DTU Wind and Energy Systems

Data-driven models for wind farm operation and monitoring

Iason Tsaklis

DTU Wind-M-0978
July 2025

Author:
Iason Tsaklis

Title:
Data-driven models for wind farm operation and monitoring

DTU Wind-M-0978
July 2025

ECTS: 45

Education: Master of Science

Supervisor(s):

Tuhfe Göçmen
Vasilis Pettas
Moritz Johann Gräfe
DTU Wind & Energy Systems

Simon Watson
TU Delft

Remarks:
This report is submitted as partial fulfillment of the requirements for graduation in the above education at the Technical University of Denmark

DTU Wind & Energy Systems is a department of the Technical University of Denmark with a unique integration of research, education, innovation and public/private sector consulting in the field of wind energy. Our activities develop new opportunities and technology for the global and Danish exploitation of wind energy. Research focuses on key technical-scientific fields, which are central for the development, innovation and use of wind energy and provides the basis for advanced education at the education.

Technical University of Denmark Department of Wind Energy Frederiksborgvej 399
4000 Roskilde
Denmark
www.wind.dtu.dk

Data-driven models for wind farm operation and monitoring

by

Iason Tsaklis

to obtain the degrees of

Master of Science
in Wind Energy
at Technical University of Denmark

Master of Science
in Aerospace Engineering
at Delft University of Technology

To be defended on the 11th of August 2025

Supervisors:	Simon Watson	TU Delft
	Tuhfe Göçmen	DTU
	Vasilis Pettas	DTU
	Moritz Johann Gräfe	DTU
Defence committee:	Damiano Casalino	TU Delft
	Søren Juhl Andersen	DTU
Start date:	November 1, 2024	
End date:	July 31, 2025	
Student numbers:	TU Delft:	5951267
	DTU:	s234040

An electronic version of this thesis is available at
https://repository.tudelft.nl and https://findit.dtu.dk

DTU

EWEM

TU Delft

Acknowledgments

It was the 18th of August 2023 when I first arrived in Denmark, almost two years ago, yet it feels like yesterday. I was excited and a little bit scared, moving alone in a new country and trying to get the best out of my very exciting Master's. The journey that followed (although stressful and challenging at times) cannot be summarised in a few sentences; however, one thing is certain: It gave me a lot of lessons and helped me improve, both personally and professionally. Through many courses, an internship, and company visits, this thesis is the capstone of this 2-year journey. I would like to thank everyone who was involved and supported me in any way.

First and foremost, I would like to thank my supervisors. Vasilis, who devoted a lot of his precious time and energy to assist me and thoroughly examined every small detail, helped me to greatly improve the quality of this study. Moritz, with his crucial input and creative insights, helped me in steering the work in the right direction throughout the whole thesis period. Simon, with his extensive experience, always reminded me to see the big picture and not lose sight of the broader goal. Tuhfe, with her expertise in machine learning, helped me dive into the marvelous world of machine learning models and had crucial answers to my questions when needed.

Next, I would like to thank my friends and fellow students of the EWEM programme. Sharing this intense yet rewarding experience across two countries in two years with such a special group of people made the journey not only manageable but truly memorable. Together, we built friendships that I will cherish for years to come.

I would also like to thank my girlfriend, Elli. Her unwavering support made tough moments much lighter, even at times when I thought I was overwhelmed.

Lastly, I owe the deepest gratitude to my parents, Panayota and Antonis. Your love and belief in me, through every step of this journey and beyond, are the foundation upon which everything else is built.

I sincerely hope that this thesis will contribute, even in a small way, to the global energy transition and that it will support the broader mission of making wind energy more accessible, sustainable, and impactful for our planet.

Kgs. Lyngby, July 2025

Iason Tsaklis

"The answer, my friend, is blowin' in the wind..."
- Bob Dylan

Abstract

This thesis investigates the potential of data-driven modelling to enhance wind farm monitoring, control, and operations and maintenance (O&M) strategies by leveraging the extensive SCADA data generated by offshore wind farms. Focusing on the Lillgrund offshore wind farm, the study develops regression-based surrogate models using XGBoost, artificial neural networks, and Gaussian process regression to predict two critical targets: the blade root flapwise and tower bottom fore-aft damage equivalent loads. All models are found to effectively capture the underlying patterns from the input features. Among them, XGBoost consistently outperforms the others in terms of prediction accuracy, computational efficiency, and robustness. Its superior performance is further validated across multiple preprocessing settings and operational scenarios, including its capacity to generalise across turbines and exploit spatial information. Finally, the applicability of the models is demonstrated through a simplified use case that estimates fatigue damage over a specific period. The findings underline the value of integrating machine learning-based surrogate models into operational workflows to reduce O&M costs and support decision-making in wind farm management.

Contents

List of Figures

List of Tables

Nomenclature

List of Abbreviations

AEP Annual Energy Production

ANN Artificial Neural Network

APE Absolute Percentage Error

BNN Bayesian Neural Network

CFD Computational Fluid Dynamics

DBSCAN Density-Based Spatial Clustering of Applications with Noise

DEL Damage Equivalent Load

FA Fore-aft

GMM Gaussian Mixture Model

GP Gaussian Process

GPR Gaussian Process Regression

HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise

IQR Interquartile Range

KDE Kernel Density Estimation

LR Linear Regression

LUT Look-up Table

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

MSE Mean Squared Error

NaN Not a Number

NRMSE Normalised Root Mean Square Error

OWT Offshore Wind Turbine

PCE Polynomial Chaos Expansion

PE Prediction Error

QRS Quadratic Response Surface

ReLU Rectified Linear Unit

RFC Random Forest Classifier

RMSE Root Mean Square Error

RSM Response Surface Methodology

SCADA Supervisory Control and Data Acquisition

SHAP SHapley Additive exPlanations

SVR Support Vector Regression

TI Turbulence Intensity

WFFC Wind Farm Flow Control

XGBoost eXtreme Gradient Boosting

List of Symbols

ω_{gen} Generator speed

θ_p Pitch angle

θ_{nac} Nacelle position

P_{active} Active power

R^2 Coefficient of determination

U Wind speed

CHAPTER 1

Introduction

Wind Power has been proven to be a sustainable, efficient, and profitable renewable energy source with a significantly smaller environmental impact compared to fossil fuels. The advancements in wind turbine technology, as well as the global shift towards renewable energy sources, have propelled the rapid development of wind energy in the past 25 years, rendering it one of the top five electricity production sources globally (see Fig. 1.1).

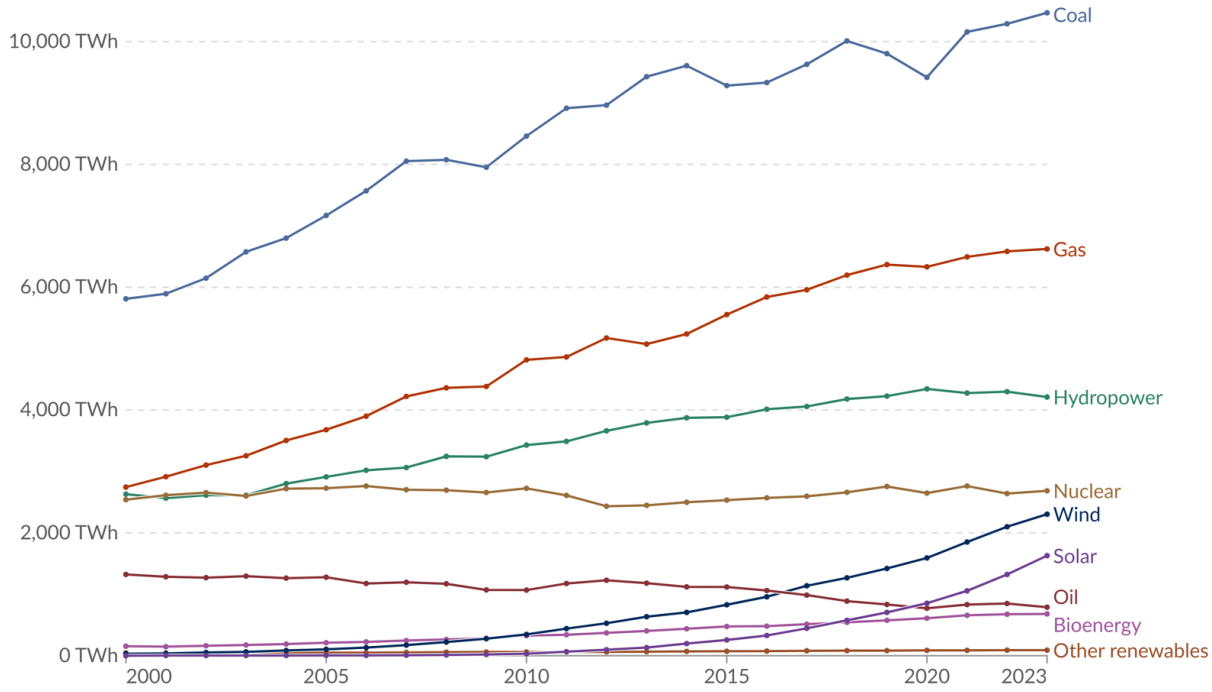


Figure 1.1: Electricity production by source, World. [1]

Denmark has been a leader in the advancement of wind energy, generating 19.41 TWh of electricity from wind in 2023, almost three times more than the second-largest electricity source (Bioenergy) [1].

In recent years, emerging trends have started to jeopardise the momentum that has been steadily built over decades. A combination of macroeconomic challenges and increasing costs of commodities and capital has led to a rise in the overall expense of wind energy projects, challenging market growth. Despite these challenges, 2024 marked yet another record year for wind energy with 117 GW of new installations worldwide [2]. This growth was mainly driven by onshore wind farms(109 GW), which present a well-established and mature technology supported by a robust global supply chain.

However, given the constraints of limited land availability, the continued expansion of onshore wind energy is expected to encounter significant obstacles in the future. Therefore,

it is essential to strongly promote the development of offshore wind energy to support the achievement of the European Union's carbon neutrality target by 2050. Nevertheless, the aforementioned market adversity in combination with immature infrastructure and a troublesome supply chain led to a decline in the installation of new offshore wind farms by 26% in 2024 compared to 2023 [2].

Thus, it is evident that reducing the operational costs, and, by extension, the levelized cost of energy (LCOE) of offshore wind farms is essential to ensure the long-term viability of the market, support its continued growth, and enable meaningful contributions to global sustainability objectives.

1.1 Topic description

This project aims to reduce the high operational costs of offshore wind farms by harnessing the extensive data they generate. The research focuses on developing data-driven methods to improve efficiency, including surrogate models that estimate key statistical properties regarding turbine response, such as damage equivalent loads (DEL), based on data from all turbines within the farm. Various machine-learning techniques are trained to predict these target values using operational data from the entire wind farm. Key learning outcomes include proficiency in large-scale data processing, feature selection, hyperparameter tuning, and quantitative model performance evaluation.

The work makes use of an existing dataset from the Lillgrund offshore wind farm [3] provided by Vattenfall as part of the TWAIN EU project [4], aiming to reveal hidden patterns that can optimise wind farm monitoring, control strategies, and operations and maintenance (O&M) planning. Through exploratory analyses, this thesis will contribute to more cost-effective and efficient offshore wind farm operations.

1.2 Aim of the thesis & Relevance

In real-world engineering, dealing with expensive and delicate machinery often makes it challenging or prohibitively costly to measure certain parameters that provide insight into performance or structural safety. This has driven significant interest in surrogate models in recent years. Surrogate models are approximate mathematical models that are designed to efficiently and cost-effectively replicate the behaviour of a system with an adequate degree of accuracy, offering an estimation of characteristics that define the wind farm performance.

Accurately predicting the DEL on wind turbine blades or towers through surrogate models, as opposed to direct measurement via strain gauges, offers the potential to significantly reduce operational costs. This is achieved by enabling predictive maintenance, facilitating lifetime extension assessments, and supporting the optimisation of control strategies. Accordingly, the objective of this thesis is to develop surrogate models capable of estimating DEL based on various measurement data (SCADA). The approach is based on regression analysis, which involves modelling the relationship between multiple input variables (SCADA data) and a single target variable (DEL) at corresponding time instances. A flowchart of this process is depicted in Fig. 1.2.

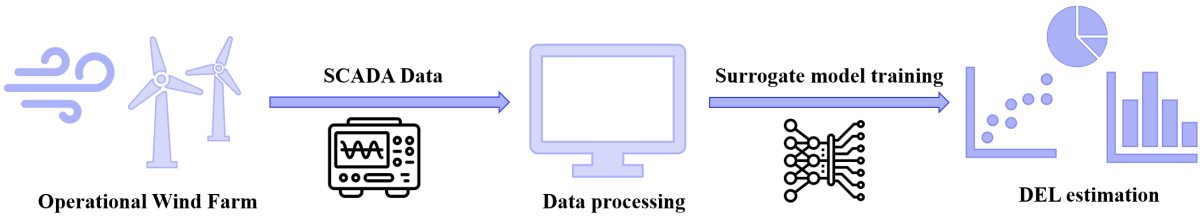


Figure 1.2: Schematic of the DEL estimation process.

Although there has been extensive research on simulated wind turbine data in the past years, applying a surrogate model to estimate key properties, such as the DEL, based on real SCADA data has not been thoroughly investigated. Thus, it would be of great interest to explore if such an application is possible and which regression methodology is the most suitable.

CHAPTER 2

Background & state-of-the-art

This chapter provides an overview of the field and presents state-of-the-art research related to surrogate modelling and regression analysis applied to wind farm data. In Chapter 1, a brief review of the current state of the wind energy sector was presented, followed by an introduction to the research topic, including the project's objectives and its broader significance. The purpose of this chapter is to establish the theoretical background and review relevant literature, thereby reinforcing the motivation and relevance of the study. Finally, the research questions addressed in this thesis are outlined.

2.1 Damage equivalent load

The assessment of the structural lifetime of a wind turbine is a critical aspect of both the design and operational phases. A key challenge in this process lies in quantifying the fatigue damage experienced by turbine components due to stochastic and highly variable loading conditions. Fatigue damage results from cyclic and varying stresses over time, which, although often below the material's ultimate strength, can lead to progressive degradation and eventual failure.

To evaluate fatigue damage, high-frequency load measurements, often from blade root moments or tower base moments, are analysed using dedicated time-series analysis techniques. The rainflow counting algorithm proposed by Endo T. & Matsuishi M. [5] is the most common method for extracting stress cycles from such load time-series. It identifies and classifies individual load cycles based on their amplitude, thereby enabling the transformation of a complex load history into a set of discrete load ranges.

From the rainflow matrix, the Damage Equivalent Load (DEL) is derived. The DEL is defined as a constant-amplitude load that, applied for a fixed number of cycles, would produce the same fatigue damage as the actual variable load history. It serves as a scalar representation of fatigue severity, facilitating load comparisons across different operational periods or turbine configurations. The DEL for a given load time series is calculated as:

$$\mathrm{DEL} = \left(\frac{\sum_{i=1}^{n} n_i S_i^m}{N_{\mathrm{ref}}} \right)^{1/m} \tag{2.1}$$

Where:

- S_i is the load range of the i^{th} cycle.
- n_i is the number of occurrences of that cycle.
- m is the Wöhler exponent (dependent on the component material).
- N_{ref} is the number of reference cycles.

To estimate the remaining useful life of a component, the Palmgren-Miner rule [6] is applied. This linear damage accumulation model states that the total damage D is given by:

$$D = \sum_{i=1}^{n} \frac{n_i}{N_i} = \frac{1}{K} \sum_{i=1}^{n} n_i S_i^m \tag{2.2}$$

Where N_i is the number of cycles to failure at load range S_i, and K is a fatigue strength coefficient related to the material. Failure is predicted to occur when $D \geq 1$.

Consequently, the DEL plays a pivotal role in fatigue analysis and structural health monitoring. It enables engineers to condense a complex load history into a single representative metric, which can be used for design certification, control optimisation, and post-construction evaluations. Notably, DEL is instrumental not only in the design phase of new turbines but also in lifetime extension studies of existing wind farms. In such studies, historical SCADA or load data is analysed to determine whether the accumulated fatigue damage allows for safe operation beyond the original design life.

In the present thesis, DEL serves as the primary target variable for predictive modelling. Specifically, the flapwise blade root moment and tower fore-aft moment DELs are estimated using SCADA data and machine learning techniques, enabling scalable and non-intrusive lifetime assessment frameworks.

The fatigue damage calculation is explained more thoroughly in [7] & [8]. A flowchart which illustrates the DEL computation is presented in Fig. 2.1.

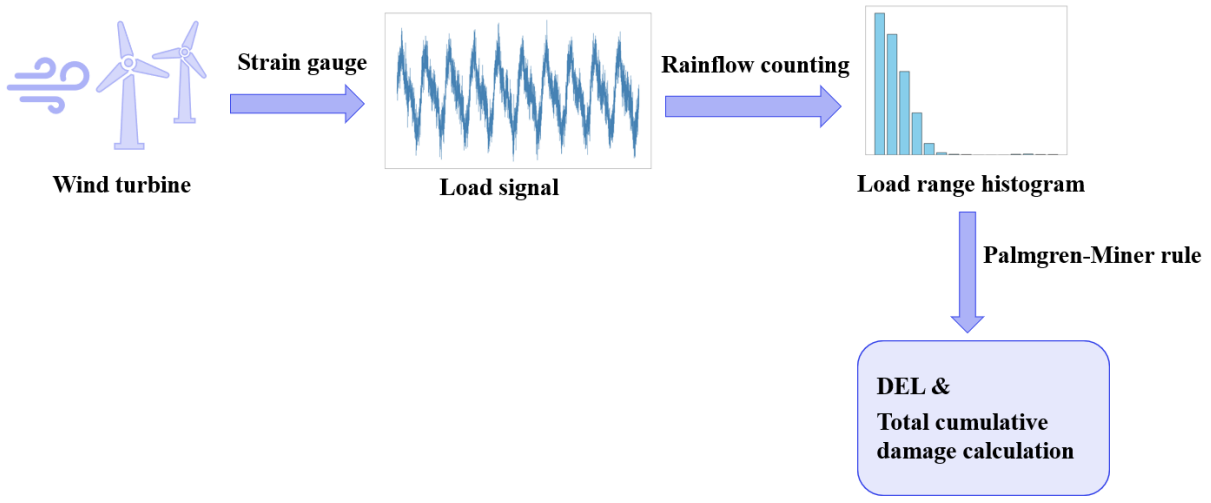


Figure 2.1: Schematic of the DEL calculation process.

2.2 Literature review

In this section, a discussion of the state-of-the-art is conducted in the field of surrogate modelling within the context of wind energy. Several peer-reviewed articles and scientific reports were examined, and the most critical findings are presented with respect to three main aspects: the objectives of the models, the modelling approaches employed, and the type of data utilised. The reviewed literature demonstrates that surrogate models have been widely adopted in wind energy applications such as wind farm layout optimisation, wake characterisation, performance monitoring, fault detection, and fatigue load estimation.

Surrogate models have shown particular promise in the field of wind turbine performance assessment and fault detection. Lyons and Göçmen [9] proposed a machine learning-based methodology that leverages SCADA data from individual turbines and their neighbours to identify performance anomalies at the wind farm scale. Their approach enables timely maintenance interventions and improves energy trading strategies through more accurate power output predictions. The methodology was validated using two days of SCADA data from the Horns Rev I offshore wind farm. Similarly, Papatheou et al. [10] utilised SCADA data from the Lillgrund wind farm to monitor significant performance events. By applying artificial neural networks (ANNs) and Gaussian Process Regression (GPR), they constructed individualised reference power curves for each turbine and successfully detected deviations using simple control charts.

In the context of fatigue load estimation, Hlaing et al. [11] investigate the use of probabilistic deep learning models, such as Bayesian neural networks (BNN), Monte Carlo dropout, and deep ensembles, for virtual load monitoring in offshore wind farms. Their approach combines SCADA and accelerometer data to predict damage equivalent loads while also quantifying predictive uncertainty. Using a two-year dataset from five turbines in a Belgian offshore wind farm, the models were trained on one year and tested on a second with different operational conditions. The study highlights the ability of these models to generalise across distributions and to signal high uncertainty under out-of-distribution scenarios. Notably, deep ensembles demonstrated the best balance between predictive accuracy and uncertainty calibration, supporting the practical value of uncertainty-aware models in industrial asset management.

Noppe et al. [12] present a virtual sensing framework for estimating the full structural load on offshore wind turbines (OWTs), combining SCADA data with accelerometer and strain measurements. Their method separates the loading contributions into quasi-static loads, predicted from SCADA-based thrust estimation models, and dynamic loads, derived using a modal decomposition and expansion technique applied to tower-mounted accelerometer data. The combination of these two components enables the reconstruction of the full strain history at fatigue-critical locations, even those below sea level and otherwise inaccessible, such as near the mudline of monopile foundations. The concept is validated on real data from a Belgian offshore wind farm, demonstrating good agreement between predicted and measured strains over a short validation period. This approach illustrates the potential of hybrid data-driven and physics-informed models for lifetime assessment and structural health monitoring in existing offshore wind assets.

Complementing this line of work, Weijtens et al. [13] propose a fleet-leader-based strategy to enable farm-wide fatigue assessment using in-situ measurements from a limited number

of instrumented turbines. By equipping select turbines in the Northwind offshore wind farm with strain gauges and accelerometers, and calculating damage equivalent loads at 10-minute resolution, the authors demonstrate that turbine-level fatigue progression can be estimated and extrapolated to non-instrumented units. The study underscores the importance of accounting for turbulence and site-specific dynamics, such as wave loading, and concludes that while the fleet-leader concept is promising, further refinement of empirical extrapolation techniques is required. This approach highlights a scalable direction for integrating physical measurements and surrogate models into long-term structural health monitoring and life extension strategies.

Additionally, several studies have highlighted the superiority of probabilistic machine learning models over conventional lookup-based approaches. For instance, a recent study by Miao et al. [14] proposed a Gaussian Process (GP) regression model to predict damage equivalent loads under yaw-misaligned conditions. Compared to traditional look-up table (LUT) methods, the GP approach achieved substantial improvements in accuracy (up to 51.87% RMSE reduction for flapwise blade root moments) due to its ability to capture the highly nonlinear relationship between yaw misalignment and fatigue loads. The model was trained using simulated data from wind farm scenarios that explicitly accounted for wake interactions, underlining the importance of including yaw-induced flow effects in DEL prediction.

Artificial Neural Networks remain a popular choice for fatigue load modelling. Schröder et al. [15] showed that a feedforward neural network with two hidden layers could estimate flapwise blade root DELs more accurately and efficiently than a polynomial chaos expansion (PCE) model using data from HAWC2 [16] simulations.

However, alternative algorithms have emerged that offer significant improvements in both accuracy and computational efficiency regarding wake modelling and power prediction tasks. For example, Extreme Gradient Boosting (XGBoost) has demonstrated superior performance over ANNs in such functions. In a study by Nakhchi et al. [17], XGBoost not only yielded lower mean deviations in power prediction (0.94% vs. 2.15% for ANN), but also achieved drastically faster training and inference times; 87.5% and 98% faster, respectively. These advantages make XGBoost highly attractive for real-time applications in yaw-controlled wind farms, where computational speed is critical.

The scalability of XGBoost is further confirmed in a comparative study by Purohit et al. [18], which evaluated Support Vector Regression, ANN, and XGBoost in predicting wake velocity and turbulence intensity. While all models achieved CFD-level precision, XGBoost distinguished itself as the most efficient and scalable, particularly suitable for large datasets. Although the study was limited to standalone turbines with uniform inflow and did not address wake interactions, it reinforces the model's practical relevance in surrogate modelling pipelines.

A direct comparison of surrogate model types specifically for DEL and power prediction is offered by Gasparis et al. [19], who evaluated Linear Regression, ANN, and Gaussian Process Regression. Their study showed that GPR consistently outperforms the other two models in predicting both short-term damage equivalent loads and electrical power. This reinforces the growing consensus around the effectiveness of GPR in fatigue-related modelling tasks due to its ability to model uncertainty and capture complex nonlinear relationships.

Alternative modelling strategies, particularly polynomial-based surrogates, have also been explored. Dimitrov et al. [20] and Murcia et al. [21] investigated polynomial regression approaches on synthetic datasets to estimate key statistical quantities such as fatigue damage and uncertainty propagation. The latter study employed Monte Carlo simulations to construct polynomial response surfaces, offering computational efficiency at the expense of reduced accuracy and flexibility compared to machine learning methods.

The applicability of the Response Surface Methodology (RSM) to structural load estimation was examined by Toft et al. [22]. Their results show that while RSMs can accurately estimate fatigue and ultimate loads for site-specific wind turbine configurations, the model uncertainty is notably higher when applied to ultimate load calculations. This underscores the need for caution when deploying surrogate models in extreme load estimation scenarios.

Lastly, a comprehensive review by Kusiak et al. [23] provides an extensive overview of recent advances in data-driven modelling for wind power systems, covering aspects such as operation, condition monitoring, and fault detection. Their work emphasises the increasing role of intelligent models in optimising wind farm operation, improving reliability, and reducing downtime.

Looking ahead, data-driven Wind Farm Flow Control (WFFC) emerges as a promising frontier in wind energy research. Göçmen et al. [24] provide a comprehensive review of this rapidly evolving field, highlighting the integration of surrogate models, reinforcement learning, and advanced analytics to improve wind farm adaptability, energy yield, and load management. The study emphasises the pivotal role of surrogate modelling in bridging the gap between high-fidelity simulations and real-time operational control. While the potential of data-driven WFFC is significant, the authors identify key challenges related to data quality, cybersecurity, and algorithm robustness that must be addressed to enable field-scale implementation. Overall, the review underscores that the synergy between data science and wind farm control not only enhances current methodologies but also paves the way toward more intelligent, resilient, and sustainable wind energy systems.

2.3 Summary and thesis motivation

Collectively, these studies underscore the growing relevance of surrogate modelling techniques in wind energy research and practice. The most relevant contributions from the reviewed literature are summarised in Table 2.1.

Table 2.1: Literature review summary of surrogate modelling techniques for estimating key performance indicators of wind turbines.

Author(s)	Data	Method	Prediction Target	Main Finding
Lyons & Göçmen [9]	SCADA	ANN	Power	• Enables timely maintenance and optimised energy trading. • Local SCADA input (neighbouring turbines) outperforms single-turbine and full-farm data.
Papatheou et al. [10]	SCADA	ANN, GPR	Power	• Possible to monitor significant events affecting turbine performance. • ANN and GPR showed similar power prediction performance.
Hlaing et al. [11]	SCADA	BNN, Deep Ensembles	DEL	• Supports the use of probabilistic deep learning models for industrial asset management.
Noppe et al. [12]	SCADA	ANN	Loads	• Demonstrates the potential of hybrid data-driven and physics-informed models for lifetime assessment and structural monitoring.
Miao et al. [14]	Generated	GPR	DEL	• GPR significantly outperforms traditional LUT methods.
Schröder et al. [15]	Generated	ANN, PCE	DEL	• ANN predicts blade DELs faster and more accurately than PCE.
Nakhchi et al. [17]	Generated	XGBoost, ANN	Power	• XGBoost is more accurate and faster in yaw-misaligned wind farm predictions.
Purohit et al. [18]	Generated	XGBoost, ANN, SVR	Wake velocity & TI	• Similar performance across models. • XGBoost is efficient and scalable for large datasets.
Gasparis et al. [19]	Generated	ANN, GPR, LR	DEL, Power	• GPR outperforms ANN and linear regression models.
Dimitrov et al. [20]	Generated	PCE, GPR, QRS	DEL	• GPR outperforms other surrogate models.
Murcia et al. [21]	Generated	PCE	Power, DEL	• Polynomial response surfaces can capture global turbine behaviour.

The overview presented in Table 2.1 reveals several noteworthy trends and gaps in the current literature on surrogate modelling in wind energy. First, it becomes evident that the majority of existing studies rely on synthetic datasets derived from low or high-fidelity simulations. While these methods offer controlled conditions for model development, they fail to capture the complexity and variability inherent in real-world turbine operation. Consequently, there remains a distinct need for studies applying surrogate models directly to SCADA data, which are more representative of operational uncertainties and real turbine behaviour.

Another key observation is the underexplored potential of XGBoost in the context of fatigue load estimation. Although XGBoost has demonstrated superior accuracy and computational efficiency in power and wake velocity prediction tasks, its application to DEL prediction remains unexamined. This creates an opportunity to assess its performance in a domain where fast inference and robust pattern recognition are especially valuable, particularly for real-time monitoring or large-scale deployment. Moreover, while power prediction and AEP optimisation dominate the focus of many surrogate modelling studies, DEL estimation is comparatively less investigated. Given the importance of DEL in turbine lifetime extension studies and predictive maintenance strategies, this gap highlights an important opportunity for further research. A model capable of accurately estimating DEL using only SCADA data would provide significant value for asset management and structural integrity assessments.

Lastly, the table shows that ANNs are widely adopted due to their flexibility and general applicability across tasks, whereas GPR consistently performs well but tends to incur higher computational costs. Polynomial-based methods seem effective, though their effectiveness must still be evaluated using real SCADA data. Evaluating the performance of these diverse methods (ANN, GPR, XGBoost, and Polynomial regression) on real SCADA data thus offers an insightful comparison of model accuracy, complexity, and generalizability under practical conditions.

2.4 Research questions

Building on the preceding discussion, this thesis addresses the key research gaps identified in the literature review, particularly those related to surrogate modelling using real SCADA data and the estimation of fatigue loads. In this context, the project aims to investigate these gaps by formulating and answering the following research questions:

1. What is the most suitable data-driven modelling approach for accurately predicting damage equivalent loads using SCADA data from operational wind turbines?
2. How do different data characteristics, such as input feature composition, data volume, and signal pre-processing, affect the performance and generalizability of surrogate models for DEL prediction?
3. To what extent can spatial information from neighbouring wind turbines enhance the accuracy of surrogate models across a wind farm?

These research questions are evaluated based on algorithms that are described in Chapter 3 and the results are reported and discussed in Chapter 4.

CHAPTER 3

Data & Methodology

In this chapter, a comprehensive analysis of the dataset and the modelling approach is presented. The chapter is structured into three main sections. The first section, Dataset, introduces the SCADA-based data used for training, validation, and testing of the models, including a detailed description of the input and target variables. The second section, Pre-processing, outlines the steps taken to clean, scale, and split the data to ensure it is suitable for modelling. This includes the application of scaling techniques, removal of outliers, and dataset splitting methods. The third section, Models, describes the machine learning algorithms employed (XGBoost, ANN and GPR) along with the configurations, architectures, and hyperparameter tuning strategies adopted for each. Additionally, a simpler Polynomial Regression model is tested to provide a baseline reference for evaluating the performance of more advanced techniques. Finally, the evaluation framework is presented, including the metrics and plots used to assess model performance and error behaviour. This chapter forms the methodological foundation for the subsequent analysis in Chapter 4.

3.1 Dataset

The dataset used in this study originates from the Lillgrund offshore wind farm, located off the southern coast of Sweden. Commissioned in 2007, Lillgrund comprises 48 Siemens wind turbines, each with a rated power capacity of 2.3 MW, resulting in a total installed capacity of 110 MW. As one of Sweden's earliest large-scale offshore wind projects, it represents a mature and operationally stable wind farm environment. The site provides a valuable real-world dataset for the development and evaluation of data-driven models aimed at understanding and predicting structural loads and performance metrics in offshore wind energy systems. A photograph of Lillgrund wind farm with the prestigious Øresund bridge in the back is shown in Fig. 3.1.

Figure 3.1: Lillgrund wind farm, picture from Vattenfall.

The geographical location of the wind farm at a macro scale, along with the detailed layout of its wind turbines, is illustrated in Fig. 3.2.

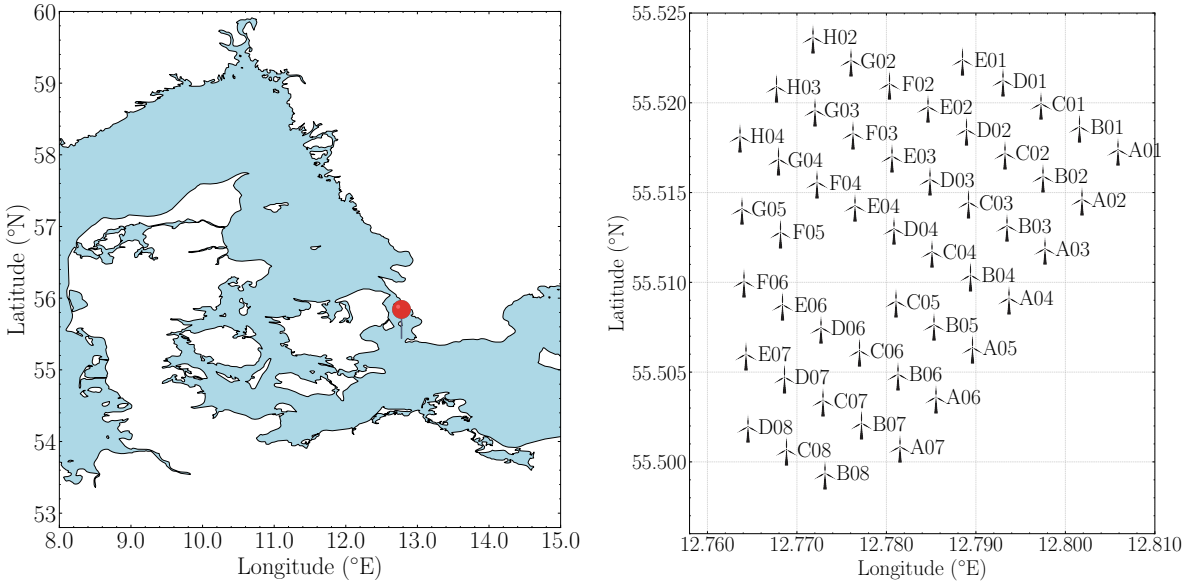


Figure 3.2: Lillgrund wind farm on the map (left) and wind turbine layout (right).

As one can see in Fig. 3.2, Lillgrund wind farm consists of a near-rectangular layout with eight rows and up to eight turbines per row. Among the 48 turbines, six are equipped with load sensors, specifically turbines B06, B07, B08, C08, D07, and D08. These are referred to as the load turbines and provide access to high-frequency structural load measurements. The available load signals include the blade root flapwise bending moment and the tower fore-aft and side-side bending moments. The dataset spans three years from 10 November 2019 to 13 November 2022 and is composed of 10-minute statistical summaries. Each data record contains a rich set of SCADA channels, including time stamps, wind speed, active power output, nacelle orientation (Nacelle Position), blade pitch angle, generator rotational speed (RPM), structural loads, and various other operational and environmental parameters.

Due to confidentiality agreements with the stakeholders, Vattenfall and Siemens, operational data related to wind turbine power and loads cannot be disclosed. Consequently, such data are presented solely in normalised form in the plots deemed essential for the analysis in this study.

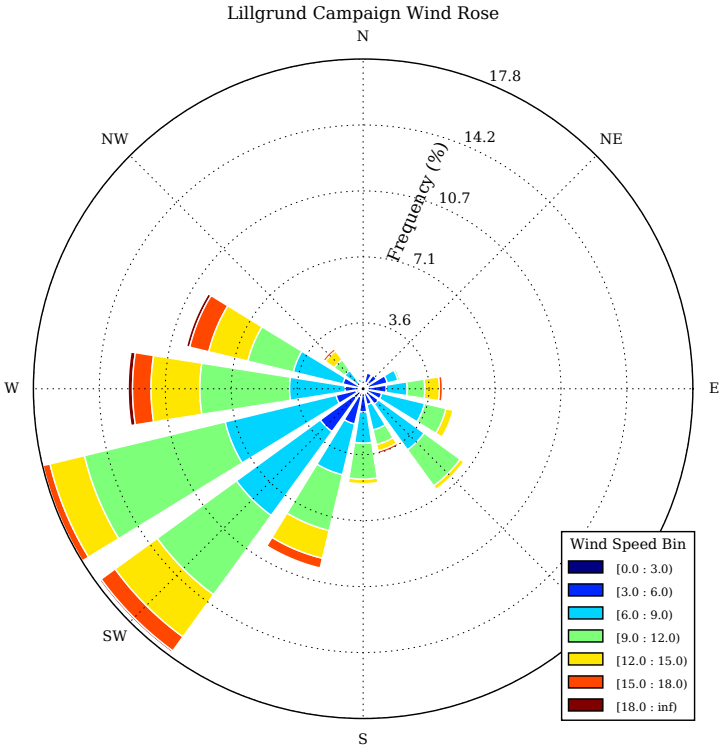


Figure 3.3: Lillgrund wind rose at hub height [25].

The wind rose presented in Fig. 3.3 illustrates the wind direction distribution, and WSW and SW as the dominant wind directions.

Having introduced the general characteristics of the Lillgrund offshore wind farm, the focus now shifts to evaluating the dataset itself. A range of statistical and visual assessments is conducted to gain insight into the quality and behaviour of the recorded signals. These include the computation of basic descriptive statistics such as the mean and standard deviation across all turbines, as well as various visualisations, including signal distributions and their relationship with wind speed. For brevity, this section presents a selection of plots for key variables that serve as indicators of data quality and consistency. Specifically, the plots against wind speed are shown for turbine B07 only, due to space limitations. These include variables such as active power, pitch angle, and structural loads. Additionally, histograms of the target variables, namely, the blade root flapwise DEL and the tower fore-aft moment DEL, are included to illustrate their distribution within the dataset.

The distribution of these variables as a function of wind speed is presented in Fig. 3.4, Fig. 3.5 & Fig. 3.6. All plots are based on SCADA 10-minute mean signals from turbine B07, with the exception of the DEL plots, which are derived from high-frequency load measurements processed into equivalent fatigue loads over each 10-minute interval. These visualisations offer a first indication of the consistency and physical plausibility of the signals across the wind speed range and serve as a basis for the subsequent modelling efforts.

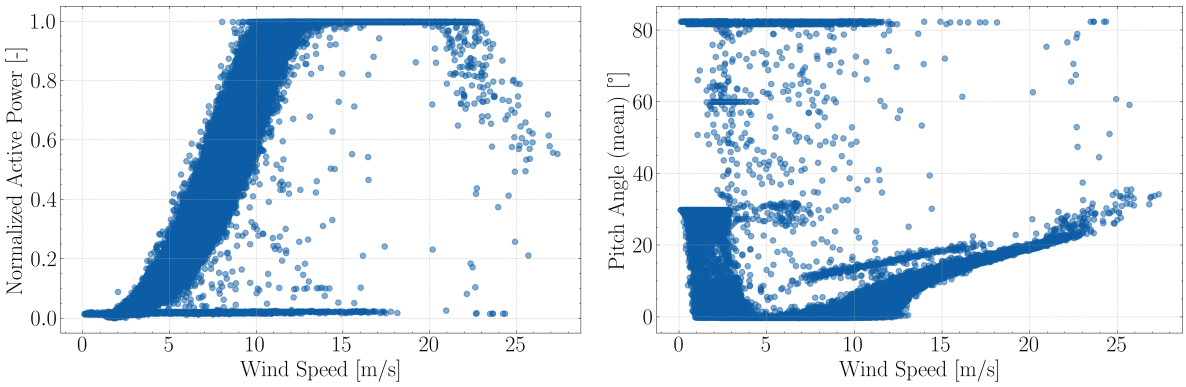


Figure 3.4: Mean active power (left) and mean pitch angle (right) against wind speed for B07.

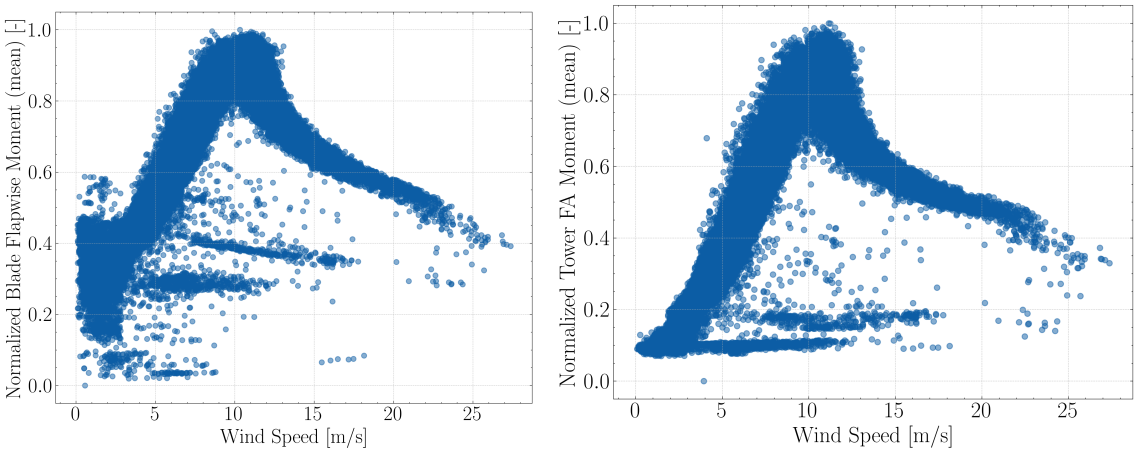


Figure 3.5: Mean blade root flapwise moment (left) and mean tower fore-aft moment (right) against wind speed for B07.

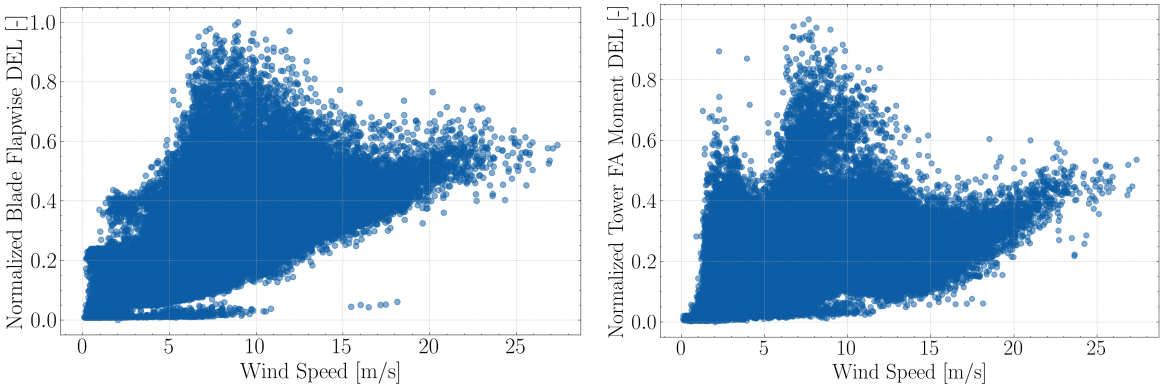
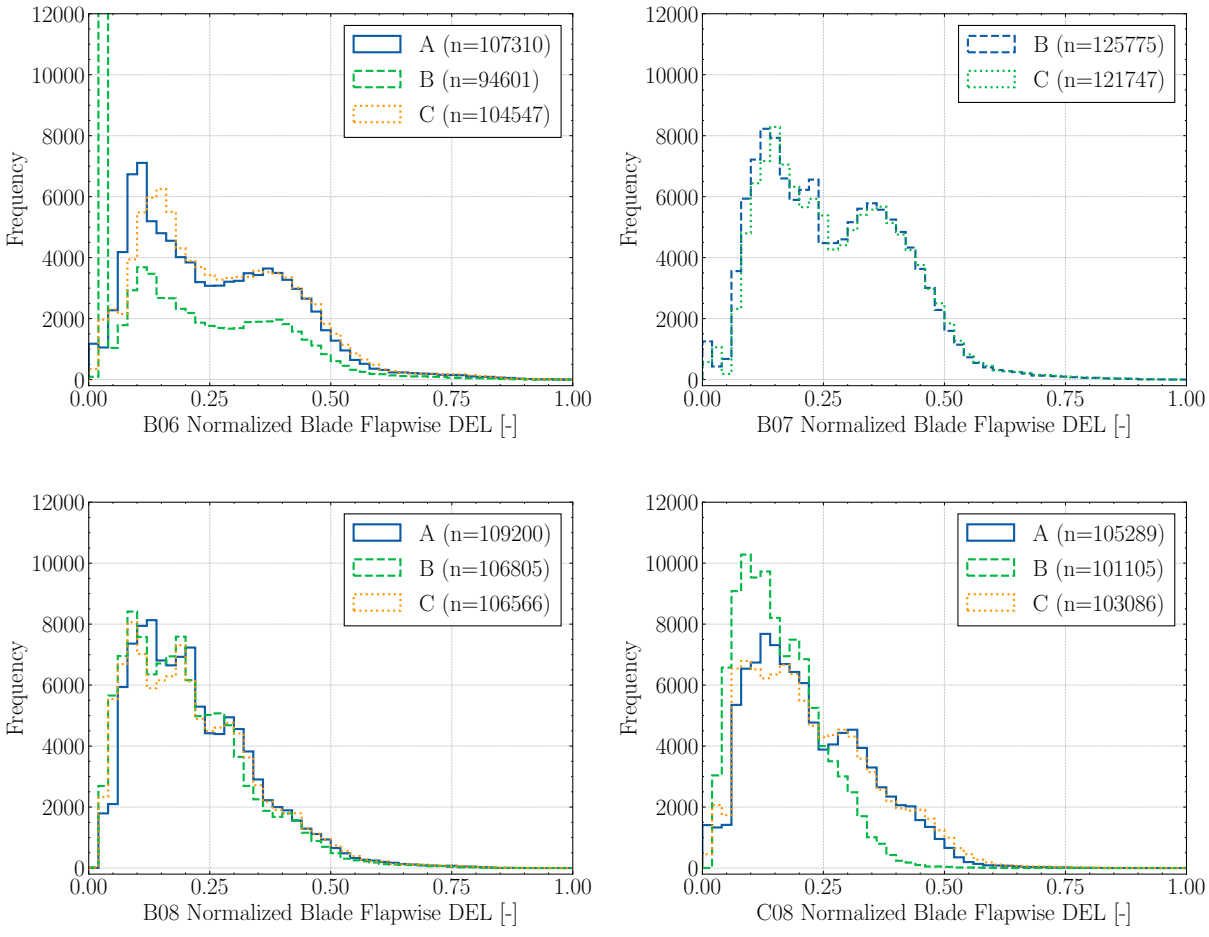


Figure 3.6: Blade root flapwise moment DEL (left) and tower fore-aft moment DEL (right) against wind speed for B07.

The plots in Fig. 3.4, Fig. 3.5 & Fig. 3.6 provide a visual overview of key operational and structural load variables as a function of wind speed for turbine B07. The active power curve follows the expected theoretical trend, increasing with wind speed until rated conditions are reached and then plateauing. However, several points deviate from the typical

power curve, including instances of zero power production at wind speeds above cut-in, and below-rated power outputs at wind speeds exceeding 20 m/s. These outliers likely correspond to periods of curtailment, implemented for load reduction or other operational constraints, and are further examined in the following section. The pitch angle curve similarly shows the anticipated trend of increasing blade pitch beyond rated wind speed. A small number of anomalous points, possibly associated with manual shutdowns or experimental campaigns, are also observed. The mean structural load trends, specifically the blade root flapwise moment and the tower fore-aft moment, are consistent with physical expectations. Both quantities increase with wind speed up to the rated region (approximately 11 m/s) and then gradually decline as pitch control reduces aerodynamic loading. Again, scattered deviations from this behaviour suggest occasional divergence from normal operation. Finally, the DELs for both the blade root flapwise and the tower fore-aft moments show a clear dependence on wind speed, generally increasing across the range. A pronounced peak appears in the rated wind speed region, consistent with elevated unsteady loading due to the control region shift. Additionally, for the tower fore-aft DEL, a secondary peak is visible in the low wind speed region, though less prominent.

To further assess the quality and distribution of the target variables used in the modelling phase, histograms of the structural loads are examined. These are shown in Fig. 3.7 & Fig. 3.8. For the blade loads, only the flapwise bending moment at the blade root is available and is therefore used as the sole target variable. In the case of the tower, both fore-aft and side-side bending moments are recorded. However, for simplicity purposes, only the fore-aft moment DEL is used as the predictive target, given its strong correlation with the side-side component.



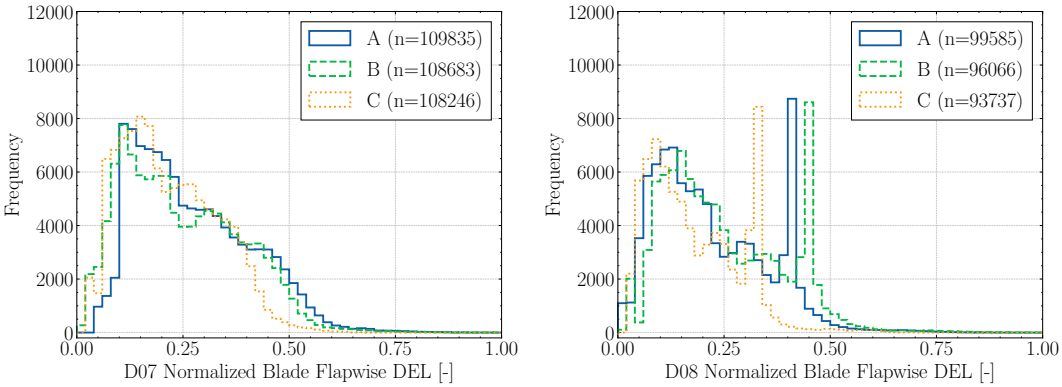


Figure 3.7: Blade flapwise DEL histogram for each wind turbine. The capital letters refer to each blade signal and n is the number of datapoints.

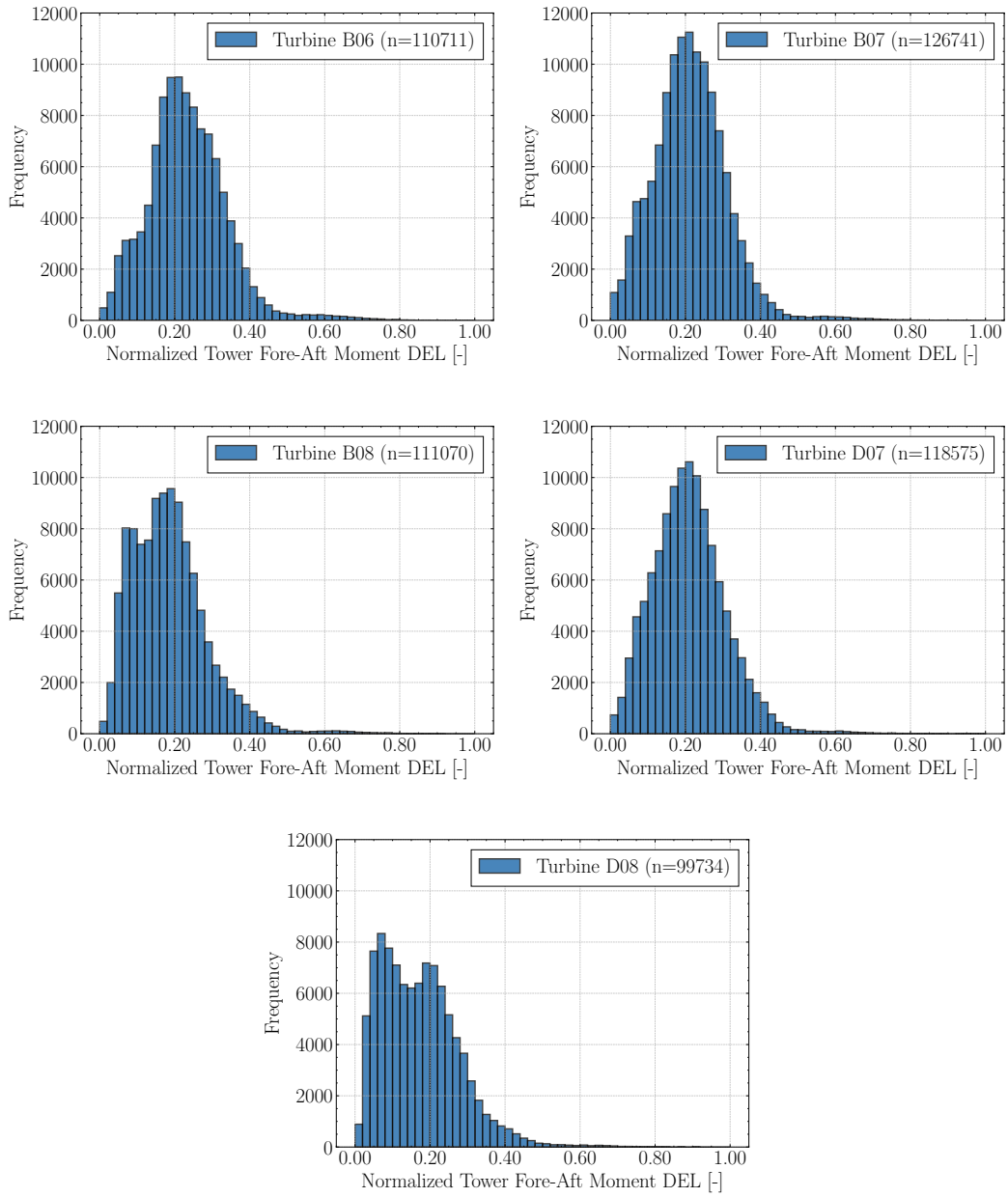


Figure 3.8: Tower FA moment DEL histogram for each wind turbine and n the number of datapoints.

The histograms provide a basic but informative sanity check, revealing the range and shape of the DEL distributions. Regarding the blade flapwise DEL for each load turbine, they are presented in Fig. 3.7, while the corresponding tower fore-aft DEL histograms are shown in Fig. 3.8. Overall, the distributions of the blade DELs are fairly consistent across the different turbines and blade signals, indicating a generally uniform loading behaviour. Nonetheless, several anomalies are observed. For instance, the B blade of turbine B06 displays a high concentration of unrealistically low DEL values, suggesting a likely strain gauge malfunction. A similar, albeit less severe, pattern is observed in the B blade of turbine C08. Furthermore, for turbine D08, all blade signals exhibit three distinct peaks in the mid-to-high DEL region. This irregularity is attributed to a prolonged shutdown period during the early months of 2022, most likely related to maintenance activities (see Fig. 3.9). Additionally, the A blade signal is missing from turbine B07; however, the remaining blade signals exhibit comparable distributions and similar data availability. Based on these observations, the A blade signal is selected as the target variable for all turbines except B07, where the B blade signal is used instead. A datapoint in this context refers to a single row in the SCADA dataset, representing the recorded measurements and signals at a specific timestamp.

Turning to the tower DEL histograms, the distributions are generally bell-shaped and centred around a dominant peak, reflecting stable load patterns across the fleet. However, variations in peak amplitude and shape are noticeable between turbines. In particular, turbine D08 again shows an irregular pattern with two distinct peaks, which can be traced back to its extended shutdown in early 2022 (see Fig. 3.9). A secondary peak is also observed in turbine B08, likely resulting from a combination of curtailment events and experimental campaigns conducted on that unit. Despite these localised irregularities, the tower DEL signals are consistent in terms of their overall shape and volume of available data, further confirming the reliability and suitability of the dataset for subsequent modelling efforts. Lastly, one can observe that the C08 tower load data is missing from the dataset.

Fig. 3.9 presents the time series and corresponding 30-day rolling averages of five key mean SCADA signals for turbine D08 over the full duration of the dataset. These include wind speed, active power, pitch angle, nacelle position, and generator speed. A distinct operational anomaly is observed at the beginning of 2022, where a prolonged period of approximately 2.5 months is characterised by zero active power output and significantly elevated pitch angles. This pattern strongly suggests a manual shutdown, likely due to maintenance or other operational constraints. The impact of this period of non-standard operation on model performance is further examined in Chapter 4.

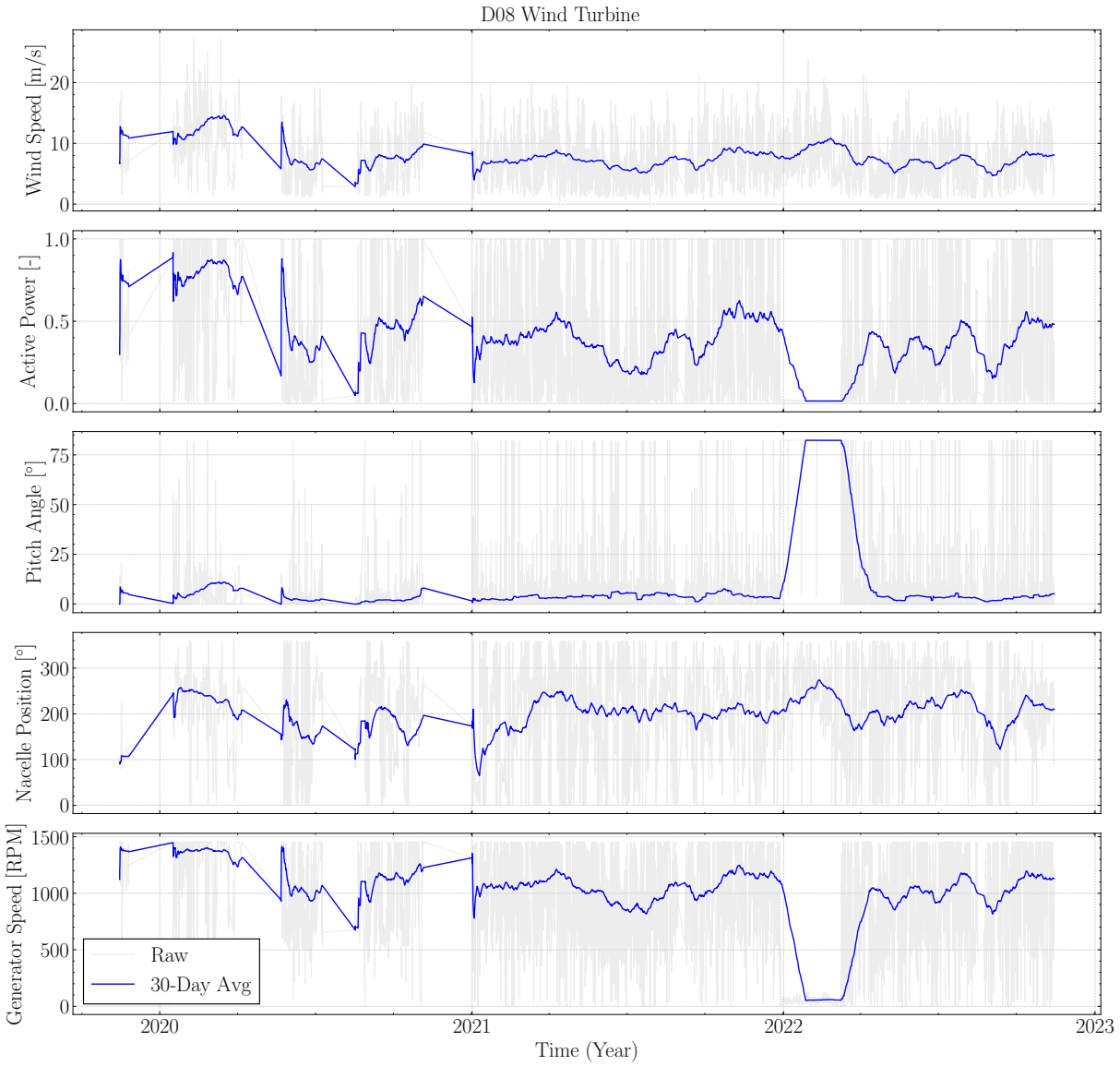


Figure 3.9: 30-Day rolling average for key mean signals of D08 SCADA.

3.2 Pre-processing

After the initial data sanity check presented in the previous section, where key operational and structural load signals were visually and statistically assessed, further data preparation steps are required before the modelling phase. This section describes the pre-processing workflow applied to the dataset to ensure its quality, consistency, and suitability for supervised learning. First, a set of filtering techniques is employed to isolate data corresponding to normal turbine operation and to exclude outliers or anomalous behaviour. Subsequently, two critical procedures are detailed in dedicated subsections: feature scaling, which scales the input variables to comparable numerical ranges, and dataset splitting, where a quantile binning technique is used to generate balanced and representative training, validation, and test sets. These steps form a crucial foundation for achieving reliable and generalizable model performance.

When developing a regression model based on real-world measurement data, it is critical to ensure that the input data is both representative and free from outliers or anomalies that

could compromise model performance. Such anomalies may appear to originate from a different underlying physical process than the one generating the bulk of the data, thereby impairing the model's ability to generalise effectively [26]. To address this, a data filtering step is essential prior to model training, aimed at identifying and removing clusters of data points associated with distinct physical behaviours. Morrison et al. [27] investigate the influence of filtering on the wind turbine power curve by comparing four anomaly detection methods. Their findings indicate that the Gaussian Mixture Model (GMM) offers slightly improved accuracy, although the optimal method ultimately depends on the specific objectives of the analysis. According to Lin et al. [28], anomalies in wind turbine SCADA data typically fall into three main groups, which any effective clustering or filtering strategy should be designed to capture:

- *Group 1*: No power production above cut-in wind speed, typically indicative of manual turbine shutdown.
- *Group 2*: Stable and continuous sub-rated power output due to curtailment strategies.
- *Group 3*: Scattered irregularities arising from sensor faults or signal noise, which deviate from the expected operational patterns.

These groups can be visualised through the B07 power curve using the raw unfiltered data in Fig. 3.10. The black circle corresponds to Group 1, while the red and green circles correspond to Groups 2 and 3, respectively.

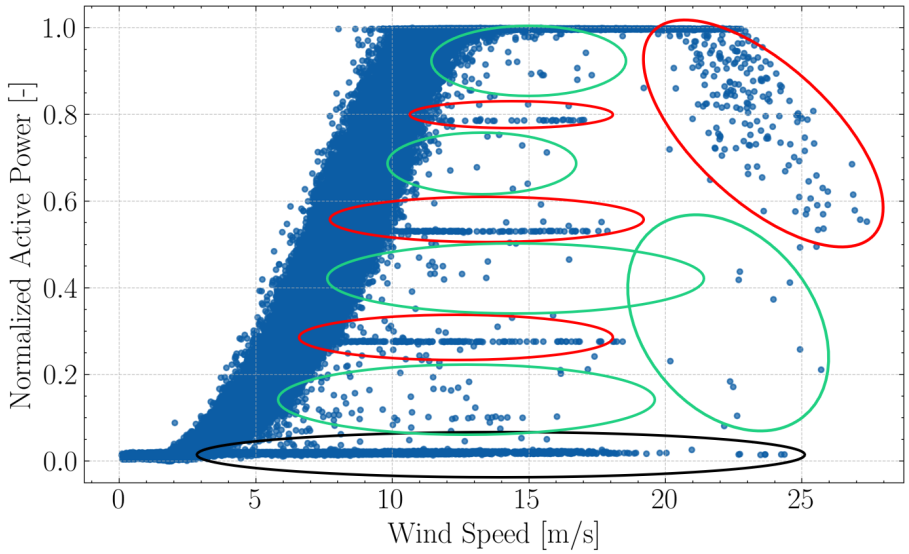


Figure 3.10: Unfiltered B07 power curve with the anomaly groups.

To ensure that the dataset used for model training accurately reflects the normal operational behaviour of the wind turbine, a filtering procedure, referred to as normal operation filtering, is applied. This process is designed to remove anomalous or non-representative data points that could adversely affect the model's predictive capability. The filtering is carried out in three sequential steps.

First, all entries containing missing values (NaNs) are excluded from the dataset. Since machine learning models cannot handle undefined values in either the input or target variables, these datapoints are unusable and thus discarded.

Second, a threshold-based filter is applied to remove datapoints with zero or negative power output. Specifically, any record with mean active power below a predefined minimal threshold is assumed to reflect turbine inactivity or measurement error and is consequently excluded.

The third and most nuanced step involves the application of a GMM [29] for unsupervised clustering. A GMM is a probabilistic approach that assumes that all the data points are generated from a mixture of several Gaussian distributions with unknown parameters and enables the identification of data clusters associated with abnormal operational states, particularly those corresponding to curtailment events (Group 2) and irregular anomalies due to sensor faults or noise (Group 3), as described by Lin et al. [28]. By detecting and removing these clusters from the dataset, the filtering process isolates the datapoints most representative of standard turbine operation, thereby likely improving the reliability and interpretability of the subsequent modelling phase. In addition to the GMM, alternative clustering methods available in the `scikit-learn` library [29], such as DBSCAN and KMeans, were also evaluated. Furthermore, the hierarchical DBSCAN (HDBSCAN) algorithm, implemented via the `hdbscan` library [30], was tested. However, none of these methods demonstrated the same level of effectiveness and robustness as GMM in capturing the relevant operational clusters.

The efficiency of GMM in identifying the anomaly groups, illustrated in Fig. 3.10, is presented in Fig. 3.11.

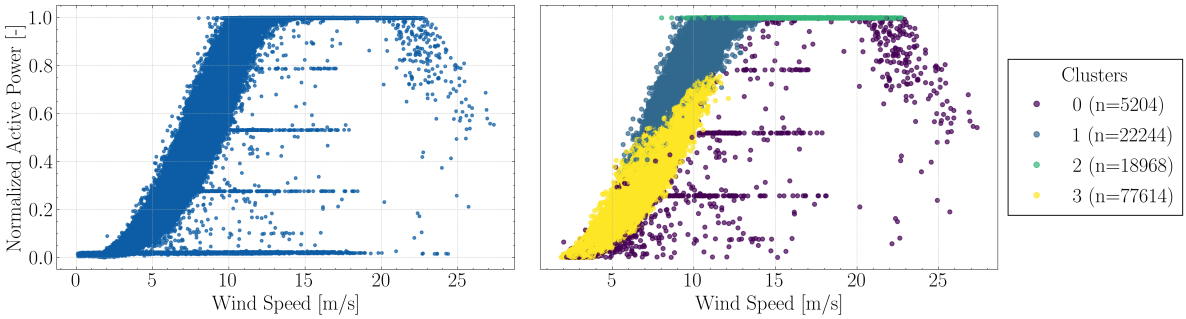


Figure 3.11: Unfiltered B07 power curve (left) and GMM clustered B07 power curve (right). n is the number of datapoints.

Fig. 3.11 compares the unfiltered (left) and clustered (right) power curves for turbine B07, as part of the GMM-based anomaly detection process. The number of clusters is the only user-defined parameter required by the GMM, while the algorithm autonomously infers the cluster structure based on the data distribution. As shown in the right subfigure, the GMM effectively isolates a distinct cluster (Cluster 0) comprising datapoints that deviate from normal operation and thus is discarded. These include instances of curtailment, signal noise, and sensor malfunction, corresponding to Groups 2 and 3 in the previously defined theoretical classification. Although these conditions arise from different physical mechanisms, grouping them poses no issue, as all such datapoints are ultimately discarded. It should also be noted that Group 1, representing zero power output above cut-in wind speed, is absent from the clustered plot, having been removed manually before the clustering step.

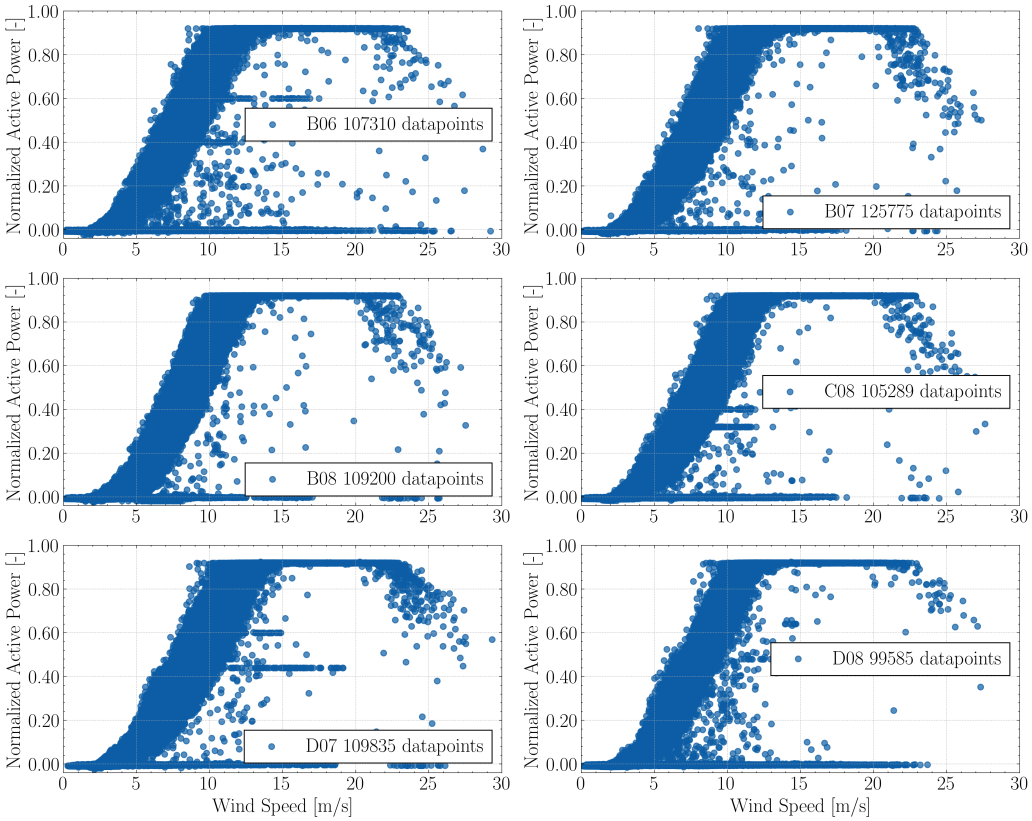


Figure 3.12: Unfiltered power curves.

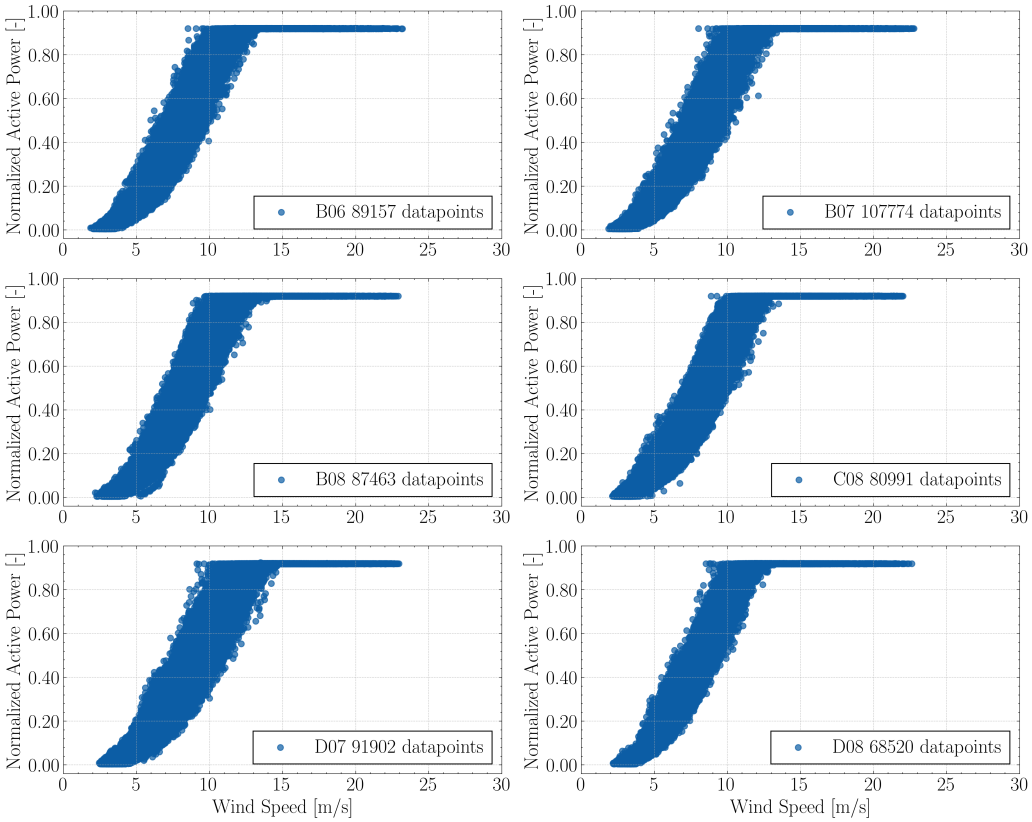


Figure 3.13: Filtered power curves.

Having established the full filtering methodology, including NaN removal, negative power filtering, and GMM-based clustering, it is now applied to all load-measuring turbines in the dataset. The resulting power curves are presented in Fig. 3.12 and Fig. 3.13. Fig. 3.12 displays the power curves after only NaN values have been removed, while Fig. 3.13 shows the cleaned power curves following the complete filtering process. Although the filtering substantially improves the clarity of the power curves, particularly by removing outliers and non-representative datapoints, a considerable spread remains in the sub-rated wind speed region, indicating persistent variability and uncertainty in turbine behaviour prior to rated operation.

It is important to note that the filtering shown here pertains specifically to the blade flapwise DEL models. The only difference compared to the tower models lies in the signals used for NaN filtering, while the resulting power curves are qualitatively consistent across both cases. Additionally, turbine C08 does not contain tower load data and is therefore excluded from the corresponding tower DEL filtering and modelling steps.

In addition to the visual assessment of the filtered power curves, a quantitative summary of the filtering process is also essential to evaluate its impact on the dataset size and consistency across turbines. Fig. 3.14 provides such a summary by illustrating both the absolute number of datapoints before and after the full filtering procedure, as well as the percentage of data removed. Here, the unfiltered datasets refer to those containing only the initial NaN filtering, while the filtered datasets have undergone all three steps: NaN removal, negative power filtering, and GMM clustering. The proportion of removed datapoints ranges from approximately 14% to 31%, with turbine D08 exhibiting the highest filtering percentage. This is consistent with the previously discussed extended shutdown period in early 2022 (see Fig. 3.9). These results confirm that a substantial and usable portion of the dataset is preserved for subsequent modelling.

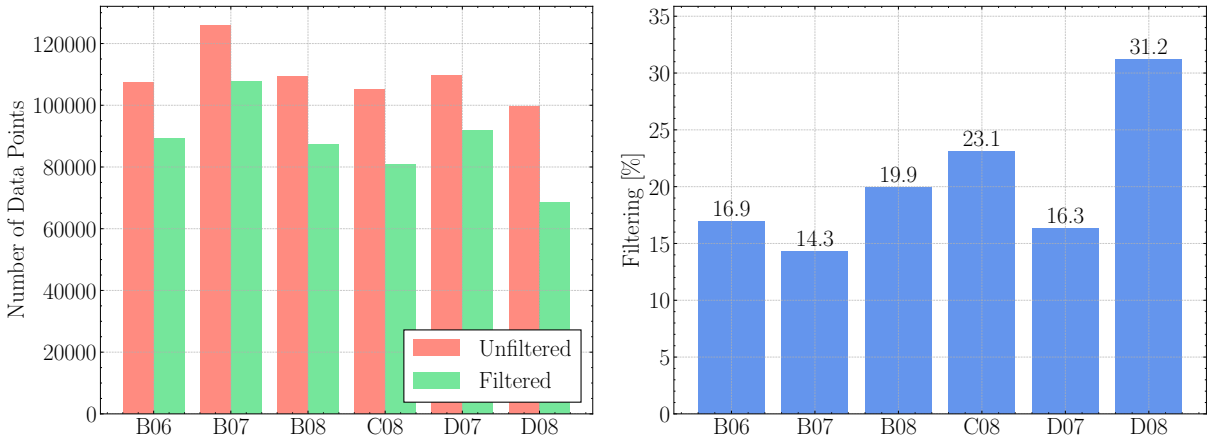


Figure 3.14: Number of data points of the unfiltered and filtered datasets (left) and filtering percentage (right) for all the wind turbines regarding the blade model datasets.

3.2.1 Dataset split

To ensure that the predictive models generalise well to unseen data and are evaluated in a statistically consistent manner, the dataset is divided into three distinct subsets: training, validation, and test sets. These subsets are created using a stratified approach based on the target variable through the use of the `KBinsDiscretizer` from the `scikit-learn`

library [29]. The goal of this approach is to preserve the distributional characteristics of the target variable across all splits and avoid training the model on operational regimes that differ significantly from those found in the validation or test data.

Specifically, the target variable is first discretised into 10 quantile-based bins. This ensures that each bin contains approximately the same number of datapoints, capturing the full range of operational loads. Stratified sampling is then performed on these bins to generate balanced splits. The data is first divided into 90% training and validation, and 10% test data. The 90% subset is then further split into 70% training and 20% validation data, resulting in an overall split of 70%-20%-10% for training, validation, and test sets, respectively. This process is repeated separately for each wind turbine and each target signal (blade or tower DEL).

To validate the effectiveness of this approach, three types of distributions are compared across the three datasets for the B07 blade model. Fig. 3.15 shows the probability density functions of the wind speed using fitted Weibull distributions for each split. The alignment of these distributions confirms that the wind conditions are similarly represented across training, validation, and test sets. In parallel, Fig. 3.15 also displays the normalised histograms of the target variable (DEL) for each split. The close agreement in the shape and range of the distributions further demonstrates that the quantile binning technique successfully balances the structural load characteristics across the datasets. Lastly, Fig. 3.16 demonstrates that the distributions of the nacelle position (proxy for wind direction) signal across the different data splits are closely aligned, indicating good agreement in the wind direction parameter as well. While the plots shown here pertain specifically to turbine B07, similar visual inspections are performed for the dataset splits of all other turbines to ensure consistency in load and wind speed distributions across the different subsets. These checks confirm that the datasets used for model training and evaluation reflect comparable physical and operational conditions, thereby supporting fair and robust model assessment.

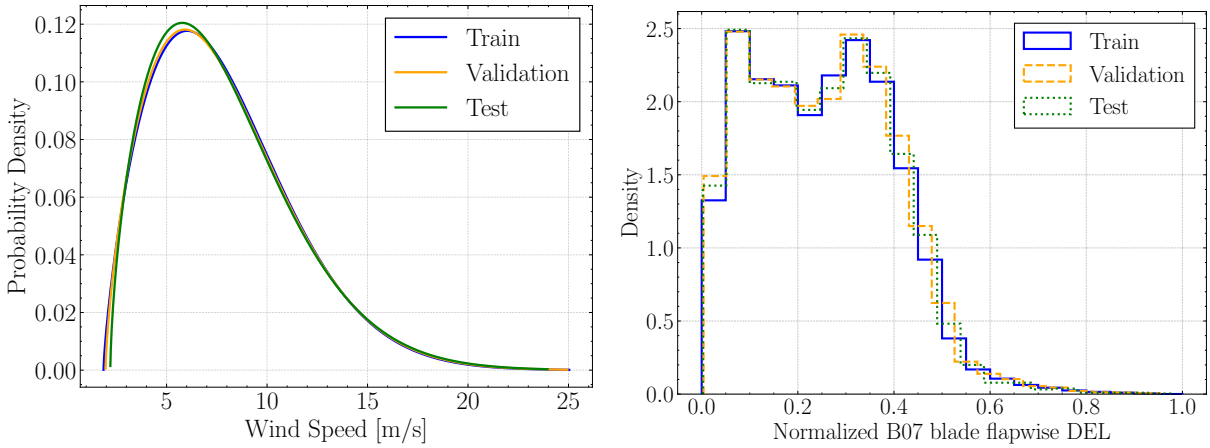


Figure 3.15: Weibull distributions (left) and target variable histograms (right) concerning B07 blade flapwise DEL dataset splitting.

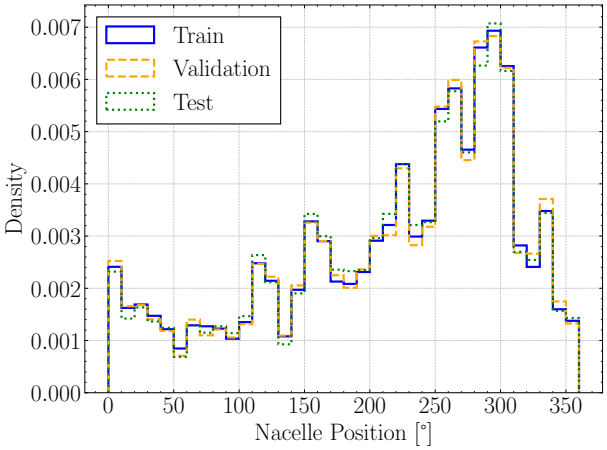


Figure 3.16: Nacelle position distribution across the B07 blade flapwise DEL datasets.

3.2.2 Scalers

Feature scaling is a fundamental step in preparing data for machine learning, particularly for algorithms sensitive to the magnitude and distribution of input variables. Applying appropriate scaling ensures that all features contribute equally to the model training process and that the model converges more efficiently during optimisation. Without scaling, variables with larger numeric ranges can dominate the learning process, potentially leading to suboptimal performance.

Several scaling techniques are evaluated using the `scikit-learn` library [29]. Each method transforms the data differently, according to its underlying assumptions and intended robustness:

- **Standard Scaler**: Transforms each feature to have zero mean and unit variance. The transformation follows:
$$x_{scaled} = \frac{x - \mu}{\sigma} \tag{3.1}$$
where μ is the sample mean and σ the standard deviation.
- **MinMax Scaler**: Scales features to a fixed range, typically [0, 1]. The transformation is given by the following two equations:
$$\sigma = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3.2}$$
$$x_{scaled} = \sigma \cdot (x_{\max} - x_{\min}) + x_{\min} \tag{3.3}$$
where $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the feature.
- **Robust Scaler**: Reduces the influence of outliers by scaling features using the median and the interquartile range (IQR):
$$x_{scaled} = \frac{x - \text{median}}{\text{IQR}} \tag{3.4}$$
making it more robust to heavy-tailed distributions or extreme values (outliers).

- **Log Scaler**: A custom scaler implemented to transform only the target variable by applying a natural logarithm:

$$x_{scaled} = \ln(x) \tag{3.5}$$

  This transformation is particularly useful for compressing the scale of skewed data with a large dynamic range, a common characteristic in structural load signals.

All scalers are tested in various combinations for both input features and target variables across the three regression models considered. The selection of the optimal scaler is based on model performance, determined by training each model and evaluating its prediction accuracy.

The final choices are as follows:

- **XGBoost**: *StandardScaler* for both input and target variables.
- **ANN**: *StandardScaler* for both input and target variables.
- **GPR**: *MinMaxScaler* for both input and target variables.

It is important to emphasise that while differences in performance were observed between the scalers, they were generally modest. This indicates that a well-trained and sufficiently flexible data-driven model should not be overly sensitive to the choice of scaler. Nonetheless, selecting a scaling method consistent with the model architecture and data characteristics can contribute to enhanced numerical stability and more efficient convergence.

3.3 Models

This section presents the regression models evaluated for predicting structural loads on wind turbines based on SCADA data. Three distinct modelling approaches are considered: Extreme Gradient Boosting (XGBoost), Artificial Neural Networks (ANN), and Gaussian Process Regression (GPR). The choice was made based on the literature review conducted in Chapter 2. In addition, a simpler Polynomial Regression model is included as a baseline reference to assess the added value of more complex machine learning algorithms. Each subsection outlines the theoretical foundations of the respective model, discusses its strengths and limitations in the context of wind turbine load prediction, and details its implementation in Python, including the libraries used and the strategy for hyperparameter tuning. The hyperparameter tuning process described in this section is performed for all case studies presented in the Results & Discussion chapter, except the Generalisability case study, where models are deliberately tested in their tuned form without further optimisation. In addition to model development, this section introduces the evaluation framework used to assess performance. Quantitative metrics are employed to compare predictive accuracy, while a set of visual tools is used to gain further insight into model reliability and potential sources of bias. Together, these components form the basis for a robust and comprehensive model comparison in the subsequent Results & Discussion chapter (Chapter 4).

3.3.1 XGBoost

XGBoost (Extreme Gradient Boosting) [31] is a tree-based ensemble learning algorithm widely recognised for its high efficiency, scalability, and robust predictive performance.

It has become a benchmark model in many regression and classification tasks due to its ability to handle large datasets, accommodate missing values, and capture complex, non-linear interactions between features. Another key strength of XGBoost lies in its ease of implementation, with well-documented libraries available in Python, including flexible support for hyperparameter tuning and early stopping.

A primary limitation of XGBoost, as with many tree-based ensemble methods, is its reduced model interpretability. While tools exist to extract feature importance or partial dependencies, the model's internal decision logic remains difficult to fully interpret, especially in high-dimensional settings.

Conceptually, XGBoost differs from traditional ensemble methods such as Random Forest Regression [32] in the way decision trees are constructed. In Random Forests, multiple trees are built independently and in parallel, with each tree trained on a sample of the data. The final prediction is then typically obtained by averaging the outputs of all trees.

In contrast, XGBoost employs a sequential learning strategy where trees are added one at a time. Each new tree is trained to predict the residual errors (i.e., the shortcomings) of the ensemble built so far. This boosting approach allows the model to iteratively refine its predictions, often leading to higher accuracy with fewer trees. Illustrative diagrams of the structural difference between XGBoost and Random Forests are presented in Fig. 3.17 & Fig. 3.18, respectively.

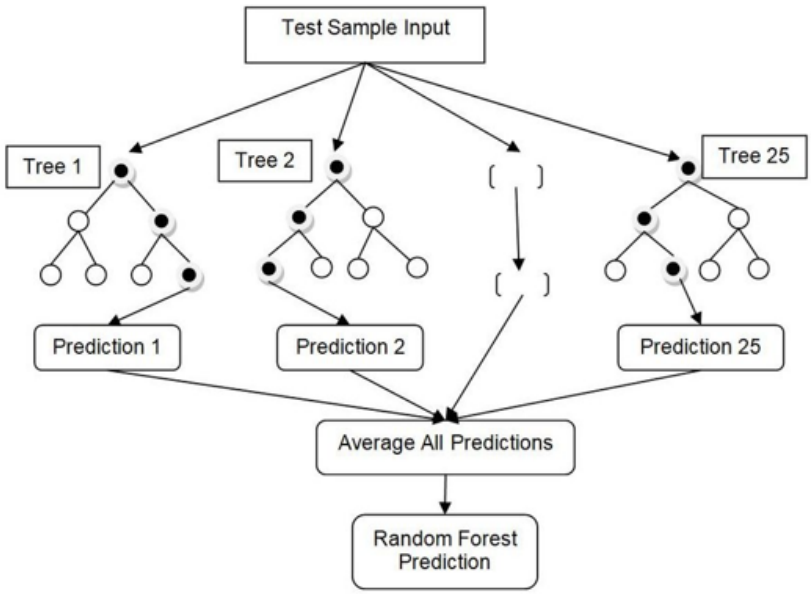


Figure 3.17: Random Forest Regressor structure [33].

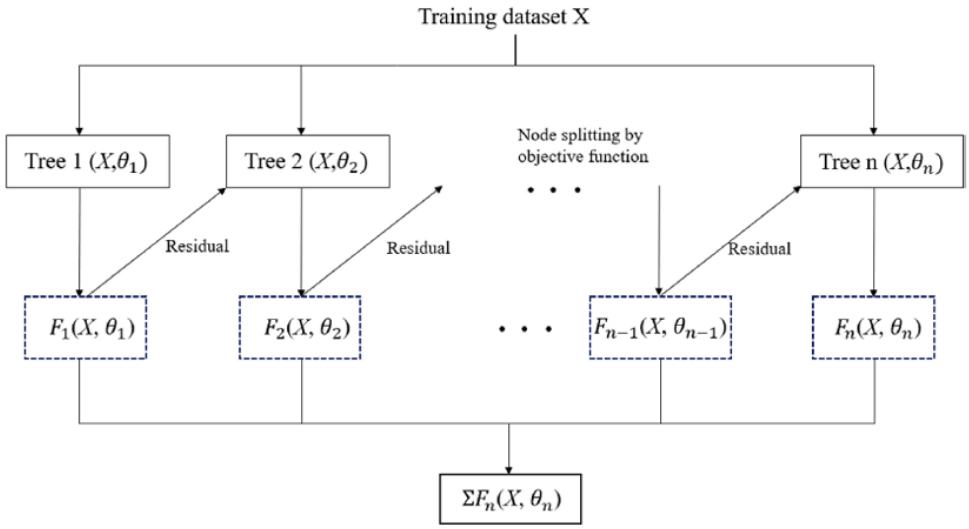


Figure 3.18: XGBoost structure [18].

In the present study XGBoost model is implemented using the `xgboost` Python library and trained on the scaled input and target variables. To optimise model performance, a hyperparameter tuning process is carried out using the `Optuna` framework [34]. Specifically, the `TPESampler` (Tree-structured Parzen Estimator) [35] is employed, which leverages a Bayesian optimisation strategy rather than exhaustive grid search.

This approach incrementally builds a probabilistic model of the objective function and uses it to identify promising regions of the hyperparameter space. Unlike traditional grid search, which explores all combinations uniformly, Bayesian optimisation focuses on evaluating hyperparameters that are more likely to improve performance based on prior trials. This significantly reduces computational cost while still yielding competitive results. However, it does introduce the risk of converging to a local minimum if the search space is highly irregular.

Model configurations are compared based on the mean absolute error (MAE) computed on the validation set. This choice of metric ensures that each configuration is evaluated with a focus on minimising the average magnitude of prediction errors without over-penalising outliers.

The following hyperparameters are included in the tuning process alongside their respective tested values:

- `colsample_bytree` *[0.6, 0.8, 1.0]*: Specifies the fraction of features to be randomly sampled for each tree; helps prevent overfitting by introducing feature-level randomness.
- `learning_rate` *[0.05, 0.1, 0.2]*: Controls the contribution of each tree to the final model; smaller values improve robustness but require more boosting rounds.
- `max_depth` *[3, 6, 9]*: Sets the maximum depth of each individual decision tree; deeper trees capture more complexity but increase the risk of overfitting.
- `alpha` *[0.0, 0.1, 1.0]*: Adds L1 regularization on leaf weights; promotes sparsity in the model and can improve generalization.

- `n_estimators` *[800, 1000, 1500]*: Determines the total number of boosting rounds (i.e., trees) to be built.
- `subsample` *[0.6, 0.8, 1.0]*: Defines the fraction of training samples used for each boosting iteration; introduces instance-level randomness to enhance generalisation.

The hyperparameter tuning process is repeated for 100 trials in parallel across 12 workers. Once the best configuration is identified, the final XGBoost model is retrained using the full training dataset and the optimal set of hyperparameters.

3.3.2 Artificial Neural Network

Artificial Neural Networks (ANNs) are a class of machine learning models inspired by the structure and functioning of the human brain. They are particularly well-suited for modelling complex, non-linear relationships in high-dimensional data. In this work, the ANN is implemented using the `TensorFlow` library [36], and it is structured as a feedforward network composed of sequential layers of neurons. The goal of the network is to approximate a continuous function that maps the input features to the target structural loads.

Each neuron performs a weighted sum of its inputs, followed by the application of a nonlinear activation function. For a single neuron, this operation can be expressed as:

$$z = \sum_{i=1}^{n} w_i x_i + b, \quad a = \phi(z) \tag{3.6}$$

where x_i are the input features, w_i are the associated weights, b is the bias term, z is the linear combination of inputs and $\phi(z)$ is a non-linear activation. These activation functions allow the network to learn complex and highly nonlinear mappings.

Fig. 3.19 illustrates a typical ANN architecture. The red nodes on the left represent the *input layer*, which receives the scaled SCADA features. The blue nodes in the middle represent the *hidden layers*, where the network processes the input data through weighted combinations and non-linear activation functions to extract patterns and relationships that are not immediately visible from the raw inputs. Finally, the green node on the right denotes the *output layer*, which produces the network's prediction; in this case, a single output representing the predicted load value (e.g., blade DEL or tower DEL).

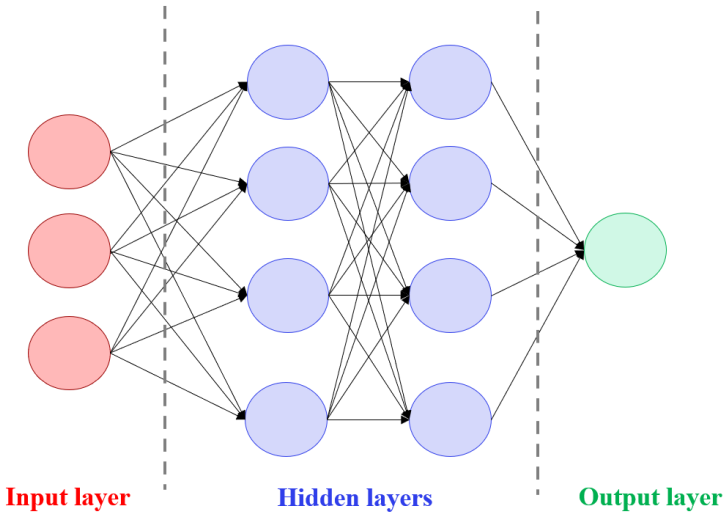


Figure 3.19: Example of an ANN architecture.

Advantages:

- Highly flexible and capable of modelling complex non-linear relationships.
- Scalable to large datasets and adaptable to a variety of input-output configurations.

Limitations:

- Computationally expensive to train, especially with deep architectures or large hyperparameter spaces.
- Sensitive to hyperparameter choices such as network depth, layer width, activation functions, learning rate, and batch size, making tuning a critical step for successful application.

The ANN used in this study is implemented using the `TensorFlow` and `Keras` libraries and is optimised using the `Optuna` framework [34], like the XGBoost. The network follows a sequential architecture comprising an input layer, three fully connected hidden layers, and a single-node output layer for regression. The number of neurons and activation functions in each hidden layer are treated as tunable hyperparameters.

The hyperparameter search space includes the following:

- `units_1`, `units_2`, `units_3` *[32, 64, 96, 128]*: Number of neurons in each of the three hidden layers, sampled in increments of 32.
- `activation_1`, `activation_2`, `activation_3` *['leakyrelu', 'tanh']*: Activation function applied after each hidden layer; either leaky ReLU for unbounded outputs or tanh for smoother, bounded transformations.
- `learning_rate` *[1e-4, 1e-2] (log-uniform)*: Learning rate of the Adam optimiser, sampled on a logarithmic scale to balance convergence speed and stability.

In contrast to the XGBoost model, where mean absolute error (MAE) is used for validation, the ANN uses mean squared error (MSE) as the validation loss function. Although MSE penalises outliers more heavily than MAE, empirical results from several turbines showed that MSE led to improved predictive performance. This choice encourages the

network to fit high-magnitude errors more carefully, which can be beneficial in load prediction tasks where certain operational conditions produce large load values.

To support efficient and stable training, two Keras callbacks are included:

- `EarlyStopping`: Stops training if the validation loss does not improve for 50 epochs, and restores the weights from the best epoch, helping to prevent overfitting.
- `ReduceLROnPlateau`: Reduces the learning rate by a factor of 0.5 after 10 stagnant epochs, allowing the model to refine its convergence in later stages of training.

The network is trained with a batch size of 512. This relatively large batch size is chosen to take advantage of computational efficiency. While large batches can result in slower convergence, this is balanced by training over up to 500 epochs, enabling the model to learn effectively over time.

The hyperparameter optimisation is carried out using the `Optuna` framework [34] over 100 trials. `TPESampler` is again used as the default Bayesian optimisation strategy. Unlike traditional grid search, which exhaustively explores all parameter combinations, the Bayesian approach incrementally models the objective function and focuses the search on promising regions of the hyperparameter space. As a result, not all configurations are tested; instead, Optuna prioritises evaluating those most likely to improve model performance based on prior outcomes. Once the optimal configuration is identified, the final model is retrained from scratch using the full training set and the selected hyperparameters, with both callbacks activated to support stable and efficient training. The model training is always assessed by plotting the training and validation loss over the number of epochs (see Fig. 3.20); if both curves decrease steadily without divergence, this indicates successful training and the absence of overfitting.

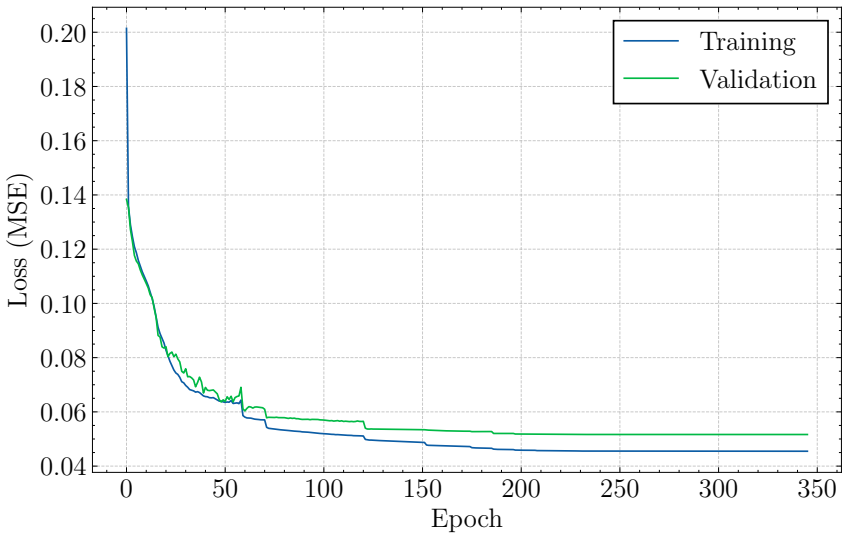


Figure 3.20: Training & validation loss of the B07 blade flapwise DEL model.

3.3.3 Gaussian Process Regression

Gaussian Process Regression (GPR) is a non-parametric, probabilistic machine learning method that defines a distribution over possible functions that fit a given dataset. Unlike parametric models, which assume a specific mathematical form for the function (such as

a linear or polynomial model), a Gaussian Process (GP) does not commit to any fixed equation. Instead, it assumes that the function could take many possible shapes and defines a probability distribution over all these possible functions, even before seeing any data. This allows GPR to capture complex and non-linear relationships while providing uncertainty estimates for its predictions.

Formally, a GP is defined such that any finite collection of function values follows a multivariate normal distribution. For a function $f(x)$, the GP is fully specified by a mean function $m(x)$ and a covariance function k(x,x'), also known as the kernel. The kernel encodes assumptions about the function's smoothness, periodicity, or other structural properties. Mathematically, this is described by the following equation:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{3.7}$$

An illustrative example of Gaussian Process regression is shown in Figure 3.21, where a sinusoidal true function is approximated using a limited number of noisy observations. The GP not only learns a smooth regression mean that fits the data, but also provides confidence intervals (quantiles) indicating the model's uncertainty. This ability to model both predictions and their associated uncertainty makes GPR especially useful in applications where understanding model confidence is critical.

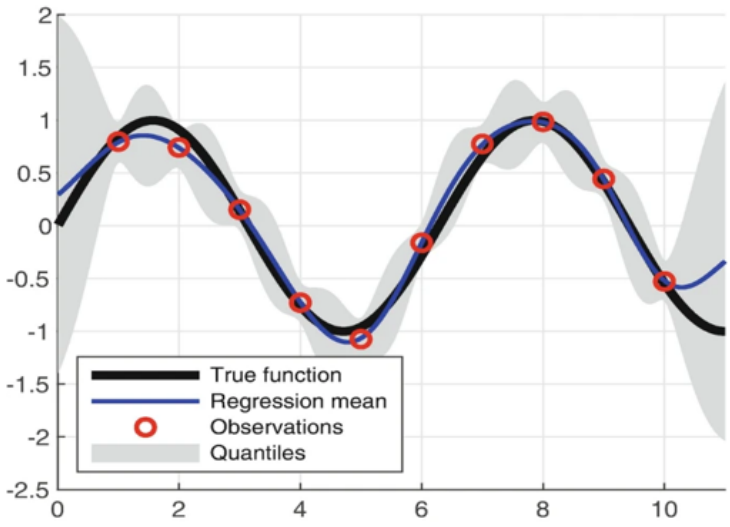


Figure 3.21: Illustration of Gaussian Process Regression [37].

Advantages:

- Straightforward implementation.
- Strong predictive accuracy, particularly in low-data regimes.
- Naturally incorporates uncertainty estimation through posterior variance.

Limitations:

- Computationally expensive, scaling as $\mathcal{O}(n^3)$ in time and $\mathcal{O}(n^2)$ in memory with respect to the number of training samples n.

- Requires careful selection and tuning of the kernel function, which significantly influences model behaviour and performance.
- Less suitable for very large datasets unless sparse approximations or scalable variants are used.

In this study, Gaussian Process Regression is implemented using the `GPy` library [38]. Due to the computational limitations of standard GPR, which scales poorly with dataset size, a Sparse Gaussian Process Regression (SGPR) model is employed. Rather than computing the full $n \times n$ covariance matrix, SGPR introduces a smaller set of representative points known as inducing points to approximate the full GP posterior. This approximation significantly reduces both the computational and memory complexity:

- Time: from $\mathcal{O}(n^3)$ to approximately $\mathcal{O}(m^2 n)$.
- Memory: from $\mathcal{O}(n^2)$ to $\mathcal{O}(mn)$.

Here, m is the number of inducing points, with $m \ll n$. The number of inducing points is a user-defined parameter; higher values yield better approximation accuracy but increase computational cost. When $m = n$, the full GPR model is recovered.

Another key design choice in GPR is the kernel function, which encodes the model's assumptions about the function it is trying to learn, such as smoothness, periodicity, or linearity. The kernel plays a central role in shaping the GP's behaviour and is arguably the most critical hyperparameter. Several kernel combinations were evaluated for turbine B07 to identify the optimal predictive accuracy:

- **RBF (Radial Basis Function)**: A highly flexible and smooth kernel that assumes infinite differentiability. While effective in capturing subtle trends, it is prone to overfitting, especially in noisy datasets.
- **Matern32**: A more realistic alternative to the RBF, allowing for less smoothness. It yields more robust performance in noisy, real-world data like SCADA, where some degree of irregularity is expected.
- **Matern32 + White**: This combination adds a **White kernel**, which explicitly models uncorrelated observation noise. This is particularly useful for SCADA data, where sensor noise and signal fluctuations are common.
- **(RBF + Linear) * Bias**: A composite kernel that combines a smooth non-linear trend (RBF) with a linear component, scaled by a bias term. This setup can model both long-term trends and fine-grained fluctuations while adjusting for global offset.

After comparative testing, the Matern32 + White kernel yielded the best overall performance, both in terms of predictive accuracy and computational efficiency. Its ability to model both structured patterns and measurement noise makes it especially suitable for the SCADA load data used in this study. For this reason, it was selected as the default kernel configuration for all wind turbines. Due to time constraints, kernel evaluations were only conducted on turbine B07, but the chosen configuration was applied consistently across the entire dataset.

The number of inducing points was set to 2500, which provided a practical trade-off between approximation quality and training time. The model was optimised for 20 iterations using the built-in optimiser in `GPy`, with a fixed random seed to ensure reproducibility.

3.3.4 Polynomial Regression

Polynomial regression is a classical regression method that extends linear regression by allowing the model to fit non-linear relationships between the input features and the target variable. Instead of modelling the target as a linear function of the inputs, polynomial regression introduces higher-order terms and interaction effects by transforming the input space. This enables the model to capture curved trends while still using a linear least-squares fitting procedure.

Formally, given an input vector $\mathbf{x} = [x_1, x_2, \ldots, x_p]$, a degree-3 polynomial regression model estimates the target variable y as:

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \sum_{j=1}^{p} \sum_{k=j}^{p} \beta_{jk} x_j x_k + \sum_{j=1}^{p} \sum_{k=j}^{p} \sum_{l=k}^{p} \beta_{jkl} x_j x_k x_l \tag{3.8}$$

Where the terms include combinations of input features raised to powers up to the chosen polynomial degree. The result is a model that remains linear in the parameters β, but nonlinear in the original input space.

In this study, polynomial regression is implemented using the `scikit-learn` library [29]. A degree-3 polynomial is used, which expands the original feature set to include all cubic terms and lower, including interactions. The modelling pipeline is constructed as follows:

- Only the mean values of the SCADA input variables are selected to reduce dimensionality and avoid excessive term proliferation.
- Input and target variables are scaled using `StandardScaler` to ensure numerical stability during fitting.
- The input matrix is expanded using `PolynomialFeatures(degree=3)` to generate the full set of polynomial terms.
- A linear regression model (`LinearRegression`) is then fitted to the transformed input matrix and scaled target.
- Predictions on the test set are made and relevant performance metrics are calculated.

Although this model does not match the flexibility or scalability of modern machine learning methods, it provides a useful baseline. Its simplicity and closed-form solution make it computationally efficient, and it helps highlight the performance improvements achieved by more advanced models such as XGBoost, ANN, and GPR.

3.3.5 Metrics

After having thoroughly discussed the models used in this study, it is essential to define the performance metrics used to evaluate their accuracy. These quantitative metrics allow for a structured comparison of model performance in Chapter 4 and help to highlight strengths and limitations across different algorithms and target variables.

Each metric offers a unique perspective on performance and comes with its own advantages and drawbacks. For instance, the Mean Absolute Percentage Error (MAPE) may become unstable or biased when actual values are close to zero, while the Normalised Root Mean

Squared Error (NRMSE) can over-penalise large outliers that may not significantly impact the operational utility of the model.

The following four metrics are employed to evaluate and compare model performance:

- **Coefficient of Determination $\mathbf{R^2}$**:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.9}$$

  Where y_i is the true value of the target variable for sample i, $\hat{y}_i$ is the predicted value from the model for sample i, and $\bar{y}$ is the mean of the actual target values across n samples. This metric indicates how well the model explains the variance in the target data. A value of 1 corresponds to perfect predictions, while a value of 0 indicates that the model performs no better than the mean of the data.

- **Mean Absolute Percentage Error (MAPE)**:

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3.10}$$

  MAPE expresses the prediction error as a percentage of the actual value, which makes it interpretable and useful for comparing across different scales. However, it becomes unreliable when the target variable exhibits low values as the percentage increases substantially.

- **Normalised Root Mean Squared Error (NRMSE)**:

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sigma_y} \tag{3.11}$$

  Where σ_y is the standard deviation of the actual values. NRMSE provides a scale-free error metric that highlights relative model accuracy, but like RMSE, it is sensitive to large errors.

- **Mean Absolute Error (MAE)**:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.12}$$

  MAE measures the average magnitude of the errors without considering their direction. It is less sensitive to outliers than RMSE or NRMSE and provides a robust measure of general prediction accuracy. In some figures (e.g., Fig. 4.10), the absolute prediction error is visualised instead. This is equivalent to MAE but shown per sample, without averaging over the entire dataset.

Additionally, some plots visualise related error forms:

- **Prediction Error (PE)**:

$$\text{PE} = y_i - \hat{y}_i \tag{3.13}$$

- **Signed Percentage Error (Signed PE)**:

$$\text{Signed PE} = \frac{y_i - \hat{y}_i}{y_i} \times 100\% \tag{3.14}$$

These quantities help assess whether the model tends to over- or under-predict and how that bias varies across operating conditions such as wind speed (see Fig. 4.9).

Although all four metrics are computed during model evaluation, only R^2, MAPE, and NRMSE are reported in the final model comparison tables in Chapter 4. This combination of metrics was chosen to balance interpretability, sensitivity to different types of error, and robustness:

- R^2 captures how well the model explains variance in the data.
- MAPE provides intuitive percentage-based errors, useful for practical interpretation.
- NRMSE gives a normalised error sensitive to large deviations, ensuring that significant under- or over-predictions are penalised appropriately.

This combination of performance metrics ensures a well-rounded evaluation of both the accuracy and reliability of the models across various wind turbines and load targets.

3.3.6 Visual evaluation

While quantitative performance metrics are essential for comparing model accuracy, they alone are not sufficient for a comprehensive evaluation. To fully understand a model's strengths and limitations, a combination of numerical metrics and visual assessments is necessary. Plots offer critical insights into aspects that metrics may obscure, such as systematic bias, over- or under-prediction, and the distribution of errors across different wind speeds or wind directions. They also help reveal how model errors vary with operational conditions and whether the predictions are well-calibrated. Such visualisations are presented in Chapter 4 to complement the metric-based evaluation and provide a more complete picture of model performance.

CHAPTER 4

Results & Discussion

This chapter presents a comprehensive evaluation of various case studies designed to address the central research questions posed in this thesis. These questions aim to assess the predictive capabilities of data-driven models for damage equivalent loads in operational wind turbines, with particular attention to the effects of data composition, model architecture, and spatial information. The analysis explores how these factors influence model accuracy, generalisability, and overall operational relevance.

The three research questions outlined in Chapter 2 serve as the backbone of this investigation:

1. What is the most suitable data-driven modelling approach for accurately predicting damage equivalent loads using SCADA data from operational wind turbines?
2. How do different data characteristics, such as input feature composition, data volume, and signal pre-processing, affect the performance and generalisability of surrogate models for DEL prediction?
3. To what extent can spatial information from neighbouring wind turbines enhance the accuracy of surrogate models across a wind farm?

To address these questions, a series of targeted case studies is conducted, each exploring a distinct aspect of the modelling problem. Table 4.1 summarises the individual case studies and the specific scientific inquiries they aim to resolve.

Table 4.1: Overview of case studies and associated scientific inquiries.

Case Study	**Scientific Inquiry**
Feature Importance	Which input features (signals) mostly influence the accuracy?
Model Comparison	What is the best model to predict the DEL on a wind turbine?
Filtering	How heavily does the filtering affect the model performance?
Dataset Size	How does the quantity of data affect the model's predictive accuracy?
Local Configuration	How does the input of the neighbouring wind turbines affect the results?
Neighbouring DEL	How much can the accuracy be improved if the DEL of a neighbouring wind turbine is used as input?
Generalisability	How does the model accuracy change when it is tested on SCADA from a different wind turbine?
Fatigue Damage Estimation	How can predicted DEL values be translated into cumulative fatigue damage over a given operational period, and how accurately do the models estimate this damage?

The first research question is tackled primarily through the **Model Comparison** and **Fatigue Damage Estimation** case studies. The former compares different regression models, namely XGBoost, ANN and GPR, to identify the most effective architecture for DEL prediction. The latter demonstrates the practical application of these models by converting predicted DELs into accumulated fatigue damage over a selected time series. This final case study illustrates how such predictions can support long-term asset integrity evaluations and decision-making regarding turbine lifetime extension.

The second research question motivates several case studies that systematically alter the data configuration, including **Feature Importance**, **Filtering**, and **Dataset Size**. These studies investigate how different preprocessing steps, input variables, and sample sizes influence model robustness and accuracy. The third research question guides the **Local Configuration**, **Neighbouring DEL**, and **Generalisability** case studies, all of which evaluate the impact of spatial information and cross-turbine generalisation.

Each case study (except for the last one) is repeated for two distinct output targets: the blade flapwise DEL and the tower fore-aft moment DEL. This dual analysis ensures that conclusions remain consistent across structurally different load channels. Except for the Model Comparison and Fatigue Damage Estimation, which necessitate cross-model evaluation, the other case studies employ the **XGBoost** algorithm due to its balance between predictive power and computational efficiency.

Operational Relevance and Motivation of Each Case Study

- **Feature Importance:** This case study identifies the most influential SCADA features affecting DEL prediction. From a business perspective, this insight can guide sensor prioritisation and data acquisition strategies, enabling operators to focus on the most informative measurements for condition monitoring and digital twin applications.
- **Model Comparison:** By comparing the performance of different machine learning models, this study determines the most suitable modelling technique for DEL estimation. The findings can inform the selection of algorithms in digital monitoring platforms, ensuring the right trade-off between accuracy, interpretability, and computational cost.
- **Filtering:** Investigating how preprocessing through filtering affects model performance, this case study directly addresses operational decisions concerning data quality control and preprocessing pipelines. Understanding its impact allows for standardising signal conditioning procedures across datasets and projects.
- **Dataset Size:** This study quantifies the sensitivity of model accuracy to the amount of training data. Its business relevance lies in answering how much historical data must be collected before a reliable surrogate model can be deployed, thus guiding data management and storage policies.
- **Local Configuration:** This case investigates whether incorporating SCADA signals from neighbouring turbines improves model performance. The results have implications for farm-wide monitoring strategies, where centralised modelling may benefit from spatially integrated data streams.
- **Neighbouring DEL:** By using the actual DEL of neighbouring turbines as an

additional input, this study examines the potential performance gain from partial load data availability. In practice, this supports hybrid monitoring solutions where high-fidelity data from a few turbines helps predict loads across the fleet.

- **Generalisability:** Evaluating model performance when applied to a different turbine's SCADA data tests the robustness of learned models across assets. This study addresses scalability and transferability, critical for developing universal models deployable across different turbines or farms.
- **Fatigue Damage Estimation:** This final case study demonstrates how predicted DELs can be converted over time to estimate cumulative fatigue damage. By applying the surrogate model to a selected test-period time series, the analysis showcases the practical end-use of DEL prediction: estimating fatigue damage and assessing whether the turbine is on track to meet, exceed, or fall short of its design life. This supports informed decisions around lifetime extension, fatigue-critical maintenance, and risk assessment.

Collectively, these case studies offer a structured framework for understanding the behaviour of surrogate DEL models under varying operational and data conditions. The findings not only answer the core scientific inquiries but also translate into actionable insights for wind farm operators seeking to implement predictive maintenance and structural health monitoring at scale.

4.1 Feature Importance

In this section, the influence of various SCADA-derived variables and their statistical descriptors on the predictive accuracy of the regression models is assessed. Given the large number of available signals (exceeding 170), it is impossible to perform a comprehensive feature importance or correlation analysis using purely mathematical methods without first narrowing down the input space. Therefore, an initial screening of variables is made based on physical insight into which operational parameters are most likely to influence wind turbine loading, either blade or tower loads. For each selected variable, key statistical metrics (minimum, maximum, mean, and standard deviation) are extracted and evaluated for their contribution to the model's performance.

The five variables selected for the feature importance study, as well as the corresponding justification for their choice, are:

- **Wind Speed** (U): Wind speed directly determines the aerodynamic forces acting on the rotor blades and tower. Higher wind speeds lead to increased loads and fatigue cycles, making it a primary driver of DEL.
- **Active Power** (P_{active}): Active power output reflects the turbine's operating condition and regime(below-above rated region) and indirectly captures the aerodynamic efficiency and control strategy. Variations in power correlate with load fluctuations, especially under partial load or curtailment scenarios.
- **Pitch angle** (θ_p): Blade pitch angle controls the aerodynamic loading on the blades. Changes in pitch, especially during dynamic events like shutdowns or gusts, significantly affect the loading of the blade and as a result, of the tower.
- **Nacelle Position** (θ_{nac}): The nacelle position indicates the orientation of the wind

turbine, effectively defining the direction in which the turbine is facing. In this study, it is used as a proxy for wind direction, as the dedicated wind direction signal was deemed unreliable based on multiple validation checks. By incorporating nacelle position, the model gains insight into the turbine's wake exposure and specifically, whether it is aligned with the wake of an upstream turbine, which can significantly affect the inflow conditions and resulting loads.

- **Generator Speed** (ω_{gen}): Although closely correlated with active power, the generator speed can possibly offer additional dynamic and transient information that reflects the turbine's control behavior and drivetrain response, both of which are relevant to understanding load fluctuations and fatigue accumulation.

To identify potential correlations among the input features and understand their contribution to the prediction of blade flapwise and tower fore-aft DELs, the Random Forest Classifier (RFC) method of the scikit-learn library [29] is employed. This approach transforms the continuous DEL targets into categorical bins, enabling the use of a classification algorithm to estimate feature importance. The `RandomForestClassifier` is an ensemble method that builds multiple decision trees using random subsets of data and features, and aggregates their predictions through majority voting. It is known for its robustness against overfitting, ability to handle high-dimensional data, and its built-in mechanism for ranking features based on their influence on the model's decisions. However, it may sometimes produce inconclusive or unstable importance rankings when features are correlated or when the target variable is noisy. The results of the feature importance analysis are visualised in Fig. 4.1, where each bar indicates the relative importance of a feature in predicting the binned DEL classes.

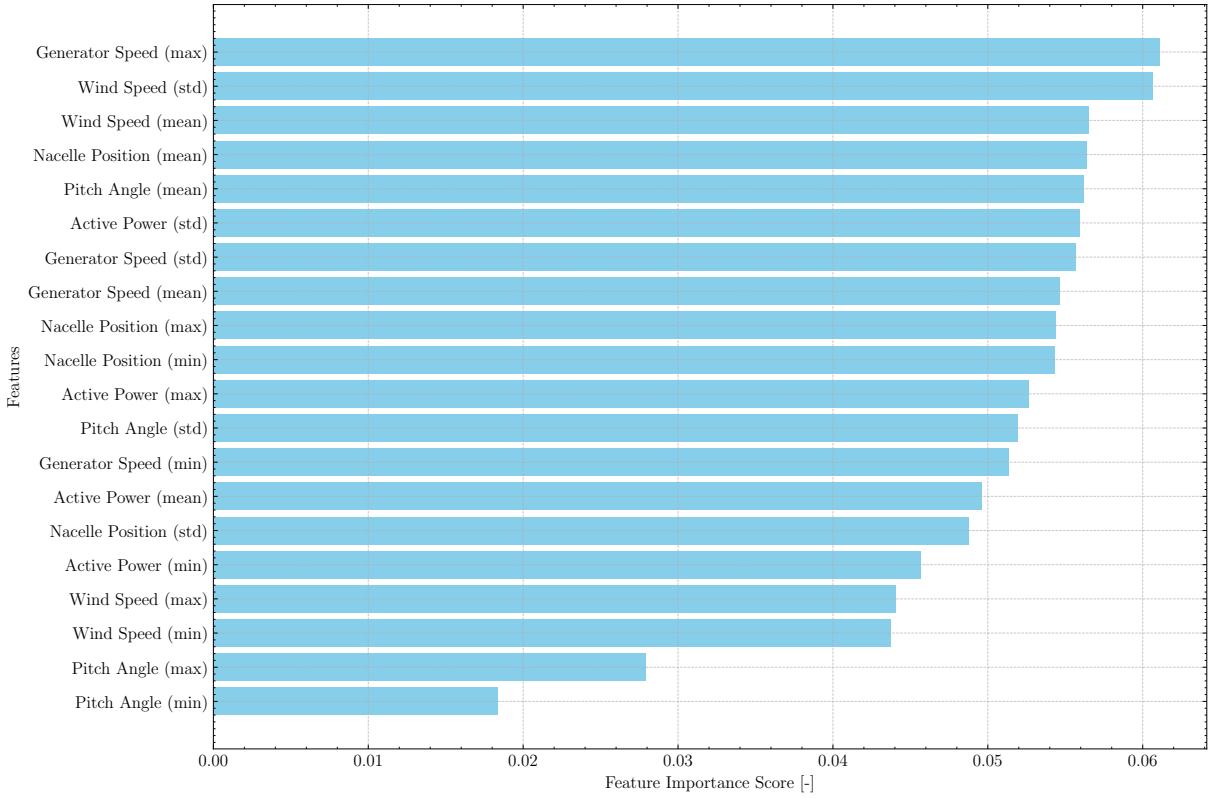


Figure 4.1: Feature Importance using RFC & targeting B07 blade flapwise DEL.

From Fig. 4.1, the results appear to be inconclusive. While Generator Speed (max) and Wind Speed (std) emerge as the most influential features, their corresponding mean values rank significantly lower in importance. This outcome is somewhat counterintuitive, as the mean and standard deviation of a variable are inherently related statistical measures. Typically, one would expect their contributions to model performance to be of comparable magnitude.

Another popular method used for feature importance in the machine learning world is the SHAP (SHapley Additive exPlanations) [39] values method. SHAP is an approach based on game theory which aims to explain the prediction of an instance(in this study, a row in the SCADA dataset) by computing the contribution of each feature to that prediction. These contributions are called SHAP values, and they sum up to the difference between the model's output for that instance and the average prediction over the training dataset (Eq. (4.1)).

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \tag{4.1}$$

where $f(x)$ is the model prediction for an input vector x, M is the number of input features, ϕ_0 is the mean prediction over all training data and ϕ_i is the SHAP value for feature i.

The SHAP value quantifies how much each feature i contributed to the prediction $f(x)$ for this instance. It is the average marginal contribution of each feature across all possible subsets of features, and it is calculated through Eq. (4.2), which is a computationally expensive equation(exponential in the number of features).

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right] \tag{4.2}$$

where F is the full set of features, S is a subset not containing i and $f_S(x_S)$ is the model trained or approximated on the subset S.

Since the **XGBoost** is used, the `shap.TreeExplainer` SHAP method is applied as it is specifically optimised for decision tree-based models. The TreeSHAP [40] is a polynomial-time algorithm which avoids enumerating all 2^M feature subsets, thereby substantially reducing computational cost. The results of the method are shown in Fig. 4.2.

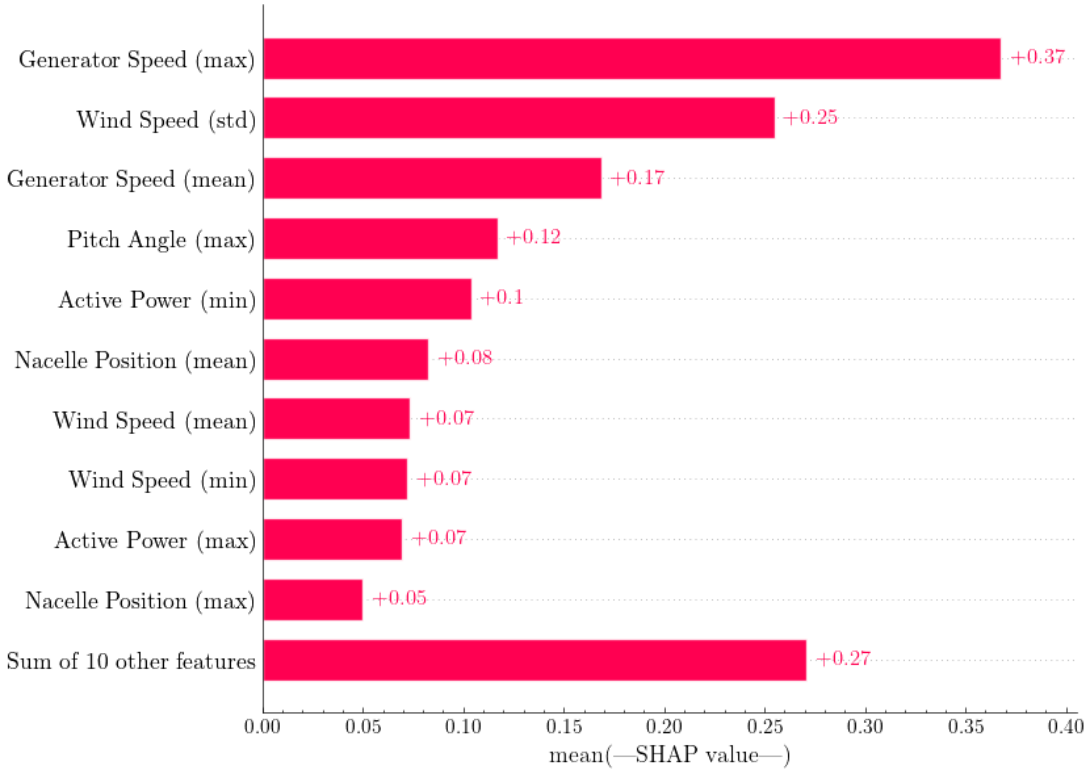


Figure 4.2: Feature Importance using TreeSHAP & targeting B07 blade flapwise DEL with XGBoost.

Fig. 4.2 depicts the 10 most influential features according to the TreeSHAP algorithm. Although the first two are in agreement with the RFC method, the rest appear in a different order, which shows inconsistency and increases the doubt regarding the validity of both approaches. Therefore, in order to safely conclude on which features should be used as input to the models, XGBoost models are trained on different input features and their performance is compared.

The feature groupings are thoughtfully selected based on physical insight, with the goal of capturing potential interactions and uncovering complex relationships among variables. The tested combinations are presented Table 4.2.

Table 4.2: Feature importance cases with XGBoost targeting B07 blade flapwise DEL and tower moment FA DEL.

Case Number	Input Features (Blade Model)	Input Features (Tower Model)
1	U (mean, std)	U (mean, std)
2	U (mean, std) P_{active} (mean, std)	U (mean, std) P_{active} (mean, std)
3	U (mean, std) θ_{nac} (mean)	U (mean, std) θ_{nac} (mean)
4	U (mean, std) P_{active} (mean, std) θ_p (mean, std)	U (mean, std) P_{active} (mean, std) θ_p (mean, std)
5	U (mean, std) P_{active} (mean, std) ω_{gen} (mean, std)	U (mean, std) P_{active} (mean, std) ω_{gen} (mean, std)
6	U (mean, std, min, max) P_{active} (mean, std, min, max)	U (mean, std, min, max) P_{active} (mean, std, min, max)
7	U (mean, std) P_{active} (mean, std) θ_p (mean, std) θ_{nac} (mean, std) ω_{gen} (mean, std)	U (mean, std) P_{active} (mean, std) θ_p (mean, std) θ_{nac} (mean, std) ω_{gen} (mean, std)
8	U (mean, std) P_{active} (std) θ_p (mean) θ_{nac} (mean, min, max) ω_{gen} (mean, std, max)	–
9	U (mean, std, min) P_{active} (min, max) θ_p (max) θ_{nac} (mean, max) ω_{gen} (mean, max)	–
10	U (mean, std, min, max) P_{active} (mean, std, min, max) θ_p (mean, std, min, max) θ_{nac} (mean, std, min, max) ω_{gen} (mean, std, min, max)	U (mean, std, min, max) P_{active} (mean, std, min, max) θ_p (mean, std, min, max) θ_{nac} (mean, std, min, max) ω_{gen} (mean, std, min, max)

One can notice in Table 4.2 that the cases are identical for the blade & tower models except for cases 8 & 9 which refer to the 10 most influential features according to the RFC and SHAP analysis, respectively. This analysis is conducted only for the blade model. The reasoning for the case selection and the corresponding investigation are presented below.

- **Case 1**: This case uses only the wind speed (mean and std), representing the most fundamental input for load prediction. Wind speed directly governs the aerodynamic forces acting on the turbine, and thus is expected to be the most influential factor based on physical understanding. This baseline case serves as a reference for evaluating the added value of other variables.
- **Case 2**: Active power (mean and standard deviation) is added to the wind speed

inputs, as it captures the energy conversion efficiency and turbine operating region. Below rated wind speed, mean power is strongly correlated with loading, while above rated, power remains constant, and variations in standard deviation become more informative. This case explores whether power output complements wind speed in improving DEL predictions.

- **Case 3**: The nacelle position mean is introduced alongside wind speed to represent the turbine's yaw alignment. Since the dedicated wind direction signal is unreliable, nacelle position provides critical insight into wake effects, specifically, whether the turbine is operating in the wake of another, which strongly influences loading.
- **Case 4**: Pitch angle (mean & std) is included to enrich the model's understanding of the turbine's control response, particularly in the above-rated region. While active power varies below rated, pitch adjustments dominate above rated, bridging an information gap in the power signal. The standard deviation may also offer insight into load cycling and variability over the averaging interval.
- **Case 5**: Generator speed is introduced to assess its marginal contribution beyond active power. While the two signals are generally correlated, generator speed may provide additional information on transient dynamics and drivetrain behaviour, which can influence fatigue loading. This case tests whether the added dynamic input improves prediction accuracy.
- **Case 6**: This case enhances the model with full range statistics (mean, std, min, and max) for wind speed and active power. Including these metrics enables the model to better capture the variability and extremes in operational conditions, which are particularly relevant for predicting load amplitudes contributing to DEL.
- **Case 7**: A balanced feature set is selected, incorporating all five physical variables, but limiting each to mean and standard deviation. This configuration aims to retain physical interpretability while reducing the input dimensionality by approximately 50% compared to the full-feature case, testing the trade-off between model complexity and performance.
- **Case 8**: In this case, the model is trained using the top 10 most influential features as determined by the feature importance scores from the RFC method. This approach serves as a data-driven alternative to the physics-based feature selection in previous cases, allowing for comparison between expert knowledge and algorithmic ranking. It helps evaluate whether a model guided by statistical importance can match or exceed the performance of physically motivated inputs.
- **Case 9**: This case uses the 10 most influential features identified via SHAP values. By comparing this model to both the RFC-based and physics-informed cases, the study explores how well this game-theoretic approach aligns with physical understanding and whether it yields superior predictive performance compared to RFC.
- **Case 10**: The most comprehensive configuration, this case includes all selected variables with their full statistical descriptors. It serves as an upper-bound scenario, designed to evaluate whether supplying the model with maximum available information yields the best predictive performance.

To balance computational cost and ensure a representative assessment of generalisability, wind turbine B07 is selected as the target for the feature importance analysis. Positioned

centrally within the wind farm and surrounded by neighbouring turbines, B07 frequently experiences waked inflow from multiple directions. Accordingly, the models predict the blade flapwise DEL in the blade-focused analysis and the tower fore-aft moment DEL in the tower-focused analysis.

4.1.1 Blade models

In this section, the feature importance results concerning the blade models are presented and discussed. The model performance for each case, described by the relevant metrics, is depicted in Table 4.3.

Table 4.3: Feature importance case results with XGBoost, targeting B07 blade flapwise DEL.

Case Number	R^2 [%]	MAPE [%]	NRMSE [–]
1	76.05	19.02	0.489
2	86.30	13.28	0.370
3	89.82	11.58	0.319
4	87.95	12.72	0.347
5	87.27	12.83	0.357
6	88.11	12.63	0.345
7	94.93	8.22	0.225
8	95.03	7.97	0.223
9	95.19	8.12	0.219
10	95.44	7.89	0.214

The feature importance analysis for the blade flapwise DEL model reveals insightful trends regarding the predictive value of different SCADA inputs. Starting with a baseline, *Case 1* demonstrates that using only wind speed statistics results in very low model accuracy. This reinforces the understanding that while wind speed is the primary external driver of aerodynamic loads, it alone is insufficient to capture the complexity of a wind turbine load response.

In *Case 2*, the addition of active power significantly improves model performance, indicating that power output, especially below rated wind speed, is a key proxy for operational state and energy capture efficiency. However, *Case 3* shows that nacelle position (used here as a proxy for wind direction) is an even more influential feature. Its inclusion leads to a substantial increase in model accuracy, underlining the importance of capturing wake effects and turbine orientation, particularly for centrally located turbines like B07.

Cases 4 & 5 investigate the added value of pitch position and generator speed, respectively. While both features contribute marginal improvements, their impact is not as pronounced, suggesting that their role may be more supportive than essential in predicting blade loading. Similarly, *Case 6*, which includes the full range (min/max) of wind

speed and power, does not yield a significant accuracy gain, implying that the mean and standard deviation are already sufficient to capture most of the relevant statistical variability.

Case 7 presents a condensed input set that includes all five core variables (wind speed, active power, pitch angle, nacelle position, and generator speed), each with its mean and standard deviation. Despite using only half the features of *Case 10* (which includes full statistical descriptors), *Case 7* achieves nearly the same level of accuracy. *Case 10* remains the most accurate overall, but the marginal performance gain does not justify the increase in input dimensionality and computational cost.

To benchmark against purely data-driven approaches, *Cases 8 & 9* use the top 10 features selected by Random Forest and SHAP-based algorithms. These models perform comparably to *Case 7*, but do not offer interpretability grounded in wind turbine physics and technology. Moreover, their reliance on statistical associations rather than causal understanding makes them less reliable across varying turbines, especially when feature interactions are present.

In conclusion, *Case 7* is selected as the optimal input configuration for the blade model in subsequent analyses. It provides a strong balance between model accuracy and computational efficiency, while maintaining physical transparency and generalisability. Compared to the exhaustive *Case 10* and the algorithmic feature selection methods (8 & 9), *Case 7* offers a more practical and interpretable foundation for operational modelling of wind turbine DELs.

4.1.2 Tower models

In this section, the feature importance results concerning the tower models are presented and discussed. The relevant performance metrics are presented in Table 4.4.

Table 4.4: Feature importance case results with XGBoost, targeting B07 tower moment FA DEL.

Case Number	R^2 [%]	MAPE [%]	NRMSE [–]
1	62.92	19.45	0.609
2	79.42	14.45	0.454
3	78.58	13.94	0.463
4	84.63	13.49	0.392
5	82.37	13.52	0.420
6	83.42	13.39	0.407
7	90.66	10.05	0.306
8	–	–	–
9	–	–	–
10	92.21	9.36	0.279

The feature importance analysis for the tower fore-aft DEL prediction follows a similar

trajectory to the blade model, though some differences in variable influence emerge. As expected, *Case 1*, which includes only wind speed statistics, performs poorly and serves as a reference baseline. The addition of active power in *Case 2* results in a substantial performance gain, reaffirming its strong correlation with structural loading under varying wind conditions.

Interestingly, while the nacelle position remains an important feature in *Case 3*, its impact on tower loading is less pronounced than in the blade model. This may reflect the tower's more global structural response, which is less sensitive to wake-induced asymmetries than the blades. In contrast, *Case 4* reveals that pitch angle plays a more prominent role in tower DEL prediction compared to blades. When added alongside wind speed and active power, pitch significantly improves model accuracy, likely due to its impact on thrust, which directly affects the tower base moments.

Generator speed in *Case 5* and statistical range descriptors in *Case 6* provide minor but consistent improvements. Notably, unlike in the blade model, the inclusion of min and max values appears more beneficial here, likely because tower loading is influenced by the amplitude of force fluctuations.

Case 7, which uses a reduced input set of mean and standard deviation for key variables, shows a marked drop in performance relative to *Case 10*. In this context, the increase in input dimensionality and computational demand in *Case 10* is justified by the significantly better model accuracy. The full statistical descriptors (mean, std, min, max) for all selected variables allow the model to more accurately capture the loading variability that drives tower fatigue.

In conclusion, *Case 10* is selected as the preferred input configuration for tower moment FA DEL prediction for the subsequent analyses. While more computationally intensive, it seems to yield the most accurate results, and its benefits outweigh the added complexity. The tower model appears to benefit more from comprehensive input data.

Interestingly, the tower models demonstrate consistently lower predictive accuracy compared to the blade models. Specifically, when evaluating *Case 10* (which includes all available input features), the tower model exhibits a 3% decrease in R^2, a 1.5% increase in MAPE, and a 30% increase in NRMSE relative to the blade model. This result is rather counterintuitive, as one might expect blade loads to be more challenging to predict due to the blades' complex dynamics, higher flexibility, and greater susceptibility to deflection and vibration-induced fluctuations. The fact that this expectation is not reflected in the results may point to lower quality or higher uncertainty in the tower load measurements, potentially stemming from sensor placement or calibration issues.

4.2 Model Comparison

As the input features have been selected based on the considerations discussed in Section 4.1, this section proceeds with a comprehensive evaluation of the models introduced in Section 3.3, examining their performance in terms of accuracy, robustness, and computational cost. The model training, validation and testing is based on the filtered dataset (see Section 3.2). The comparison relies on the evaluation metrics defined in Section 3.3.5 as well as relevant plots for visual assessment. For clarity, the analysis is structured into two subsections, focusing on the prediction of blade and tower moment DELs, respectively.

4.2.1 Blade models

In this section, the results of the blade flapwise DEL models, along with the relevant plots, are presented and discussed. The performance metrics for each model are displayed in Table 4.5, Table 4.6 & Table 4.7.

Table 4.5: R^2 scores [%] of blade models using internal data. Green highlights the best model, red the worst.

Wind Turbine	XGBoost	ANN	GPR	Polynomial
B06	95.31	94.65	95.20	85.62
B07	94.93	94.85	94.85	83.06
D07	93.43	93.01	93.02	77.11
D08	95.01	94.82	95.02	87.13
C08	94.52	94.68	94.48	81.95
B08	94.80	94.59	94.44	81.46

Table 4.6: MAPE [%] of blade models using internal data. Green highlights the best model, red the worst.

Wind Turbine	XGBoost	ANN	GPR	Polynomial
B06	8.77	9.41	8.95	16.82
B07	8.22	8.50	8.47	16.25
D07	8.39	8.78	8.60	16.41
D08	8.81	8.86	8.90	15.38
C08	8.57	8.63	8.74	16.11
B08	8.77	9.19	9.19	17.50

Table 4.7: NRMSE [-] of blade models using internal data. Green highlights the best model, red the worst.

Wind Turbine	XGBoost	ANN	GPR	Polynomial
B06	0.217	0.231	0.219	0.380
B07	0.225	0.227	0.227	0.412
D07	0.256	0.264	0.264	0.478
D08	0.223	0.228	0.223	0.359
C08	0.234	0.231	0.235	0.425
B08	0.228	0.233	0.236	0.431

Three key observation emerge from examining Table 4.5, Table 4.6 & Table 4.7:

- **Polynomial regression** is insufficiently expressive, as evidenced by its consistently lower performance across all turbines. This supports the hypothesis that the underlying relationships in the data are highly non-linear and complex, justifying the use of more sophisticated data-driven models.

- All advanced models (**XGBoost**, **ANN**, **GPR**) demonstrate strong predictive capability, with comparable performance across turbines. Their similar metrics indicate that the information contained in the data is captured sufficiently by all three algorithms.
- **XGBoost** achieves the best performance in almost all of the wind turbines and performance metrics, slightly outperforming the other models in terms of accuracy while also offering lower computational cost. Its robustness and ease of implementation further reinforce its suitability for this task.

The average and the range of the performance metrics across all wind turbines are summarised in Table 4.8, which confirms the marginally better performance of XGBoost against the other two models.

Table 4.8: Model performance summary for blade flapwise DEL prediction across all wind turbines.

Metric	**Statistic**	**XGBoost**	**ANN**	**GPR**
R^2 [%]	Mean	94.67	94.43	94.50
	Range	93.43 – 95.31	93.01 – 94.85	93.02 – 95.20
MAPE [%]	Mean	8.59	8.90	8.80
	Range	8.22 – 8.81	8.50 – 9.41	8.47 – 9.19
NRMSE [-]	Mean	0.231	0.236	0.234
	Range	0.217 – 0.256	0.227 – 0.264	0.219 – 0.264

After presenting the quantitative performance metrics, the analysis now turns to a visual inspection of model behaviour. Relevant plots are used to assess predictive accuracy, detect potential biases, and better understand the models' strengths and weaknesses. The plots corresponding to the B07 wind turbine are presented and discussed. As B07 is situated centrally within the wind farm, it consistently operates under waked conditions, making its results more representative of general turbine behaviour.

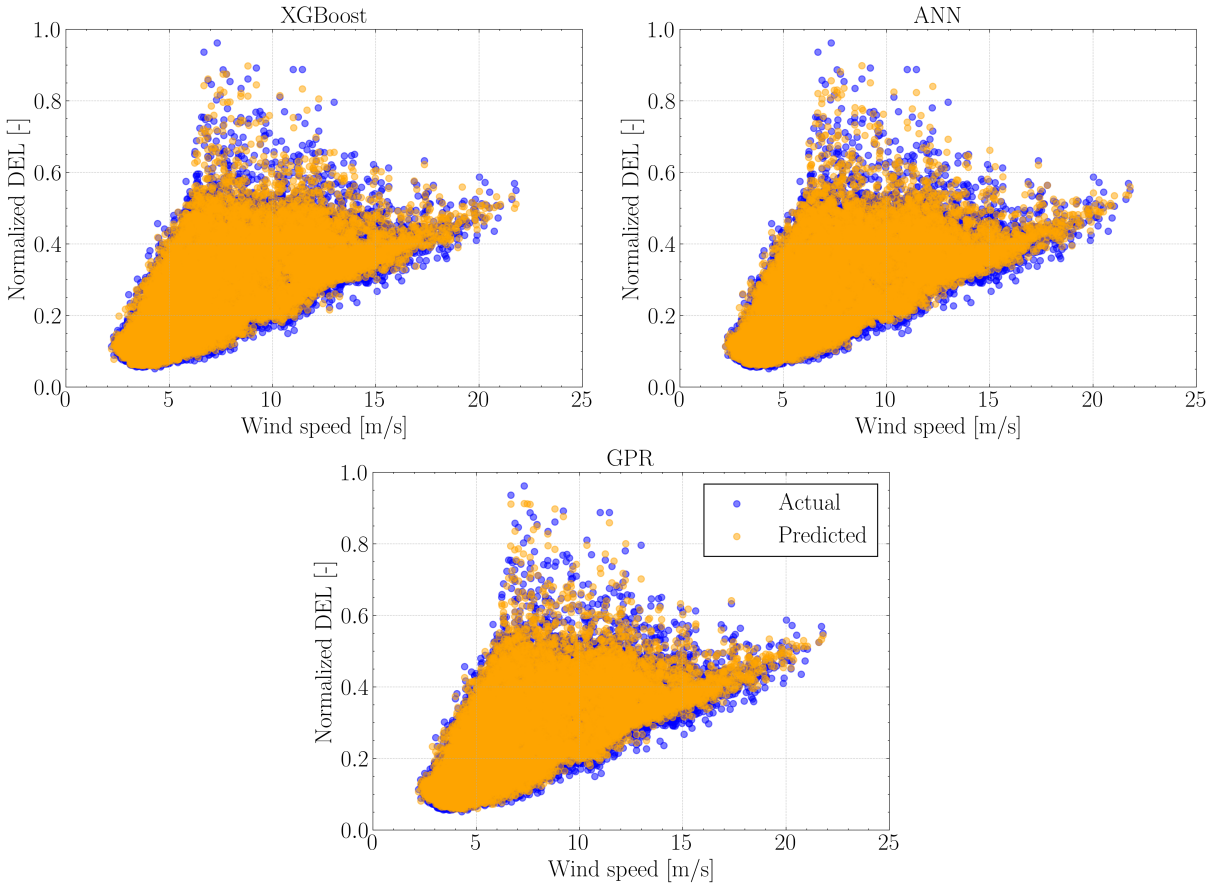


Figure 4.3: B07 blade flapwise actual & predicted DEL against wind speed based on the three models.

Fig. 4.3 illustrates the predictive performance of the three models across the full range of wind speeds. Overall, the models demonstrate strong agreement with the actual values, successfully capturing the majority of the variance and trends in the data. However, a slight tendency towards overprediction is observed, which is noticeable at the lower part of the graphs where the predicted values (orange) consistently appear above the actual values (blue). This subtle bias suggests a small systematic error that may warrant further investigation or model refinement.

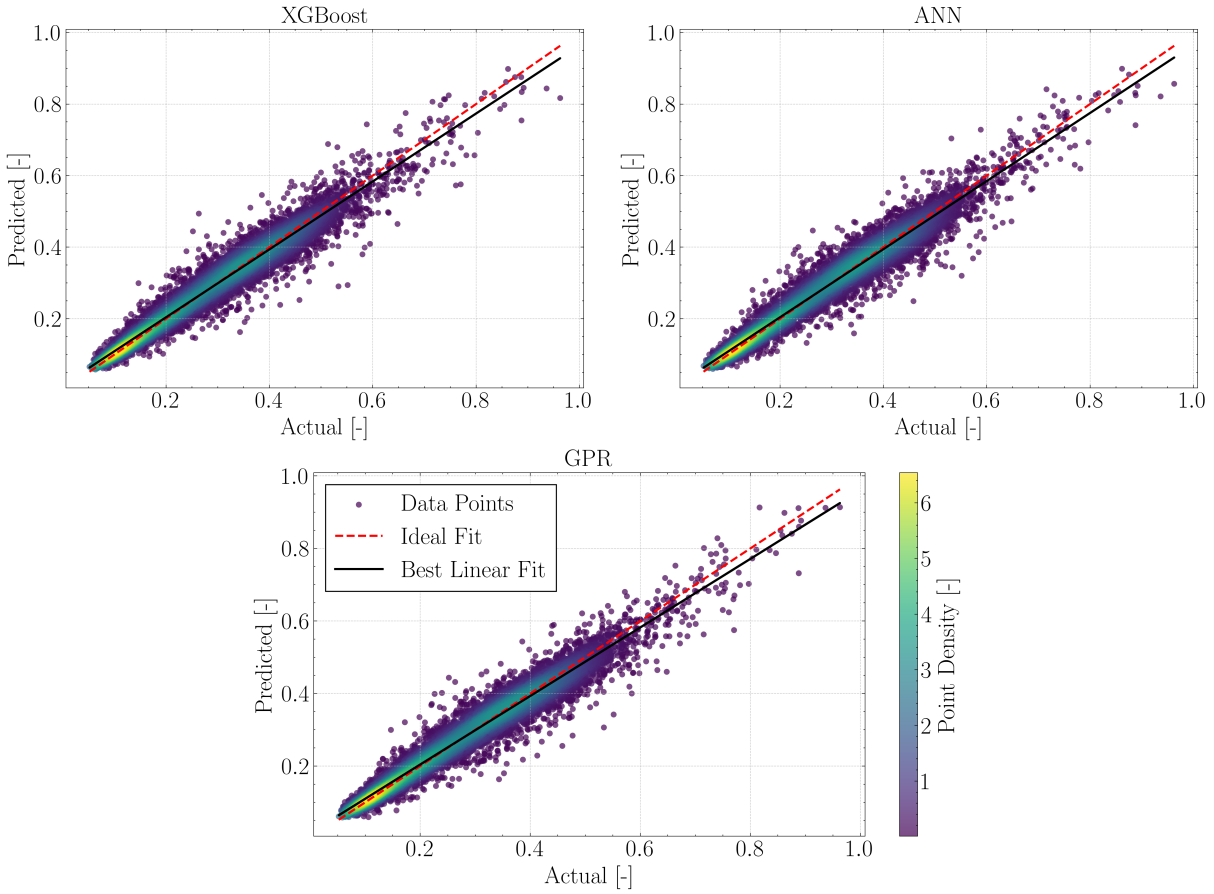


Figure 4.4: B07 blade flapwise predicted vs actual DEL based on the three models with point density colorbar using the Gaussian KDE method.

Fig. 4.4 further supports the models' ability to capture the underlying variance in the data, as indicated by the relatively tight clustering of points around the ideal 1:1 fit line. This alignment reflects a strong overall agreement between predicted and actual values. Nonetheless, a consistent trend is observed across all models: a tendency to overpredict for lower DEL values (below approximately 0.2 normalised DEL) and underpredict for higher DELs. This pattern is visualised by the Best Linear Fit line initially lying above and then dipping below the ideal fit line, suggesting a slight non-linearity in model behaviour across the DEL range.

The colorbar represents point density, computed using the `gaussian_kde` method from the `scipy.stats` library [41], which estimates the underlying probability density function based on the spatial distribution of data points. The kernel density estimation provides a smooth, continuous approximation of the data concentration across the two-dimensional parameter space. While the density values are unitless, higher values indicate regions with greater clustering of data points. However, the Point Density score displayed on the colorbar axis does not represent raw point counts, nor does it maintain strict proportional relationships between values. Due to the inherent smoothing properties of the Gaussian kernel, the density metric should be interpreted qualitatively rather than quantitatively. Consequently, while higher density scores reliably indicate regions of greater data concentration, direct numerical comparisons are not valid. The density estimates are therefore best utilised for identifying relative patterns and trends in the data distribution rather than for precise quantitative analysis.

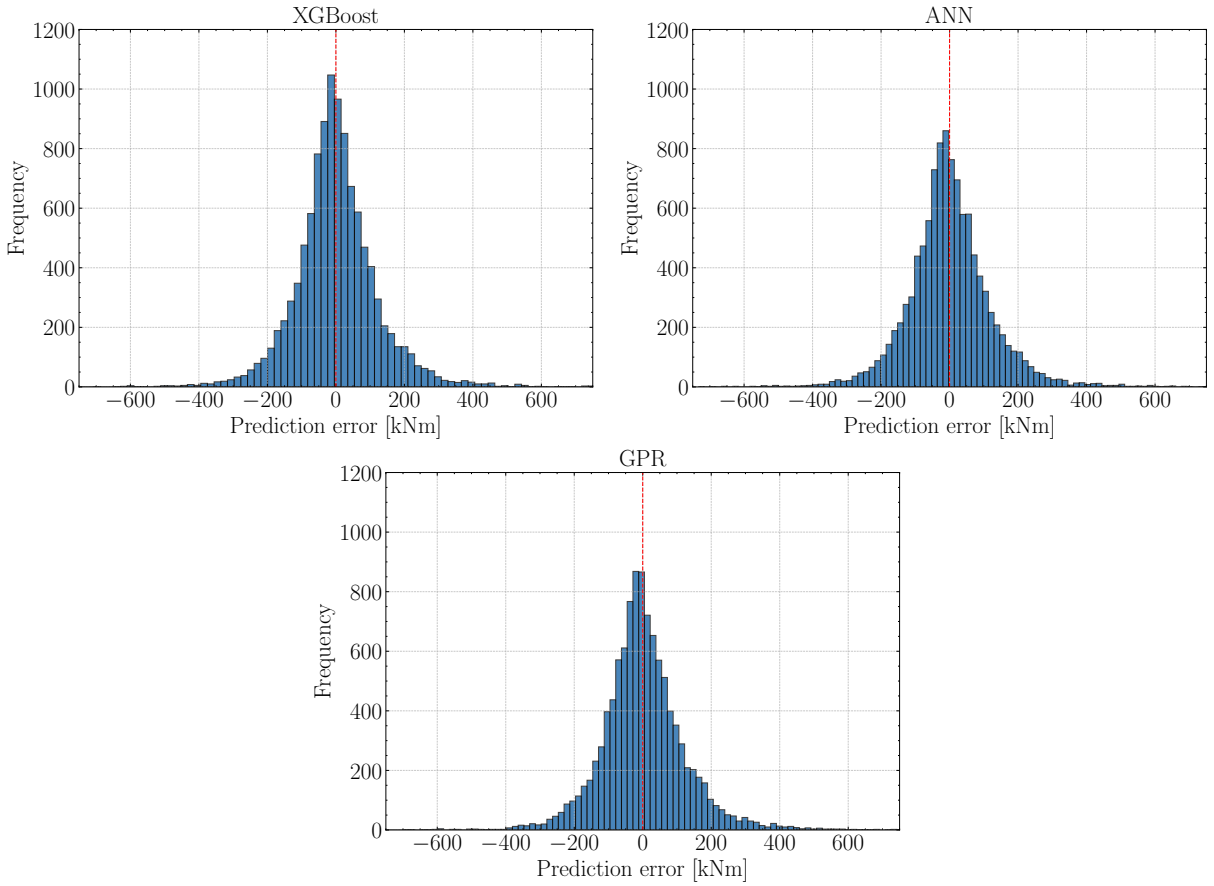


Figure 4.5: B07 blade flapwise DEL error distribution based on the three models.

Fig. 4.5 illustrates the distribution of prediction errors for the three models. All models exhibit a symmetric, bell-shaped distribution centred around zero, which closely resembles a normal distribution and indicates generally unbiased predictions. However, the XGBoost model shows a noticeably denser concentration of errors near zero (reaching bin frequencies close to 1000) compared to around 800 for the other two models. This observation aligns with its superior error metrics discussed previously. As defined in Section 3.3.5, negative prediction errors indicate overprediction, while positive errors correspond to underprediction.

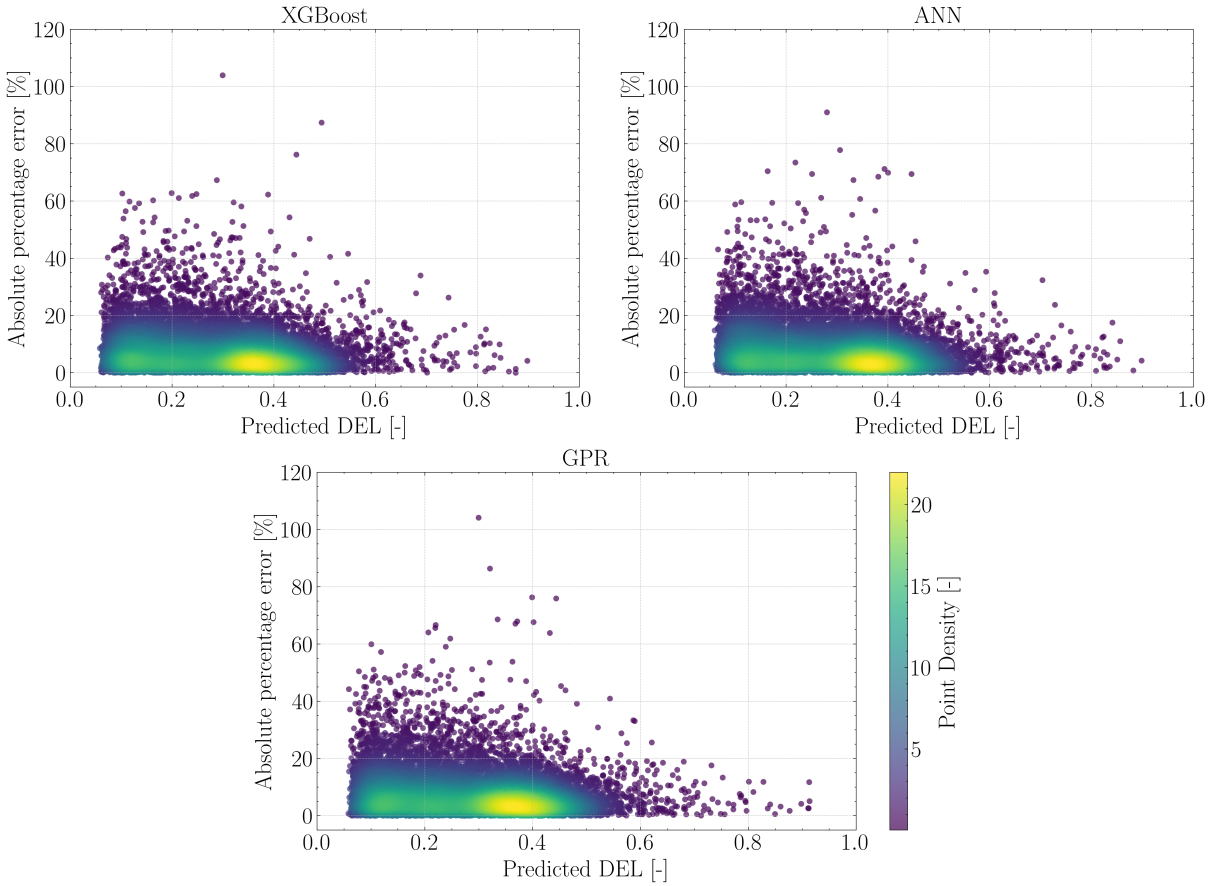


Figure 4.6: B07 blade flapwise DEL absolute percentage error against predicted DEL based on the three models with point density colorbar using the Gaussian KDE method.

Fig. 4.6 presents the relationship between the absolute percentage error (APE) and the predicted DEL values. A clear trend emerges: APE tends to decrease as the predicted DEL increases, indicating that low-magnitude predictions contribute disproportionately to the overall MAPE. This supports the hypothesis that, while MAPE is an intuitive and widely used metric, it can be misleading in scenarios with small target values. Consequently, MAPE should not be the only error metric used in model performance assessment; it is essential to consider at least one complementary metric(such as NRMSE or MAE) for a more balanced evaluation of model performance.

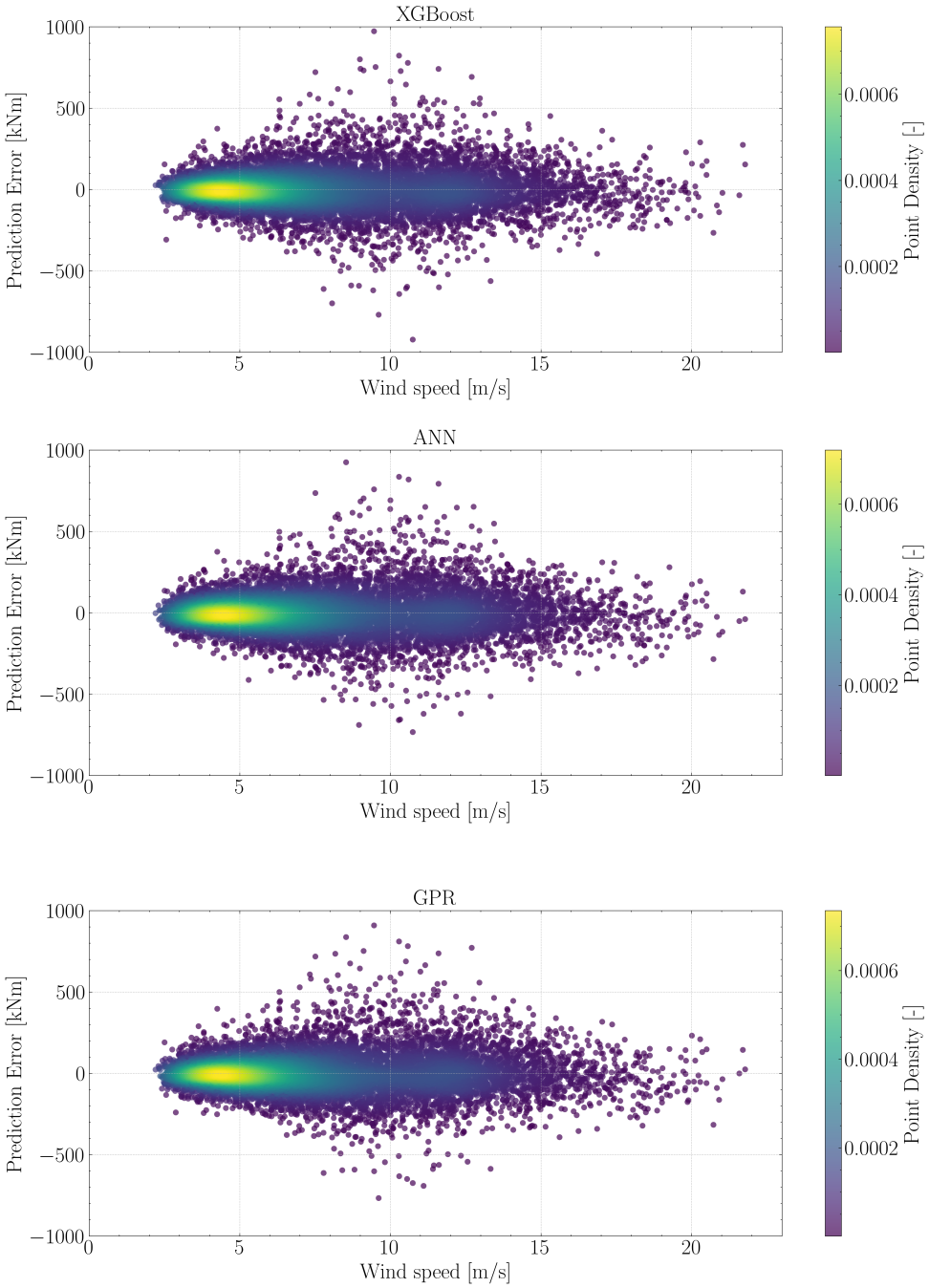


Figure 4.7: B07 blade flapwise DEL prediction error against wind speed with point density colorbar using the Gaussian KDE method.

While Fig. 4.6 illustrates the magnitude of percentage errors, it does not reveal how these errors vary with wind speed. Fig. 4.7 addresses this limitation by displaying the prediction errors as a function of wind speed, enhanced with a point density colorbar. The largest absolute errors are observed around 10 m/s, which aligns with the transition into the rated operating region(where turbines begin regulating power and loads), introducing greater system complexity and increasing model error. In contrast, errors at lower wind speeds (below approximately 7-8 m/s) are generally smaller in magnitude and tightly clustered around zero, as reflected by the high-density band near the center.

Fig. 4.4 & Fig. 4.3 suggest a subtle model bias across different wind speed ranges. To

investigate this further, it is important to examine how the statistical properties of the prediction errors vary with wind speed. However, as seen in Fig. 4.7, the number of data points (number of rows in the SCADA dataset) becomes sparse at higher wind speeds, limiting the statistical reliability of any conclusions drawn in those regions. To address this, Fig. 4.8 presents the number of test samples available within each wind speed bin, defined in 1m/s intervals (e.g., 3.5-4.5m/s, 4.5-5.5m/s etc.), to assess the data coverage and support the interpretation of wind speed-dependent error trends.

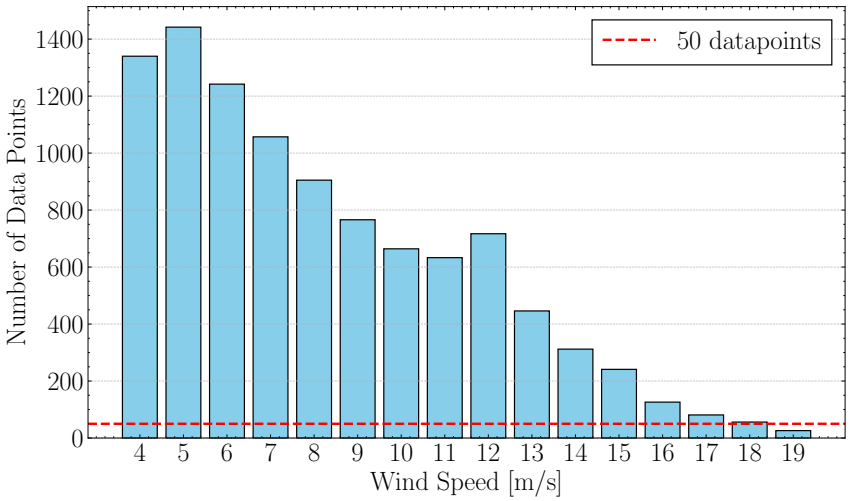
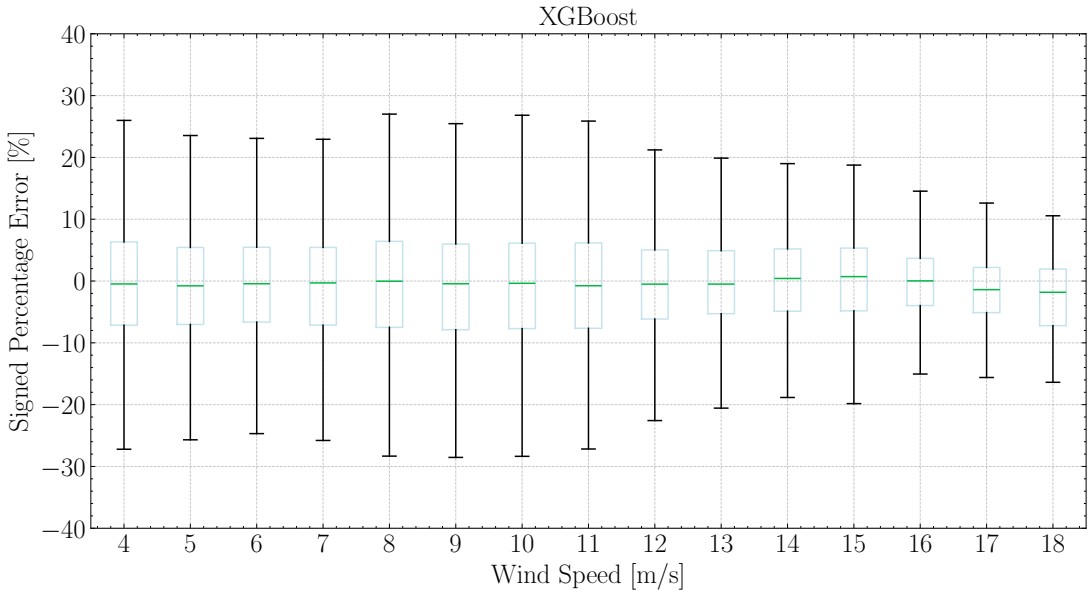


Figure 4.8: Datapoints per wind speed bin for the B07 blade test dataset.

Fig. 4.8 clearly illustrates the decreasing number of data points with increasing wind speed. To ensure statistical reliability in the analysis, a threshold of 50 datapoints per bin is set as the minimum requirement. Based on this criterion, the boxplots in Fig. 4.9 are generated only up to the 18 m/s wind speed bin.



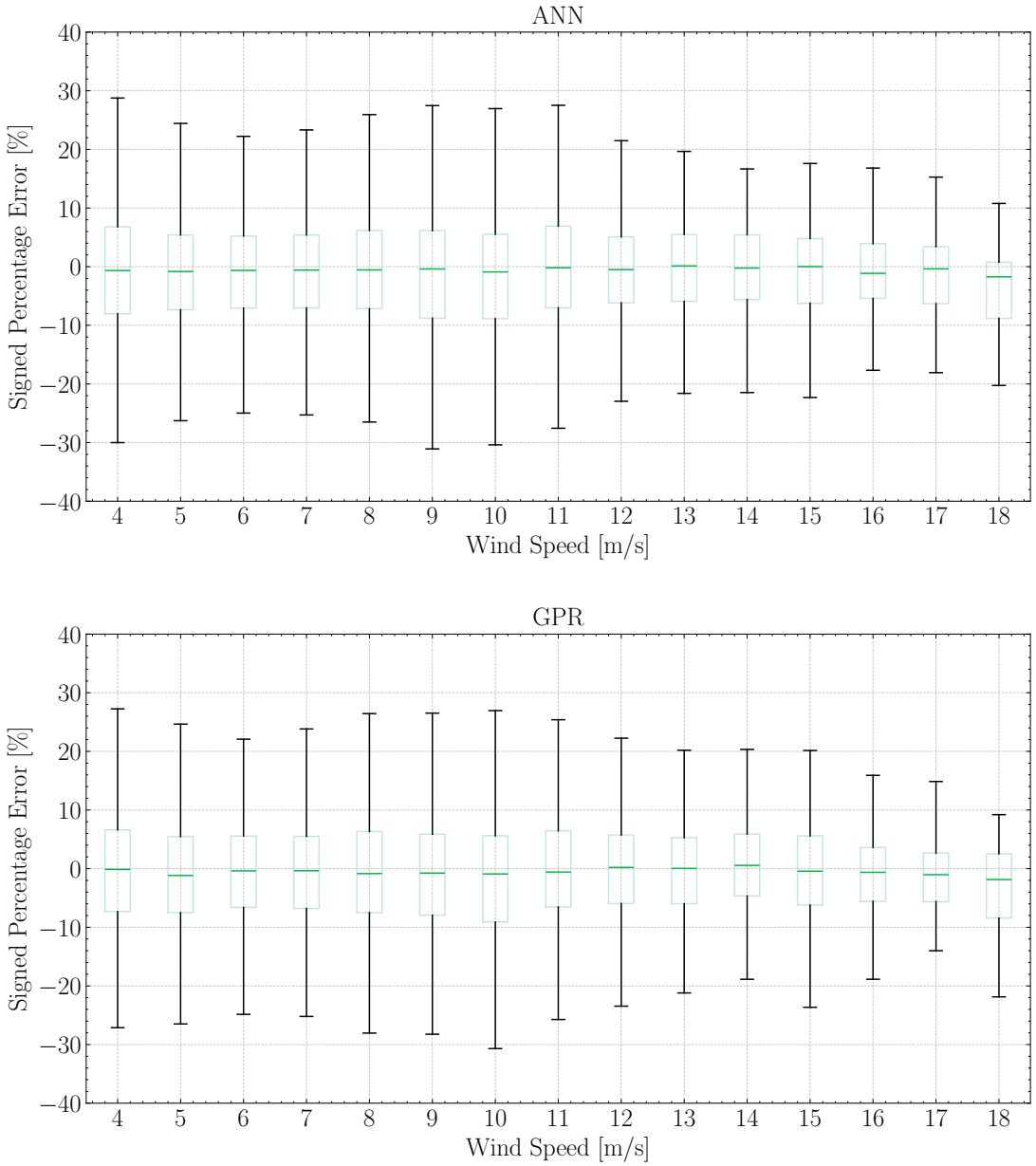


Figure 4.9: B07 blade flapwise DEL signed percentage error against wind speed using boxplots based on the three models. The central line in each box indicates the median signed percentage error, while the box spans the interquartile range (25th to 75th percentiles). Whiskers extend to 1.5 times the interquartile range, and outliers which exist outside this range are not shown.

The boxplots in Fig. 4.9 reveal no evidence of strong model bias across the wind speed range. However, a subtle tendency toward overprediction is observed in all three models, as indicated by the median error line lying slightly below zero for most wind speed bins, particularly those below 13 m/s. This confirms the initial hypothesis of mild overprediction bias based on Fig. 4.4 & Fig. 4.3. The overall signed percentage errors remain within the range of -30% to +30% for all models, reinforcing their comparable performance as previously quantified by error metrics. Notable spreads in error distribution are observed near the rated region (8-11 m/s), due to the increased complexity related to the power & load regulation regime change. A second region of increased spread appears at low wind

speeds, not due to large absolute errors, but rather due to the low DEL values in this range, as supported by the density plot in Fig. 4.7.

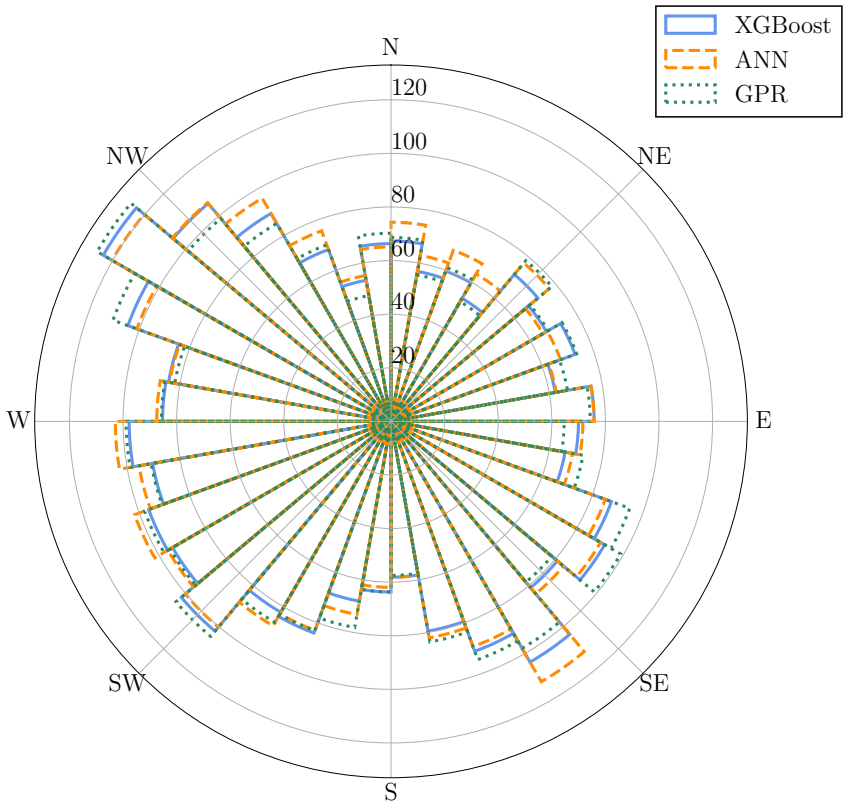


Figure 4.10: B07 blade flapwise DEL error distribution against wind direction(10° bin) based on the three models.

Lastly, Fig. 4.10 illustrates the absolute prediction error of the three models as a function of wind direction, aiming to assess potential directional bias. The absolute prediction error varies between approximately 70 and 120 kNm across different wind sectors, likely reflecting the influence of upstream turbines inducing wake effects on B07 and complicating the blade flapwise DEL prediction. Despite this variation, all three models exhibit consistent performance trends across all directions, indicating no strong directional bias. XGBoost maintains a slight but consistent advantage, achieving the lowest errors across all wind sectors and thereby reaffirming its superior predictive accuracy.

In summary, all three models demonstrate comparable predictive accuracy for blade flapwise DEL, with R^2 values around 94.5% and closely aligned MAPE and NRMSE scores. This indicates that the underlying patterns in the data are effectively captured by the models. Among them, XGBoost emerges as the most practical choice for subsequent case studies due to its slightly superior performance, combined with its robustness, simplicity, low computational cost, and fast inference time, making it more suitable for repeated application across multiple business scenarios. The visual analyses further support these findings, revealing no signs of significant bias in any model, but rather a subtle and consistent tendency toward overprediction across the wind speed range for all the models.

4.2.2 Tower models

This section presents and discusses the results of the tower fore-aft moment DEL models, accompanied by the corresponding plots. In addition to discussing model-specific insights, this section also highlights key differences in model behaviour between the blade and tower cases, offering further perspective on the suitability of each regression methodology. The performance metrics for each model are summarized in Table 4.9, Table 4.10 & Table 4.11. As discussed in Section 3.1, **C08** is excluded from this analysis as it lacks the relevant tower sensor data.

Table 4.9: R^2 scores [%] of tower models using internal data. Green highlights the best model, red the worst.

Wind Turbine	XGBoost	ANN	GPR	Polynomial
B06	92.45	91.92	91.48	76.65
B07	92.21	91.21	90.89	71.84
D07	92.24	91.74	91.20	70.19
D08	94.24	93.41	93.29	80.00
B08	93.73	93.25	92.95	70.83

Table 4.10: MAPE [%] of tower models using internal data. Green highlights the best model, red the worst.

Wind Turbine	XGBoost	ANN	GPR	Polynomial
B06	8.70	9.38	9.66	16.62
B07	9.36	10.15	10.52	20.53
D07	9.07	9.40	9.97	21.39
D08	9.67	10.66	10.85	20.60
B08	8.93	9.46	9.81	23.33

Table 4.11: NRMSE [-] of tower models using internal data. Green highlights the best model, red the worst.

Wind Turbine	XGBoost	ANN	GPR	Polynomial
B06	0.275	0.284	0.292	0.483
B07	0.279	0.297	0.302	0.531
D07	0.279	0.287	0.297	0.546
D08	0.240	0.257	0.259	0.447
B08	0.250	0.260	0.265	0.540

The results of the tower models, summarised in Table 4.9, Table 4.10 & Table 4.11 reflect trends similar to those observed in the blade model analysis. Polynomial regression again proves inadequate, confirming the need for more expressive models to capture the complex, non-linear relationships in the data. All advanced models(XGBoost, ANN, and GPR) demonstrate strong and comparable predictive performance, suggesting that the input features are informative across turbines. Notably, XGBoost consistently outperforms the

other models with a wider performance margin than in the blade model case, further highlighting its robustness, efficiency, and suitability for this task. However, the overall accuracy of the tower models is slightly lower than that of the blade models, which may indicate higher uncertainty or reduced quality in the tower moment measurements, possibly due to sensor placement or calibration limitations, as discussed in Section 4.1.2.

Table 4.12 presents the average as well as the range of each metric, representing the tower model performance across all wind turbines, further confirming the consistently superior performance of XGBoost compared to the other two models.

Table 4.12: Model performance summary for tower fore-aft moment DEL prediction across all wind turbines.

Metric	**Statistic**	**XGBoost**	**ANN**	**GPR**
R^2 [%]	Mean	92.97	92.31	91.96
	Range	92.21 – 94.24	91.21 – 93.41	90.89 – 93.29
MAPE [%]	Mean	9.15	9.81	10.16
	Range	8.70 – 9.67	9.38 – 10.66	9.66 – 10.85
NRMSE [-]	Mean	0.245	0.277	0.283
	Range	0.240 – 0.279	0.257 – 0.297	0.259 – 0.302

As in the blade model comparison, the analysis proceeds with a visual inspection to assess model accuracy and potential biases. Plots for the B07 turbine are presented, as its central location in the farm exposes it to consistent wake effects, making it representative of general turbine behaviour.

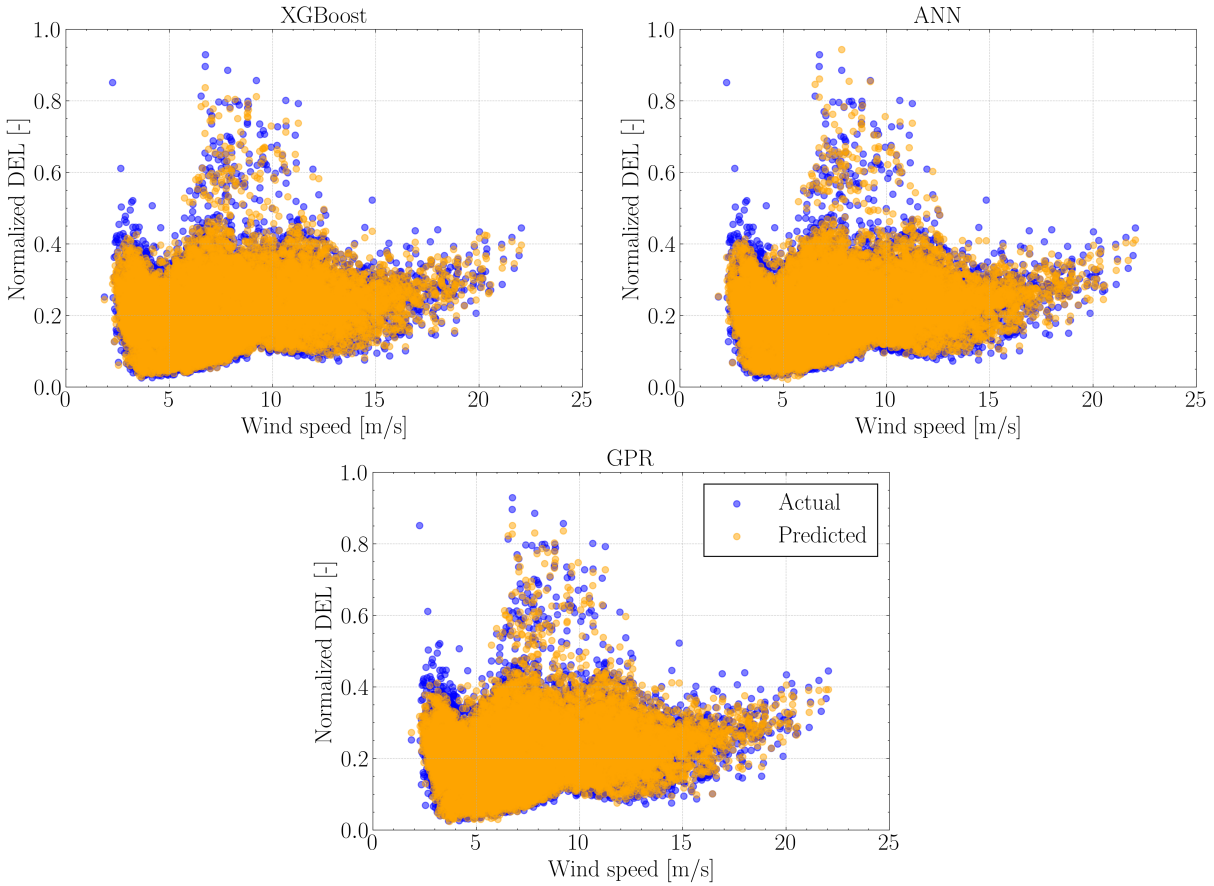


Figure 4.11: B07 tower fore-aft moment actual & predicted DEL against wind speed based on the three models.

As before, Fig. 4.11 is used for an initial visual assessment of the tower fore-aft moment DEL predictions. All three models seem to successfully capture the variance in the data across the full wind speed range. However, a slight underprediction is observed at very low wind speeds (around 4 m/s), a pattern that is examined in more detail later in this section.

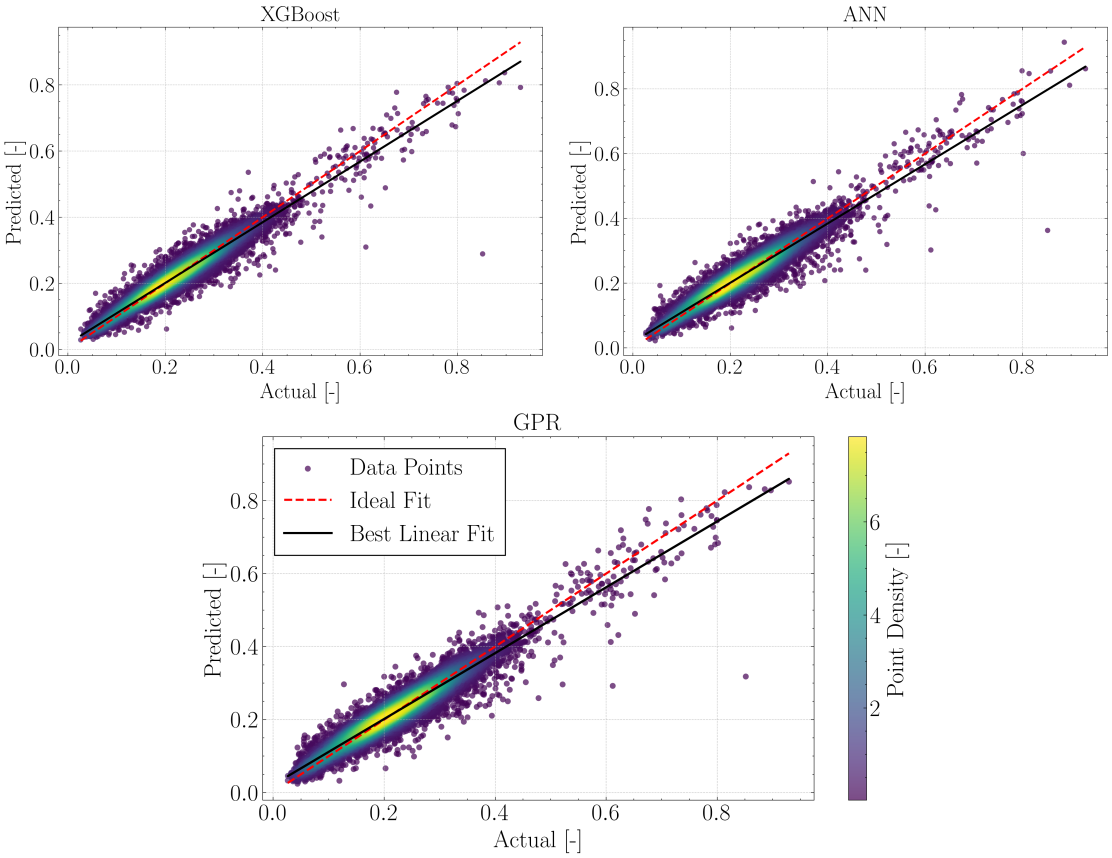


Figure 4.12: B07 tower fore-aft predicted vs actual DEL based on the three models using the Gaussian KDE method.

The observations from Fig. 4.12 plots closely mirror those made for the blade models in Fig. 4.4. The predicted values are generally well-aligned with the actual values, as indicated by the clustering of points around the 1:1 fit line. A consistent pattern is observed across all models: slight overprediction at low DEL values and underprediction at higher DELs, reflected by the Best Linear Fit line lying above the ideal fit line initially and falling below it at higher values.

The error distribution plots in Fig. 4.13 for the tower models show similar patterns to those observed in the blade analysis. All three models produce error distributions that resemble a normal distribution, centered around zero and exhibiting a clear bell-shaped form. XGBoost once again demonstrates superior performance, with a sharper concentration of errors near zero. Its distribution peak reaches approximately 1600 datapoints, compared to 1200-1400 for ANN and GPR, indicating a higher frequency of low-error predictions.

Fig. 4.14 depicts the absolute percentage error against predicted DEL and shows a familiar trend: APE decreases with increasing DEL, while relatively high APE values, and even some quite large outliers, are observed at lower DEL magnitudes. This behaviour is consistent with earlier observations and raises the question of whether the elevated APE is primarily due to the small magnitude of DELs or genuinely poor model performance in the low wind speed region. This issue is investigated further in Fig. 4.15.

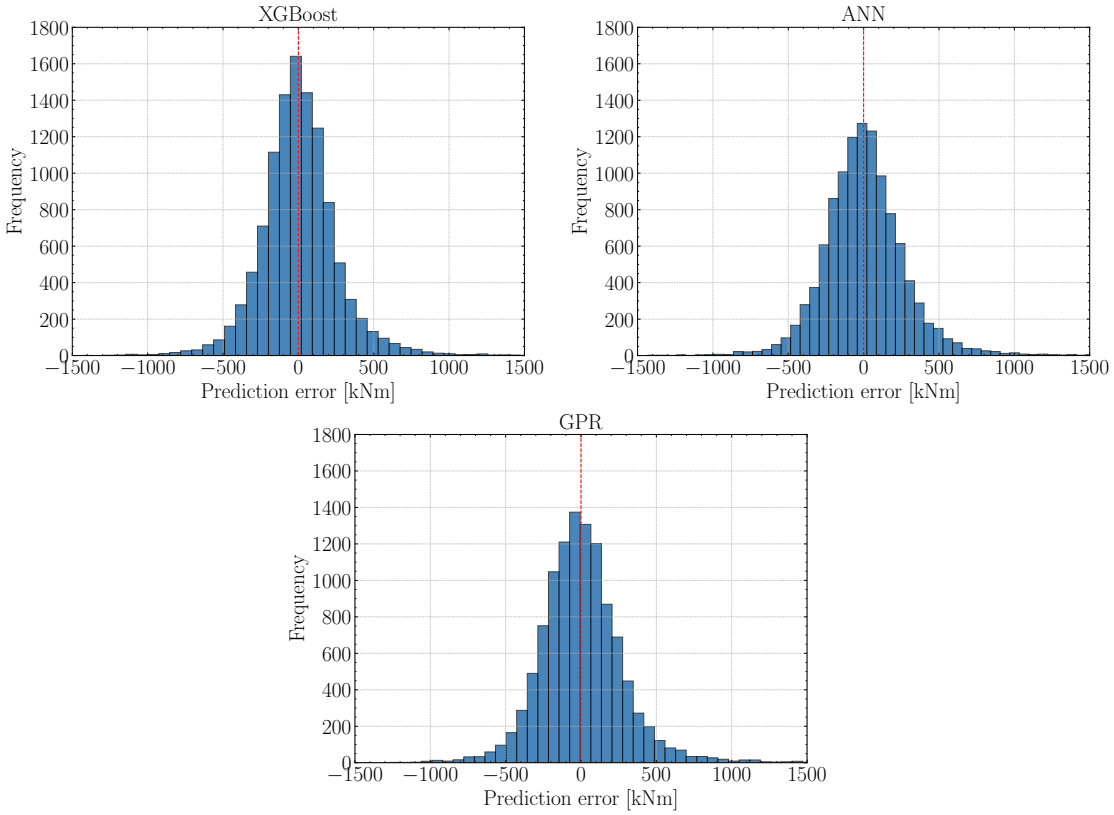


Figure 4.13: B07 tower fore-aft moment DEL error distribution based on the three models.

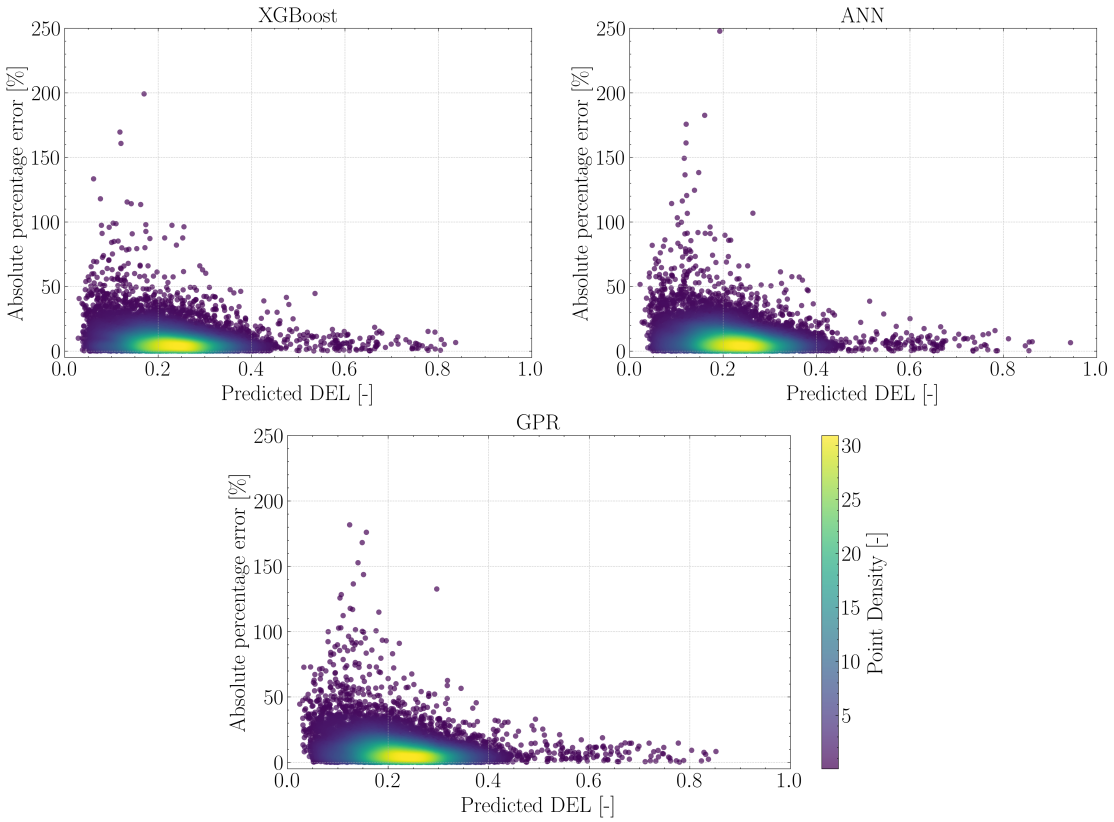


Figure 4.14: B07 tower fore-aft moment DEL absolute percentage error against predicted DEL based on the three models using the Gaussian KDE method.

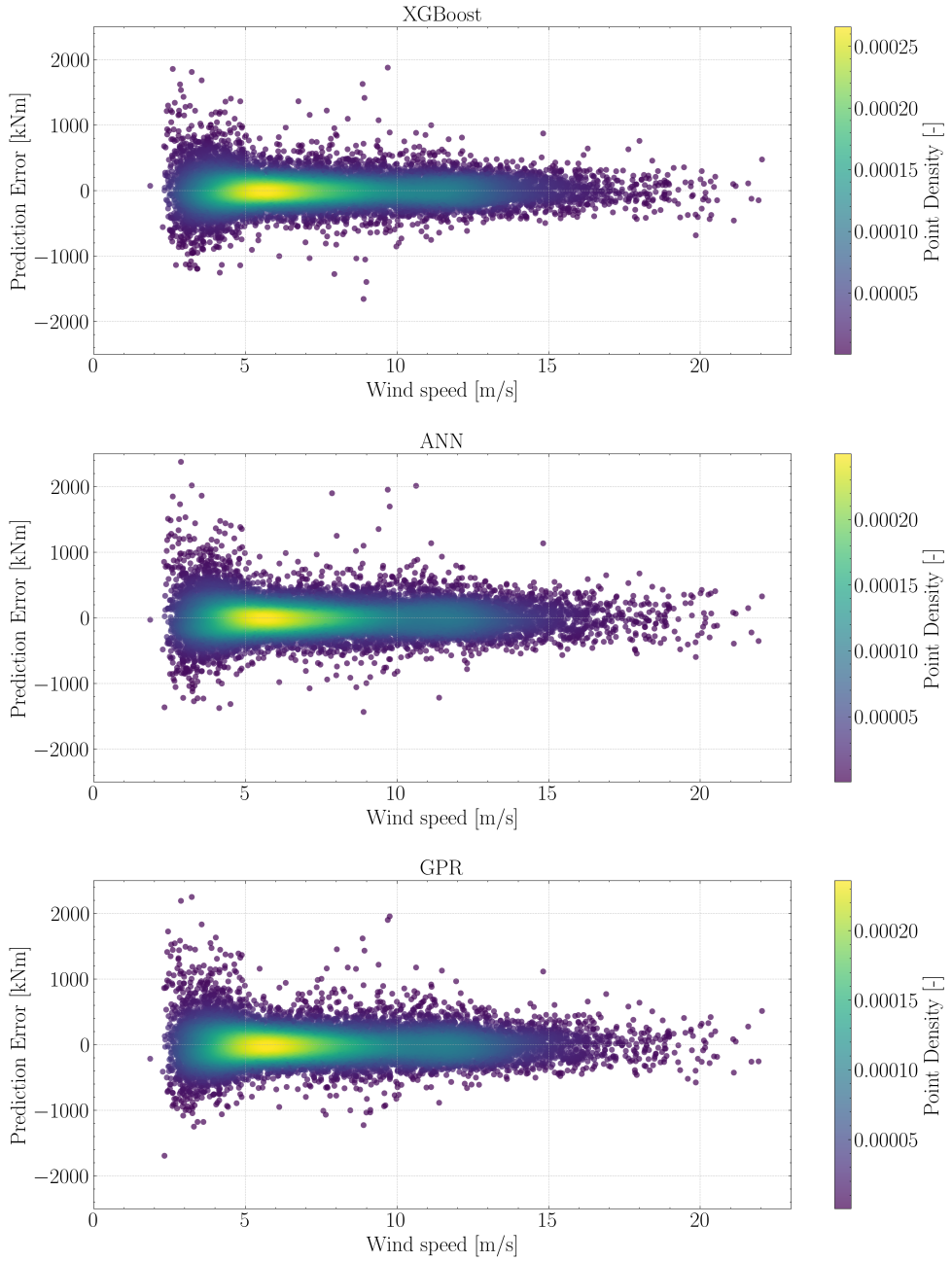


Figure 4.15: B07 tower fore-aft moment DEL prediction error against wind speed with point density colorbar using the Gaussian KDE method.

Fig. 4.15 offers a lot of insight. Similar to the blade model case, the majority of prediction errors in the error scatter plot are concentrated around zero. However, unlike Fig. 4.7, the largest spread here occurs at low wind speeds, with a smaller spread observed near the rated region. This indicates that the tower models struggle more in the low wind speed regime, likely due to transient loading conditions associated with turbine start-up or shutdown. Consequently, the high APE observed at low DEL values in Fig. 4.14 is not merely a result of small magnitudes, but rather reflects genuinely higher prediction errors in this region.

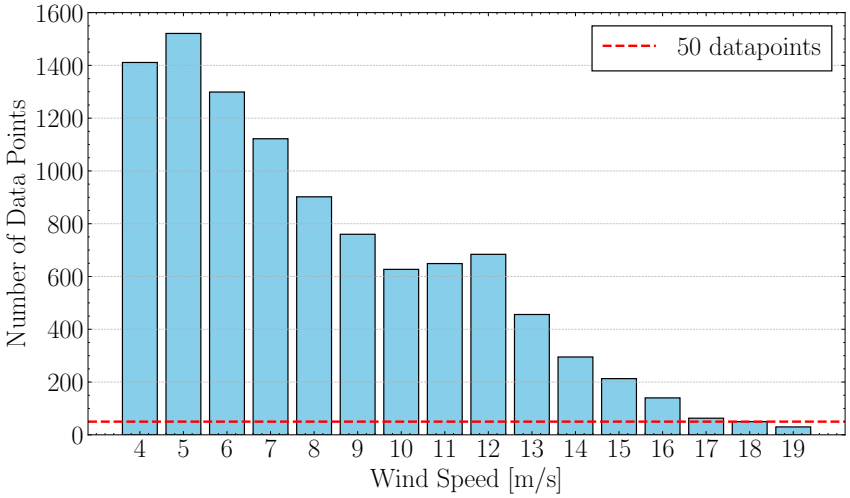
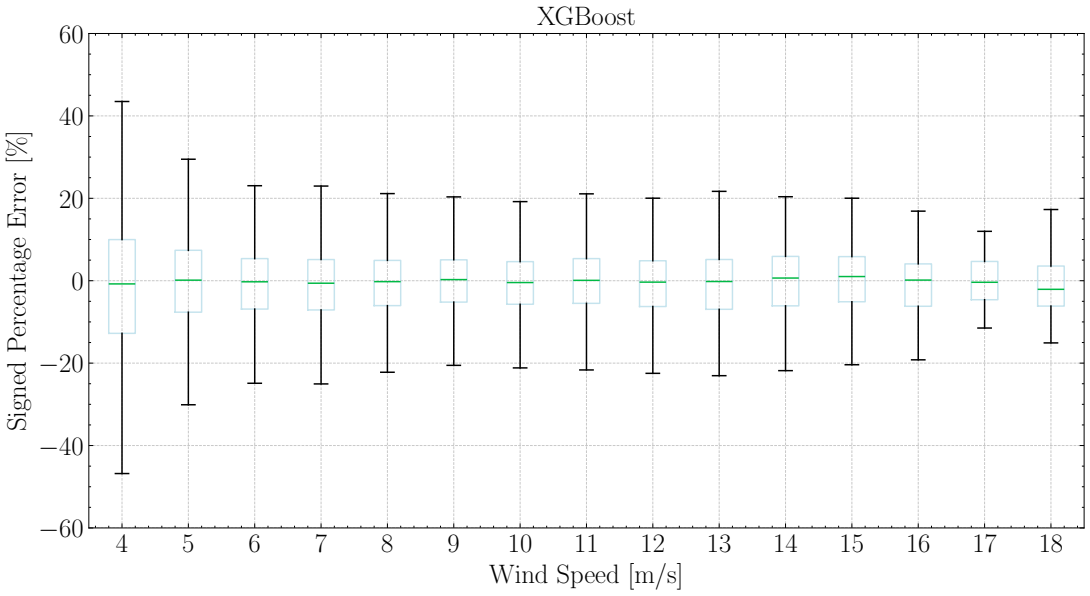


Figure 4.16: Datapoints per wind speed bin for the B07 tower test dataset.

Fig. 4.16 shows the number of datapoints against the wind speed for every wind speed bin. As with the blade case, wind speeds up to 18 m/s satisfy the threshold of at least 50 datapoints per bin. To further investigate the potential bias identified in Fig. 4.11 & Fig. 4.12, boxplots of signed percentage error versus wind speed are generated and presented in Fig. 4.17.

The boxplots in Fig. 4.17 reveal no clear indication of systematic bias in any of the models. The median error lines remain close to zero and exhibit no consistent tendency toward over- or underprediction. As observed in previous plots, the largest errors occur at low wind speeds, particularly around 4 m/s, where the percentage error values range from 50% to +50%. Above 5 m/s, the prediction errors decrease substantially, generally falling within the 30% to +30% range across all models. XGBoost once again demonstrates superior performance, exhibiting a narrower error spread compared to ANN and GPR. Additionally, the increased variability near the rated wind speed region (around 11 m/s) is more pronounced in the ANN and GPR models.



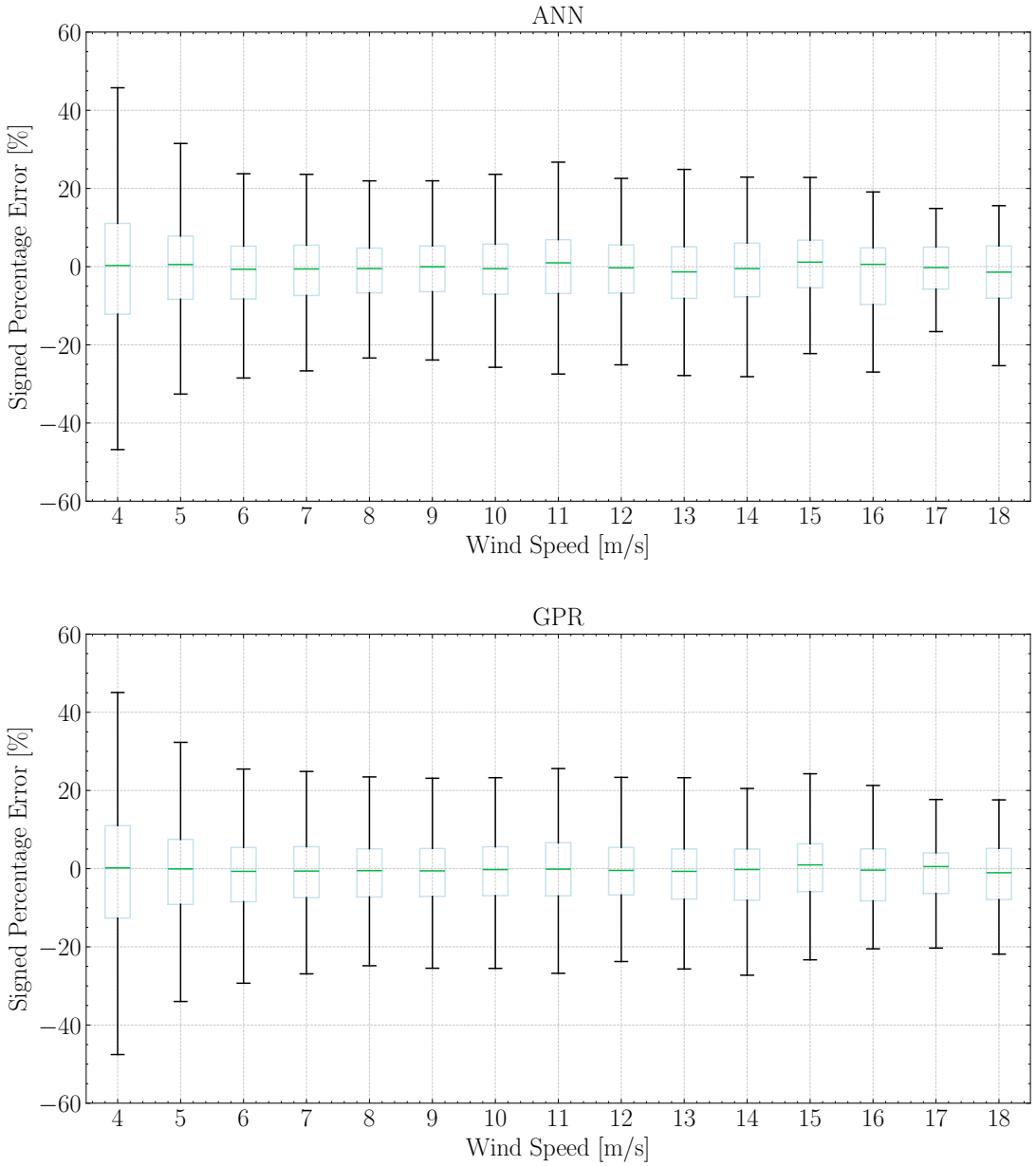


Figure 4.17: B07 tower fore-aft moment DEL signed percentage error against wind speed using boxplots based on the three models. The central line in each box indicates the median signed percentage error, while the box spans the interquartile range (25th to 75th percentiles). Whiskers extend to 1.5 times the interquartile range, and outliers which exist outside this range are not shown.

The plot of absolute prediction error against wind direction in Fig. 4.18 reveals trends similar to those observed in the blade model case. All three models exhibit comparable prediction error patterns across wind directions, with XGBoost consistently achieving the lowest error values in each sector. A notable difference from Fig. 4.10 is that, in this case, the north-west (NW) and south-east (SE) directions stand out as the sectors associated with the highest prediction errors across all models. This suggests a stronger directional dependence of error in the tower moment predictions.

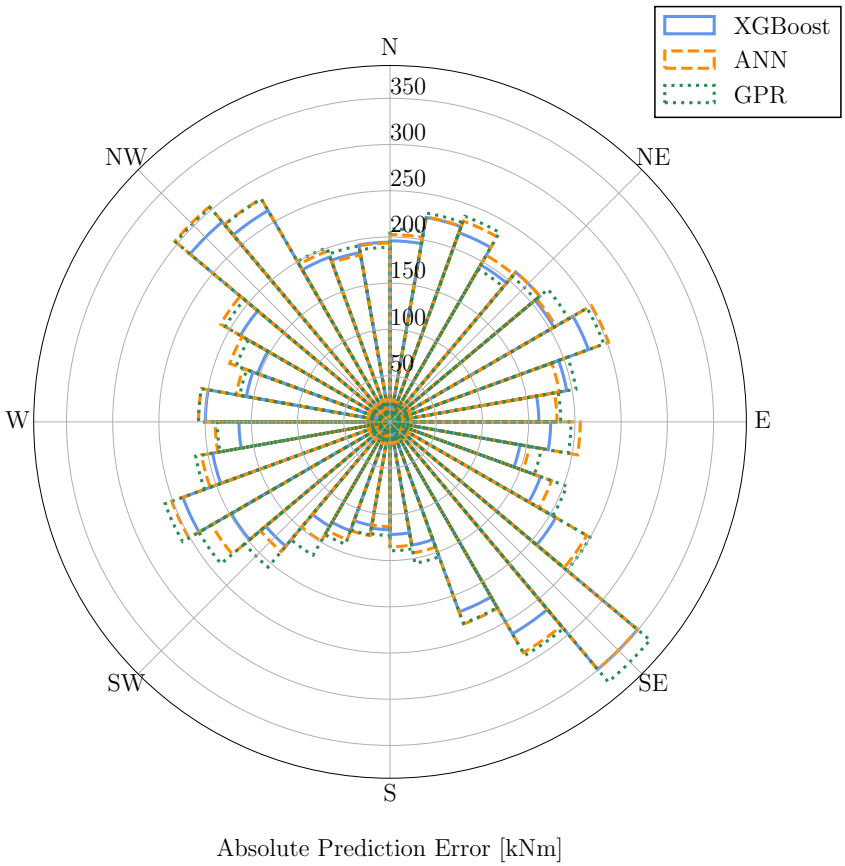


Figure 4.18: B07 tower fore-aft moment DEL error distribution against wind direction(10° bin) based on the three models.

In conclusion, the tower moment models demonstrate strong and consistent performance across all three modelling approaches, with R^2 values in the range of 92-93% and MAPE between 9-10%, supported by similarly close NRMSE scores. These results confirm that the models effectively capture the underlying patterns in the data. Among the candidates, XGBoost consistently outperforms the others across all metrics and visual diagnostics, reinforcing its selection for the subsequent case studies due to its high accuracy, robustness, low computational cost, and ease of implementation. Notably, no systematic bias is observed in any of the models, affirming their reliability across varying operational conditions. When compared to the blade model results, the tower models exhibit lower accuracy, which is a somewhat counterintuitive outcome given that blade loads are typically more dynamic and complex due to structural flexibility. Furthermore, while the blade models tend to struggle near the rated wind speed region, the tower models show greater prediction errors in the low wind speed regime, highlighting varying challenges in load estimation between the two cases.

4.3 Filtering

As analysed in Section 4.2, XGBoost was identified as the most suitable model for predicting both the flapwise blade and tower fore-aft moment DEL signals. Consequently, it is selected as the primary model for conducting the subsequent exploratory case studies. This section investigates the impact of data filtering on model performance, focusing on how different levels of signal preprocessing influence prediction accuracy. Three distinct filtering scenarios are considered:

(a) **Unfiltered** case, where only rows with missing values in the input or target variables are removed.

(b) **Normal Operation filtering**, which excludes datapoints with negative power and applies Gaussian Mixture Model clustering, as described in Section 3.2.

(c) **Start-up & Shutdown filtering**, which introduces an additional criterion: datapoints are discarded if the mean active power is greater than zero while the minimum active power is zero, thereby capturing periods when the turbine was either starting up or shutting down. This condition is likely to affect particularly the tower model robustness in low wind speed conditions, where the errors are significantly high as described in Section 4.2.2.

To ensure a consistent and fair comparison between the three filtering scenarios, a common test dataset is employed across all cases. This test dataset is derived by applying a 70-20-10 train-validation-test split to the Start-up & Shutdown filtered dataset, which represents the strictest level of filtering. The timestamps corresponding to this test split are then used to identify and remove the same datapoints from the Unfiltered and Normal Operation datasets. This approach guarantees that no test samples, either fully or partially, are included in the training or validation sets of any scenario, thus avoiding data leakage. After the removal of these test instances, the remaining data in the Unfiltered and Normal Operation cases is repartitioned using an 80-20 train-validation split. This methodology allows for an unbiased evaluation of the impact of each filtering strategy on model performance.

For clarity, the section is divided into two subsections, the Blade & Tower models, discussing the results of the blade flapwise & tower fore-aft moment DEL respectively.

4.3.1 Blade Models

This section presents and discusses the results obtained for the blade flapwise DEL prediction models. Table 4.13 reports the performance metrics of the XGBoost models for each individual wind turbine under all three filtering scenarios, while Table 4.14 summarises the average performance across all turbines per scenario. Additionally, to investigate the extent of data reduction introduced by each filtering method, the number of training samples (rows) retained in each case is also reported in Table 4.13.

Table 4.13: Filtering case results using XGBoost targeting the blade flapwise DEL. The datapoints refer to the number of rows in the respective training dataset.

Wind Turbine	Unfiltered	Normal Operation Filtering	Start-up & Shutdown Filtering
B06	R^2 = 94.83 % MAPE = 8.73 % NRMSE = 0.227 Datapoints: 79,224	R^2 = 94.92 % MAPE = 8.64 % NRMSE = 0.225 Datapoints: 64,702	R^2 = 94.82 % MAPE = 8.74 % NRMSE = 0.228 Datapoints: 57,948
B07	R^2 = 94.38 % MAPE = 8.15 % NRMSE = 0.237 Datapoints: 92,494	R^2 = 94.38 % MAPE = 8.13 % NRMSE = 0.237 Datapoints: 78,093	R^2 = 94.38 % MAPE = 8.16 % NRMSE = 0.237 Datapoints: 71,094
D07	R^2 = 93.03 % MAPE = 8.44 % NRMSE = 0.264 Datapoints: 80,994	R^2 = 92.97 % MAPE = 8.45 % NRMSE = 0.265 Datapoints: 66,648	R^2 = 92.96 % MAPE = 8.48 % NRMSE = 0.265 Datapoints: 60,140
D08	R^2 = 95.01 % MAPE = 8.67 % NRMSE = 0.223 Datapoints: 74,442	R^2 = 94.89 % MAPE = 8.74 % NRMSE = 0.226 Datapoints: 49,590	R^2 = 94.88 % MAPE = 8.84 % NRMSE = 0.226 Datapoints: 45,722
C08	R^2 = 94.33 % MAPE = 8.36 % NRMSE = 0.238 Datapoints: 78,123	R^2 = 94.36 % MAPE = 8.32 % NRMSE = 0.237 Datapoints: 58,684	R^2 = 94.42 % MAPE = 8.37 % NRMSE = 0.236 Datapoints: 53,443
B08	R^2 = 94.58 % MAPE = 8.67 % NRMSE = 0.233 Datapoints: 80,652	R^2 = 94.62 % MAPE = 8.64 % NRMSE = 0.232 Datapoints: 63,263	R^2 = 94.62 % MAPE = 8.68 % NRMSE = 0.232 Datapoints: 58,686

Table 4.14: Average performance metrics across all turbines for each filtering scenario using XGBoost, with blade flapwise DEL as the target variable.

Metric	Unfiltered	Normal Operation Filtering	Start-up & Shutdown Filtering
R^2 [%]	94.36	94.36	94.35
MAPE [%]	8.50	8.49	8.55
NRMSE [-]	0.237	0.237	0.237

As shown in Table 4.13 & Table 4.14, the application of filtering, despite resulting in a substantial reduction in the number of training datapoints, has a negligible impact on the predictive performance of the XGBoost models. Across all filtering scenarios, the

performance metrics remain remarkably consistent, with variations in R^2, MAPE, and NRMSE being marginal. These results further support the findings from the previous sections regarding the robustness of the XGBoost algorithm. The model demonstrates a strong capacity to generalise, maintaining high accuracy even when trained on datasets that include points potentially outside of normal operating conditions, such as those encountered during start-up or shutdown or various curtailment scenarios. This suggests that the inclusion of such uncommon data does not degrade performance, nor does it confuse the learning process of the model.

4.3.2 Tower Models

This section presents and discusses the results for the tower fore-aft moment DEL prediction models. Table 4.15 displays the performance metrics of the XGBoost models for each wind turbine across the three filtering scenarios, while Table 4.16 provides the corresponding average metrics over all turbines. To assess the impact of each filtering approach on the training dataset size, the number of training samples retained after cleaning is also included in Table 4.15.

Table 4.15: Filtering case results using XGBoost targeting the tower moment FA DEL. The datapoints refer to the number of rows of the respective training dataset.

Wind Turbine	Unfiltered	Normal Operation Filtering	Start-up & Shutdown Filtering
B06	R^2 = 93.65 % MAPE = 8.13 % NRMSE = 0.252 Datapoints: 81,608	R^2 = 93.49 % MAPE = 8.14 % NRMSE = 0.255 Datapoints: 68,436	R^2 = 93.52 % MAPE = 8.16 % NRMSE = 0.254 Datapoints: 60,893
B07	R^2 = 93.43 % MAPE = 8.77 % NRMSE = 0.256 Datapoints: 93,098	R^2 = 93.51 % MAPE = 8.71 % NRMSE = 0.256 Datapoints: 80,068	R^2 = 93.35 % MAPE = 8.85 % NRMSE = 0.258 Datapoints: 72,569
D07	R^2 = 93.35 % MAPE = 8.47 % NRMSE = 0.258 Datapoints: 87,283	R^2 = 93.44 % MAPE = 8.43 % NRMSE = 0.256 Datapoints: 73,791	R^2 = 93.26 % MAPE = 8.56 % NRMSE = 0.260 Datapoints: 66,292
D08	R^2 = 94.98 % MAPE = 9.17 % NRMSE = 0.224 Datapoints: 74,232	R^2 = 94.93 % MAPE = 9.15 % NRMSE = 0.225 Datapoints: 52,856	R^2 = 94.88 % MAPE = 9.15 % NRMSE = 0.226 Datapoints: 48,604
B08	R^2 = 95.25 % MAPE = 8.52 % NRMSE = 0.218 Datapoints: 82,224	R^2 = 95.17 % MAPE = 8.56 % NRMSE = 0.220 Datapoints: 63,032	R^2 = 95.17 % MAPE = 8.54 % NRMSE = 0.220 Datapoints: 58,022

Table 4.16: Average performance metrics across all turbines for each filtering scenario using XGBoost, with tower moment FA DEL as the target variable.

Metric	Unfiltered	Normal Operation Filtering	Start-up & Shutdown Filtering
R^2	94.13	94.11	94.04
MAPE	8.61	8.60	8.65
NRMSE	0.242	0.242	0.244

The results Table 4.15 & Table 4.16 show that the choice of filtering scenario has only a marginal effect on the predictive performance of the tower fore-aft moment DEL models. The differences in R^2, MAPE, and NRMSE across the three cases are minimal, confirming the robustness and consistency of XGBoost when applied to tower data. These findings align with the blade model results, further highlighting the model's ability to handle variability in operational conditions without loss of accuracy.

Notably, the extended shutdown period observed for turbine D08 in early 2022 (Chapter 3) does not appear to hinder the model's ability to predict either the blade or tower DEL. The D08 models demonstrate comparable performance even in the unfiltered case.

4.4 Dataset size

Following the analysis of the filtering effect on model performance in Section 4.3, this section provides an investigation of how the volume of training data influences predictive accuracy. This case study addresses the scientific inquiry: How does the quantity of data affect model performance? Due to time constraints, the analysis focuses exclusively on turbine B07 and employs the XGBoost model. Although Section 4.3 indicated that data filtering has a limited impact, the present study adopts the normal operation filtered dataset. This decision stems from the need to ensure that small datasets (5000 & 10,000 datapoints) are not disproportionately influenced by rare operational conditions such as curtailments or manual shutdowns, which are more prevalent in unfiltered data. The methodology followed in this section consists of the following steps:

1. The normal operation dataset for B07 is split into training (70%), validation (20%), and test (10%) sets using quantile binning.
2. The test set is fixed and reused across all dataset size cases to ensure consistency in evaluation.
3. The remaining data is shuffled and subsets of specific sizes are selected using quantile binning to construct the combined training-validation set ($X_{train,val}$).
4. Each $X_{train,val}$ set is then split into 80% training and 20% validation, and the model is subsequently trained and evaluated. The dataset splits for each case are reported in Table 4.17.

The results of the study, expressed by the performance metrics, are reported in Table 4.17. The dataset size influence is visualised in Fig. 4.19 where the R^2 & MAPE of each dataset

size scenario is plotted against the number of datapoints corresponding to the training dataset.

Table 4.17: Performance metrics for different dataset sizes using XGBoost, targeting B07 blade flapwise and tower moment FA DEL. The dataset split refers to training-validation-test percentages.

Training datapoints	Blade Model (B07)	Tower Model (B07)	Blade Split [%]	Tower Split [%]
5,000	R^2 = 92.58 % MAPE = 9.83 % NRMSE = 0.272	R^2 = 88.60 % MAPE = 11.49 % NRMSE = 0.338	40-10-50	43-10-47
10,000	R^2 = 93.27 % MAPE = 9.43 % NRMSE = 0.259	R^2 = 89.08 % MAPE = 10.83 % NRMSE = 0.330	54-13-33	55-14-31
20,000	R^2 = 93.77 % MAPE = 9.18 % NRMSE = 0.250	R^2 = 90.06 % MAPE = 10.36 % NRMSE = 0.315	64-16-20	66-16-18
40,000	R^2 = 94.18 % MAPE = 8.72 % NRMSE = 0.241	R^2 = 90.98 % MAPE = 9.83 % NRMSE = 0.300	71-18-11	72-18-10
60,000	R^2 = 94.51 % MAPE = 8.54 % NRMSE = 0.234	R^2 = 91.35 % MAPE = 9.63 % NRMSE = 0.294	74-18-8	75-18-7
80,000	R^2 = 94.81 % MAPE = 8.32 % NRMSE = 0.228	R^2 = 91.67 % MAPE = 9.50 % NRMSE = 0.289	75-19-6	76-19-5

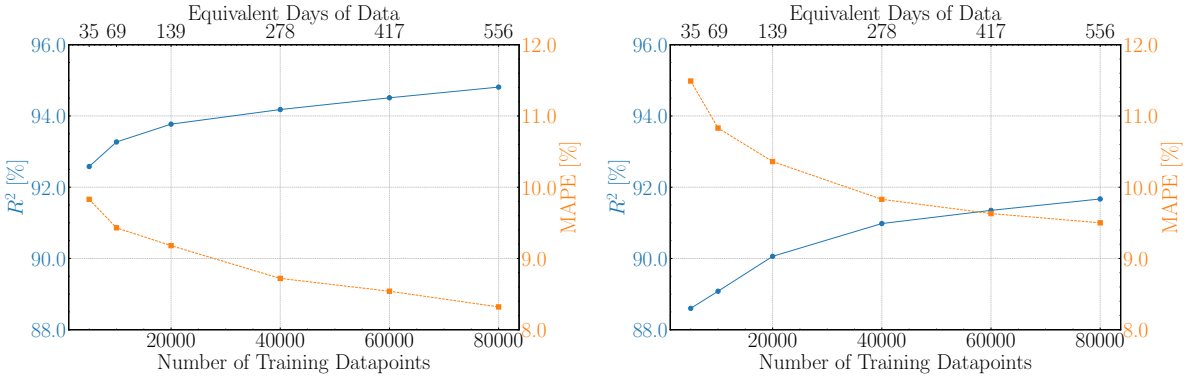


Figure 4.19: R^2 & MAPE against number of datapoints for the blade model (left) and for the tower model (right). The plots refer to the B07 normal operation data using XGBoost.

The results presented in Table 4.17 & Fig. 4.19 confirm the expected trend that increasing the training dataset size leads to improved model performance for both the blade and tower models. In particular, a more pronounced improvement is observed in the smaller dataset sizes, where the gradients of R^2 and MAPE are steeper. Nevertheless, even when

the training set exceeds a full year of operational data, the models continue to benefit from additional information, albeit with diminishing gains. Importantly, the inclusion of more data does not appear to introduce noise or confusion into the learning process, suggesting that the XGBoost model effectively leverages the added samples to enhance generalisation.

4.5 Local Configuration

The analysis proceeds with the Local Configuration case, which investigates the question: How does the input from neighbouring wind turbines influence model performance? The motivation stems from the hypothesis that directional effects in the wind field impact the accuracy of load predictions. Specifically, it is suspected that increased turbulence from waked inflow, which is associated with certain wind directions, complicates the estimation of DEL. This hypothesis is supported by the directional absolute prediction error distribution shown in Fig. 4.20, derived from the model comparison results in Section 4.2.1 for turbines B07 and B08. The figure reveals that while both models perform similarly when the wind originates from the NW to E sector, significant discrepancies arise in the E to NW sector: B07 exhibits much higher prediction errors compared to B08. This difference can be attributed to the wake-free inflow conditions of B08 in those directions, in contrast to B07, which consistently operates under waked conditions, as illustrated in Fig. 4.21.

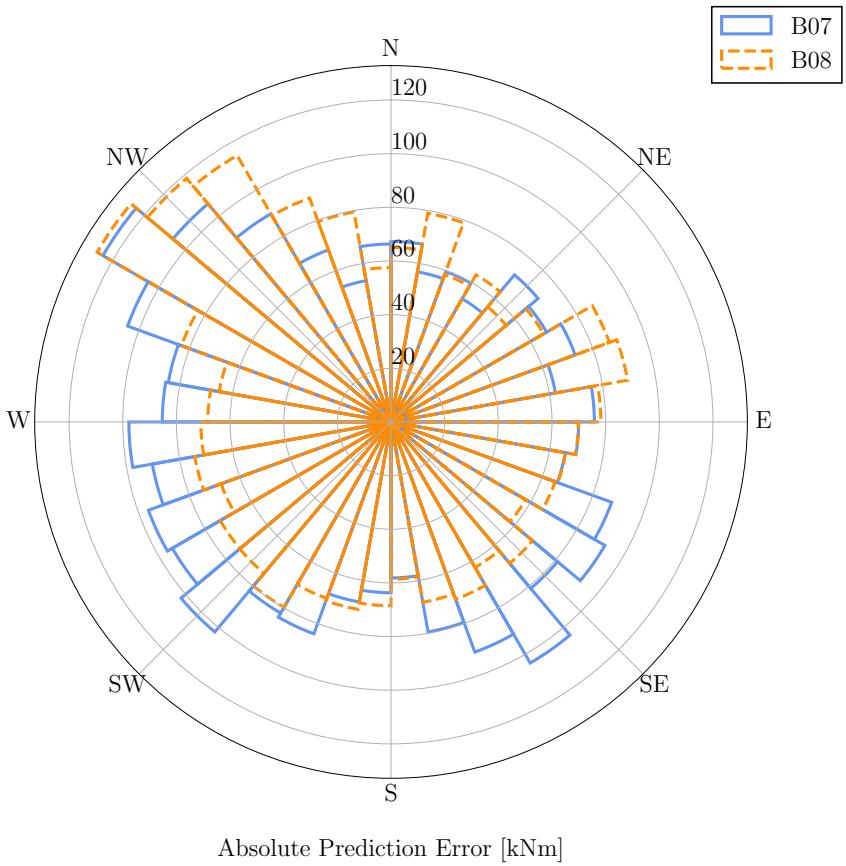


Figure 4.20: B07 & B08 blade flapwise DEL error distribution against wind direction(10° bin) based on XGBoost.

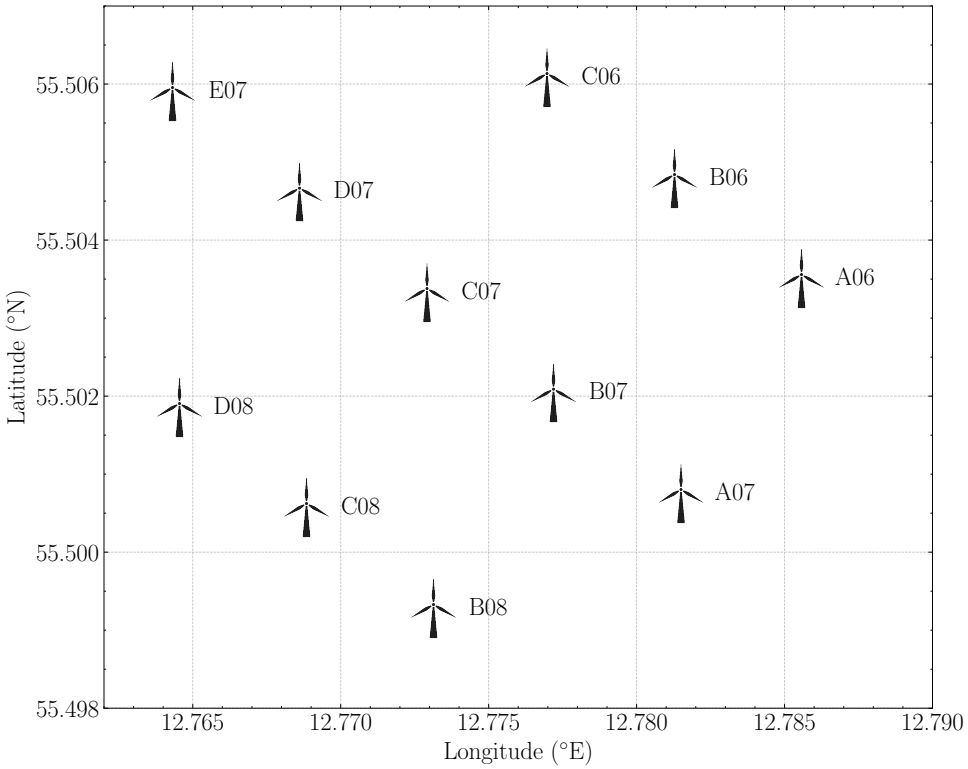


Figure 4.21: Lillgrund offshore wind farm zoomed in on the wind turbines with DEL data.

To assess the influence of neighbouring wind turbines on model performance, a targeted local configuration analysis is conducted. Given the time constraints, it is not feasible to perform this analysis for all turbines with available DEL data. Consequently, turbine B07 is selected as the representative case. B07 is centrally located within the wind farm layout (Fig. 4.21), allowing for the assessment of upstream interactions from multiple wind directions. This configuration is particularly advantageous for identifying directional dependencies and for examining the potential impact of wake-induced turbulence, especially in directions aligned with the prevailing wind, as suggested by the site's wind rose in Fig. 3.3.

The input features for B07 remain consistent with those defined in the Feature Importance analysis in Section 4.1. For the blade load models, the selected internal features include the mean and standard deviation of Wind Speed, Active Power, Nacelle Position, Pitch Angle, and Generator speed. In contrast, the tower load models also incorporate the minimum and maximum signals of those variables.

When incorporating signals from neighbouring turbines, only a subset of variables is considered: the mean values of Wind Speed, Active Power, Nacelle Position, and Pitch Angle. This feature selection is primarily motivated by the need to reduce computational cost. To achieve this, the dataset is limited to variables identified as most influential in the Feature Importance study (Section 4.1), thereby excluding Generator Speed and higher-order statistical descriptors.

The unfiltered dataset for all the wind turbines is used in order to increase the generalisability of the results while maintaining predicting accuracy (see Section 4.3).

The case study is divided into two subsections, Blade & Tower models, each addressing

the results for the blade flapwise and tower fore-aft moment DEL, respectively.

4.5.1 Blade models

In this section, the results regarding the performance of the models targeting the B07 blade flapwise DEL are reported and discussed. Table 4.18, Table 4.19 & Table 4.20 present the performance of the local configuration models for predicting the B07 blade flapwise DEL compared to the B07 internal model while Table 4.21 displays the total number of datapoints used across the training, validation and test sets for the internal and local configuration models respectively.

Table 4.18: R^2 scores [%] for B07 blade flapwise DEL prediction using internal and local configuration models with XGBoost.

Turbine Added	B07 Internal	Local Configuration	ΔR^2 [%]
D07	95.49	96.09	0.60
C08	95.49	96.38	0.89
B08	95.49	96.14	0.65
A07	95.49	96.06	0.57
A06	95.49	95.72	0.23
B06	95.49	95.91	0.42
D08	95.49	96.38	0.89
All	95.49	96.47	0.98

Table 4.19: MAPE [%] for B07 blade flapwise DEL prediction using internal and local configuration models with XGBoost.

Turbine Added	B07 Internal	Local Configuration	ΔMAPE [%]
D07	8.28	8.03	-0.25
C08	8.28	7.92	-0.36
B08	8.28	8.02	-0.26
A07	8.28	8.53	0.25
A06	8.28	8.54	0.26
B06	8.28	8.24	-0.04
D08	8.28	7.84	-0.44
All	8.28	8.21	-0.07

Table 4.20: NRMSE [-] for B07 blade flapwise DEL prediction using internal and local configuration models with XGBoost.

Turbine Added	**B07 Internal**	**Local Configuration**	**ΔNRMSE** [%]
D07	0.212	0.198	-6.60
C08	0.212	0.190	-10.38
B08	0.212	0.196	-7.55
A07	0.212	0.198	-6.60
A06	0.212	0.207	-2.36
B06	0.212	0.202	-4.72
D08	0.212	0.190	-10.38
All	0.212	0.188	-11.32

Table 4.21: Number of datapoints used in training, validation and test sets for each local configuration compared to the internal B07 model.

Turbine Added	**B07 Internal**	**Local Configuration**
D07	125,777	106,760
C08	125,777	103,309
B08	125,777	106,253
A07	125,777	51,524
A06	125,777	53,798
B06	125,777	104,935
D08	125,777	98,116
All	125,777	41,717

The first column in each table specifies the local configuration scenario, where each row corresponds to the inclusion of input data from a specific neighbouring turbine. The final row ('All') represents the case in which input from all listed turbines is simultaneously incorporated. The second column reports the performance metric or number of datapoints of the internal model using only B07's own input signals. The third column provides the corresponding performance metrics for each local configuration. The final column in Table 4.18, Table 4.19 & Table 4.20 indicate the relative improvement or deterioration of the model's predictive performance compared to the internal model. Positive values in ΔR^2 and negative values in ΔMAPE or ΔNRMSE suggest improved accuracy with the inclusion of neighbouring turbine data.

The results in Table 4.18, Table 4.19 & Table 4.20 demonstrate that incorporating input data from neighbouring wind turbines enhances the predictive performance of the blade flapwise DEL model. Notably, the most substantial improvements are observed when data from upstream turbines, particularly those aligned with the prevailing wind direction (C08 and D08), are included. These turbines likely provide additional information about the incoming flow conditions before it reaches B07, enabling the model to better capture the effects of wake-induced turbulence and directional variability. The configuration that includes data from all neighbouring turbines yields the highest overall performance, suggesting that the combined spatial context significantly enriches the model's understanding

of the operational situation. This highlights the importance of local flow conditions and upstream interactions in accurately estimating structural loads.

In addition, it is important to address the observation that MAPE does not consistently improve across the local configuration models. As previously discussed (Section 3.3.5), MAPE is a biased metric, as it is strongly influenced by the magnitude of the actual DEL values and tends to be inflated in cases of low-magnitude predictions. Therefore, relying solely on MAPE may not accurately reflect improvements in model performance. To gain a more solid understanding of the impact of local inputs, it is necessary to examine the distribution of errors with respect to wind direction, as well as the distribution of prediction errors across the full range of DEL values. A comparative analysis between the internal and local B07 models, addressing these aspects, is presented in Fig. 4.22.

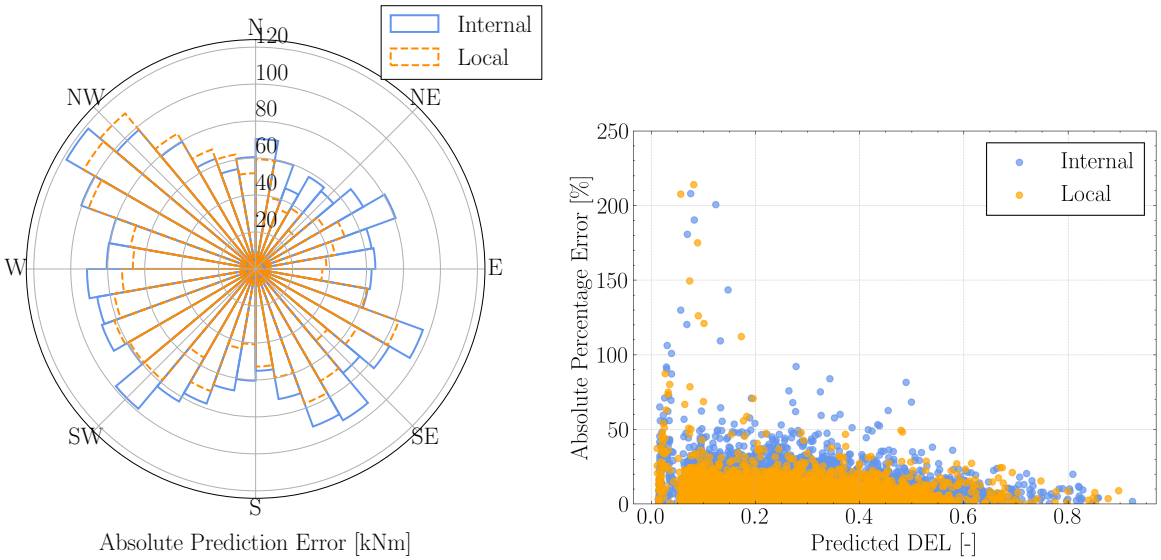


Figure 4.22: Directional absolute error distribution (left) and absolute percentage error vs. predicted DEL (right) for the B07 internal and local ('All') blade models using XGBoost.

The results illustrated in Fig. 4.22 further validate the superiority of the local configuration model. The left subplot shows that the elevated prediction errors previously observed in the E to NW wind directions are significantly mitigated when incorporating input from neighbouring turbines. This confirms that upstream information enhances the model's ability to handle directional variability, particularly in regions affected by wake interactions. Moreover, the right subplot demonstrates that the local model consistently yields lower percentage errors than the internal model, indicating improved accuracy and generalisation.

This raises the question of why the overall MAPE does not reflect this improvement. The most plausible explanation lies in the differing sizes of the test datasets (see Table 4.21). The internal model is evaluated on approximately 12,500 test datapoints, while the local model's test set contains around 4,000. This discrepancy may result in a skewed MAPE for the local model, particularly if it includes a higher proportion of low-DEL values(where percentage errors are naturally inflated) and lacks sufficient representation in the mid-to-high DEL range, which would otherwise dampen the overall MAPE.

Thus, while MAPE is often used as an intuitive performance metric, it is sensitive to the distribution of the target variable and should be interpreted cautiously, especially when

comparing models trained or tested on datasets of different sizes or compositions. In this context, the percentage error distribution plot provides a more reliable assessment of model behaviour, reinforcing the superior performance of the local configuration.

4.5.2 Tower models

This section presents and discusses the results for the models predicting the tower fore-aft moment DEL of turbine B07. Table 4.22, Table 4.23 & Table 4.24 present the performance of the local configuration models based on XGBoost.

Table 4.22: R^2 scores [%] for B07 tower FA moment DEL prediction using internal and local configuration models with XGBoost.

Turbine Added	B07 Internal	Local Configuration	ΔR^2 [%]
D07	91.64	92.21	0.57
B08	91.64	92.58	0.94
A07	91.64	92.74	1.10
A06	91.64	92.80	1.16
B06	91.64	92.16	0.52
D08	91.64	92.87	1.23
All	91.64	92.61	0.97

Table 4.23: MAPE [%] for B07 tower FA moment DEL prediction using internal and local configuration models with XGBoost.

Turbine Added	B07 Internal	Local Configuration	ΔMAPE [%]
D07	11.09	10.54	-0.55
B08	11.09	10.12	-0.97
A07	11.09	10.82	-0.27
A06	11.09	11.46	0.37
B06	11.09	10.73	-0.36
D08	11.09	10.18	-0.91
All	11.09	10.13	-0.96

Table 4.24: NRMSE [-] for B07 tower FA moment DEL prediction using internal and local configuration models with XGBoost.

Turbine Added	B07 Internal	Local Configuration	ΔNRMSE [%]
D07	0.289	0.279	-3.46
B08	0.289	0.272	-5.88
A07	0.289	0.269	-6.92
A06	0.289	0.268	-7.27
B06	0.289	0.280	-3.11
D08	0.289	0.267	-7.61
All	0.289	0.272	-5.88

Table 4.25: Number of datapoints used in training, validations and test sets for each local configuration compared to the internal B07 model.

Turbine Added	B07 Internal	Local Configuration
D07	126,742	112,580
B08	126,742	107,065
A07	126,742	50,330
A06	126,742	52,911
B06	126,742	107,041
D08	126,742	97,284
All	126,742	38,239

Similar to the blade model results, a notable improvement in predictive performance is observed when incorporating input from neighbouring turbines. To support the error analysis, the directional distribution of prediction errors and the percentage error variation across the full range of predicted DEL values are presented in Fig. 4.23.

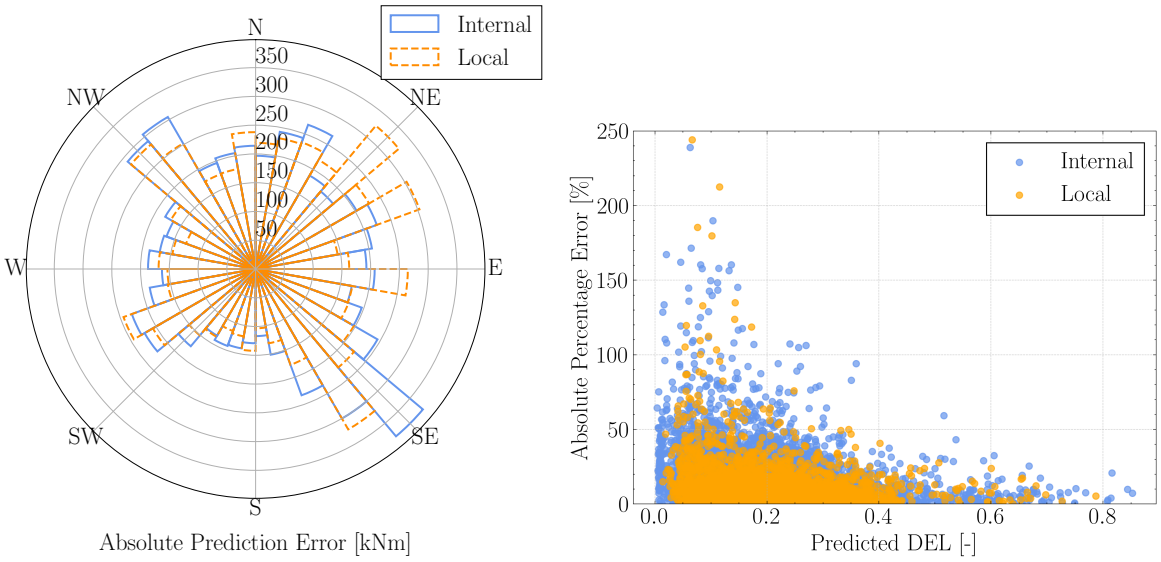


Figure 4.23: Directional absolute error distribution (left) and absolute percentage error vs. predicted DEL (right) for the B07 internal and local ('All') tower models using XGBoost.

Fig. 4.23 presents a nuanced view of the performance differences between the internal and local tower models. The directional absolute error distribution (left) indicates that both models exhibit comparable error magnitudes across most wind directions, with no substantial advantage observed for the local configuration. This suggests that, in terms of absolute error, the inclusion of neighbouring turbine data does not lead to significant improvements.

However, the right subplot, depicting the absolute percentage error against the predicted DEL, reveals a more favourable outcome for the local model. The percentage errors are consistently lower across the majority of the DEL range, corroborating the slightly improved MAPE values reported in Table 4.23. Nevertheless, it is important to note the evident absence of datapoints in the low-DEL region of the local model's test dataset. Given

that percentage error is highly sensitive to small denominators, this under-representation can artificially lower the average MAPE.

This observation reinforces the earlier discussion in the blade model section regarding the limitations of MAPE as a standalone metric. It highlights the need to consider dataset composition when interpreting performance metrics, particularly when comparing models tested on datasets of different sizes or distributions. Overall, while the local model shows some signs of improved relative accuracy, the findings must be contextualised within the underlying data characteristics.

4.6 Neighbouring DEL

In this section, the potential of using DEL data from neighbouring wind turbines to enhance the prediction accuracy of damage equivalent loads (DELs) is explored. The motivation stems from a practical operational scenario: in a wind farm equipped with load sensors on each turbine, sensor malfunctions are inevitable. When a sensor fails, data-driven models become essential to estimate the missing DEL values. The central question addressed here is whether incorporating DEL measurements from adjacent turbines as input features can improve the predictive performance of such models. By analysing the extent of improvement introduced by this additional spatial information, the study aims to assess the practical value of leveraging turbine interdependencies in real-time condition monitoring and predictive maintenance strategies.

The methodology adopted in this case study focuses on predicting the DEL for turbine B06. For each experiment, the model is trained using B06's internal SCADA features, specifically wind speed, active power, nacelle position, pitch angle, and generator speed, augmented with the DEL measurement from a neighbouring turbine. The last case ('All') refers to the selection of the DEL of all the previously reported turbines as input. The objective is to quantify the added predictive value of this additional input by comparing model performance metrics across different neighbour selections. Importantly, the unfiltered dataset (see Section 3.2) is used consistently to secure a fair model comparison. The selection of B06 is deliberate: it is the most spatially isolated turbine among those with available load data (see Fig. 4.21). This positioning enables a meaningful investigation into how the spatial configuration within the wind farm, including the distance between turbines, influences the predictive utility of neighbouring DEL information. To enhance clarity, the analysis is divided into two distinct subsections, Blade & Tower models, allowing a focused discussion of each result set.

4.6.1 Blade models

In this section, the results regarding the investigation of the blade flapwise DEL prediction models are reported and discussed. The performance metrics are presented in Table 4.26, Table 4.27 & Table 4.28.

Table 4.26: R^2 scores [%] for B06 blade flapwise DEL prediction using internal and neighbouring turbine DEL input models with XGBoost.

Neighbouring Turbine	B06 Internal	Neighbour DEL Input	ΔR^2 [%]
B07	96.23	96.72	0.49
D07	96.23	96.45	0.22
B08	96.23	96.67	0.44
C08	96.23	96.45	0.22
D08	96.23	96.47	0.24
All	96.23	96.94	0.71

Table 4.27: MAPE [%] for B06 blade flapwise DEL prediction using internal and neighbouring turbine DEL input models with XGBoost.

Neighbouring Turbine	B06 Internal	Neighbour DEL Input	ΔMAPE [%]
B07	8.94	8.45	-0.49
D07	8.94	8.61	-0.33
B08	8.94	8.65	-0.29
C08	8.94	8.65	-0.29
D08	8.94	8.85	-0.09
All	8.94	8.31	-0.63

Table 4.28: NRMSE [-] for B06 blade flapwise DEL prediction using internal and neighbouring turbine DEL input models with XGBoost.

Neighbouring Turbine	B06 Internal	Neighbour DEL Input	ΔNRMSE [%]
B07	0.194	0.181	-6.70
D07	0.194	0.188	-3.09
B08	0.194	0.183	-5.67
C08	0.194	0.188	-3.09
D08	0.194	0.188	-3.09
All	0.194	0.175	-9.79

The results from the blade models in Table 4.26, Table 4.27 & Table 4.28 demonstrate that incorporating neighbouring turbine DEL as an additional input consistently enhances the predictive performance of the model targeting B06 blade flapwise DEL, albeit with modest gains. Even turbines positioned relatively far from B06, such as C08 and D08, which, unlike B06, operate mostly under free-wake conditions, contribute positively to the model's accuracy. Notably, the best performance is achieved when the DELs from all neighbouring turbines are included simultaneously. This suggests that the additional spatial load information enriches the model's understanding of the overall loading conditions within the wind farm, leading to improved predictions without introducing noise or overfitting.

4.6.2 Tower models

This section presents and discusses the results of the tower fore-aft moment DEL prediction models. The corresponding performance metrics are reported in Table 4.29, Table 4.30 & Table 4.31.

Table 4.29: R^2 scores [%] for B06 tower FA moment DEL prediction using internal and neighbouring turbine DEL input models with XGBoost.

Neighbouring Turbine	B06 Internal	Neighbour DEL Input	ΔR^2 [%]
B07	91.83	92.79	0.96
D07	91.83	93.48	1.65
B08	91.83	92.44	0.61
D08	91.83	93.55	1.72
All	91.83	94.02	2.19

Table 4.30: MAPE [%] for B06 tower FA moment DEL prediction using internal and neighbouring turbine DEL input models with XGBoost.

Neighbouring Turbine	B06 Internal	Neighbour DEL Input	ΔMAPE [%]
B07	11.06	9.91	-2.05
D07	11.06	10.05	-1.01
B08	11.06	9.93	-1.13
D08	11.06	10.23	-0.83
All	11.06	9.46	-1.60

Table 4.31: NRMSE [-] for B06 tower FA moment DEL prediction using internal and neighbouring turbine DEL input models with XGBoost.

Neighbouring Turbine	B06 Internal	Neighbour DEL Input	ΔNRMSE [%]
B07	0.286	0.268	-6.30
D07	0.286	0.255	-10.80
B08	0.286	0.275	-3.80
D08	0.286	0.254	-11.20
All	0.286	0.244	-14.70

Similar to the blade models, the tower fore-aft moment DEL prediction models consistently benefit from the inclusion of neighbouring turbine DEL as an additional input. However, the magnitude of improvement varies more substantially in this case. The observed gains, from one turbine DEL input, in R^2 range from 0.61 to 1.72 percentage points, while the NRMSE reductions span from 3.8% to 11.2%. Interestingly, turbines D07 and D08 yield greater improvements with regard to NRMSE than B07 and B08, although no definitive spatial or operational pattern explains this discrepancy. One plausible explanation lies in the previously discussed limitations associated with tower load measurements, which may introduce additional noise or variability into the learning process. Nevertheless, the combined case, where all neighbouring DELs are included, once

again provides the best performance, reinforcing the conclusion that incorporating spatial load information enhances the model's capacity to capture the underlying loading dynamics.

4.7 Generalisability

The next part of the Results & Discussion chapter investigates the generalisability of data-driven models across different wind turbines within the same farm. Specifically, it aims to address the question: How does model accuracy change when tested on SCADA data from a different turbine than the one it was tuned for? This scenario is highly relevant in practical settings where data availability or computational constraints might only allow hyperparameter tuning on a single representative turbine.

As an initial attempt, the XGBoost model configuration optimised using normal operation data from turbine B07 (see Section 4.3) was directly applied to the test datasets of other turbines without any retraining. The corresponding results are shown in Table 4.32. As expected, the performance deteriorates significantly, with R^2 drops exceeding 30% and relative errors (MAPE and NRMSE) increasing by more than 100% in some cases. This poor transferability is primarily due to spatial variability in operating conditions across the wind farm. For example, B07 typically operates under wake conditions, while turbines like B08, C08, and D08 experience more frequent free-stream exposure (see Fig. 4.21). Consequently, a model trained on wake-affected data tends to overestimate loads in free-stream cases. Moreover, since the scalers were fitted solely on the B07 dataset, discrepancies in input and output distributions between turbines also lead to unreliable predictions. These findings clearly demonstrate that directly reusing a trained model across turbines without any retraining is not viable.

Table 4.32: Generalisability case results with B07 XGBoost blade flapwise DEL prediction model (no retraining).

Wind Turbine	**Internal Model**	**B07 Model (No Retrain)**	**ΔR^2 [%]**	**ΔMAPE [%]**	**ΔNRMSE [%]**
B06	R^2 = 95.31 % MAPE = 8.77 % NRMSE = 0.217	R^2 = 80.93 % MAPE = 19.98 % NRMSE = 0.437	-14.38	11.21	101.38
D07	R^2 = 93.43 % MAPE = 8.39 % NRMSE = 0.256	R^2 = 59.32 % MAPE = 23.98 % NRMSE = 0.638	-34.11	15.59	149.22
D08	R^2 = 95.01 % MAPE = 8.81 % NRMSE = 0.223	R^2 = 70.93 % MAPE = 25.86 % NRMSE = 0.539	-24.08	17.05	141.70
C08	R^2 = 94.52 % MAPE = 8.57 % NRMSE = 0.234	R^2 = 63.03 % MAPE = 24.60 % NRMSE = 0.608	-31.49	16.03	159.83
B08	R^2 = 94.80 % MAPE = 8.77 % NRMSE = 0.228	R^2 = 70.92 % MAPE = 24.56 % NRMSE = 0.539	-23.88	15.79	136.40

To address this issue, a more pragmatic strategy was evaluated. The B07 XGBoost model configuration was reused, but the model itself was retrained individually on each turbine's internal SCADA and DEL dataset. By fixing the hyperparameters, the computationally intensive tuning process is avoided, while still allowing the model to adapt to turbine-specific data distributions. Table 4.33 and Table 4.34 summarise the performance results for flapwise blade and tower fore-aft moment DEL predictions, respectively. In both cases, the retrained B07 model achieves nearly identical performance to the internal model trained and tuned specifically for each turbine. Differences in R^2, MAPE, and NRMSE are negligible; typically within $\pm 0.1\%$.

Table 4.33: Generalisability case results with B07 XGBoost blade flapwise DEL prediction model.

Wind Turbine	**Internal Model**	**B07 Model Retrained**	**ΔR^2 [%]**	**ΔMAPE [%]**	**ΔNRMSE [%]**
B06	$R^2 = 95.31$ % MAPE = 8.77 % NRMSE = 0.217	$R^2 = 95.37$ % MAPE = 8.73 % NRMSE = 0.215	0.06	-0.04	-0.92
D07	$R^2 = 93.43$ % MAPE = 8.39 % NRMSE = 0.256	$R^2 = 93.40$ % MAPE = 8.40 % NRMSE = 0.257	-0.03	0.01	0.39
D08	$R^2 = 95.01$ % MAPE = 8.81 % NRMSE = 0.223	$R^2 = 94.99$ % MAPE = 8.81 % NRMSE = 0.224	-0.02	0.00	0.45
C08	$R^2 = 94.52$ % MAPE = 8.57 % NRMSE = 0.234	$R^2 = 94.55$ % MAPE = 8.55 % NRMSE = 0.233	0.03	-0.02	-0.43
B08	$R^2 = 94.80$ % MAPE = 8.77 % NRMSE = 0.228	$R^2 = 94.79$ % MAPE = 8.81 % NRMSE = 0.228	-0.01	0.04	0.00

Table 4.34: Generalisability case results with B07 XGBoost tower FA moment DEL prediction model.

Wind Turbine	**Internal Model**	**B07 Model Retrained**	**ΔR^2 [%]**	**ΔMAPE [%]**	**ΔNRMSE [%]**
B06	$R^2 = 92.45$ % MAPE = 8.70 % NRMSE = 0.275	$R^2 = 92.41$ % MAPE = 8.69 % NRMSE = 0.276	-0.04	-0.01	0.36
D07	$R^2 = 92.24$ % MAPE = 9.07 % NRMSE = 0.279	$R^2 = 92.28$ % MAPE = 9.09 % NRMSE = 0.278	0.04	0.02	-0.36
D08	$R^2 = 94.24$ % MAPE = 9.67 % NRMSE = 0.240	$R^2 = 94.17$ % MAPE = 9.69 % NRMSE = 0.241	-0.07	0.02	0.42
B08	$R^2 = 93.73$ % MAPE = 8.93 % NRMSE = 0.250	$R^2 = 93.81$ % MAPE = 8.91 % NRMSE = 0.249	0.08	-0.02	-0.40

These results suggest that once an optimal configuration is found, it can be effectively reused across the wind farm by simply retraining the model with turbine-specific data. This approach drastically reduces computational cost, as hyperparameter tuning, which is approximately 120 times more time-consuming than training, only needs to be performed once. On a standard laptop, training the XGBoost model takes only about 20 seconds. Overall, this generalisability case further reinforces the robustness and consistency of the XGBoost algorithm, confirming its suitability for deployment in large-scale operational settings with minimal additional effort.

4.8 Fatigue Damage Estimation

To complement the data-driven modelling results, this section demonstrates how the predicted DEL values can be used to estimate fatigue damage on wind turbine components. As discussed in Chapter 2, both DEL and fatigue damage (denoted as D) play a central role in assessing structural integrity and potential failure over time. Fatigue damage serves as an index that, when accumulated linearly over the turbine's operational life, may indicate failure once $D \geq 1$. The theoretical formulation is given in Eq. (2.2), where D depends on material properties and structural dimensions through a proportionality coefficient K, which in this study is unknown.

In the absence of specific material data, a qualitative evaluation can still be conducted by normalising $K = 1$, which preserves the relative comparison and does not alter the overall discussion. Given that the rainflow counting method has been applied and a DEL time series is available, Eq. (2.2) simplifies to:

$$D = N_{\text{eq}} \cdot \text{DEL}^m \tag{4.3}$$

Where N_{eq} represents the equivalent number of cycles (set to 600 for 10-minute data time-series) and m is the Wöhler exponent characterising the material's fatigue sensitivity. This formulation enables a consistent and illustrative fatigue damage estimation across the dataset.

The following analysis presents an illustrative example to demonstrate how fatigue damage estimation can be approached using model-predicted DEL values. It is important to note that this is not a rigorous fatigue assessment. A comprehensive evaluation would require long-term, continuous DEL measurements, careful filtering of specific operational conditions, and a deeper understanding of the component material properties, all of which fall beyond the scope of the present study.

Instead, this example serves to highlight the methodology and identify key factors influencing fatigue damage estimation. A test period of seven consecutive days, from 25-01-2021 to 01-02-2021 (dd-mm-yyyy), is selected from the dataset for this purpose. The remaining data are used for model development, with an 80-20% split between training and validation. The target variable is the tower fore-aft moment DEL. Once trained and validated, the models are evaluated on the extracted test set. The resulting relative errors in DEL prediction for this test period are illustrated in Fig. 4.24, where the x-axis corresponds to the day of January 2021. The relative error in DEL prediction is computed as:

$$\text{Relative Error} = \frac{y - \hat{y}}{y} \times 100\% \tag{4.4}$$

Where y & $\hat{y}$ are the actual and predicted DEL values, respectively. Accordingly, negative values indicate instances where the model overpredicts the actual DEL.

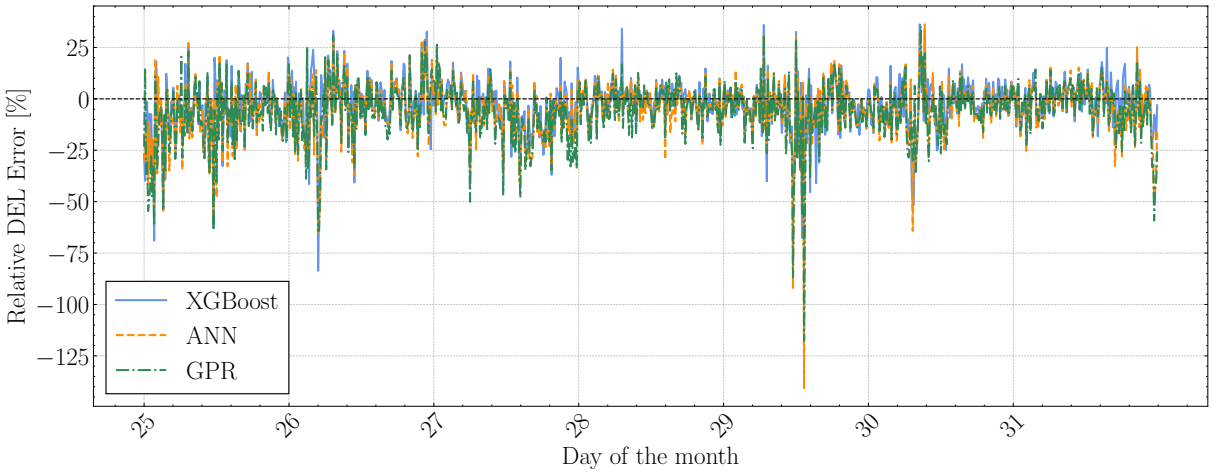


Figure 4.24: Tower FA DEL relative error against time.

Using the formulation provided in Eq. (4.3), the DEL time series can be converted into fatigue damage estimates. Based on this transformation, the relative error in damage prediction is computed for each model. The evolution of the damage relative error over the test period is shown in Fig. 4.25, providing insight into how model prediction errors propagate into fatigue damage estimation.

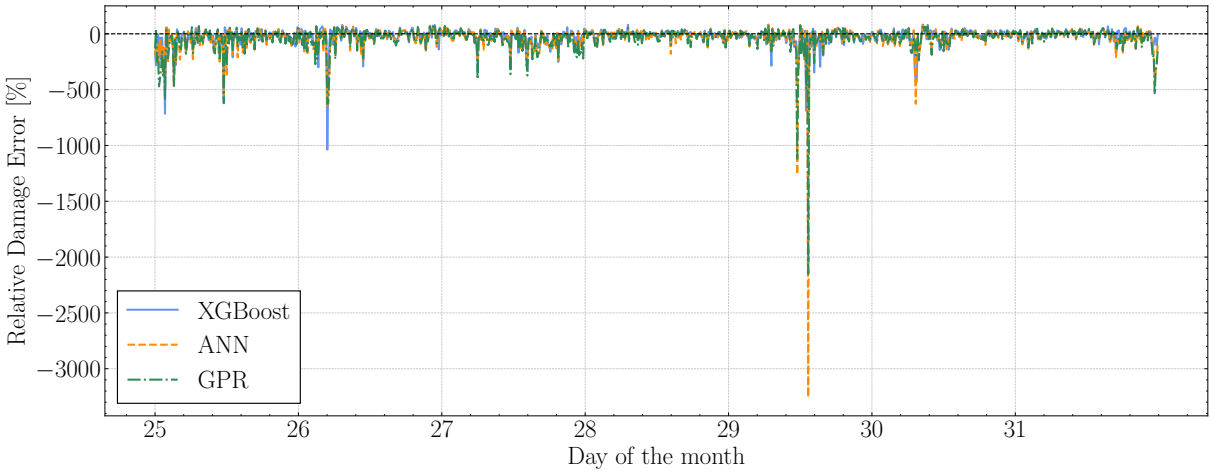


Figure 4.25: Tower FA Damage relative error against time.

A comparison of the two plots reveals that errors in DEL prediction are significantly amplified when converted to damage. This amplification is particularly evident in the pronounced negative peaks observed on the 26th and 29th of January. Such behavior is expected, as the DEL values are raised to the power of the Wöhler exponent m, which is set to 4 for the tower moment. Consequently, even moderate errors in DEL estimation can lead to disproportionately large deviations in damage prediction.

It is also important to note that this effect would be even more severe if the same analysis were performed on blade root moments, where the Wöhler exponent is typically taken as $m = 10$. This higher exponent would magnify prediction errors to an even greater extent, making accurate DEL estimation particularly critical for blade fatigue analysis.

To assess the overall impact of these deviations, the cumulative damage over the test period is shown in Fig. 4.26. This plot provides insight into whether short-term overpredictions result in significant overestimation of cumulative fatigue damage.

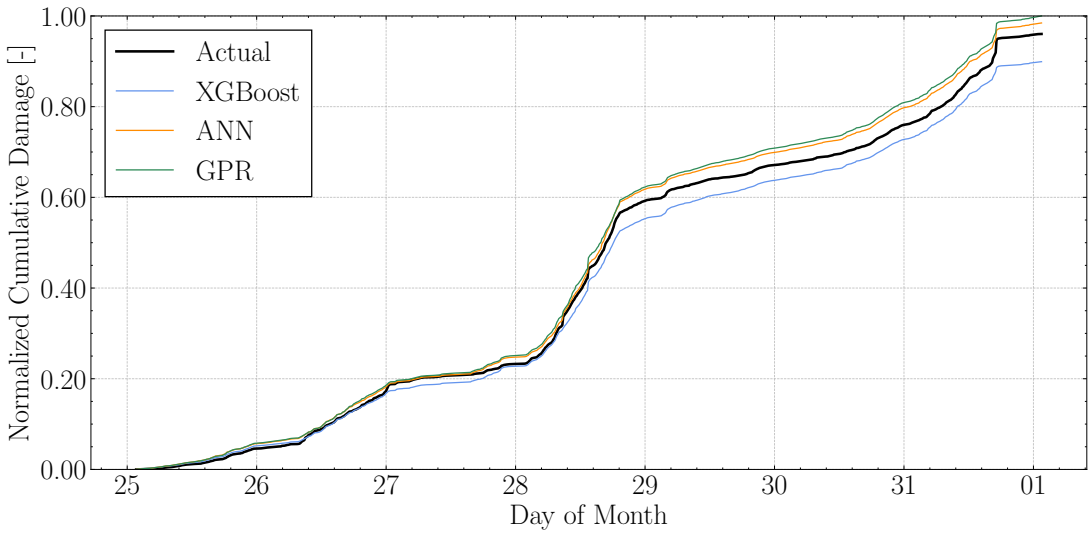


Figure 4.26: Tower FA Cumulative Damage tower against time.

As shown in Fig. 4.26, the ANN and GPR models slightly overestimate the cumulative damage over the test period, whereas XGBoost tends to underestimate it. Notably, the pronounced negative peaks observed in Fig. 4.25 do not appear to significantly affect the cumulative damage. This is because these large relative errors occur at time steps where the actual DEL values are low. Although the percentage errors are high, their absolute contribution to cumulative damage is minimal.

In contrast, other prediction errors that are smaller in relative terms but larger in absolute magnitude can have a more substantial impact on cumulative damage. Such deviations are not easily visible in Fig. 4.25 but become evident in the cumulative curve. Two instances of this behaviour can be seen on the 28th and 31st of January, where the predicted damage curves deviate more noticeably from the actual.

To better illustrate this, the time series of predicted and actual DEL values for 31 January 2021 is presented in Fig. 4.27. This plot highlights how local discrepancies in DEL predictions translate into differences in cumulative damage over time.

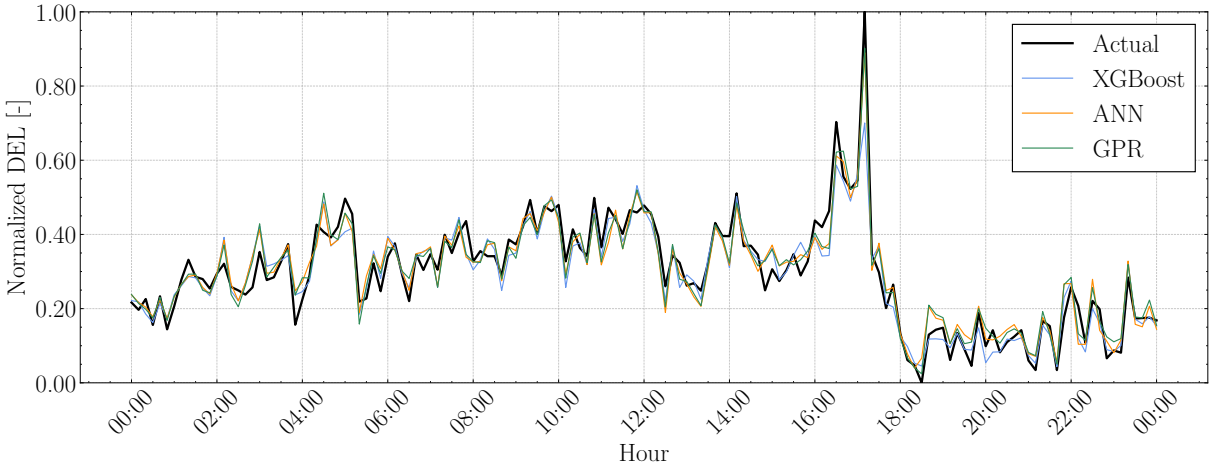


Figure 4.27: Tower FA DEL time-series against time for the 31-Jan-2021.

Fig. 4.27 provides a noteworthy insight into the temporal accuracy of the model predictions. The large positive DEL peak occurring around 17:00 is well captured by both the ANN and GPR models, while XGBoost significantly underpredicts it by approximately 40% (the percentage is not calculated based on the plot since it is normalised). Given the high magnitude of this DEL value, its contribution to fatigue damage is strongly amplified by the fourth power relationship defined by the Wöhler exponent. This substantial underestimation largely accounts for the elevated cumulative damage prediction error observed for XGBoost on 31 January. In contrast, the closer alignment of ANN and GPR with the actual DEL results in improved damage estimates for that day and in total.

The overall relative error in cumulative damage prediction across the test period for each model is summarised in Table 4.35.

Table 4.35: Relative error in cumulative damage prediction for the tower FA moment.

Model	**Relative Error [%]**
XGBoost	−6.37
ANN	2.53
GPR	4.14

The results presented in Table 4.35 indicate that ANN and GPR outperform XGBoost in predicting cumulative fatigue damage for the tower FA moment during the selected test period. This observation contrasts with earlier findings in this chapter, where XGBoost generally demonstrated superior performance across conventional error metrics. However, it is important to emphasise that this evaluation is based on a limited one-week dataset and is intended solely to illustrate the methodology rather than to provide a definitive comparison. Despite this, the exercise yields several valuable insights.

First, it highlights that if fatigue damage estimation is the ultimate objective, model evaluation should be conducted with this metric in mind. This requires selecting representative test periods and evaluating performance based on damage-related outcomes rather than standard regression metrics alone. Second, the analysis underscores the sensitivity of fatigue damage calculations to high DEL values, which are disproportionately amplified by the Wöhler exponent. This sensitivity would be even greater in blade fatigue

analysis, where the exponent is typically higher. Ensuring accurate model predictions at these high-load events is therefore critical. One possible improvement for XGBoost could be the use of mean squared error (MSE) as the validation loss to increase sensitivity to extreme values.

Lastly, the importance of robust data preprocessing cannot be overstated. Outliers caused by sensor faults, calibration issues, or erroneous data processing can significantly skew the DEL values and, by extension, the damage estimates. Ensuring data integrity is therefore a prerequisite for any reliable fatigue analysis.

This illustrative case study thus emphasises the need for task-specific evaluation criteria, careful handling of high-load events, and rigorous data quality assurance when using data-driven models for fatigue damage estimation.

4.9 Discussion

The case studies presented in this chapter constitute a comprehensive and insightful exploration of the factors that influence data-driven load estimation in wind turbines. By systematically investigating model performance, input data characteristics, and turbine-specific configurations, the analysis has provided valuable perspectives on both the predictive capacity of machine learning methods and the operational challenges inherent to wind farm monitoring. Notably, these case studies highlight the critical role of data quality, input selection, and model robustness in determining reliable outputs, as well as the need to interpret model results in light of their intended engineering application.

The Feature Importance analysis demonstrates that purely data-driven techniques can effectively identify relevant input signals, even in the absence of physical interpretability. However, the relative importance of these features is dependent on the prediction target. For instance, tower load models appear more sensitive to the range of input variables, a finding that may reflect the comparatively lower accuracy of the tower dataset used in this study. This distinction underscores the importance of considering the data fidelity associated with different structural components when developing prediction models.

In the Model Comparison, advanced machine learning models generally exhibit strong performance across all targets, with no visible signs of bias. Interestingly, the tower models achieve lower accuracy despite the relative simplicity of tower dynamics compared to blades. This outcome contradicts physical expectations and again points to the influence of data quality. Among the models evaluated, XGBoost consistently outperforms others in terms of accuracy, efficiency, and robustness, making it the most promising candidate for fatigue load prediction tasks.

The Filtering experiments reveal that model performance remains largely unaffected by the inclusion or exclusion of extreme values and outliers, indicating that XGBoost, in particular, maintains strong generalisation capabilities under varying data conditions. This robustness is further reinforced by the Dataset Size study, which shows that prediction accuracy continues to improve with additional data, even beyond one full year. These findings advocate for maximising data availability when training machine learning models for fatigue estimation.

In the Local Configuration study, incorporating input from neighbouring wind turbines aids in reducing prediction errors and enhances overall model accuracy across all wind

directions. This highlights the benefit of including location-specific measurements to capture turbine-specific dynamics and localised wind conditions. Similarly, the Neighbouring DEL case confirms that information from adjacent turbines can enhance model performance, suggesting that spatially correlated load patterns exist and can be leveraged in a multi-turbine modelling framework.

The Generalisability analysis provides evidence that retrained XGBoost models perform comparably to models that underwent hyperparameter tuning using internal wind turbine data. This finding is important for operational scalability, as it implies that frequent retuning for each wind turbine is not necessary, thereby reducing computational cost and complexity in intra-farm applications.

Finally, the Fatigue Damage Estimation study brings the analysis closer to practical implementation. It illustrates how DEL predictions can be used to estimate cumulative fatigue damage, aligning the modelling efforts with the ultimate goal of condition monitoring and lifetime assessment. Importantly, this case highlights that even accurate models can yield misleading damage estimations if they under- or over-predict high-load events. A more detailed examination of fatigue estimation uncertainties would be a valuable direction for future work.

In conclusion, the case studies reveal that model accuracy, data quality, and results interpretability are deeply interconnected. High-quality measurements and careful calibration are essential to produce reliable load predictions. Moreover, interpreting model performance must always be aligned with the end-use objective, whether it be anomaly detection, load monitoring, or fatigue estimation. Ultimately, a thorough evaluation using appropriate metrics and visualisations tailored to each task is necessary to ensure meaningful and actionable insights from data-driven models in wind energy applications.

CHAPTER 5

Conclusions & Recommendations

This study presented an exploratory analysis using real measurement data from the Lillgrund offshore wind farm, with the objective of predicting blade root flapwise and tower bottom fore-aft moment DELs under various operational conditions. The research addressed three central questions. First, it evaluated and compared different data-driven modelling approaches to identify the most suitable method for predicting DEL from SCADA data in operational wind turbines. Second, it explored how variations in data characteristics (including input feature composition, data volume, and signal pre-processing) affect model performance and generalisability. Third, it investigated the potential of incorporating spatial information from neighbouring turbines to enhance predictive accuracy across the wind farm. The findings offered substantive answers to all three questions and contributed to a better understanding of the challenges and opportunities associated with deploying surrogate models in realistic wind farm settings.

Building upon prior work in the field, which relied primarily on synthetic datasets, this thesis demonstrated the applicability of machine learning models in operational wind farms. Among the evaluated methods, all advanced models captured underlying data patterns effectively, with XGBoost emerging as the most accurate and robust approach, offering favourable trade-offs between predictive performance, computational cost, and resilience to noise.

A key finding was the decisive role of data quality in model performance. Models trained to predict blade root loads performed significantly better than those targeting tower bottom moments. This discrepancy was attributed to differences in data accuracy; the blade loads were derived from calibrated outputs provided directly by the manufacturer (Siemens), whereas the tower load estimates relied on in-house calibration procedures that involved several assumptions, likely introducing systematic errors.

The Fatigue Damage Estimation case study demonstrated the practical application of the trained models for estimating fatigue accumulation and conducting lifetime assessments. This use case highlighted both the benefits and limitations of integrating such models into an asset management framework, particularly under realistic operating conditions.

Overall, the findings indicated that combining high-quality SCADA data with well-tuned machine learning models can support the estimation of DELs and contribute meaningfully to the operation and monitoring of wind farms. In particular, the results underscored the potential of data-driven approaches to serve as virtual sensors, offering an alternative to long-term deployment of strain gauges, which require continuous maintenance. This capability is especially valuable in offshore environments, where operational and maintenance costs are considerably higher. The study thus established a solid foundation for integrating data-driven methods into condition monitoring systems, supporting more cost-effective and scalable wind farm management.

Future Work

Due to the time constraints of this thesis project, the work had to be concluded at a stage where several promising research directions remained unexplored. Nonetheless, the findings and limitations of this study point toward a number of compelling opportunities for further investigation.

- **Cross-Farm Validation and Model Robustness**: A natural next step is to test the developed models on datasets from other wind farms, particularly those equipped with larger and more modern turbines, and operating under different environmental conditions and layout configurations. Such a study would enable a broader evaluation of model robustness and generalisation capacity, and could offer insights into the transferability of trained surrogate models across varying operational contexts.
- **Expanded Model Evaluation Across Case Studies**: While this thesis applied XGBoost to all case studies, the application of GPR and ANN was limited to the Model Comparison case. Extending these models to all case studies would provide a more comprehensive comparison of their robustness, adaptability, and performance under different operational and spatial configurations.
- **Enhanced Hyperparameter Optimisation for ANNs**: The results highlighted the sensitivity of ANN performance to hyperparameter choices. Future work could expand the hyperparameter search space, incorporate more complex architectures, and explore advanced tuning methods. This was not feasible during the thesis due to the high computational cost, but remains an important area for performance improvement.
- **Direct Optimisation Based on Fatigue Damage**: Given that the ultimate goal of DEL prediction is fatigue damage estimation, future studies should consider integrating damage computation directly into the model development process. This could involve redefining the loss function or validation metric to minimise the error in fatigue damage rather than in DEL. Such an approach may yield models that are more aligned with the needs of structural lifetime assessments.
- **Incorporating Spatial Configuration as Input Features**: Although the local turbine configurations were partially explored, future work could explicitly include spatial features such as inter-turbine distance and relative angular positioning as model inputs. These features may help capture wake interactions and directional dependencies, enabling the models to assign greater importance to relevant upstream turbines under certain wind conditions.
- **Exploration of Uncertainty-Aware Modelling Techniques**: To further improve model reliability, future research should explore uncertainty-aware approaches such as Bayesian Neural Networks (BNNs), deep ensembles, or quantile regression. These methods not only have the potential to enhance predictive accuracy, as suggested by Hlaing et al. [11], but also to provide uncertainty estimates, which are critical for decision-making in operational asset management and maintenance planning.

Bibliography

[1] Ember and Energy Institute. *Statistical Review of World Energy (2024)*. Ember; Energy Institute. 2024. URL: https://ourworldindata.org/energy.

[2] Council(GWEC), Global Wind Energy. *Global Wind Report 2025*. 2025. URL: www.gwec.net.

[3] Vattenfall. *Lillgrund Offshore Wind Farm*. URL: https://powerplants.vattenfall.com/lillgrund/.

[4] EU. *TWAIN project*. URL: https://twainproject.eu/.

[5] Endo, T. and Matsuishi, M. "Fatigue of metals subjected to varying stress". In: *Japan Society of Mechanical Engineering* (1968).

[6] Miner, Milton A. "Cumulative Damage in Fatigue". In: *Journal of Applied Mechanics* 12.3 (Mar. 2021), A159–A164. ISSN: 0021-8936. DOI: 10.1115/1.4009458. URL: https://doi.org/10.1115/1.4009458.

[7] L Hansen, Martin O. *Aerodynamics of Wind Turbines*. Tech. rep.

[8] Hauge Madsen, Peter, Dekker, Jos WM, Erik Thor FFA, Sven, Ken McAnulty, Sweden, Matthies Germanischer Lloyd, Hermann, and Germany Robert Thresher, W W. *RECOMMENDED PRACTICES FOR WIND TURBINE TESTING 3. FATIGT.]E LOADS 2. EDITION I9!X) In cooperation with*. Tech. rep.

[9] Lyons, John Thomas and Göçmen, Tuhfe. "Applied machine learning techniques for performance analysis in large wind farms". In: *Energies* 14.13 (July 2021). ISSN: 19961073. DOI: 10.3390/en14133756.

[10] Papatheou, Evangelos, Dervilis, Nikolaos, Maguire, Andrew Eoghan, Antoniadou, Ifigeneia, and Worden, Keith. "A Performance Monitoring Approach for the Novel Lillgrund Offshore Wind Farm". In: *IEEE Transactions on Industrial Electronics* 62.10 (Oct. 2015), pp. 6636–6644. ISSN: 02780046. DOI: 10.1109/TIE.2015.2442212.

[11] Hlaing, Nandar, Morato, Pablo G, De, Francisco, Santos, Nolasco, Santos, Francisco De N, Weijtjens, Wout, and Devriendt, Christof. *Virtual Load Monitoring of Offshore Wind Farms via Uncertainty-Aware Deep Learning Models*. Tech. rep. URL: https://www.researchgate.net/publication/392924095.

[12] Noppe, N., Iliopoulos, A., Weijtjens, W., and Devriendt, C. "Full load estimation of an offshore wind turbine based on SCADA and accelerometer data". In: *Journal of Physics: Conference Series*. Vol. 753. 7. Institute of Physics Publishing, Oct. 2016. DOI: 10.1088/1742-6596/753/7/072025.

[13] Weijtens, Wout, Noppe, Nymfa, Verbelen, Tim, Iliopoulos, Alexandros, and Devriendt, Christof. "Offshore wind turbine foundation monitoring, extrapolating fatigue measurements from fleet leaders to the entire wind farm". In: *Journal of Physics: Conference Series*. Vol. 753. 9. Institute of Physics Publishing, Oct. 2016. DOI: 10.1088/1742-6596/753/9/092018.

[14] Miao, Yizhi, Soltani, Mohsen N., and Hajizadeh, Amin. "A Machine Learning Method for Modeling Wind Farm Fatigue Load". In: *Applied Sciences (Switzerland)* 12.15 (Aug. 2022). ISSN: 20763417. DOI: 10.3390/app12157392.

[15] Schröder, Laura, Krasimirov Dimitrov, Nikolay, Verelst, David Robert, and Sorensen, John Aasted. "Wind turbine site-specific load estimation using artificial neural networks calibrated by means of high-fidelity load simulations". In: *Journal of Physics: Conference Series*. Vol. 1037. 6. Institute of Physics Publishing, June 2018. DOI: 10.1088/1742-6596/1037/6/062027.

[16] Larsen, Torben J. and Hansen, Anders Melchior. *How 2 HAWC2, the user's manual*. English. Denmark. Forskningscenter Risoe. Risoe-R 1597(ver. 3-1)(EN). Risø National Laboratory, 2007. ISBN: 978-87-550-3583-6.

[17] Nakhchi, M. E., Win Naung, S., and Rahmati, M. "Wake and power prediction of horizontal-axis wind farm under yaw-controlled conditions with machine learning". In: *Energy Conversion and Management* 296 (Nov. 2023). ISSN: 01968904. DOI: 10.1016/j.enconman.2023.117708.

[18] Purohit, Shantanu, Ng, E. Y.K., and Syed Ahmed Kabir, Ijaz Fazil. "Evaluation of three potential machine learning algorithms for predicting the velocity and turbulence intensity of a wind turbine wake". In: *Renewable Energy* 184 (Jan. 2022), pp. 405–420. ISSN: 18790682. DOI: 10.1016/j.renene.2021.11.097.

[19] Gasparis, Georgios, Lio, Wai Hou, and Meng, Fanzhong. "Surrogate models for wind turbine electrical power and fatigue loads in wind farm". In: *Energies* 13.23 (Dec. 2020). ISSN: 19961073. DOI: 10.3390/en13236360.

[20] Dimitrov, Nikolay, Kelly, Mark C., Vignaroli, Andrea, and Berg, Jacob. "From wind to loads: Wind turbine site-specific load estimation with surrogate models trained on high-fidelity load databases". In: *Wind Energy Science* 3.2 (2018), pp. 767–790. ISSN: 23667451. DOI: 10.5194/wes-3-767-2018.

[21] Murcia, Juan Pablo, Réthoré, Pierre Elouan, Dimitrov, Nikolay, Natarajan, Anand, Sørensen, John Dalsgaard, Graf, Peter, and Kim, Taeseong. "Uncertainty propagation through an aeroelastic wind turbine model using polynomial surrogates". In: *Renewable Energy* 119 (Apr. 2018), pp. 910–922. ISSN: 0960-1481. DOI: 10.1016/J.RENENE.2017.07.070.

[22] Toft, Henrik Stensgaard, Svenningsen, Lasse, Moser, Wolfgang, Sørensen, John Dalsgaard, and Thøgersen, Morten Lybech. "Assessment of wind turbine structural integrity using response surface methodology". In: *Engineering Structures* 106 (Jan. 2016), pp. 471–483. ISSN: 18737323. DOI: 10.1016/j.engstruct.2015.10.043.

[23] Kusiak, Andrew, Zhang, Zijun, and Verma, Anoop. *Prediction, operations, and condition monitoring in wind energy*. Oct. 2013. DOI: 10.1016/j.energy.2013.07.051.

[24] Göçmen, Tuhfe et al. *Data-driven wind farm flow control and challenges towards field implementation: A review*. Mar. 2025. DOI: 10.1016/j.rser.2025.115605.

[25] Sood, I, Simon, E, Vitsas, A, Blockmans, B, Larsen, G C, and Meyers, J. "Comparison of large eddy simulations against measurements from the Lillgrund offshore wind farm". In: *Wind Energy Science* 7.6 (2022), pp. 2469–2489. DOI: 10.5194/wes-7-2469-2022. URL: https://wes.copernicus.org/articles/7/2469/2022/.

[26] Prasad, Nadipuram R., Almanza-Garcia, Salvador, and Lu, Thomas T. "Anomaly detection". In: *Computers, Materials and Continua* 14.1 (2009), pp. 1–22. ISSN: 15462218. DOI: 10.1145/1541880.1541882.

[27] Morrison, Rory, Liu, Xiaolei, and Lin, Zi. "Anomaly detection in wind turbine SCADA data for power curve cleaning". In: *Renewable Energy* 184 (Jan. 2022), pp. 473–486. ISSN: 18790682. DOI: 10.1016/j.renene.2021.11.118.

[28] Lin, Zi, Liu, Xiaolei, and Collu, Maurizio. "Wind power prediction based on high-frequency SCADA data along with isolation forest and deep learning neural networks". In: *International Journal of Electrical Power and Energy Systems* 118 (June 2020). ISSN: 01420615. DOI: 10.1016/j.ijepes.2020.105835.

[29] Pedregosa, F et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[30] McInnes, Leland, Healy, John, and Astels, Steve. "hdbscan: Hierarchical density based clustering". In: *The Journal of Open Source Software* 2.11 (Mar. 2017). DOI: 10.21105/joss.00205. URL: https://doi.org/10.21105%2Fjoss.00205.

[31] Chen, Tianqi and Guestrin, Carlos. "XGBoost: A scalable tree boosting system". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol. 13-17-August-2016. Association for Computing Machinery, Aug. 2016, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785.

[32] Breiman, Leo. *Random Forests*. Tech. rep. 2001, pp. 5–32.

[33] Pangarkar, Darshan Jagannath, Sharma, Rajesh, Sharma, Amita, and Sharma, Madhu. "Assessment of the Different Machine Learning Models for Prediction of Cluster Bean (Cyamopsis tetragonoloba L. Taub.) Yield". In: *Advances in Research* (Aug. 2020), pp. 98–105. DOI: 10.9734/air/2020/v21i930238.

[34] Akiba, Takuya, Sano, Shotaro, Yanase, Toshihiko, Ohta, Takeru, and Koyama, Masanori. "Optuna: A Next-generation Hyperparameter Optimization Framework". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2019.

[35] Bergstra, James, Bardenet, Rémi, Bengio, Yoshua, and Kégl, Balázs. *Algorithms for Hyper-Parameter Optimization*. Tech. rep.

[36] Martin, Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. URL: http://tensorflow.org/.

[37] Särkkä, Simo. "The Use of Gaussian Processes in System Identification". In: *Encyclopedia of Systems and Control*. London: Springer London, 2019, pp. 1–10. ISBN: 978-1-4471-5102-9. DOI: 10.1007/978-1-4471-5102-9{_}100087-1. URL: https://doi.org/10.1007/978-1-4471-5102-9_100087-1.

[38] GPy. *GPy: A Gaussian process framework in python*. 2012. URL: http://github.com/SheffieldML/GPy.

[39] Lundberg, Scott M and Lee, Su-In. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[40] Lundberg, Scott M et al. "From local explanations to global understanding with explainable AI for trees". In: *Nature Machine Intelligence* 2.1 (2020), pp. 2522–5839.

[41] Virtanen, Pauli et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2. URL: https://doi.org/10.1038/s41592-019-0686-2.