, the 1PC model shows a faster response to anomaly and is computationally more efficient, rendering it the superior option between the two considered here. The 1PC response time can range from a mere 6% to 70% of the duration between the inception of the anomaly and the blade-tower strike based on the sampling frequency as shown in Figure ??. The MPC range under the same sampling frequency variations were between 89% and 225%. Which means that, if the sampling frequency is not small enough, the anomaly can have detrimental effects before it is even detected if multiple-PC models were being used for that detection. With the two models tested with healthy data under different operating conditions , the number of anomalies detected was small. If the $\Delta E \& E$ anomaly conditions were selected, both false positives and negatives counts for healthy data were zero.~~

~~As for the incident dataset ,~~ As for the incident dataset, $\mathcal{D}_3$, since the ~~experiment has also different operating conditions, the~~ turbine was operating under different conditions than the training dataset for $\mathcal{M}_1$ model, the error was high. Therefore, using the error derivative as an additional criterion helps with detecting the true positives. This is an important aspect of a good anomaly detection model because most of the time, the anomaly will most likely occur under conditions that have not been seen before.

It is eventually up to the developer/operator to gather certain amount of anomaly points before activating an alerting system or before acting upon it to limit disturbance to the main testing campaign. It could also be developed such that the model can trace back to which of the channels contributing to this anomaly based on the correlation matrix calculated during ~~the~~ PC analysis.

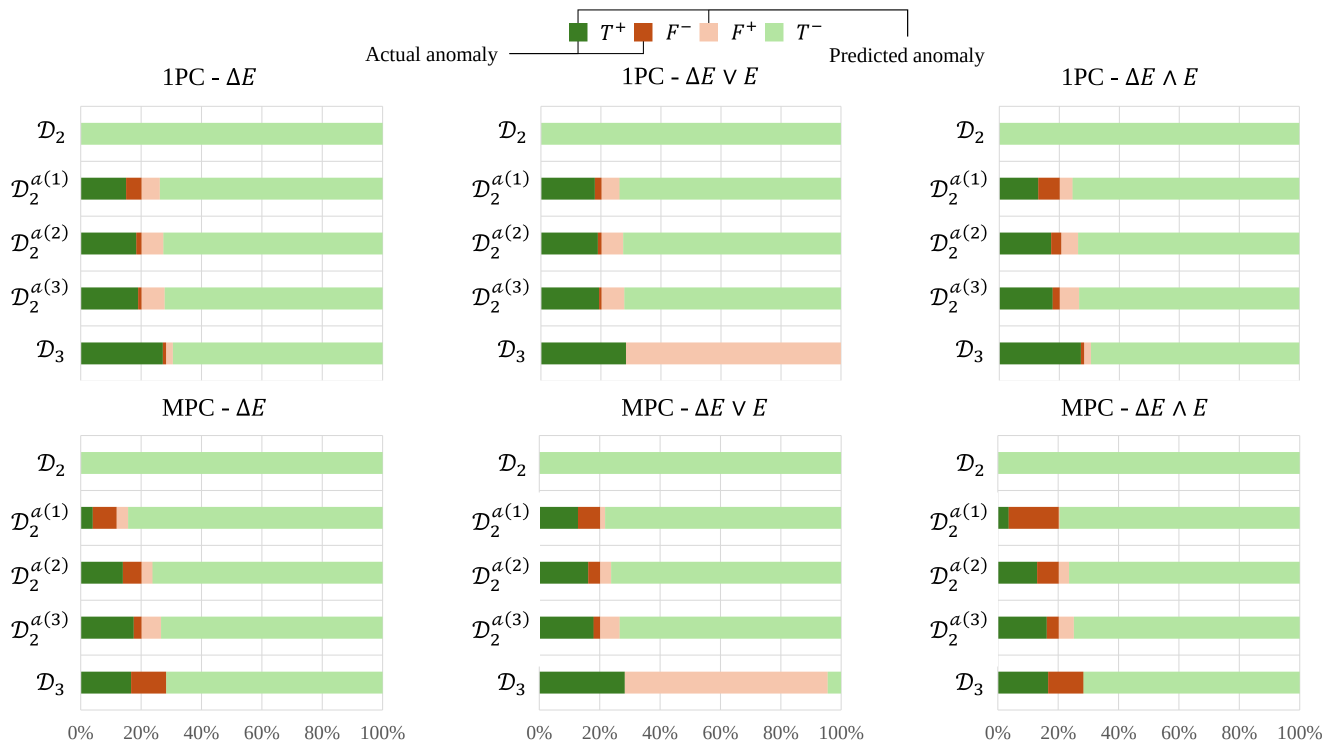


Figure 15. Percentage of true positives, false negatives, false positive, and true negatives occurrence when testing $\mathcal{M}_1$ model for the various testing datasets.

6 Conclusions

This ~~research highlights the potential for multi-variate long short term memory (MLSTM~~study demonstrates the feasibility and effectiveness of a multivariate long short-term memory (LSTM)-based ~~model to enhance the safety and reliability of~~

~~on- and offshore wind experimental campaigns at lab-scale, with the possibility of extension to ocean and commercial scale deployments. After proper data preprocessing, scaling, and their covariance computed to project onto principal components, the healthy data were trained and two criteria for anomaly detection to occur was studied~~ reconstruction model for early anomaly detection in scaled wind turbine experiments. By leveraging principal component projections of healthy operational data, the model enables robust monitoring through reconstruction error and its derivative. Various detection criteria were evaluated, including threshold exceedance of ~~error or error derivative or a combinationof both . When combining both criteria, the anomaly detection was observed to be more accurate and~~ the error E, its time derivative ΔE, and their logical combination. Two models were trained on distinct datasets, each evaluated using both single and multiple principal component (1PC and MPC) variations.

The results indicate that the criterion combining both error and error derivative ($\Delta E \wedge E$) yields the most consistent and accurate anomaly detection performance across test cases. The 1PC model offers superior responsiveness and recall, making it well-suited for identifying abrupt anomalies, though at the expense of precision. In contrast, the ~~anomaly response time short. The model with a single principal component showed to be the more superior option in terms of response time, computational time, and the prediction quality. When healthy data are tested, no erroneous anomalies were observed for the model that combined both criteria of error and its derivative. This work lays the foundation for future~~ MPC model exhibits greater robustness to false alarms due to its conservative aggregation of reconstruction error. Importantly, both model variations demonstrated zero false positives when applied to healthy test data, underscoring their reliability for real-time deployment.

This work serves as a proof of concept that ~~such easy to establish~~ simple, interpretable, and computationally efficient techniques can be ~~used as a safety precaution during future campaign testings to reduce human error and equipment malfunction in the laboratory, in particular when innovative technologies and control strategies are being tested~~deployed to enhance safety and operational awareness during laboratory-scale wind turbine testing. The approach holds promise for extension to ocean-based and full-scale wind energy systems, where early anomaly detection is critical for preventing equipment failure and improving system reliability during experimental campaigns and operational phases.

7 Competing interests

One co-author (Dr. Amir Nejad) is a member of the editorial board of Wind Energy Sicence Journal.

8 Author contributions

KH, RK, and BH were responsible for the acquisition of the financial award for the project alongside SG and RH from the University of Rhode Island and coordinated this research activity. AV and AN initiated a research exchange between the University of Maine and the Norwegian University of Science and Technology that equipped YA and IA with the methodologies and tools used in this research and ultimately led to this publication. BH and IA designed the experiments and IA and YA carried them out. IA managed data collection, while YA with vital input from AN and SG developed the models and performed the analysis. KH, AV, and RH supervised over the experiment and the analysis carried out. YA prepared the manuscript with the contribution from all co-authors.

Acknowledgements. The authors ~~would like to acknowledge the~~ gratefully acknowledge funding from the US Department of Energy ~~for funding this research under project name: Ocean_193360_1, project grant~~ - Office of Science, under grants #: DE-SC0022103 and # DE-SC0024295 (DOE-EPSCOR program), awarded to the University of Rhode Island and the University of Maine.

References

Wind/Wave Basin Verification of a Performance-Matched Scale-Model Wind Turbine on a Floating Offshore Wind Turbine Platform, vol. Volume 9B: Ocean Renewable Energy of *International Conference on Offshore Mechanics and Arctic Engineering*, https://doi.org/10.1115/OMAE2014-24166, 2014.

Alkarem, Y. R., Huguenard, K., Kimball, R. W., Hejrati, B., Goupee, A., Albertson, S., and Dahl, J.: Digital twin model development for mitigating floating offshore wind turbine motions due to wave actions, in: ISOPE International Ocean and Polar Engineering Conference, pp. ISOPE–I, 2023.

Alkarem, Y. R., Huguenard, K., Kimball, R. W., Hejrati, B., Ammerman, I., Nejad, A. R., and Grilli, S.: On Building Predictive Digital Twin Incorporating Wave Predicting Capabilities: Case Study on UMaine Experimental Campaign-FOCAL, in: Journal of Physics: Conference Series, vol. 2745, p. 012001, IOP Publishing, 2024.

Ammerman, I., Alkarem, Y., Kimball, R., Huguenard, K., and Hejrati, B.: Experimental validation of a Kalman observer using linearized OpenFAST and a fully instrumented 1:70 model, Wind Energy, 2024.

Association, E. W. E.: The economics of wind energy, EWEA, 2009.

Cao, Q., Chen, Y., Zhang, K., Zhang, X., Cheng, Z., and Wen, B.: Design approach of thrust-matched rotor for basin model tests of floating straight-bladed vertical axis wind turbines, Journal of Renewable and Sustainable Energy, 15, 063 308, https://doi.org/10.1063/5.0176064, 2023.

Dibaj, A., Nejad, A. R., and Gao, Z.: A data-driven approach for fault diagnosis of drivetrain system in a spar-type floating wind turbine based on the multi-point acceleration measurements, in: Journal of Physics: Conference Series, vol. 2265, p. 032096, IOP Publishing, 2022.

Dibaj, A., Valavi, M., and Nejad, A. R.: Unsupervised anomaly detection of permanent magnet offshore wind generators through electrical and electromagnetic measurements, Wind Energy Science Discussions, pp. 1–32, 2024.

Fowler, M., Lenfest, E., Viselli, A., Goupee, A., Kimball, R., Bergua, R., Wang, L., Zalkind, D., Wright, A., and Robertson, A.: Wind/Wave Testing of a 1:70-Scale Performance-Matched Model of the IEA Wind 15 MW Reference Wind Turbine with Real-Time ROSCO Control and Floating Feedback, Machines, 11, https://doi.org/10.3390/machines11090865, 2023.

ISO: Mechanical vibration-evaluation of machine vibration by measurements on non-rotating parts, ISO/IS, 10816, 1996.

Le, H. X. and Matunaga, S.: A residual based adaptive unscented Kalman filter for fault recovery in attitude determination system of microsatellites, Acta Astronautica, 105, 30–39, 2014.

Leahy, K., Hu, R. L., Konstantakopoulos, I. C., Spanos, C. J., and Agogino, A. M.: Diagnosing wind turbine faults using machine learning techniques applied to operational data, in: 2016 IEEE International Conference on Prognostics and Health Management (ICPHM), pp. 1–8, IEEE, 2016.

Lu, B., Dibaj, A., Gao, Z., Nejad, A. R., and Zhang, Y.: A class-imbalance-aware domain adaptation framework for fault diagnosis of wind turbine drive trains under different environmental conditions, Ocean Engineering, 296, 116 902, 2024.

Mehlan, F. C. and Nejad, A. R.: On the Uncertainty of Digital Twin Models for Load Monitoring and Fatigue Assessment in Wind Turbine Drivetrains, Wind Energy Science Discussions, pp. 1–28, 2024.

Miele, E. S., Bonacina, F., and Corsini, A.: Deep anomaly detection in horizontal axis wind turbines using graph convolutional autoencoders for multivariate time series, Energy and AI, 8, 100 145, 2022.

Nejad, A. R., Odgaard, P. F., and Moan, T.: Conceptual study of a gearbox fault detection method applied on a 5-MW spar-type floating wind turbine, Wind Energy, 21, 1064–1075, 2018.

Parker, A.: Development and Testing of a 1:70 Scale Model Wind Turbine of the IEA Reference 15 MW Floating Offshore System, Master's thesis, University of Maine, https://digitalcommons.library.umaine.edu/etd/3548, electronic Theses and Dissertations, No. 3548, 2022.

Peng, H., Li, S., Shangguan, L., Fan, Y., and Zhang, H.: Analysis of wind turbine equipment failure and intelligent operation and maintenance research, Sustainability, 15, 8333, 2023.

Soares-Ramos, E. P., de Oliveira-Assis, L., Sarrias-Mena, R., and Fernández-Ramírez, L. M.: Current status and future trends of offshore wind power in Europe, Energy, 202, 117 787, 2020.

Wang, Y., Yoshihashi, R., Kawakami, R., You, S., Harano, T., Ito, M., and Naemura, T.: Unsupervised anomaly detection with compact deep features for wind turbine blade images taken by a drone, IPSJ Transactions on Computer Vision and Applications, 11, 1–7, 2019.

Wang, Z., Guo, Y., and Wang, H.: Review on monitoring and operation-maintenance technology of far-reaching sea smart wind farms, Journal of Marine Science and Engineering, 10, 820, 2022.

Zhou, J. and Zhu, Y.: On the optimality of Kalman Filter for Fault Detection, arXiv preprint arXiv:2301.11573, 2023.