Master Thesis

Data-driven Surrogate Models for Predicting Wind Turbine Wake Effects

A COMPARATIVE ANALYSIS

Niccolò Morabito

Supervisors

Erik Quaeghebeur - Eindhoven University of Technology e.quaeghebeur@tue.nl
Laurens Bliek - Eindhoven University of Technology l.bliek@tue.nl

Eindhoven, August 2023

Abstract

Wind turbine wake modelling plays a crucial role in optimizing wind farm layouts and maximizing wind energy production. This thesis addresses the critical challenges of this problem through the development of innovative and original data-driven surrogate models, supported by machine learning techniques and thorough analytical exploration.

The study develops novel data-driven surrogate models for the widely-used Ainslie model, effectively overcoming the data scarcity limitations inherent in this domain. These models encompass a range of approaches, including traditional methods like Decision Trees and Multilayer Perceptrons, as well as the innovative application of Neural Radiance Fields to wind turbine wake modelling. These surrogate models exhibit a remarkable increase in computational efficiency, achieving up to 6 to 7 orders of magnitude improvement, all while maintaining an impressive level of accuracy. Additionally, in-depth analysis and comprehensive investigation of model behaviours offer a nuanced understanding of these surrogate models, enhancing their applicability in diverse scenarios.

The findings underscore the considerable potential of data-driven surrogate models in predicting wind turbine wake effects. The study emphasizes the critical significance of strategically acquiring a sufficient volume of sparse data, encompassing a wide range of relevant parameters. This process includes the assessment of model suitability and skilful navigation of the trade-offs between accuracy and model complexity. Collectively, the contributions presented in this work pave the way for more efficient, accurate, and insightful wind turbine wake modelling strategies. These advancements hold immense promise for the progressive enhancement of wind farm layout optimization and the elevation of energy production capabilities.

Keywords: *Wind Turbine Wake Modelling, Ainslie Model, Surrogate Models, Machine Learning, Neural Radiance Fields*

Acknowledgments

First and foremost, I would like to express my deepest gratitude to the Big Data Management and Analytics (BDMA) consortium for selecting me to be part of this esteemed program. The opportunity to pursue this exceptional journey has been an immense privilege.

I extend my sincere appreciation to my two supervisors, Erik Quaeghebeur and Laurens Bliek, for their guidance, support, and inspiration throughout this thesis. Erik's generous sharing of sources, inputs, and insights ushered me into a subject I am deeply passionate about, wind energy and wake modelling. His material about the Ainslie model has significantly contributed to the foundation of this report. Laurens' guidance in the realm of machine learning, including his groundbreaking work on Fourier features, has been instrumental in shaping the analytical aspects of this study. Their insightful suggestions have been invaluable in shaping this research.

I am also deeply grateful to my dear family and friends. Their unwavering encouragement, steadfast support, and empathetic understanding have been instrumental in shaping my journey and made me persevere through challenges and strive for excellence.

Finally, I want to acknowledge the contributions of my fellow colleagues and friends who have journeyed alongside me in this program. Collaborating with each of them has been an illuminating experience in itself. The collective work, the constant exchange of diverse perspectives, and the shared moments of both triumph and challenge have undeniably enriched my learning journey. Each interaction, discussion, and collaboration has contributed to a deeper understanding of the subjects at hand, making the entire academic pursuit more meaningful and insightful.

Contents

Chapter 1

Introduction

1.1 Context

Renewable energy sources have been experiencing significant growth in the past two decades, and this trend is expected to continue in the coming years in order to meet the world's future energy needs and mitigate the effects of climate change. Many countries have set ambitious targets for low greenhouse gas and pollutant emissions in line with international climate agreements. To prevent global average temperatures from rising 1.5°C above pre-industrial levels, it is projected that renewable energy must increase from 20% to 67% of global energy production from 2018 to 2040 [31]. Additionally, it is estimated that renewables will account for 43% of global electricity generation by 2030, up from the current level of 28% [3].

Wind energy is considered as a viable solution to meet a substantial part of the world's electricity needs and it is experiencing fast and continuous growth in recent years. In the last decade, the energy produced from wind farms has increased by almost 300% [2]. Total global wind power capacity is now up to 837 GW, helping the world avoid over 1.2 billion tonnes of CO2 annually and 557 GW of new capacity is expected to be added in the next five years under current policies [1]. However, as larger wind turbine diameters and wind farms are required to meet increased energy demand, the issue of the **wake effect**, also known as the wind turbine wake, has emerged as a significant challenge [64].

The term wake in the context of wind energy refers to the low momentum and highly turbulent region located downstream of an operating wind turbine due to the interaction between the spinning blades of the turbine and the surrounding air. The wake can extend for several rotor diameters behind the turbine and can have a significant impact on the performance of a wind farm. Figure 1.1 provides a visualization of the flow in a large wind farm, obtained by running very complex simulations on advanced supercomputers. The blue regions indicate the low-velocity wind-speed regions (wakes) formed behind the turbines.

When wind turbines are closely spaced in a wind farm, the wake generated by upstream turbines can cause decreased wind speed and increased turbulence for downstream turbines [68, 73]. This can result in reduced power generation efficiency [10] and increased fatigue load on downstream turbines, which can shorten their operational lifespan [21]. While this study exclusively addresses the former concern, it is important to note that the wake effect's impacts extend beyond just efficiency, encompassing both economic and environmental considerations. Research has revealed that the power production efficiency of a wind farm can experience reductions ranging from 10% to 20% due to the wake effect [11, 37].

Therefore, wind farm layout optimization is commonly employed during the design phase to minimize the wake effect and consequently improve the cost-efficiency of the

wind turbines within the wind farm boundaries. This approach takes into consideration the wake characteristics generated by upstream wind turbines, as a good understanding and accurate prediction of turbine wakes can significantly improve the efficiency of wind energy conversion in large-scale wind farms and achieve an optimal turbine layout scheme [50, 25, 72].

As a result, wake prediction is of primary importance for the successful design of a wind farm, as it contributes to reducing wake interference effects and enhancing the performance of each wind turbine within the farm. By studying the wake characteristics of upstream turbines, efficient wind farm designs can be developed, resulting in improved overall performance and reduced wake effects, thus maximizing the potential of wind energy conversion in a sustainable manner.

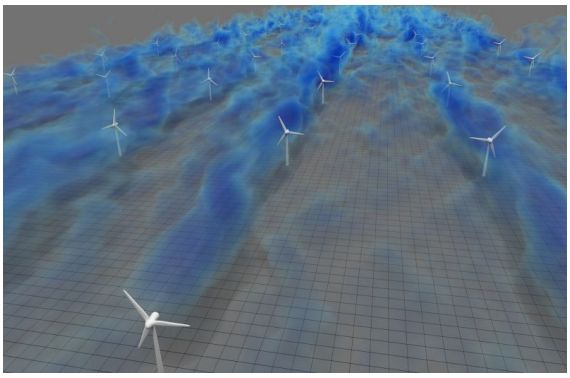


Figure 1.1: Simulation visualization depicting the wake effect, represented by the blue regions [17].

1.2 Classification of Wake Models

In the last decades, several approaches have been proposed to model the interaction between the flow and the wind turbines, following an increasing order of complexity and trading off fidelity and efficiency. We distinguish two main categories of models, the **analytical** and the **computational models**.

1.2.1 Analytical Models

Analytical wake models, which are also known as engineering models, employ semi-empirical analytical formulas to calculate wake effect properties in a very efficient way [46]. This is why these models are often used for wind farm layout optimization and wake control. However, their accuracy is limited as they rely on a number of assumptions and neglect certain physics. In particular, these static models cannot capture unsteady wakes and are not appropriate for designing wind farm control systems.

According to the assumptions on the profile of the wake, we can further distinguish two main subcategories of analytical models: *linear* and *Gaussian*. The former postulate "top-hat" profile of the wake, assuming that the velocity distribution changes linearly along the axial direction of incoming flow while the radial velocity is constant [34, 38, 26]. However,

more recent and sophisticated studies have shown that the velocity deficit in the wake is much closer to a bell-shaped profile, making the Gaussian model the preferred choice in recent studies [13, 71, 28].

Nevertheless, the wake profile is nearly Gaussian in shape only in the far wake region ($x/D \geqslant 2$ or 3, where $D$is the diameter of the wind turbine rotor) because of factors such as blade geometry and its aerodynamics, vortex shedding at the blade's tip and root region, 3D effects, and stalled flow, which make the near-wake profile non-Gaussian in structure. Thus, another relevant limitation of the Gaussian analytical models is that they cannot be used to describe the critical near-wake characteristics.

Moreover, although the existing analytical wake models perform well when compared to some experimental data in the far-wake region, there is still a significant error (even more than 20% in terms of power estimation) due to the simplifications that make it impossible to fully consider the characteristics of the local environment and inflow conditions [8]. Factors like wind speed and wind direction, temperature and atmospheric stability, and turbulence intensity are challenging to simulate individually, let alone couple with others.

Finally, in all the aforementioned models, there are one or more adjustable parameters or empirical constants that determine the accuracy of analytical models to a large extent. These parameters should be calibrated through experiments or fluid dynamics simulations.

1.2.2 Computational Models

The computational wake models are based on the Computational Fluid Dynamics (CFD) simulations and they are extensively used to get a more accurate characterisation of flows and prediction of the velocity and turbulence intensity in the wake. These models can capture complex and unsteady wakes, making them suitable for the design of wind farm control systems. However, CFD models require significant computational resources, making them less practical for wind farm layout optimization.

These techniques can be based either on the *Reynolds-Averaged Navier-Stokes* (RANS) equations or on *Large Eddy Simulations* (LES). The former can integrate different numerical techniques of basic fluid equations to fully characterize the flow around the wind turbine, providing a high level of detail for the wake and turbulence, and they take several hours to run [63]. The latter provides very high-resolution computations of the flow around the wind turbines, by directly computing large eddies and parametrizing their impact with volumes less than the grid spacing. However, LES simulations can take several weeks to complete, even when using large computer clusters [35, 52, 76].

1.2.3 Comparison Between Analytical and Computational models

To summarize, analytical models for wake analysis do not consider the entire profile variation of the turbulence intensity in the radial wake direction, accounting only for its average or outlier, and they can lead to significant discrepancies in both the near and far wake regions [46]. But they are very simple and have a low computational burden. On the other hand, CFD models can provide more detailed results than analytical wake models. Nevertheless, their use is quite difficult due to the volume of the input data analyzed in their processes, leading sometimes to an unacceptable computational time cost, especially for wind farm research with massive wake interaction computations.

To bridge this gap between fidelity and computational efficiency, surrogate models emerge as a promising solution to approximate mathematical representations that mimic the behaviour of CFD simulations while being computationally cheaper to evaluate.

1.3 Surrogate Models

Real-world optimization problems often require significant computational resources, and traditional metaheuristic algorithms tend to be effective only for smaller-scale problems. These problems typically involve a high number of dimensions, and the fitness functions developed for them can be extremely time-consuming, sometimes requiring several hours or even days for a single evaluation. To face this problem, surrogate models have gained significant and increasing attention in the field of artificial intelligence and data-driven modelling as powerful tools for approximating complex systems and mitigating the runtimes of expensive computational tasks [40].

A **surrogate model**, also known as metamodel or emulator, is a mathematical or statistical model that serves as a substitute for computationally expensive simulations or experiments. Its aim is to capture the essential characteristics and behaviour of the original system while offering faster and more efficient evaluations. By constructing a surrogate model, one can obtain rapid predictions and insights into the system's behaviour without the need for time-consuming simulations or costly experiments.

Surrogate models find applications in various domains, and they are particularly valuable in situations where the evaluation of the original system is time-consuming, resource-intensive, or impractical. In fact, one of the key advantages of surrogate models is their ability to explore the design space efficiently to perform extensive sensitivity analyses, optimization studies, or uncertainty quantification without the need for repeated evaluations of the computationally expensive system.

Surrogate models follow a **data-driven approach**, focusing on the input-output relationship rather than the intricate details of the simulation code, this is why they can be built using a variety of Machine Learning techniques, such as Linear Regression [27], Neural Networks [78], Gaussian Processes [60], or other regression models, depending on the nature of the problem and available data.

Prior studies have demonstrated the effectiveness of surrogate models across various

	Analytical Models	**Computational Models**	**Data-driven Surrogate Models**
Examples	Jensen [34], Park [38], Frandsen [26], ...	OpenFOAM [49], OpenFAST [47] ...	Models in 3.1 or in this work
Based on	Flow observations	RANS equations	Real or synthetic datasets
Method	Flow analysis	CFD simulations	Machine learning
Speed	Fast	Slow	Fast
Accuracy	Low	High	Moderate/High

Table 1.1: Classification of Wake Models

domains, including mechanical engineering [32], chemical engineering [15], biomedical engineering [54], architecture [74], aerospace engineering [77], evolutionary computation [36], swarm optimization [65], etc. In the field of wind simulations, Chapter 3 will provide an overview of relevant research that has employed surrogate models for wind turbines energy prediction and wake modelling, which is the primary focus of this work.

The use of surrogate models has consistently demonstrated competitive results while requiring significantly less computational resources and exhibiting lower run times [40]. These models offer a promising approach for addressing computationally demanding problems and enabling efficient data-driven decision-making. Table 1.1 provides a comparison between surrogate models and other wake models. In the subsequent chapters of this thesis, we will explore the specific application of surrogate models within the context of our research focus, highlighting their advantages and demonstrating their effectiveness in addressing the challenges at hand.

1.4 Motivation and Contributions

In the given context and scenario, the purpose and main objective of this thesis is to build a **data-driven surrogate model** for the calculation of wind speed deficits in wind farms. In particular, we consider the problem of modelling the wake effect (in terms of wind deficit) behind the rotor of a single horizontal-axis wind turbine facing a free-stream wind with a single direction at the rotor plane and a fixed speed at hub height. The benchmark used as ground truth for the data-driven approach is the **Ainslie model** [4, 5], detailed in Section 2.1.4. The primary goal is to develop a dependable and precise alternative to this model while significantly enhancing computational efficiency.

Using some input variables, we want to predict the wind deficit at any point in the modelled wake, which typically consists of all points downstream (more than a certain minimal axial distance $x_{\min}$ from the rotor plane) that are defined according to the radial and axial dimensions. More information about these concepts and a more detailed and complete definition of the problem can be found in Sections 2.1 and 4.2.3.

According to the literature review in Section 3.1, the current state of the art, and the investigations carried out within this thesis, several critical **challenges** emerge that must be addressed to effectively tackle this complex problem. Foremost, the scarcity of authentic ground truth data to facilitate model training and validation poses a significant challenge (as discussed in Section 4.2.1). The intricate nature of wakes, stemming from their inherent instability, intricate interactions with atmospheric turbulence, and intricate mutual effects, adds to the complexity. Additionally, the absence of computationally efficient models suitable for wake-induced loads analysis and wind-farm production assessment presents a notable hurdle. Moreover, the task of establishing a robust surrogate model encounters another pivotal challenge: the demanding trade-off between model accuracy and computational efficiency. As a whole, these challenges underscore the multifaceted nature of wind turbine wake modelling and emphasize the need for innovative and adaptive approaches to achieve accurate and practical predictions.

Based on the primary research objective, this report aims to address the following **research questions and subquestions**:

- What is the most effective way to gather data, taking into account the methodologies

used in related investigations and the computational expenses associated with the most accurate data generation methods? How does the Ainslie model compare to these alternatives?

  - Which Ainslie attributes have the most significant influence on its predictive performance in wind turbine wake modelling?
  - What is the optimal extent and diversity of data essential for training a robust surrogate model for wind turbine wake prediction? Furthermore, what strategy for exploring input parameter space should be employed to achieve this?

- Which types of Machine Learning models are most appropriate for wind turbine wake prediction when a substantial amount of data is available? Considering different aspects of suitability, the following subquestions will be explored:

  - In terms of accuracy, which model yields the most appropriate outcomes for wake prediction?
  - What is the visual reliability of different surrogate models in representing the shape and appearance of predicted wake fields?
  - How do the surrogate models trade-off between accuracy and computational efficiency?
  - What unique and particular behaviours emerge from each surrogate model in relation to wake prediction?

- What potential lies in surrogate models to supplant conventional tools for wind farm design and wake estimation?

This work presents several significant **contributions** that collectively advance the field of wind turbine wake prediction.

Firstly, we address a substantial gap in the existing literature by focusing on the analysis of large datasets. Our work takes on the challenge of constructing a surrogate model for the Ainslie method, which significantly amplifies simulation capabilities. This advancement facilitates the generation of a significantly larger number of simulations, leading to the creation of surrogate models that exhibit an impressive enhancement in computational efficiency, improving by several orders of magnitude. Importantly, this efficiency boost is achieved with only a minimal sacrifice in accuracy. In essence, this groundbreaking achievement introduces a highly efficient alternative for wind turbine wake modelling.

Secondly, taking a comprehensive approach, we conduct in-depth analyses of the behaviours exhibited by the developed surrogate models. These analyses not only provide a nuanced understanding of the models themselves but also lay the foundation for cautious generalization of insights to diverse data sources. This contribution enhances the potential for transferring gained knowledge to various scenarios, including further studies with different data generation methods.

Collectively, these contributions, accompanied by practical recommendations and additional insights, significantly propel the utilization of surrogate models in wind turbine wake prediction. Furthermore, they hold the potential to reshape traditional wind farm design approaches by introducing innovative tools for improved efficiency and accuracy.

1.5 Structure of the Report

To address the research questions outlined earlier, this report is structured into distinct chapters, each building upon the previous one to provide a comprehensive understanding of the research journey.

Chapter 2 lays the foundational **background** by delving into the essential context and fundamentals needed to comprehend the challenges faced in wind turbine wake prediction and the subsequent solutions proposed. The **related work** explored in Chapter 3 serves as a bridge between existing knowledge and the innovations presented in this work. Drawing inspiration from a variety of papers, this chapter sets the stage for how this study contributes to the wider field.

In Chapter 4, our **methodology and approach** are detailed. This chapter not only unveils the overarching methodology adopted, but also provides insights into the data collection and generation strategies, experimental setup, and the selection of models under consideration.

Chapter 5, dedicated to **experiments**, takes readers through the practical implementation of the surrogate models. Different experiments are presented and thoroughly discussed, highlighting the performances and behaviours of the models in various contexts.

The journey culminates in Chapter 6, where **conclusions and future work** are presented. Here, the report circles back to the research questions, drawing insightful conclusions from the experiments and suggesting directions for future research.

Chapter 2

Background

This chapter provides the necessary context and foundations to comprehend the problem statement, proposed approach, experiments, and analyses presented in subsequent sections. The aim is to furnish readers with sufficient theoretical grounding across the mathematical terminology, notation and concepts of wind modelling and machine learning to fully grasp the intricacies of the proposed methodology and appreciate the novelty of the approach presented in the following chapters.

In the first part (Section 2.1), the key concepts pertaining to wind turbine wake modelling are elucidated, including characteristics of wind, turbine models, and the assumptions and the mathematical models using differential equations. At the end of the first part, the Ainslie Eddy Viscosity model is defined and described, as it forms the basis for the data generation process.

In the second part (Section 2.2), prominent machine learning techniques for regression tasks are explored, specifically linear regression, tree-based models, Gaussian processes and neural networks. For each method, the underlying principles and mathematical foundations are delineated alongside strengths, limitations, and relevance to the wind wake prediction problem. Brief explanations of recent innovations like Neural Radiance Fields and Fourier feature mappings provide useful connections to the topic at hand.

2.1 Wake Modelling

As mentioned in Chapter 1, the wind turbine wake refers to the highly turbulent airflow trailing behind an operating wind turbine, characterized by a reduced wind speed. In wind farms, where multiple turbines are grouped together, the downstream turbines can be affected by the wakes generated by the upstream turbines, leading to a decrease in their inflow wind speed. Consequently, this reduction in wind speed hampers the power production of the downstream turbines, resulting in a phenomenon known as wake loss [23]. Wake loss represents a significant energy reduction in a wind farm, causing the overall energy output to be lower compared to individual turbines operating in undisturbed conditions at the same location.

This is the reason why the **optimization of the wind farm layout** plays a crucial role: the arrangement of wind turbines in a wind farm has a direct impact on the extent of wake losses experienced by the turbines. Varying layouts can result in different degrees of wake effects, influencing not only the power production but also other factors such as the length of electrical cables required for interconnection or the fatigue load on downstream turbines. Consequently, the optimization should trade off between wake losses, electrical losses and installation costs.

2.1.1 Wind and Wake Characteristics

In this study, the primary focus is to analyze the impact of a *single wake*, specifically examining the influence of the wake generated by an upstream wind turbine on the performance of downstream turbines within a wind farm. Therefore, the analysis is conducted without considering the interactions with other turbines, isolating the effects of a single wake.

The wake can be defined and described using several quantities [41, 73]:

- **Wind deficit** within the wake plays a crucial role. It represents the decrease in wind speed compared to the undisturbed flow, indicating the impact of the turbine on the surrounding air movement.
- **Turbulence intensity** refers to the chaotic and irregular fluctuations in the wind flow, which can significantly affect the stability and behaviour of the wake.
- **Air pressure variations** also contribute to defining the wake characteristics, as they reflect the changes in air density and flow patterns caused by the presence of the wind turbine.

Both velocity and turbulence should be represented as vectors in the wake field, but they can be simplified to scalar numbers. Moreover, considering the radial symmetry (described in 2.1.1.1), the final output can be considered a two-dimensional space representing the wind velocity. If the space is discredited, it can even be modelled as a matrix or a single-channel image (see Section 4.2.3 for dataset details).

This project focuses primarily on studying the wind velocity dimension for two key reasons. Firstly, the data generation model employed in this project, as explained in Section 4.2.3, primarily generates the wind velocity within the wake. Secondly, while acknowledging the importance of other contributing factors, wind speed directly influences the energy production of wind farms, making it a critical factor to consider, as further discussed in Section 2.1.2.1.

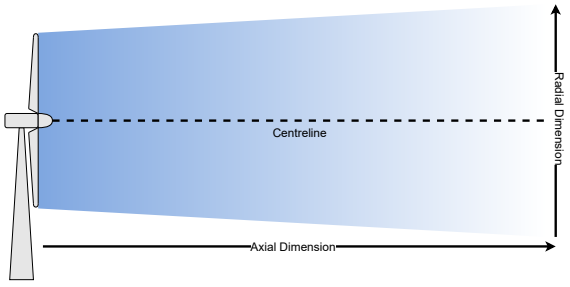


Figure 2.1: Schematic illustration of the wake effect on a single wind turbine. The centreline denotes the axis where radial symmetry is observed.

2.1.1.1 Radial Symmetry

In the context of wake analysis, a specific class of wake models is generally considered, i.e. the models that define the velocity distribution within the wake by combining an axial

wind speed U along the axial direction and a radial wind speed V that remains constant in all directions perpendicular to the axial direction. The wake is assumed to exhibit axisymmetric or radial symmetry around the axial direction.

Radial symmetry refers to the characteristic flow pattern or structure observed in the wake, where the flow or turbulence is symmetrically distributed in a circular or radial pattern around the object. This assumption allows for a simplified representation of the wake in two dimensions, specifically the axial and radial dimensions. By rotating this 2D representation around the centreline (see Figure 2.1), the full three-dimensional wake can be reconstructed.

The velocity components in the wake are functions of the axial distance from the rotor plane and the radial distance from the centreline, typically normalized by the rotor diameter to obtain dimensionless variables (x/D and y/D), where D is the turbine's rotor diameter. The **axial dimension** refers to the free-stream direction of the wind flow, while the **radial dimension** refers to the direction perpendicular to the wind flow. Figure 2.2 shows a schematic drawing of the main elements of axial-radial wake models.

This approach, which considers the wake in two dimensions and assumes radial symmetry, offers a practical framework for investigating wake dynamics and describing various aspects of wake characteristics within wind farm scenarios. By concentrating on the upper rectangle within the space divided by the centreline and rotating it around the axis, a complete 3D representation can be generated. This approach enables efficient analysis and enhances the comprehension of wake features, thereby supporting wind farm layout optimization and downstream turbine performance assessment. However, for the purpose of this project, two mirrored rectangles will be considered to study also how the models learn symmetry.

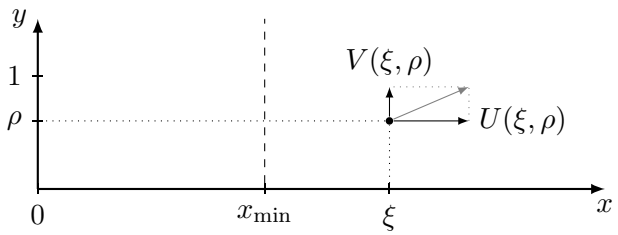


Figure 2.2: A schematic drawing of the main elements of axial-radial wake models. The y-axis represents the radial dimension, while the x-axis corresponds to the axial dimension.

2.1.2 Turbine Model

To understand the context of this project, it is essential to introduce the concept of a **turbine model**, which serves as the foundation for studying the relationship between wind speed and energy generation. Turbines used in wind energy systems are designed to convert the kinetic energy of the wind into electrical power. A turbine consists of various components, including rotor blades, a nacelle, a generator, and control systems. The turbine diameter refers to the diameter of the turbine's rotor blades, which determines the area swept by the blades and influences the amount of wind energy captured.

The findings and insights obtained from this study have the potential to be applied

to various types of Horizontal-axis Wind Turbines (HAWTs), as the generated data is not dependent on specific turbine models or characteristics (see Section 4.2.3). However, it should be noted that the results are not applicable to Vertical-axis Wind Turbines (VAWTs) due to the significant differences in wake characteristics: VAWTs exhibit non-radial symmetry in their wake profiles, which contradicts a key assumption utilized in many approaches for analyzing HAWTs (see Subsection 2.1.1.1). As shown in Figure 2.3, HAWTs (Figure 2.3a) and VAWTs (Figure 2.3b) have distinct designs that result in different wake characteristics. The scope of the analysis exclusively focuses on studying the wake behaviour associated with HAWTs.

In order to comprehensively understand and characterize the wake phenomenon, multiple influential factors need to be considered, impacting its shape and behaviour. These factors encompass various atmospheric conditions, including wind speed, wind direction, and temperature. Additionally, wind turbine characteristics, such as rotor diameter, hub height, and performance parameters like the thrust coefficient and yaw angle[1], play a crucial role. The following sections will delve into two important aspects related to the turbine: the power and the thrust curves.

(a) HAWTs (b) VAWTs

Figure 2.3: Examples of Horizontal Axis Wind Turbines (HAWTs) on the left and Vertical Axis Wind Turbines (VAWTs) on the right.

2.1.2.1 Power Curve

In order to better understand the aim of this project, it is important to introduce the concept of the power curve of a wind turbine to study the relationship between wind speed and the generation of energy. The power output of a wind turbine varies with the wind speed, and every wind turbine has a characteristic power performance curve. The **power curve** gives the electrical power output as a function of the wind speed at hub height, without considering the technical details of its various components [41].

Figure 2.4 shows an example of a power curve for a hypothetical wind turbine, where the dimensions involved are the wind speed (measured in m/s) and the power output (measured in kW or MW) [61].

[1] In wind turbine engineering, the term "yaw" denotes the rotation of the turbine around its vertical axis, enabling it to face the wind optimally and enhance energy capture. Yaw control systems are commonly employed in wind turbines to maintain optimal energy capture by adjusting the rotor's orientation to be perpendicular to the incoming wind direction.

The generation depends on three key points on the velocity scale:

- **Cut-in speed** ($v_{\text{cut-in}}$): the minimum wind speed at which the machine will deliver useful power;
- **Rated wind speed** (v_{rated}): the wind speed at which the rated power (i.e. the maximum power output of the electrical generator) is reached;
- **Cut-out speed** ($v_{\text{cut-out}}$): the maximum wind speed at which the turbine is allowed to deliver power, generally set by engineering design and safety constraints.

These values delineate four distinct regions accordingly, as the figure shows. The former and the latter represent the phases when the turbine is deactivated and the overall electrical generation is zero. The cut-in speed is around 3–4 m/s for most turbines and the cut-out at 25 m/s [41]. Region 2 shows an increase in power output as the wind speed increases from the cut-in value until the rated wind speed, generally following a cubic relationship. In the rated region, instead, P is constant and it reaches its maximum value (i.e. *Rated Power*, or P_{rated}).

The wake effect, being an inherent phenomenon influenced by wind direction variability, poses a challenge in wind farm design. While there are typically one to three dominant wind directions, the wind can approach from any direction, necessitating strategies to mitigate the probability of wind speeds falling below the rated region as this would result in diminished power output.

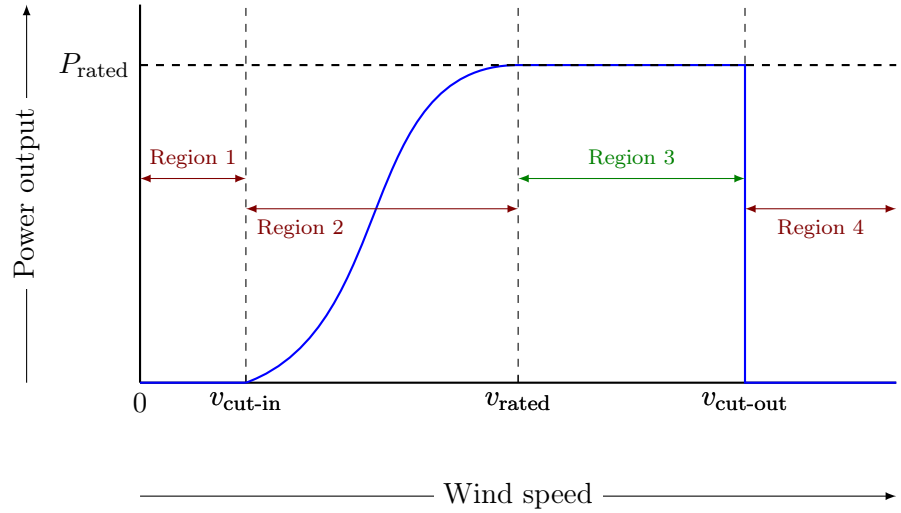


Figure 2.4: Power Curve of a Wind Turbine: Illustrative Performance Profile

2.1.2.2 Thrust Curve

The thrust curve of a wind turbine represents the relationship between the rotor's thrust force and the wind speed. The thrust force is the force exerted by the wind on the rotor blades perpendicular to the direction of the wind. The thrust force is crucial in determining the mechanical loads on the turbine structure and the power output of the turbine. The

thrust force is typically measured in newtons (N), and the wind speed is measured in meters per second (m/s).

The thrust curve plots the dimensionless thrust coefficient (C_T) as a function of the wind speed, providing valuable insights into the behaviour of the turbine under different wind conditions. It plays a crucial role in optimizing the turbine design and performance, ensuring the structural integrity of the turbine while maximizing power production. The shape of the thrust curve is influenced by factors such as blade geometry, rotor diameter, wind turbine controller settings, and the aerodynamic properties of the blades. By analyzing the thrust curve, engineers can make informed decisions regarding the design and operation of wind turbines to achieve optimal performance and efficiency. An example of a thrust curve for the Vestas V80 turbine is in Figure 2.5.

While some physical aspects of what parameters contribute to the Ainslie model will be mentioned in Section 2.1.4, more details about the used input variables will be provided in Section 4.2.3.

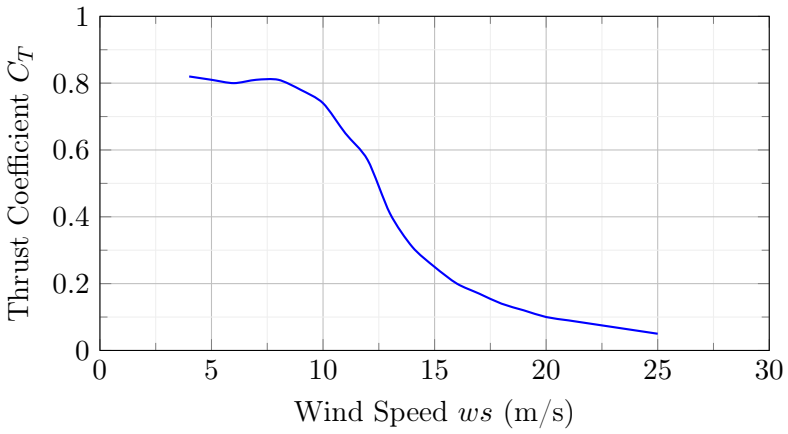


Figure 2.5: Thrust curve for the Vestas V80 turbine.

2.1.3 Modelling Using Navier-Stokes and RANS Equations

Considering the context established in the previous sections, to estimate wake losses accurately and to guide the layout design optimization, it is necessary to employ a reliable **wake model** that can effectively predict the characteristics of the wake. The use of such a model becomes crucial given the need to evaluate numerous layout options, requiring both accuracy and computational efficiency. Thus, the development of a predictive wake model plays a significant role in optimizing wind farm layouts and ensuring efficient power generation.

As mentioned in Section 1.2, various mathematical tools are employed to describe and model the behaviour of the wind wake. One such tool is the differential equation, which arises when it is easier to describe change rather than absolute amounts [30].

The **Navier-Stokes equations** are a set of partial[2] differential equations that describe the motion of viscous fluid substances, including airflow and wind. These equations

[2]Differential equations can be classified into two types: Ordinary Differential Equations (ODEs), which involve functions with a single input and a finite collection of values changing over time, and Partial

capture the fundamental principles of fluid dynamics and account for the conservation of mass, momentum, and energy in the fluid flow [67]. In the context of wind modelling, the Navier-Stokes equations provide a mathematical framework to understand the behaviour and properties of the airflow. Solving the Navier-Stokes equations allows for a complete characterization of the wind flow. However, it is important to note that finding analytical solutions to the Navier-Stokes equations is generally limited to simplified cases due to their inherent complexity.

The **Reynolds-averaged Navier-Stokes** (RANS) equations are a time-averaged form of the Navier-Stokes equations, primarily used to describe turbulent flows [6]. RANS equations incorporate statistical averaging to represent the average behavior of turbulent flows, taking into account the time-averaged quantities. However, RANS equations rely on *turbulence closure* models, which introduce assumptions and simplifications to close the equations, leading to potential inaccuracies.

Eddy viscosity models are a class of turbulence models used to calculate the Reynolds stress term in the equations [22]. These models introduce an eddy viscosity term that represents the turbulent transport of momentum in the fluid. Eddy viscosity models are employed in RANS simulations to capture the effects of turbulence in a computationally efficient manner, and the Ainslie model is one of them [4, 5].

In contrast to RANS simulations based on differential equations, engineering wake models, such as the Jensen model [34], describe the wake using simplified mathematical functions. They are defined by direct analytical expressions for the wake. Such engineering models, therefore, have a computational efficiency advantage over the RANS simulation, which, however, are considered to be more accurate as they provide a more detailed representation of the fluid flow.

To provide insights into the data used and the methodology applied in this project, it is important to introduce the Ainslie model, a RANS-based wake model that will be utilized to generate the data for our study. The following section provides a comprehensive understanding of the Ainslie model, elucidating its fundamental principles and dependencies.

2.1.4 Ainslie (Eddy Viscosity) Model

J.F. Ainslie developed a wake model that combines an axial-radial formulation with simplifications of the Reynolds Averaged Navier-Stokes equation [4, 5].

The Ainslie wake model has gained significant recognition in the industry for its effectiveness in estimating power production and optimizing wind farm layouts. This widespread acceptance has influenced the choice of utilizing the Ainslie model as the basis for constructing a surrogate model, as explained in Section 4.2.3. In the following paragraphs, instead, the fundamental principles and dependencies of the model will be elucidated to provide a comprehensive understanding of its behaviour and the variables or dimensions upon which it relies.

This model, generally referred to in the literature as the **Eddy Viscosity Model**, incorporates heuristic components derived from empirical results to describe the wake

Differential Equations (PDEs), which involve multiple inputs and a continuum of values changing over time, such as the velocity of a fluid at every point in space.

dynamics (that are out of the scope of this project) and relies on the following assumptions [24], which will be also valid for the surrogate model:

- the wake is axisymmetric and fully turbulent, without circumferential velocities;
- the flow field is assumed to be stationary in time;
- beyond the first few rotor diameters downstream/in the far wake region, the gradient of mean quantities in the radial direction is much greater than that in the axial direction (thin shear approximation).

The parabolic Reynolds averaged Navier-Stokes (RANS) equation resulting from the assumptions above is:

$$U\frac{\partial U}{\partial x} + V\frac{\partial U}{\partial y} = -\frac{1}{y}\frac{\partial(y\overline{u'v'})}{\partial y} \tag{2.1}$$

where U is the axial velocity, V is the radial velocity, x is the coordinate of the axial direction (from the wind turbine), and y is the coordinate of the radial direction (from the wake centre).

$\overline{u'v'}$ is the Reynolds-averaged shear stress term, and it is defined as follows:

$$\overline{u'v'} = -\varepsilon\frac{\partial U}{\partial y} \tag{2.2}$$

introducing the *eddy viscosity equation* ε defined by:

$$\varepsilon = l_w(x)U_w(x) + \varepsilon_a \tag{2.3}$$

where l_w and U_w are the length scale and velocity scale describing the wake shear layer respectively and ε_a is the ambient turbulence contribution to the eddy viscosity (e.g. vortexes).

Ainslie considers l_w and U_w to be proportional to the wake width b and the velocity difference $U_o - U_c$ across the wake shear layer ($y = 0$) and considers ε_a to be given by the eddy diffusivity of momentum K_M, therefore the eddy viscosity can be re-written as:

$$\varepsilon = F[k_1 b(U_0 - U_c) + K_M] \tag{2.4}$$

where k_1 is a dimensionless constant whose value has been empirically set to 0.015, U_0 is the free-stream wind velocity, U_c is the centreline velocity and F is the *filter function*, which is a parameter used to simulate the effect of the eddy's momentum on the near-wake region:

$$F = \begin{cases} 0.65 + 0.35 \cdot \sqrt[3]{x - 4.5} & x < 5.5 \\ 1 & x \geqslant 5.5 \end{cases} \tag{2.5}$$

This function's definition is particularly effective in introducing another crucial concept associated with this type of wake models - **the differentiation between the near and far wake regions**. In fact, the equation for the eddy viscosity in the *near wake* (that Ainslie empirically identified as extending up to 5.5 rotor diameters downstream of a turbine) requires modification due to the lack of equilibrium between the mean velocity field and the turbulence field in this region. To address this issue, the treatment of the near

wake region incorporates turbulence data that indicates the accumulation of turbulence in the shear layer of turbine wakes through a filter function that varies based on distance. It is important to note that the limitations in the near-wake region are not specific to the Ainslie model, but constitute a common challenge faced by various wake models due to the complex dynamics of the wake region.

Moreover, in Equation 2.4 b is the dimensionless initial wake width:

$$b = \sqrt{\frac{3.56 C_T}{8 d_M (1 - 0.5 d_M)}} \tag{2.6}$$

and for the initial centreline wake deficit d_M the following empirical expression is used:

$$d_M = C_T - 0.05 - \frac{(1.6 \cdot C_T - 0.5) \cdot I_0}{10} \tag{2.7}$$

where C_T represents the rotor's **thrust coefficient**, and I_0 denotes the **ambient turbulence intensity**. In the original paper [5], turbulence intensity is expressed as a percentage. However, for simplicity and clarity, in this report, it is represented using the decimal notation. The thrust coefficient is a non-dimensional number that compares the axial force exerted on the flow by the turbine to the incoming momentum of the flow.

Finally, K_M in Equation 2.4 is the eddy viscosity momentum, a constant that in neutral atmospheric conditions is expressed as:

$$K_M = \frac{\kappa}{\alpha} \cdot I_0 \cdot U_0 \tag{2.8}$$

with the Von Kármán constant $\kappa = 0.40$ and $\alpha \approx 2.4 \approx \frac{1}{\kappa}$.

In order to solve the original differential equation (2.1), an initial Gaussian wake deficit profile must be defined as the boundary condition at $x_{\min} = 2$ (i.e. it is applied to 2 diameters downstream, as before the dynamics are strongly nonlinear and inconsistent):

$$1 - \frac{U(2, y)}{U_0} = d_M \cdot \exp\left(- 3.56 \cdot \frac{y^2}{b^2} \right). \tag{2.9}$$

Moreover, to force the radial symmetry mentioned in subsection 2.1.1.1, it is also necessary to have the radial velocity to be zero on the centreline:

$$V(x, 0) = 0. \tag{2.10}$$

In conclusion, taking into account the expressions presented in this section and the defined boundary conditions, it is important to note that the Ainslie model relies on the following input parameters:

- the free-stream wind speed U_0;
- the ambient turbulence intensity I_0;
- the thrust coefficient of the turbine C_T.

However, the solution of Ainslie equations demands a numerical integration scheme that, to ensure unobstructed wake expansion, necessitates an integration domain an order of magnitude larger than the rotor diameter. Consequently, the solution process can be time-consuming. and alternative approaches have been devised to address this challenge.

The Anderson Solution A widely used and well-known simplified solution to the Ainslie equations is proposed by Anderson [7]. This approach simplifies the Ainslie model by exploiting the self-similarity of the wake profile in the streamwise direction. By reducing the dimensions from two (streamwise and radial distance) to one (streamwise distance), the equations become significantly easier and faster to solve. It has been demonstrated that this simplification does not sacrifice accuracy or precision to a significant extent, making it widely adopted in the industrial field, including the framework used for generating the data in this work.

The governing equation for the dimensionless wind deficit rate of change in the downstream direction is given by:

$$\frac{\partial \tilde{u}_c}{\partial \tilde{x}} = \frac{16\,\tilde{\varepsilon}\left(\tilde{u}_c^3 - \tilde{u}_c^2 - \tilde{u}_c + 1\right)}{\tilde{u}_c\, C_T}, \tag{2.11}$$

where $\tilde{u}_c$ represents the dimensionless streamwise component of the wind centreline velocity in the wake (U/U_0), $\tilde{x}$ denotes the non-dimensional downwind distance from the rotor (x/D), $\tilde{\varepsilon}$ represents the non-dimensional wake eddy viscosity, and C_T is the thrust coefficient.

Assuming a Gaussian-shaped wake deficit profile, the final equation becomes:

$$\tilde{u}_d(\tilde{x}, \tilde{y}) = \tilde{u}_{d_c}(\tilde{x})\, \exp\left(-3.56\left(\frac{\tilde{y}}{\tilde{w}}\right)^2\right), \tag{2.12}$$

where $\tilde{y}$ represents the non-dimensional radial distance from the wake centreline (y/D), and $\tilde{w}$ is the non-dimensional wake width (w/D), which provides an estimate of the full width of the wake.

2.2 Foundational Machine Learning Regression Techniques

In this section, we explore the theoretical foundations of the machine learning models utilized in this study. By delving into the core concepts and definitions, we develop a deeper understanding of how these models operate and their inherent capabilities. While we will not provide an exhaustive overview of all basic machine learning concepts, our focus will be on some of the most renowned techniques for predicting continuous variables. We will delve into the intricacies of prominent machine learning algorithms, such as linear regression, decision trees, and neural networks. Throughout our exploration, we will examine their respective strengths, weaknesses, and the underlying mathematical principles that underpin their predictive prowess.

2.2.1 Linear Regression

Linear regression is a fundamental and widely used machine learning algorithm for predicting continuous variables. It is based on the assumption that there exists a linear relationship between the input features and the target variable. The goal of linear regression is to find the best-fitting line that minimizes the difference between the predicted and actual values [27].

In its simplest form, linear regression can be represented by the mathematical formula:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \tag{2.13}$$

where y is the target variable, $x_1, x_2, \ldots, x_n$ are the input features, $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the coefficients to be estimated (representing the intercept and slopes of the linear regression line), and ε is the error term.

Linear regression models are computationally efficient and relatively straightforward to interpret, as the coefficients of the features indicate the magnitude and direction of their impact on the target variable. Additionally, linear regression models are less prone to overfitting, a phenomenon where a model learns to perform extremely well on the training data but fails to generalize effectively to new, unseen data. This issue arises particularly when a model becomes excessively complex or when it captures noise in the training data rather than the true underlying patterns. Linear regression models tend to exhibit less overfitting compared to more complex models, especially when the underlying relationship between the features and the target variable is approximately linear.

However, it is important to note that linear regression assumes a linear relationship between the features and the target variable, which may not hold in all cases. Nonlinear relationships and interactions among features are not captured by linear regression, limiting its flexibility in capturing complex patterns. Furthermore, linear regression models are sensitive to outliers and can be influenced by the presence of influential observations.

In the context of wind turbine wake prediction, as we will see in the following chapters, we do not expect a linear relationship between the input features and the wind deficit. Therefore, it is important to explore more sophisticated algorithms to capture nonlinear relationships and interactions effectively.

2.2.2 Tree-Based Regression Models

The decision tree algorithm is a powerful and versatile machine learning technique that can be applied both for classification and regression problems [19]. It is based on a hierarchical structure resembling a tree, where each internal node represents a test on a particular feature, each branch corresponds to the outcome of the test, and each leaf node represents the final prediction or value [62].

Regression tree Regression tree works by recursively partitioning the input space into smaller regions, aiming to minimize the overall prediction error. At each internal node, a feature is selected based on a criterion that measures the ability of the feature to split the data effectively. The most commonly used criterion is the mean squared error (MSE), which quantifies the average squared difference between the predicted and actual values.

Once a feature is chosen, the decision tree algorithm determines the optimal threshold or value for splitting the data into two subsets. This splitting process continues recursively until a stopping criterion is met, such as reaching a maximum depth or a minimum number of samples required to split a node. At this point, the algorithm assigns the predicted value to the leaf node.

One of the main advantages of regression trees is their ability to capture nonlinear relationships between features and the target variable. Unlike linear regression models, decision trees can handle complex interactions and nonlinearities without the need for

explicit feature engineering. Furthermore, decision trees are highly interpretable, as the decision path from the root to a leaf node can be easily understood and visualized.

In the context of wind turbine wake prediction, the interpretability of the regression tree makes it a valuable tool for understanding the impact of different features on wake characteristics. By analyzing the decision nodes and the splitting of the feature space, it is also possible to gain insights into the relationships between these features and the predicted wake values and assess the accuracy of its predictions based on the available data and evaluate the likelihood of overfitting. This interpretability aspect enhances our ability to perform in-depth analyses and make informed decisions regarding wake modelling and optimization strategies.

However, decision trees are prone to overfitting, especially when the tree grows too deep or when the training data contains noise or outliers. Overfitting occurs when the model captures the noise or idiosyncrasies of the training data, resulting in poor generalization to unseen data. To mitigate overfitting, various techniques can be employed, such as limiting the tree depth [48], setting a minimum number of samples per leaf, or using ensemble methods like Random Forest [18, 29] or eXtreme Gradient Boosting (XGBoost) [20].

Random Forest Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It offers improved performance and robustness compared to individual decision trees. Random Forest operates by creating a collection of decision trees, where each tree is trained on a different subset of the training data. This process, known as bootstrapping, introduces randomness in data sampling and feature selection, reducing the risk of overfitting and enhancing the generalization ability of the model. By aggregating the predictions of all individual decision trees, Random Forest provides more accurate and robust predictions, making it less prone to overfitting compared to a single decision tree.

XGBoost XGBoost (eXtreme Gradient Boosting) is another popular ensemble learning method that also builds multiple decision trees. It is known for its excellent predictive performance and has been widely used in various machine learning tasks. XGBoost employs gradient boosting, which involves building weak learners sequentially, each attempting to correct the errors made by its predecessor. By leveraging multiple weak learners, XGBoost can effectively capture complex relationships and interactions between input features and the target variable. This allows for more accurate and robust predictions, making XGBoost a popular choice in various machine learning applications.

The advantages of ensemble methods like Random Forest and XGBoost lie in their ability to reduce bias and variance, leading to improved generalization performance. By combining multiple decision trees, these methods achieve a more robust and accurate prediction. However, the interpretability of ensemble methods may be compromised compared to individual decision trees.

2.2.3 Gaussian Processes

Gaussian Processes (GPs) are a non-parametric and flexible probabilistic modelling approach widely used for regression tasks [60]. Unlike the previously mentioned linear

regression and neural networks, GPs do not assume a fixed functional form and can model complex relationships between input and output variables without explicitly defining the model's structure. Instead, GPs define a distribution over functions and are fully specified by a mean function and a covariance function (kernel) that captures the similarity between inputs.

Given a set of training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the input feature and $y_i \in \mathbb{R}$ is the corresponding output (target) value, a GP models the relationship between inputs and outputs as follows:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{2.14}$$

where $f(\mathbf{x})$ is the latent function mapping inputs $\mathbf{x}$ to outputs y, $m(\mathbf{x})$ is a mean function, and $k(\mathbf{x}, \mathbf{x}')$ is a covariance function (or kernel). The mean function models the expected value of the latent function $f(\mathbf{x})$, and the covariance function measures the similarity between pairs of inputs $\mathbf{x}$ and $\mathbf{x}'$. The kernel function captures the prior beliefs about the underlying function, such as its smoothness and noise characteristics.

To make predictions on new, unseen data points, the GP generates a joint distribution over the observed data and the new data, and then conditions on the observed data to infer the distribution of the latent function at the new points. The predictive distribution of a GP is typically Gaussian, allowing for uncertainty quantification in predictions.

One of the main advantages of Gaussian Processes is their ability to provide a rich representation of uncertainty. In contrast to other models that yield a single-point estimate, GPs generate predictive distributions incorporating the model's uncertainty about the predictions. This is particularly useful in practical applications where uncertainty quantification is critical.

However, one of the main limitations of GPs is their computational complexity. Inverting the covariance matrix requires $\mathcal{O}(N^3)$ operations, making GPs computationally expensive for large datasets. To address this challenge, approximations and sparse GP methods have been developed to reduce computational complexity while preserving the predictive performance.

Approximated Gaussian Processes (AGPs) are a class of methods that approximate GPs using random Fourier features [55, 16, 59]. Random Fourier features approximate the kernel function in GPs using a random feature map, enabling GPs to handle large datasets efficiently. These methods exploit the equivalence between certain kernels and the feature maps obtained from random Fourier features, allowing GPs to scale effectively to large datasets.

In the context of wind turbine wake prediction, AGPs can offer an efficient and flexible alternative to GPs when dealing with bigger datasets. By employing random Fourier features, AGPs can capture complex relationships between input features and target values, providing accurate predictions while overcoming the computational burden of exact GPs.

Overall, Gaussian Processes and Approximated Gaussian Processes provide valuable alternatives to linear regression and neural networks by offering rich uncertainty quantification and flexibility in modelling complex relationships. Their incorporation into the study further enriches the range of methodologies employed and contributes to the overall robustness of the data-driven surrogate models for wind turbine wake modelling.

2.2.4 Feedforward Artificial Neural Networks

In this section, we provide a comprehensive overview of feedforward artificial neural networks, specifically focusing on **Multilayer Perceptron** (MLP), one of the fundamental architectures used in predictive modelling tasks [78]. In this family of neural networks, the information flows in a unidirectional manner, from the input layer through hidden layers to the output layer.

At the core of an MLP are interconnected layers of artificial neurons, known as perceptrons. The perceptrons in each layer receive inputs, apply a weighted sum of the inputs, and pass the result through an activation function to produce an output. The output of each perceptron becomes the input for the subsequent layer:

$$\mathcal{L}^{(l)}\left(\mathbf{t}^{(l-1)}\right) := f_{\text{act}}^{(l)}\left(\mathbf{w}^{(l)}\mathbf{t}^{(l-1)} + \mathbf{b}^{(l)}\right) \tag{2.15}$$

where $\mathbf{t}^{(l-1)}$ is the output of layer l, characterized by the weights $\mathbf{w}^{(l)}$ and the biases $\mathbf{b}^{(l)}$, and $f_{\text{act}}^{(l)}$ is the activation function of the l-th layer.

And this process is repeated until the final output layer, which generates the predictions. If the input $\mathbf{t}^{(0)} \in \mathbb{R}^d$ and the output $\mathbf{t}^{(L)} \in \mathbb{R}^k$, where L is the number of layers, the complete representation of the neural network is:

$$F\left(\mathbf{t}^{(0)}; \mathbf{w}, \mathbf{b}\right) = \left(\mathcal{L}^{(L)} \circ \mathcal{L}^{(L-1)} \circ \cdots \circ \mathcal{L}^{(1)}\right)\left(\mathbf{t}^{(0)}\right) \tag{2.16}$$

where $F : \mathbb{R}^d \mapsto \mathbb{R}^k, t \mapsto t^{(L)} = F(t^{(0)}; (\mathbf{w}, \mathbf{b}))$. A schematic representation can be found

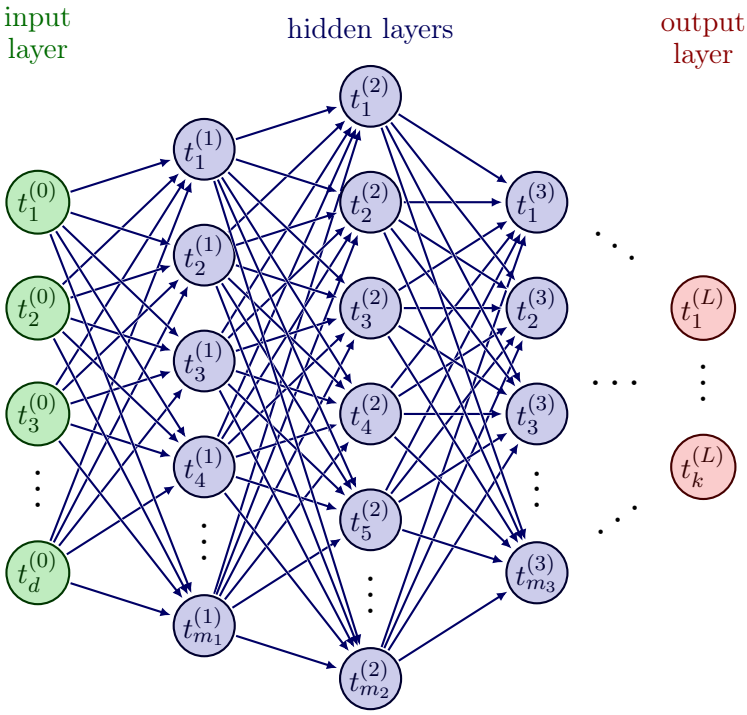


Figure 2.6: Schematic representation of a generic neural network with d inputs, k outputs, L layers, and m_h units in the h-th hidden layer.

in Figure 2.6. The presence of multiple hidden layers with multiple perceptrons allows the MLPs to capture and represent intricate patterns and interactions in the input data.

Training an MLP involves an iterative process of *forward propagation* and *backpropagation*. During forward propagation, input data is fed through the network, and predictions are generated. The difference between the predicted output and the true target values is quantified using a loss function. Backpropagation then calculates the gradient of the loss function with respect to the network parameters and updates the weights and biases using optimization algorithms, such as gradient descent, to minimize the loss.

Neural networks have been successfully utilized in a wide range of practical applications, including image classification, natural language processing, and time series forecasting. They have shown remarkable performance in capturing intricate patterns and achieving state-of-the-art results in various domains thanks to their ability to learn complex nonlinear relationships within the data. Other advantages are that they can handle large-scale datasets efficiently, thanks to their parallel computing capabilities and the availability of optimized implementations using graphics processing units (GPUs). Moreover, MLPs exhibit generalization ability, allowing them to make accurate predictions on unseen data by capturing underlying patterns and avoiding overfitting.

However, designing an MLP requires making decisions about the number of hidden layers, the number of perceptrons in each layer, and the choice of activation functions. These design choices can have an impact on the network's performance and generalization ability and may require careful tuning.

Over the years, various advancements have been made in MLP architectures, activation functions, regularization techniques, and optimization algorithms. Different strategies, such as dropout and weight decay, have been developed to mitigate overfitting and improve generalization. Additionally, techniques like batch normalization and adaptive learning rate schedules have been introduced to enhance training stability and convergence.

It is important to note that there are many different and specific types of neural networks, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which excel in specific domains. The choice of network architecture depends on the nature of the data and the specific requirements of the prediction task. For the sake of this project, for the reasons mentioned in 4.4, there is only another neural network-based architecture that has to be mentioned, and it will be defined and described in the following section.

Neural Radiance Field and Random Fourier Features Despite the remarkable ability of neural networks to approximate any function, as demonstrated by the universal approximation property, some studies have observed a phenomenon known as "spectral bias" in MLPs [56, 12]. Spectral bias refers to the challenge faced by MLPs in accurately representing the high-frequency content found in natural images and scenes due to the rapid falloff of frequencies. In their work, Mildenhall et al. [43] experimentally discovered that a heuristic sinusoidal mapping of input coordinates, known as "positional encoding", enables MLPs to effectively represent higher frequency content.

This positional encoding technique is a particular case of Fourier features [57] and serves as the foundation for a novel approach that has garnered significant attention in the fields of computer graphics and computer vision: **Neural Radiance Fields** (NeRF).

NeRF is a powerful technique that enables the realistic synthesis of complex scenes and the generation of high-fidelity images.

While most recent studies in NeRF have primarily focused on 3D scene modelling and volume rendering [43, 42, 39], this project diverges from that scope. As explained in Section 4.2.3.4, the data used in this project is two-dimensional. However, it is worth noting that the wake field can be considered a monochrome image consisting of a single channel, which represents the wake deficit value. Therefore, the work of Tancik et al. [66] is highly relevant for wake modelling, and it will be discussed in Section 3.2.

Chapter 3

Related Work

The landscape of wind turbine wake modelling and prediction has witnessed a surge of interest in data-driven surrogate models, machine learning techniques, and innovative methodologies. In this chapter, we delve into the relevant body of work that forms the foundation for our research, highlighting the distinctive contributions that have informed and shaped our approach. The first section (Section 3.1) provides an overarching view of the diverse data-driven surrogate models, categorizing them based on their use of real-world data or synthetic datasets. This comprehensive survey underscores the strides made in this field and elucidates the prevailing trends in wind turbine wake modelling. The second section (Section 3.2) takes a more specific focus, delving into works outside of wake modelling that have inspired the experimental design. In particular, the influential work that explores the impressive capacity of Neural Radiance Fields (NeRF) to surmount the spectral bias challenge.

3.1 Data-driven Surrogate Models

The field of surrogate models for the wake effect has undergone significant research and development, with the goal of improving the accuracy and computational efficiency of modelling approaches. Various studies have focused on employing the Ainslie model and its solutions [4, 5, 7], as well as other analytical or computational methods, to approximate the behaviour of wake phenomena. These models have been widely embraced in the wind energy industry due to their effectiveness in estimating power production and optimizing wind farm layouts.

However, in recent years, there has been a growing interest in exploring data-driven surrogate models for the wake effect. These models leverage machine learning techniques and algorithms to capture complex patterns and relationships within wake data, enabling more precise predictions and efficient evaluations.

Despite the notable advancements in the field, it is essential to highlight that the attention given to data-driven surrogate models for the wake effect has been relatively limited compared to the extensive efforts invested in analytical, experimental, and numerical studies of wake flows. The majority of the existing literature has primarily focused on traditional mathematical and engineering models, with only limited exploration of machine learning-based approaches.

To address this gap, this chapter aims to review and analyze the existing research on data-driven surrogate models for the wake effect. It emphasizes the potential benefits and challenges associated with adopting these models. Throughout the following sections, we will delve into the literature to explore the current state of research in data-driven surrogate models, examining the methodologies, techniques, and applications employed.

By comprehensively understanding the advancements in this area, we can identify the potential contributions that data-driven surrogate models offer in enhancing the accuracy and efficiency of wake modelling in the wind energy industry. This analysis will also shed light on the opportunities for further development and improvement in this field.

The chosen approach for categorizing the papers in the related work involves distinguishing between studies that utilize real data and those that rely on synthetic data. This classification method provides a comprehensive perspective on the sources of data used in the research, enabling a thorough examination of the associated challenges and limitations.

Furthermore, it is worth noting that, to the best of our knowledge, no prior research has conducted a comparative analysis of the different data-driven surrogates for wake prediction that use datasets obtained from various models or sources. It is well-established that the simplifications and heuristics applied in certain fluid dynamics simulations can significantly impact the profile and intensity of the wake, deviating from real-world observations. Considering the potential variations among these models, it is reasonable to anticipate that disparities in the models or their parameters employed for data generation may further impact the wake's characteristics. This notion is reinforced by the significant dissimilarities observed among the generated and utilized datasets, as summarized in Table 3.3. Additionally, it is important to mention that the differences in the features and outputs make direct comparisons between the different approaches challenging or even unfeasible.

3.1.1 Using Real Data

Despite the challenges in accessing real data that accurately describes the wake effect in operational wind farms (as discussed in more detail in 4.2.1), there are a few studies that have utilized real datasets, and it is essential to provide an overview of them first. This overview will offer valuable insights into the availability and characteristics of the data that will serve as the basis for approximation using the selected data generation method.

Ashwin Renganathan et al. [9] utilized LiDAR measurements for data-driven wind turbine wake modelling. The dataset was collected between August 2015 and March 2017 at a wind farm in North Texas, where a WindCube 200S sensor was installed on one of the 25 wind turbines. The LiDAR data consists of 2501 dimensions, providing detailed information on the wind flow. In addition to LiDAR measurements, meteorological and SCADA (Supervisory Control and Data Acquisition) data were continuously recorded throughout the campaign, capturing 10-minute mean and standard deviation values of wind speed, as well as parameters such as wind direction, temperature, atmospheric pressure, active power, RPM, and blade pitch angle. After the preprocessing steps (filtering out points with a low carrier-to-noise ratio and realignment with the wind direction, estimation of the horizontal equivalent velocity and correction for vertical variability due to wind shear), the dataset consisted of 6654 quality-controlled, re-aligned, and non-dimensional LiDAR scans. Each instance represents a spatiotemporal measurement of the wind flow characteristics, capturing the wake behaviour of the wind turbines.

The authors' approach uses deep neural networks (in particular, convolutional autoencoders) to achieve a drastically compressed latent-space representation of the high-dimensional LiDAR data. They then use multilayered perceptrons (MLPs) and Gaussian

processes (GP) to learn the input parameter-latent-space map. Moreover, an alternative approach is proposed to address the well-known tractability issues with exact-inference Gaussian Processes [60] by using variational sparse Gaussian processes.

The results show that the predictive capability of all the machine learning models is somewhat similar. However, Gaussian process regression with exact inference resulted in the least root mean square error (RMSE), indicating the best prediction.

Similarly, also the prediction time is comparable among the different approaches as they all take $O(1)$ second of wall-clock time to evaluate, showing that the sensitivity of the wake flow field to the input parameters can be analyzed in real-time, compared to the unrealistic CPU hours necessary for high-fidelity simulations.

However, there are some limitations that must be acknowledged regarding the used dataset:

- The presence of outliers in the dataset (approximately 10% of the overall dataset) poses a challenge in terms of data quality and may influence the modelling results.
- The LiDAR measurements are corrupted by unknown noise, which is not captured by the models. Although the Gaussian process (GP) models employed in this approach can address some aspects of the noise, they make simplifying assumptions such as independent and identically distributed noise, which may not fully capture the complexity of the real noise structure.
- The raw measurements have missing elements, which are imputed using a local interpolation method. It is important to note that this imputation approach may introduce biases in the dataset, potentially affecting the accuracy of the modelling results.
- The number of instances in the dataset is relatively small, and their partial coverage further limits the generalizability of the findings.

To the best of our knowledge, this is the only surrogate-modelling approach that exclusively utilizes data collected from wind farms, offering valuable insights into the real-world behaviour of wind turbines' wake effect.

Nai-Zhi, Ming-Ming, and Bo [45] still use real data, but they combine them with analytical models. The authors of this work focus on predicting power generation, and their approach's optimization problem centres around minimizing the difference between the predicted and calculated power generation using the wake model. Therefore, the next paragraphs will solely analyze the results achieved in the wind speed prediction (as a single scalar value). They integrate a basic analytical wake model (proposed by Bastanjhah and Porté-Agel [13]) with SCADA data for capturing meteorological and operational information from each wind turbine. However, only the features recorded by the upwind turbine are considered in the analysis. The meteorological data (wind speed, wind direction, turbulence intensity and temperature) serve as input features for the model. Meanwhile, the operational data (yaw angle and power output) is combined with the analytical wake model.

The authors propose a new model based on Random Forest called the Data-Driven Analytical Wake Model (DDAWM) to learn the relationship between local inflow information

and wake expansion features. Two different datasets are leveraged to test the resulting model (A and B). The first dataset includes 2766 time points of simpler SCADA data, including direction and wind speed as inflow features. The second dataset consists of more complex instances with additional inflow features and operational information about the yaw angles.

To verify the performance of the proposed DDAWM, the basic analytical wake model is used to predict the wind velocity at the same sample points. The results, summarized in Table 3.1, demonstrate that DDAWM achieves a significant improvement in the R^2 score for velocity calculation compared with the traditional model. The DDAWM model exhibits good generalization ability and prediction accuracy on the new dataset.

Another work utilizing real data is worth mentioning, although the available results solely pertain to estimated power production. Hence, the focus of this project is not on the applied machine learning models, training processes, or specific results. The inclusion of this work aims to complete the overview of the available real dataset in the literature research. Japar et al. [33] leveraged performance data from the Horns Rev offshore wind farm in Denmark to predict wake losses using machine learning, combined with nearby meteorological masts to obtain wake-free conditions (wind speeds and direction). It is important to note that this dataset is limited in size, consisting of only three different wind speeds and three wind directions. Additionally, the dataset solely provides information on the power generated (in kW) by five or eight turbines, limiting the definition of wake losses to the gradients between two consecutive turbines in power generation.

In conclusion, it is important to consider the limitations concerning the quantity and quality of the data collected from real wind farms, which are applicable to all the previously mentioned works. In [45], the authors had to compensate by combining data from other models, but it should be noted that the chosen model belongs to the class of engineering models mentioned in Subsection 1.2.1, which are recognized for their restrictions in terms of reliability and accuracy.

Considering these factors, the subsequent studies presented in the following section will exclusively focus on synthetic datasets generated primarily through more accurate computational fluid dynamics (CFD) simulations.

Case	**Model**	**Training Set R^2**	**Test Set R^2**
Case A	Basic Model	0.496	0.508
	DDAWM Model	0.749	0.744
Case B	Basic Model	0.729	0.728
	DDAWM Model	0.863	0.852

Table 3.1: Comparison of R^2 scores for Basic Model and DDAWM Model in [45]. The data used in Case A comes from a flat terrain and simpler environment, while the data used in case B is more comprehensive and complex, with diverse surrounding environments and complete SCADA data.

3.1.2 Using Synthetic Data

As mentioned in Subsection 1.2.2, computational fluid dynamics (CFD) simulations are not limited to the study of wind turbine wake alone. The Navier-Stokes equations, upon which most CFD simulations are based, are applicable to various Newtonian fluids. Consequently, it is relatively easier to find relevant works in the broader field of fluid dynamics, including those specifically addressing wind-related applications, e.g. in the aviation industry.

These simulations often utilize similar underlying models, with differences lying primarily in the specific objects being studied. For instance, in [14], compressible Reynolds-averaged Navier-Stokes (RANS) equations (as mentioned in Subsection 2.1.3) are solved to simulate the wind flow around three types of airfoils. The output of the simulation is a two-dimensional velocity field that represents the velocity distribution around the airfoil.

In the context of wind turbine wake modelling, some approaches have emerged that leverage synthetic datasets to simulate and analyze wind turbine wakes. These studies serve as valuable contributions to the field, offering insights into the behaviour and characteristics of wakes under controlled conditions. The following paragraphs provide an overview of some notable works in this area and highlight their key contributions.

Zhang and Zhao [79] run a series of large eddy simulations to generate a high-fidelity CFD database capturing the wind flow around turbine rotors. In a manner similar to the approach taken in [9], the authors employed dimensionality reduction techniques to process the original flow field data. Their comparative analysis revealed that singular value decomposition outperformed other methods such as auto-encoders and independent component analysis in this particular context.

Considering that the simulations also include the temporal dimension (1110s simulations are carried out with a time step of 0.02s), the authors approached the problem as a time-series prediction task. In particular, in addition to the inflow features (like velocity and the control parameters), also the flow field at time T is required as the input in order to predict the flow field at time step $T + 1$. Consequently, they opted for a Long-Short Term Memory (LSTM) network due to its ability to model temporal dependencies. However, since the time component is not a focal point in this project (as the Ainslie model used for data generation generates static simulations), the specific results derived from the LSTM network are not particularly relevant to the present investigation.

On the contrary, the flow field extracted from a simulation for a single turbine at a specific timestamp closely resembles the one generated for this study (refer to Section 4.2.3). This extracted flow field is a two-dimensional representation that undergoes discretization and interpolation to form a uniform grid consisting of $N_x = 50 \times 30 = 1500$ cells, effectively capturing the velocity distribution behind a single turbine.

It is worth noting that the generation phase required significant computational resources. Specifically, 180 distinct flow scenarios, each comprising 710 discrete time instants, demanded approximately 7×10^5 CPU hours on local high-performance computing (HPC) clusters.

Wilson, Wakes, and Mayo [75] conducted experiments using the CFD software "ANSYS Fluent" to simulate the 3D wake effect of a wind turbine. Eight simulations were conducted with different wind speed values ranging from 5.5 m/s to 17.5 m/s. Each simulation resulted in approximately 72,800 instances of data, and each data instance includes the x,

y, and z coordinates along with the corresponding wind velocity. It is important to note that no meteorological features or inflow parameters were used in this study. The primary objective of these experiments was to train models to predict the velocity at specific 3D points based on training data from different locations or wind speed values.

Various regression algorithms were tested in these experiments, including Mean, Linear Regression, M5, Random Forest, and Multi-Layer Perceptron. The experiments were conducted in the following settings to explore the generalization capabilities of the developed models:

- *Interpolation*: Five out of the eight datasets were used for both training and testing, examining the model's interpolation capability.
- *Extrapolation*: Five datasets were used for training, while the remaining three datasets were used for testing. This setup evaluated the model's extrapolation capability.

As a sub-experiment within the extrapolation setting, a Multi-Layer Perceptron was also trained, using different numbers of hidden nodes (through empirical analysis, it was determined that using 32 hidden nodes yielded the best results). The Multi-Layer Perceptron was trained on the same features as the other regression algorithms, along with additional non-linear transformations of the Cartesian coordinates (e.g. $\frac{1}{x}$, x^2, $\frac{1}{x^2}$, xy, etc.).

These experiments aimed to assess the performance of the regression algorithms in predicting the wake velocity at specific 3D points, considering both interpolation and extrapolation scenarios, and the best results are summarized in Table 3.2.

The results indicate that the MLP model with 32 hidden layers performs the best in terms of extrapolation, with the lowest MAE of 0.1749 m/s. However, it is worth noting that all the machine learning models demonstrate significantly better performance in the interpolation setting, which is relatively simpler to predict.

Also Ti, Deng, and Yang [69] have used RANS simulations to provide a massive dataset of wake flows for training, testing, and validation of the Artificial Neural Network model.

The input variables are inflow hub-height velocity and turbulence intensity. The generation of the flow field in each simulation requires therefore a combination of these two variables. The ranges for both are decided respectively according to the operation wind speed ranges of the tested wind turbine Vestas V80 2MW (from 5 to 20 m/s, with a step of 0.5 m/s) and the turbulence intensity reported in offshore wind farms (from 2% to 26%, with a step of 2%).

Model	**Mean**	**LR**	**M5'**	**RF**	**MLP**
Interpolation	0.716	0.081	0.022	**0.012**	–
Extrapolation	4.4494	**0.195**	1.144	1.451	**0.1749**

Table 3.2: Comparison of Mean Absolute Error (MAE) in m/s of the wind velocities predicted by the different machine learning algorithms in the interpolation and extrapolation settings of [75].

The total number of simulations is therefore $31 \times 31 = 961$, out of which 20 are reserved for testing. The resulting flow field is interpolated into an $x \times y \times z = 40 \times 250 \times 24$ matrix $\mathbf{V}$, representing a 3D flow data with 240,000 cells. Each cell contains the wake velocity deficit and the added turbulence kinetic energy (TKE), although the TKE is not relevant to the scope of this project. To train an artificial neural network that can predict 240,000 elements, the authors sliced the output variable v into 2,000 partitions, each of which contains 120 elements. Each partition is trained independently in a sub-model with 10 neurons and a single hidden layer using parallel computing. The comparison between the ANN predictions and RANS/ADM-R simulations shows that the ANN-based wake model can accurately predict the velocity of a standalone turbine with errors of less than 5% in most of the far-wake regions, demonstrating superior performance compared to classical analytical wake models such as Jensen [34] or Bastankhah-Porté-Agel [13]. These results have also been further analyzed and extended for power prediction applications in [70], demonstrating the model's ability to generalize to broader problems and be applied to real-world wind farms.

Other experiments involving complex RANS simulations have been conducted in [80] and [53]. In the former study, the authors utilized a deep convolutional conditional generative adversarial network (DC-CGAN) and achieved a prediction error of 0.102 m/s for the streamwise velocity (the velocity component parallel to the primary flow direction) and 0.045 m/s for the spanwise velocity (the velocity component perpendicular to the primary flow direction). For the latter, the implicit relationship between inflows and wake flows is established using Support Vector Regression, Artificial Neural Networks and Extreme Gradient Boosting (XGBoost) techniques. The results, both in terms of R^2 score and Mean Squared Error (MSE), are almost aligned for neural networks and support vector regression, reaching an MSE error of 0.0015 and 0.0016 respectively in the test set.

3.1.3 Final Remarks and Insights

This chapter has provided an in-depth analysis of the existing research on data-driven surrogate models for the wake effect. The review has highlighted the increasing but still immature interest in leveraging machine learning techniques to capture complex patterns and relationships within wake data. While traditional mathematical and engineering models have been widely used in the wind energy industry, data-driven approaches offer the potential for more accurate predictions and efficient evaluations.

One key distinction observed in the related work is the utilization of either real or synthetic datasets. Studies employing real data face limitations in terms of the quantity and quality of the collected data from wind farms. These limitations necessitate compensatory measures, such as combining data from other models. On the other hand, studies using synthetic datasets generated through CFD simulations offer several advantages. Synthetic datasets allow for precise control over the underlying physics, the generation of larger volumes of data, and the absence of limitations associated with real-world data collection (such as installing sensors at different locations, data interpolation, detection noise, etc.), providing more insights into wake characteristics, dynamics, and predictive capabilities.

However, some significant limitations are also prevalent among the works carried on synthetic datasets.

Small dataset Real datasets often face limitations in terms of the number of instances available due to challenges in data collection from wind farms or issues with spatial distribution. On the other hand, synthetic datasets generated through CFD simulations can provide a large volume of data, but their size is still restricted by the high computational cost of complex CFD software. As a result, the number of simulations is generally limited, and even the resolution of the wake field may be constrained, although some studies have managed to perform 3D simulations.

Small feature space Synthetic datasets, as mentioned in Section 3.1.2, typically exhibit a limited number of input features. For instance, in [75], only wind speed, coordinates, and position relative to the turbine hub are considered. This limitation arises due to the constrained number of simulations that can be conducted within a reasonable timeframe, preventing the generation of wake fields with numerous parameters. Additionally, wake models, in general, tend to simplify the representation of parameters to enable quick computation. Interestingly, real datasets often incorporate more inflow features as they are combined with meteorological information from actual meteorological masts.

Regarding the size of the output space, i.e., the wake field, there is considerable variation among different studies, with the number of cells ranging from 1,500 in [79] to 240,000 in [69]. The number of cells can be controlled by the models used to achieve sparser or finer grids, depending on the desired precision and available computational and storage resources. It is worth noting that simulations enable the generation of multi-dimensional and complete wake fields, while real datasets either predict single values or rely on physical sensors installed at specific locations for spatial representation, necessitating interpolation

Paper	Data Origin	Wake Represent.	Size
Ashwin Renganathan et al. [9]	Real	Static 2D field	6654 instances
Nai-Zhi, Ming-Ming, and Bo [45]	Hybrid	Dynamic single value	2766 instances (dataset A)
Zhang and Zhao [79]	Synthetic	Dynamic 2D field	180 simulations of 710 instants and 1,500 cells each
Wilson, Wakes, and Mayo [75]	Synthetic	Static 2D field	8 simulations of ≈9,100 cells each
Ti, Deng, and Yang [69]	Synthetic	Static 3D field	961 simulations of 240k cells each
Purohit, Ng, and Kabir [53]	Synthetic	Static 2D field	9 simulations of ≈1,150 cells each

Table 3.3: Comparison of related work papers based on dataset characteristics. The size refers to the instances after any pre-processing or cleaning. Static and dynamic refer to the presence of the time dimension.

for the rest of the domain.

Lack of comparability The following points highlight some key challenges that contribute to the difficulty of comparing and drawing meaningful conclusions in the field of data-driven surrogate models for wake modelling:

- *Differences in dataset characteristics*: Comparing machine learning methods becomes challenging when datasets differ in terms of source or quantity. However, establishing a direct comparison becomes practically impossible when datasets differ not only in terms of input features and dimensions of the output space but even in the output variable itself, as observed in cases where turbulence intensity or power generation is predicted. Even in terms of wake representation, as shown by Table 3.3, there are significant discrepancies.

- *Inconsistency in evaluation metrics*: Adding to the complexity, variations extend beyond the datasets and encompass the metrics used for evaluation. Different papers employ diverse error measures, such as Mean Squared Error or Absolute Error, while others focus on the R^2 score in the wind deficit. This lack of uniformity in metrics further hinders the ability to compare and draw meaningful conclusions across various studies.

- *Absence of conventions and synchronization*: The absence of established conventions or synchronization in the field of data-driven surrogate models for wake modelling contributes to these disparities. With most papers being relatively recent and published within a short timeframe, comprehensive reviews or surveys have not yet been conducted to standardize the methodologies or metrics employed. As a result, researchers have operated within their own frameworks and preferences, leading to divergence in approaches and evaluation methods.

These discrepancies are the primary reason for the lack of comparability between the current work and the state of the art, especially from a quantitative perspective. Nevertheless, a qualitative comparison will still be conducted in 4.4.3 to gain insights into the overall performance and effectiveness of the proposed approach compared to existing methods.

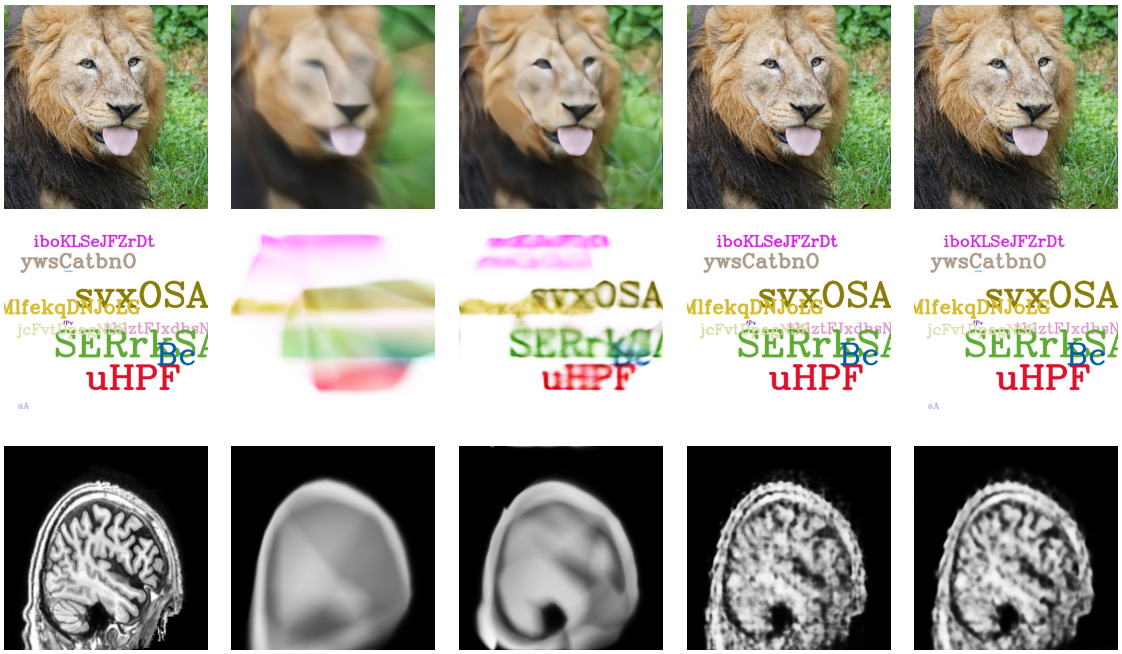
In conclusion, the analysis of real and synthetic data-driven surrogate models has provided valuable insights into the progress and constraints of this field. It is evident that further investigation is necessary to fully harness the potential of data-driven approaches and address the existing challenges and complexities. Building upon these limitations, the next chapter will present a detailed methodology for data generation and surrogate model training. This approach aims to overcome the data scarcity issue by utilizing a less complex model than CFD simulations, yet more reliable than engineering models, to generate a large number of instances. Each instance will encompass comprehensive information about a two-dimensional wind deficit field. By addressing these limitations, the proposed methodology seeks to enhance the accuracy and efficacy of wake modelling in the wind energy industry.

3.2 Fourier Features for 2D Image Regression

In this section, we delve into a study that explores the application of Fourier feature mappings in coordinate-based MLPs for regression tasks. This investigation intriguingly aligns with the methodology adopted in this project, offering valuable insights that have significantly influenced our experimental approach and decision-making. This follows the overview of Neural Radiance Fields (NeRF) provided in Section 2.2.4. While our central focus remains on wind turbine wake prediction through data-driven surrogate models, delving into the findings of this paper casts a broader light on the landscape of neural network applications. These insights have inspired our decision-making process and contributed to the meticulous design of our experiments.

The paper [66] investigates the benefits of using Fourier feature mappings in coordinate-based MLPs, which aligns with the approach employed in this project (see Sections 4.3.2 and 4.4.1). The ensuing results offer valuable insights that transcend the confines of standard image regression. While the authors investigate various regression tasks, this project specifically focuses on two related tasks: 2D image regression and 2D computed tomography (CT). In both cases, the model takes a 2D pixel coordinate as input and predicts the corresponding RGB value for images or the volume density for CT data. This approach is akin to what has been undertaken in this work for wake modelling, as further elaborated in Section 4.3.2.

The study compares the performance of coordinate-based MLPs with no input mapping



(a) Ground truth (b) No mapping (c) Basic (d) Positional enc. (e) Gaussian

Figure 3.1: Reconstruction results of [66] for the 2D image regression and 2D CT tasks. Top rows: Two images from the Natural dataset. Third row: One image from the Text dataset. Bottom row: One image from the 2D CT dataset.

and with three variants of random Fourier feature (RFF) mappings: Basic, Positional Encoding, and Gaussian. All tasks utilize L2 loss and a ReLU MLP architecture with 4 layers and 256 channels. For the 2D image regression tasks, 512×512 resolution images are employed, where a sub-sampled grid of 256×256 pixels is used for training and an offset grid of the same size is used for testing.

The experimental results demonstrate that all Fourier feature mappings improve the performance of coordinate-based MLPs compared to using no mapping at all. Specifically, the Gaussian RFF mapping achieves the best results across all experiments. Furthermore, the results reveal an intriguing enhancement in image quality when Fourier feature mappings are employed (see Figure 3.1). The fuzziness and blurriness observed in images without mapping (b) are significantly mitigated, particularly by Positional Encoding (d) and Gaussian (e) mappings, which closely resemble the original images. This observation implies that Fourier feature mappings can substantially enhance the suitability of coordinate-based MLPs for modelling functions in low dimensions, effectively overcoming the spectral bias inherent in such models.

The insights gleaned from [66] resonate with our quest to enhance wind turbine wake prediction through data-driven surrogate models. While the contexts differ, the conceptual parallels between our objectives and the findings of this paper have informed our experimental decisions and added depth to our understanding of neural network behavior.

Chapter 4

Methodology and Approach

In this chapter, we present a comprehensive methodology and approach utilized to develop and evaluate the data-driven surrogate models for wind turbine wake predictions. The general pipeline of our work comprises several key components and stages, which will be briefly introduced in Section 4.1.

The foundation of our methodology lies in the data collection process, serving as the core starting point. In Section 4.2, we meticulously examine the various possibilities for wind turbine wake datasets, understanding their characteristics and limitations, and provide a detailed description of the data generation process. Subsequently, in Section 4.3, we focus on the experimental setup, delineating all the settings and processing steps necessary to conduct the experiments presented in the following chapters.

Finally, Section 4.4 delves into the modelling aspect, exploring the architectures and hyperparameters of the most promising data-driven surrogate models, along with the metrics used for quantitative evaluation. Additionally, this section contains considerations regarding similarities and differences with related work to provide a comprehensive understanding of their characteristics and performance.

4.1 General Pipeline

Figure 4.1 provides a visual representation of the general pipeline followed in this work, offering an overview of the subsequent sections of this report.

The first segment of the pipeline focuses on the data generation process, elaborated in Section 4.2.3. Given the scarcity of real-world datasets for wake predictions, we employ the Ainslie model as a technique to generate synthetic data that effectively mimics the wake characteristics of real wind turbines. This step is pivotal, as it directly impacts the performance and generalization capabilities of the data-driven surrogate models to be developed and evaluated. The Ainslie implementation utilizes specific inflow parameters (e.g., TI and C_T) to generate the actual wake fields, which serve as the ground truth for our experiments.

The subsequent part of the figure addresses some aspects of the experimental setup, including data processing (detailed mainly in 4.3.3 and 4.3.6) and data splitting strategies (refer to 4.3.4). Dataset splitting strategies play a crucial role in our experimental setup, serving both classical machine learning requirements and the data reduction step. We introduce different splitting strategies, each corresponding to a distinct experiment, providing deeper insights into our models' performance under various scenarios, including interpolation and extrapolation settings. Additionally, in Section 4.3, we provide insights into the hardware used for our experiments, ensuring efficient training and evaluation of our models. Furthermore, we discuss the distinction between univariate and multivari-

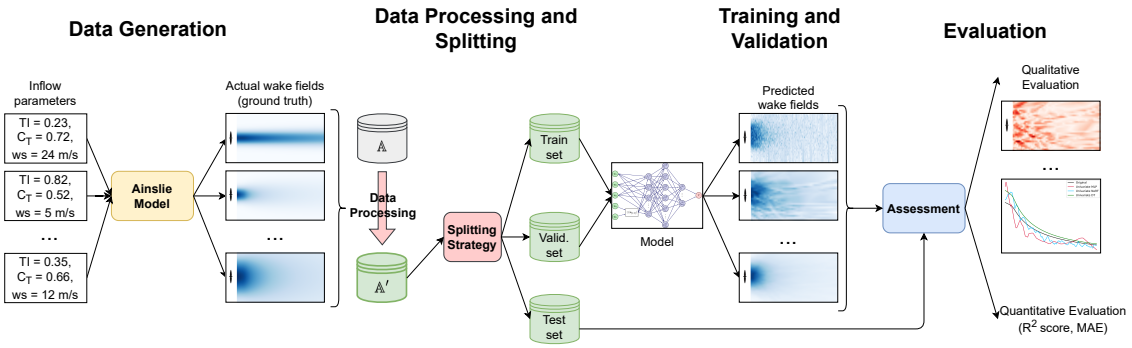


Figure 4.1: Schematic representation of the pipeline.

ate modelling approaches, each posing unique challenges and advantages in capturing the intricate wake interactions.

Moving forward, the third segment of the figure depicts the modelling aspect, which is discussed properly in Section 4.4. In this section, we explore the architectures and hyperparameters of the most promising data-driven surrogate models. Each model's architecture is carefully designed to capture the complex relationships within the wind turbine wake data. Additionally, we conduct a comparison with related work to provide a comprehensive understanding of our models' strengths and limitations. Finally, we introduce the performance metrics utilized for quantitative evaluation. The last part of the Figure is dedicated to the evaluation, discussed in Chapter 5.

Throughout this chapter, we systematically address the various components of our methodology and approach, ensuring the robustness and reliability of our data-driven surrogate models for wind turbine wake predictions. The combination of data generation, experimental setup, and model exploration enables us to gain valuable insights and draw meaningful conclusions from our evaluation.

4.2 Data Overview

Before delving into the details of the process for generating the dataset that has been used for all the experiments, it is essential to provide a broader overview of the available options in terms of real datasets and other generation techniques. Indeed, while the primary objective of the original thesis has consistently been the investigation of the Ainslie model's surrogate, significant attention has also been devoted to acquiring real data from diverse sources and generating synthetic data through alternative computational models. The purpose behind this endeavour was to harness a multitude of data sources for the purpose of testing and comparing the models. Recognizing that variations in the modelling techniques employed may yield valuable insights and implications for the overall analysis, and exploration of these differences may become an area of potential interest and relevance.

4.2.1 Lack of Real Datasets

Leveraging data directly from wind farms holds the potential to provide a realistic and practical foundation for understanding the complexities of wake modelling and enhancing the applicability of findings to real-world wind farm scenarios. However, the scarcity of real-world data poses a significant challenge in studying the wake effect, as mentioned in Sections 1.4 and 3.1.1. Through a comprehensive review of the existing literature (refer to Chapter 3), it became evident that a majority of the studies relied on synthetic datasets, further emphasizing the limitations associated with accessing and utilizing real-world data. Even for the studies that utilized real data, important limitations and challenges were identified, underscoring the complexities and constraints involved in collecting and utilizing such data.

Despite these challenges and the limitations in terms of scope and scale, extensive efforts were made to obtain and collect real data, in order to compare and contrast real data with synthetic data, with the aim of deriving comprehensive and insightful conclusions. This involved thorough online research in identifying available open or shared datasets, as well as direct communication with the authors of relevant papers, requesting access to the datasets used in their studies. However, unfortunately, responses to these inquiries remained unanswered, making it impossible to incorporate actual operational data from wind farms into this study.

4.2.2 Techniques to Generate Synthetic Datasets

Following the wake model classification presented in the introductory chapter (see Section 1.2), we can identify two primary categories of models that are currently utilized for estimating the wake effect and that could be used to generate synthetic datasets: analytical and computational models.

Without reiterating their characteristics and differences, it becomes readily apparent that focusing on the first category in the context of surrogate models lacks rationality. Their operational efficiency fails to rationalize the investment in surrogate model construction, given that their principal objective is the provision of a computationally expedient estimation for intricate models or systems. The trade-off between accuracy loss and speed gain would not make the resulting surrogates more interesting than the alternatives currently available.

On the other hand, CFD simulations are often used by the literature as a data generation method [14, 53, 69, 75, 79, 80] and there are some software implementing very complex mathematical models. Most of the time, these software are closed source and they are also very expensive since they mainly target big energy companies. However, it is also possible to find some open-source alternatives: OpenFOAM [49] and OpenFAST [47] are two examples.

Nevertheless, these software applications are typically unsuitable for use on personal machines due to their demanding computational requirements. Even with access to large clusters or similar infrastructure, running a large number of simulations can be challenging due to the significant computational costs involved. For example, in [79], each simulation took 46 hours to complete on a high-performance computing (HPC) cluster with 256 CPUs. Even with simpler simulations, the time required for each simulation can still be

significant, typically in the order of hours as seen in [75] (8 hours per simulation).

Continuing with the exploration of data generation methods, it is important to note that while the aforementioned challenges and limitations exist, there have been proposed simplified solutions specifically tailored to the Ainslie model. These simplified solutions address the computational cost concerns associated with other complex computational models, making them more amenable for use on private machines. The decision to focus on the Ainslie model as the primary data generation method arises from its ability to strike a balance between the simplicity of engineering models and the computational complexity of other models. This choice offers a favourable compromise, enabling a larger number of simulations to be executed (as will be detailed in the next section), thus facilitating the training of complex models and the thorough testing of surrogates.

Moreover, the selection of the Ainslie model as the primary data generation method is driven by its inherent significance and widespread utilization in the industrial field. In practical applications, the Ainslie model has emerged as a preferred choice due to its ability to provide more accurate estimations compared to simplistic engineering models by offering a computationally efficient alternative to expensive and time-consuming CFD simulations, which are often impractical even for industrial contexts. This is particularly relevant for tasks such as power production estimation and wind farm layout optimization. By leveraging the strengths of the Ainslie model, we can delve into a comprehensive exploration of surrogate models while ensuring a realistic and practical approach. In the next subsection, further details regarding the process of data generation through the utilization of the Ainslie model will be provided. This will contribute to a comprehensive understanding of the specific techniques and methodologies employed in harnessing the capabilities of the Ainslie model for generating the dataset used in this study.

However, the choice of the Ainslie model as the primary data generation method comes with some drawbacks that need to be considered when studying the learning processes and the results of the implemented models. Firstly, the Ainslie model provides simpler wake fields, which may lead to less precise estimations compared to more complex and detailed computational models. This simplicity is achieved by smoothing out the complex effects and capturing smoother behaviours, potentially oversimplifying the actual wake phenomena. Secondly, the use of the thrust coefficient (C_T) to calculate the rotor-averaged actions on the inflow restricts the Ainslie model from capturing the specific effects induced by the complex geometry of turbine blades on the wakes (refer to Section 2.1.2.2). Lastly, the input space for the Ainslie model is relatively small, as only three parameters (as explained in Section 2.1.4) are considered in the predictions. This limited input space may restrict the model's ability to capture the full complexity of wake behaviour. It is important to note that the last two limitations are not unique to the Ainslie model but are shared by other CFD-based approaches as well, as mentioned in 3.1.1.

Despite these limitations, the Ainslie model offers a practical and computationally efficient alternative to more expensive and time-consuming CFD simulations, making it a suitable choice for generating synthetic datasets for this study.

4.2.3 Dataset Generation

In this subsection, we will provide an in-depth exploration of the data generation process using the Ainslie model, offering a comprehensive understanding of the specific techniques and parameters employed to harness its capabilities for generating the dataset used in this study.

After conducting extensive research on the available implementations of the Ainslie model, it became apparent that the **PyWake** software package [51] was the most suitable choice. The decision was almost a necessity, as there were no other freely accessible options available that offered the Ainslie model's implementation. While the Ainslie implementation is not yet merged into the PyWake repository at the time of writing this thesis, the authors and maintainers have confirmed its viability for our study, as the principles and insights from the Ainslie model remain consistent. The implemented solution follows Anderson's approach [7], elaborated upon in Section 2.1.4. Moreover, PyWake's versatility and power in executing automated computations for wind simulations, along with its inclusion of the Anderson solution of the Ainslie model, made it the ideal candidate for fulfilling the requirements of this project.

By leveraging the PyWake software package, we can effectively employ the Ainslie model to simulate wake fields and generate the necessary dataset for our study. The utilization of this established and validated implementation ensures the reliability and accuracy of the generated data.

4.2.3.1 Simulation Settings

As discussed in Section 2.1.4, the input parameters of the surrogate models consist of the free-stream wind speed (from now on referred to as ws), ambient turbulence intensity (referred to as TI), and the thrust coefficient C_T. These parameters will be examined in Section 4.2.3.2. Additionally, in the *univariate setting*, the coordinates x and y (or more precisely, the normalized values x/D and y/D) are included in the input space (refer to Sections 4.2.3.3 and 4.3.2).

This approach allows for an investigation focused on understanding the behaviour of the wake in relation to the wind conditions, while other factors such as turbine type may also influence wake characteristics. For the purposes of this project, modelling different types of turbines will primarily involve varying the thrust coefficient, as the Ainslie model does not incorporate parameters such as hub height, turbine diameter, or geometry.

Moreover, for the purposes of this project, the yaw angle of the wind turbine, representing the angle between the rotor axis and the direction of the incoming wind, will be assumed to be aligned with the incoming wind direction. This simplifying assumption allows for a focused analysis and examination of the wake behaviour in relation to wind speed, enabling an investigation of the impact of wind conditions on wake characteristics.

As a convention, the direction of the flow was set to be 270° (from West to East), and the wind turbine was positioned in the geometrical centre of the wind farm as the grid origin (at the coordinates $[0, 0]$). However, since the simulations are run for a single wind turbine, these choices uniquely impact the grid spacing (see 4.2.3.3) and future visualizations, and do not affect the analysis of wake characteristics.

4.2.3.2 Inflow Parameters

The ranges and the step values for the inflow parameters are summarized in Table 4.1. The range of the free-stream wind speed ws is determined based on the average cut-in and cut-out speeds (as mentioned in Section 2.1.2.1). Selecting a range from 4 m/s to 25 m/s ensures that only wind speeds within the operational range of the wind turbines, capable of generating energy, are considered.

The turbulence intensity TI is expressed as a percentage and represents the fluctuation or variability of wind speed. Typically the turbulence intensity ranges between 5% and 20%, and rarely exceeds 50%. However, for the purpose of gaining additional insights (e.g. about the extrapolation ability) a wider range of turbulence intensities from 0% to 100% was used, with a step size of 0.01 (1%).

The dimensionless thrust coefficient (C_T) starts from 0.1 as lower values did not yield interesting results. The upper limit is set to 0.96, adhering to the limit defined by Moriarty and Hansen [44]. Also for this parameter, the step size is 0.01.

To summarize, we have the following lists of input parameters:

$$\begin{aligned} \mathcal{W} &= \langle 4, 5, \ldots, 24, 25 \rangle, \quad |\mathcal{W}| = 16, \\ \mathcal{T} &= \langle 0, 0.01, \ldots, 0.99, 1 \rangle, \quad |\mathcal{T}| = 100, \\ \mathcal{C} &= \langle 0.1, 0.11, \ldots, 0.95, 0.96 \rangle, \quad |\mathcal{C}| = 86. \end{aligned}$$

4.2.3.3 Grid Spacing

Another crucial aspect in conducting a simulation is the determination of the **grid spacing**, which refers to the *discretization* of the downstream space, specifically the region beyond the wind turbine. The grid spacing plays a vital role in capturing the spatial resolution of the simulation and delimitating the relevant part of the wakefield.

Following the conventions to ensure consistency and comparability across different wind turbine setups, both distances in the simulation are normalized with respect to the turbine diameter D, and therefore naturally give rise to the dimensionless variables x/D and y/D.

This normalization approach allows for a standardized representation of distances and facilitates the generalization of the wake characteristics regardless of the specific turbine dimensions. By normalizing the coordinates, the focus can be placed on the relative positioning and behaviour of the wake, independent of the actual physical size of the wind turbine.

The inner grid cell is sized to $D/8$, while the extension of the inner grid was defined as follows:

- for the axial dimension x, the range goes from $2D$ to $30D$, from West to East;
- for the radial dimension y, the range goes from $-2D$ to $+2D$, from South to North.

A summarized overview is provided in Table 4.1.

The grid resolution has been carefully selected to strike a balance between resolution and computational cost. A finer grid allows for higher resolution and precision in the simulation results but comes at the expense of increased computational time and resource requirements.

Parameter	Range	Step	Description
ws (m/s)	[4, 25]	1	Wind speed at the turbine location
TI	[0, 1]	0.01	Turbulence intensity in the incoming wind (percentage)
C_T	[0.1, 0.96]	0.01	Thrust coefficient of the wind turbine
x/D	[2, 30]	1/8	Axial distance from the turbine in terms of rotor diameters
y/D	[-2, +2]	1/8	Radial distance from the turbine in terms of rotor diameters

Table 4.1: Overview of the inflow parameters and wake field coordinates for the data generation. All the variables are dimensionless, except for ws.

In order to explore the impact of grid spacing on the simulation, several values have been tested. The subsequent spacings examined were $D/2$, $D/4$, $D/8$, and $D/16$. It was determined that the first two values, $D/2$ and $D/4$, were too coarse for the specific objectives of this project. This is especially significant considering that the turbines being studied typically have diameters ranging from approximately 100 to 200 meters, with a minimum diameter of 50 meters. [41]. On the other hand, the $D/16$ spacing resulted in computationally expensive simulations in terms of both generation time and storage requirements.

After careful consideration, a grid spacing of $D/8$ was deemed the most suitable choice for this project. This spacing strikes a reasonable balance between resolution and computational efficiency, providing a satisfactory level of detail in the wake simulation without incurring excessive computational costs. Further experiments will explore coarser grids for training to study the interpolation capabilities.

The parameters controlling the extension of the inner grid were determined according to two primary considerations. Firstly, it was essential to exclude the near-wake region, as outlined in 2.1.4. The Ainslie model is not applicable for values of $x < 2D$ since it does not account for pressure gradients, which significantly influence the flow near the rotor. Therefore, the grid extension was designed to start beyond this critical region.

The second factor considered was to limit the grid extension to the region where the wake still has a noticeable impact on the wind deficit. To achieve this, the wind deficit (as defined in Equation 4.1) was analyzed in various simulations. A threshold of 1/100 was employed to identify the cells that are considered interesting and still influenced by the wake among the different simulations. The grid extension was then determined based on this criterion, ensuring that it encompasses the relevant part of the wake field. By combining these considerations, the chosen parameters for the extension of the inner grid strike a balance between excluding the near-wake region where the Ainslie model is invalid and capturing the significant portion of the wake that affects the wind deficit, such that the influence of wake is negligible outside the defined domain.

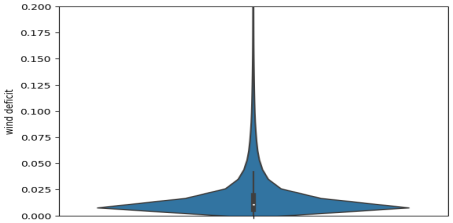
From this point forward, to avoid excessive notation, the coordinates x/D and y/D will be simply referred to as x and y, respectively, assuming that these coordinates are normalized by the diameter D. The resulting statistics showed that on average the maximum value of x is 30,989. For y since the wind direction is stably behind the turbine, the range is definitely smaller, and among all the simulations the wind deficit never went below the

threshold in the interval $[-2, 2]$. Therefore, we define the following lists of coordinates:

$$\mathcal{M} = \langle 2, 2.125, \ldots, 29.875, 30 \rangle, \quad |\mathcal{M}| = 224,$$
$$\mathcal{N} = \langle -2, -1.75 \ldots, 1.875, 2 \rangle, \quad |\mathcal{N}| = 32.$$

Metric	Min	Max	Mean	Median
wd	0	0.9	0.0195	0.0105

(a) Summary table



(b) Violin plot

Figure 4.2: Overview of the output variable wd. On the left (a), a summary table presents key statistics for the output variable, including the minimum, maximum, mean, and median values. On the right (b), a violin plot displays the distribution of wind deficit values, showing the range and density of data points. The visualization is vertically truncated at 0.2 to enhance visibility of the distribution, excluding the less informative region.

4.2.3.4 Output and Final Dataset

We define the set of simulations as the Cartesian product of the inflow parameter lists: $\mathcal{S} = \mathcal{W} \times \mathcal{T} \times \mathcal{C}$, with $|\mathcal{S}| = |\mathcal{W}| \times |\mathcal{T}| \times |\mathcal{C}| = 16 \times 100 \times 86 = 137,600$ simulations.

For each simulation $i \in \mathcal{S}$, the wake field output is generated by considering a specific combination of $ws^{(i)}$, $TI^{(i)}$, and ${C_T}^{(i)}$ values and represents the effective wind speed $\overline{ws}$ in the space downstream of the turbine. The wake field is discretized into a two-dimensional matrix $\mathbf{E}^{(i)}$ as explained in 4.2.3.3. Each cell in the resulting matrix, denoted as $\mathbf{E}^{(i)}_{j,k}$, corresponds to the effective wind speed $\overline{ws}^{(i)}_{j,k}$ at a specific location (j, k) within the i-th wake field. The coordinates $j \in \mathcal{M}$ and $k \in \mathcal{N}$ refer to the x and y directions, respectively. Therefore, the size of every $\mathbf{E}^{(i)}$ is determined as $|\mathbf{E}^{(i)}| = |\mathcal{M}| \times |\mathcal{N}| = 224 \times 32 = 7,168$.

To ensure comparability and independence from the wind speed ws, we introduce the concept of the **wind deficit** wd for the output variable. The dimensionless wind deficit $wd^{(i)}_{j,k}$ is defined as follows:

$$wd^{(i)}_{j,k} = 1 - \frac{\overline{ws}^{(i)}_{j,k}}{ws^{(i)}} \tag{4.1}$$

The wind deficit quantifies the reduction in wind speed caused by the wake at each specific (j, k) location, relative to the free-stream wind speed $ws^{(i)}$ in the i-th simulation. By using the wind deficit as the output variable, we ensure that the predictions are normalized and not directly influenced by the varying wind speeds across different simulations. This normalization is essential for meaningful and comparable wake predictions across different wind conditions. This variable ranges from 0 to 1, but it is significantly skewed towards 0, as shown by Figure 4.2. The table provides an overview of the minimum, maximum, mean, and median values of the wind deficit obtained from our experiments, while the violin plot shows the density of the data points in the lower parts.

Similarly, we define a matrix $\mathbf{D}^{(i)}$ as the final output to predict in a simulation i, where $\mathbf{D}^{(i)}_{j,k} = wd^{(i)}_{j,k}$ for each cell (j,k). Figure 4.3 shows a formal matrix and a random visual representation. The blue color indicates the wind deficit value's intensity within the corresponding cell. This approach ensures consistent comparison and analysis of wake characteristics across different wind conditions. Therefore, the final output to be predicted consists of 7,168 regression values for each simulation i. Other works based on 2D fields typically involve around 1,000 cells per simulation [53, 79], whereas only [75] employs a larger number, approximately 9,100 cells per simulations. In the next section, two possible problem settings will be proposed to predict these values.

In conclusion, the final dataset $\mathbb{A}$ generated through the Ainslie model can be defined as follows:

$$\mathbb{A} = \left\{ \left(ws^{(i)}, TI^{(i)}, {C_T}^{(i)}, j, k, \mathbf{D}^{(i)}_{j,k} \right) \middle| \; i \in \mathcal{S}; j \in \mathcal{M}; k \in \mathcal{N} \right\}.$$

Therefore, it consists of $|\mathbb{A}| = |\mathcal{S}| \times |\mathcal{M}| \times |\mathcal{N}| = 137,600 \times 224 \times 32 = 986,316,800$ instances. However, for the experiments detailed in Chapter 5, only a part of this data will be used, as explained in Sections 4.3.3 and 5.1.1. To our knowledge, there is no existing approach that encompasses such an extensive number of instances.

$$\mathbf{D}^{(i)} = \begin{bmatrix} wd^{(i)}_{1,1} & wd^{(i)}_{1,2} & \dots & wd^{(i)}_{1,|\mathcal{M}|} \\ \vdots & \vdots & \ddots & \vdots \\ wd^{(i)}_{|\mathcal{N}|,1} & wd^{(i)}_{|\mathcal{N}|,2} & \dots & wd^{(i)}_{|\mathcal{N}|,|\mathcal{M}|} \end{bmatrix} \rightarrow$$

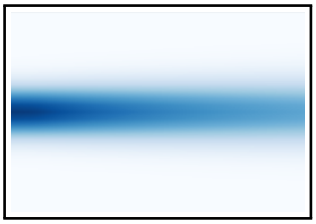


Figure 4.3: Formal matrix representation (on the left) and visual representation (on the right) of the wake field resulting from a generic simulation i.

4.3 Experimental Setup

In this section, we outline the experimental setup, which serves as a crucial foundation for the methodology and experiments presented in the subsequent chapter. Understanding the characteristics of the dataset, the distinct approaches for predicting wind turbine wake effects, and the reasons associated with data splitting and reduction techniques is essential for comprehending the rationale behind our research and the decisions made during model development and evaluation of the next chapter. The insights gained from this analysis will not only provide context for the methodologies employed but also shed light on the experiments and comparisons to be discussed in the following chapters.

The complete code for all aspects of this research, including data generation, data splitting strategies, model development, experimentation, and plotting, can be found in the dedicated repository [1].

[1] https://github.com/NiccoloMorabito/Ainslie-surrogate

4.3.1 Experimental Hardware

The experiments were conducted on a Macbook Air with an Apple M1 chip, featuring 8 GB of RAM and a unified memory design. The Apple M1 chip is an advanced ARM-based system-on-a-chip (SoC) designed by Apple, which incorporates an eight-core CPU, an integrated GPU, and other components, providing a high-performance computing environment on a single chip.

Using the same hardware platform for all experiments ensured consistency and reproducibility, allowing for fair and meaningful comparisons between different models and methodologies. The combination of efficient hardware and unified memory design facilitated smooth and efficient execution of the experiments, enabling us to explore various models and techniques while maintaining computational performance and reliability.

4.3.2 Univariate and Multivariate

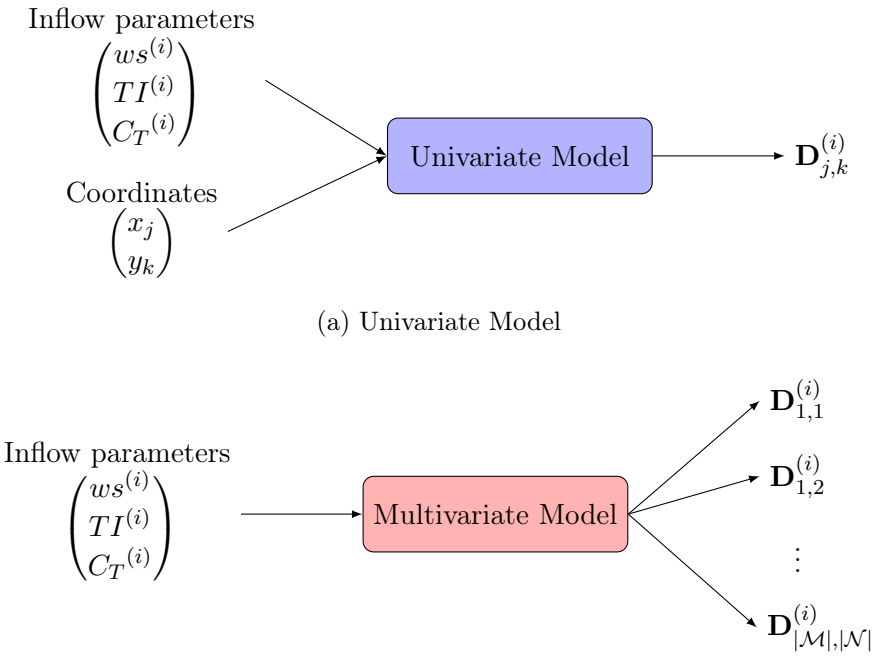
Considering the generated dataset $\mathbb{A}$ described in the previous section, different approaches can be employed to predict the wind deficit at a specific location (j, k) for each simulation i. These approaches significantly impact both the input and output space, thereby influencing the architecture of the corresponding surrogate models. In this context, we consider the following two distinct settings:

- **Univariate (Coordinate-based) Approach**: In this approach, the surrogate model predicts one wind deficit value (i.e. one cell $\mathbf{D}^{(i)}_{j,k}$) at a time. The input variables for this approach include also the coordinates x and y.

- **Multivariate Approach**: Conversely, the multivariate approach involves predicting the entire wind deficit matrix $\mathbf{D}^{(i)}$ given a combination of input variables $ws^{(i)}$, $TI^{(i)}$, and ${C_T}^{(i)}$. In this approach, the coordinates x and y are not considered as part of the input space. The model's objective is to provide predictions for all cells in the wind deficit matrix simultaneously, resulting in a multivariate output. Each prediction consists of $|\mathbf{D}^{(i)}| = 7,196$ values, capturing the wind deficit distribution across the entire domain.

It is important to note that the concepts of input and output vectors change according to the setting. Considering an instance $t = (\mathbf{u}_t, \mathbf{v}_t)$, in the univariate setting it represents the value of a single cell in the wake field. $\mathbf{u}_t = \left(ws^{(i)}, TI^{(i)}, {C_T}^{(i)}, j, k\right) \in \mathbb{R}^5$ is the input vector, containing the wind speed, turbulence intensity, thrust coefficient, and the x and y coordinates; and $\mathbf{v}_t = \left(\mathbf{D}^{(i)}_{j,k}\right) \in \mathbb{R}^1$ is the output vector, representing the wind deficit value at cell (j, k) in the wake field.

On the other hand, for the multivariate case, $\mathbf{u}_t = \left(ws^{(i)}, TI^{(i)}, {C_T}^{(i)}\right) \in \mathbb{R}^3$ is the input vector, containing the wind speed, turbulence intensity, and thrust coefficient, and $\mathbf{v}_t = \left(\mathbf{D}^{(i)}\right) \in \mathbb{R}^{|\mathcal{M}|\times|\mathcal{N}|}$ is the output vector, representing the entire wind deficit matrix in the wake field. Refer to Figure 4.4 for a visual representation that further illustrates these concepts.



(a) Univariate Model

(b) Multivariate Model

Figure 4.4: Comparison of the two settings for a model prediction.

4.3.3 Wind Speed Input Variable

The dataset $\mathbb{A}$, as discussed in Section 4.2.3.4, consists of nearly one billion instances, which makes it impractical to handle the entire dataset on a local machine, even after splitting it into training, validation, and test sets. To address this challenge, a standard reduction factor of 4 has been applied to the other input variables (TI and C_T) when wind speed ws is used as an input. This reduction helps reduce the data volume while preserving the integrity of the analysis. Specifically, the step size for TI and C_T has been effectively increased from 0.01 to 0.04.

However, most of the experiments have been conducted without considering wind speed as an input feature. Instead, a default value of wind speed ($ws^{(def)} = 12$) has been chosen, and the corresponding subset of the dataset, denoted as $\mathbb{A}'$, includes all the TI-C_T combinations for all coordinates x and y. The reasons for this approach will be further explained in 5.1.1.

$$\mathbb{A}' = \left\{ \left(ws^{(def)}, TI^{(i)}, C_T{}^{(i)}, j, k, \mathbf{D}^{(i)}_{j,k} \right) \middle| \; i \in \mathcal{S}'; j \in \mathcal{M}; k \in \mathcal{N} \right\}.$$

where $\mathcal{S}'$ is the number of simulations involving only TI and C_T, i.e. $\mathcal{S}' = \mathcal{T} \times \mathcal{C}$. The total number of instances for the experiments which do not consider ws as input variable is then: $|\mathbb{A}'| = |\mathcal{S}'| \times |\mathcal{M}| \times |\mathcal{N}| = 61,644,800$.

These numbers are still considerably large, especially when compared with those of related works (refer to 3.1). This abundance of data eliminates any concern about a lack of data for this project. In fact, in the following sections and chapters, it will be shown how reduction techniques were necessary to avoid making this analysis trivial. However, it is

important to note that this abundance of data had to be balanced with its complexity and reliability, leading to a trade-off that will be carefully considered in the further analyses.

4.3.4 Dataset Splitting Strategies

Different strategies have been tested to split the complete dataset $\mathbb{A}$ into training, validation, and test subsets, in order to carry out more experiments on the available data. The key consideration in most of the experiments was to ensure that the simulations remained intact and unsplittable.

This means that each simulation was assigned entirely to one subset, regardless of whether the model was univariate or multivariate. By adopting this approach, every model had access to all coordinates' values within a given simulation, ensuring the comparability of results between the univariate and multivariate models.

In addition to this approach, to broaden the scope of the analyses, additional experiments were conducted by training the models with partial wake fields. This involved reducing the number of coordinates included in the training set. This variation will be referred to as "Interpolation on coordinates", and it refers to a *regularly-spaced grid*, similar to what has been explained in 3.2 for [66]. The rationale behind this approach will be explained in detail later.

Below is a recap of the different data reduction strategies that have been applied to create subsets for training, validation, and testing purposes. Only Strategy 4.1 has been applied to the complete dataset $\mathbb{A}$, while the others refer to the subset $\mathbb{A}'$. However, it is important to note that even the subset $\mathbb{A}'$ contains a substantial number of instances. Therefore, even small percentages correspond to a significant amount of data, exceeding of many times the number of simulations used in many similar works. Additionally, considering the large size of the wake field, the number of instances grows exponentially. A large training set affects both computational complexity and training time, and it may also increase the likelihood of overfitting. Consequently, conventional splitting percentages cannot be strictly followed in this case.

Strategy 4.1 (Randomly sampling the ws values) This technique divides the dataset $\mathbb{A}$ based on wind speed values. For each wind speed, all the TI-C_T combinations are assigned to that specific value. A variable percentage of the wind speed categories are randomly selected for the training set, while the remaining is equally split among validation and testing sets. However, due to the reasons elaborated upon in Section 5.1.1, this technique will not be further discussed in this report.

Strategy 4.2 (Randomly sampling the TI-C_T combinations) In this technique, a variable percentage of the combinations of TI and C_T is randomly sampled for the training set. The remaining combinations are equally and randomly divided between the validation and testing sets.

Strategy 4.3 (Uniformly sampling the TI-C_T combinations) Here, reduction factors (r_{TI} and r_{C_T}) are chosen for TI and C_T respectively. The training set is created by selecting specific combinations of TI and C_T values based on these reduction factors. For example, the chosen values for TI are at indices 1, $1 \times r_{TI}$, $2 \times r_{TI}$, and so on. All other combinations are equally and randomly split between the validation and testing sets. In

order to get the percentage of instances that end up in the training set, it is possible to use the following formula:

$$\text{Train Percentage}(4.3) = \left\lceil \frac{|\mathcal{T}|}{r_{TI}} \right\rceil \times \left\lceil \frac{|\mathcal{C}|}{r_{C_T}} \right\rceil \times \frac{100}{|\mathcal{T}| \times |\mathcal{C}|} \tag{4.2}$$

this number corresponds also to the percentage of simulations that are considered in this case, as each wake field is kept entirely in one of the subsets.

Strategy 4.4 (Regularly-spaced wake field) This technique employs a similar approach of 4.3, but in addition to r_{TI} and r_{C_T} values, reduction factors r_x and r_y are also chosen for the coordinates. The training set is constructed by selecting coordinates at regular intervals based on these reduction factors. For example, if r_x and r_y are both 2, every second coordinate in both the horizontal (x) and vertical (y) dimensions is included in the training set. The remaining coordinates are distributed randomly between the validation and testing sets. This technique allows for the investigation of the model's performance when trained on a sparser set of coordinate values. It is important to notice that this is equivalent to reducing the grid spacing defined in Section 4.2.3.3 for the training set. Similarly to 4.3, the percentage of the instances used for training can be computed as:

$$\text{Train Percentage}(4.4) = \left\lceil \frac{|\mathcal{T}|}{r_{TI}} \right\rceil \times \left\lceil \frac{|\mathcal{C}|}{r_{C_T}} \right\rceil \times \left\lceil \frac{|\mathcal{M}|}{r_x} \right\rceil \times \left\lceil \frac{|\mathcal{N}|}{r_y} \right\rceil \times \frac{100}{|\mathcal{T}| \times |\mathcal{C}| \times |\mathcal{M}| \times |\mathcal{N}|}. \tag{4.3}$$

Strategy 4.5 (Reducing the ranges of TI and C_T) In this technique, specific intervals $I_{TI} = [a_{TI}, b_{TI}]$ and $I_{C_T} = [a_{C_T}, b_{C_T}]$ are chosen for TI and C_T respectively to define the training set. Both a_v and b_v should be in the range $[\min_v, \max_v]$ of the corresponding variable v, with $a < b$. Only the simulations that involve combinations of TI and C_T within these chosen ranges are used for training. The remaining simulations are equally and randomly distributed between the validation and testing sets. In this case, the percentage of the instances used for training can be computed as:

$$\text{Train Percentage}(4.5) = \frac{b_{TI} - a_{TI}}{\max_{TI} - \min_{TI}} \times \frac{b_{C_T} - a_{C_T}}{\max_{C_T} - \min_{C_T}} \times 100 \tag{4.4}$$

where $\min_{TI}$=0, $\max_{TI} = 1$, $\min_{C_T} = 0.1$, $\max_{C_T} = 0.96$, as shown in Table 4.1.

From this, it becomes clear that the aim of these techniques goes beyond splitting the dataset into disjoint sets. They are applied to reduce the amount of available training data in different ways, allowing us to explore how the models perform in both interpolation and extrapolation settings.

4.3.5 Interpolation and Extrapolation

The challenges and computational limitations associated with large-scale computational fluid dynamics (CFD) simulations emphasize the importance of **data reduction**. Given the significant disparity between the number of simulations possible through CFD and the volume of data generated in this project, the exploration of diverse data reduction techniques becomes indispensable. Additionally, the value of such exploration is echoed by

other related works. For instance, also [75] delved into similar investigations, underlining the importance of exploring both interpolation and extrapolation capabilities. Thus, exploring various data reduction techniques becomes crucial in facilitating thorough analysis and paving the way for further investigations.

Interpolation techniques focus on estimating values within the range of the known data, while **extrapolation** techniques aim to predict values outside the range of the known data. Both settings provide valuable insights into the generalization capability of the models and contribute to the foundation for future work.

In the context of wind turbine wake surrogate models, interpolation plays a significant role as it reflects the generation of data using a grid-like structure. In research and industrial settings, data is often generated by systematically varying the input variables with a specific step size to cover a wide range of cases. Consequently, uniformly sampling the TI-C_T combinations (Strategy 4.3) aligns with this expected approach for generating data and allows for exploring the performance of the models within this framework. It is worth noting that, initially, some experiments were conducted with random sampling (Strategy 4.2) and yielded comparable results to uniform sampling. However, to ensure consistency and adhere to the expected data generation approach, uniform sampling has been extensively explored in this study, even when considering the incorporation of coordinates for interpolation (Strategy 4.4). This deliberate choice ensures that our findings and evaluations remain reliable and relevant, consistent with the typical data generation practices employed in this field.

On the other hand, the rationale for extrapolation in wind turbine wake modelling differs from that of interpolation, since we expect the generation of the data to be carried out in the widest way possible (i.e. to cover the entire range of possible values for each inflow parameter). Nevertheless, the extrapolation capacity of these models provides valuable insights for better understanding the model's behaviour and capacity and exploring scenarios beyond the known data range. Furthermore, it is important to note that extrapolation can also be relevant when real-world data is used. In such cases, the available data may only cover a certain range of input variables, representing the conditions observed during the data collection period. Understanding the behaviour and limitations of the models in extrapolation scenarios is crucial for robust and reliable predictions, even with limited real-world research. These are the reasons to leverage Strategy 4.5.

Finally, Strategy 4.4, even if it still explores the interpolation capability, serves a slightly different purpose compared to the previous approaches. Firstly, it allows for the investigation of the model's behaviour when trained with a coarser grid. By reducing the density of coordinate values in the training set, the model is exposed to a sparser representation of the wake field. This analysis provides insights into how the model performs when trained on a less refined grid, potentially leading to computational efficiency gains. Secondly, also this technique can be valuable in reducing the computational cost of simulations as it allows for a reduction in the required output size and resolution. Lastly, studying the behaviour of models trained on a coarser grid provides insights into their performance under more realistic conditions because, in practice, it is often not feasible to obtain a complete and finely resolved representation of the wake field.

In the next chapter, we will revisit these concepts and further analyze the performance of the models in different data reduction scenarios, providing additional insights.

4.3.6 Scaling

Scaling is a crucial preprocessing step in the context of this wind turbine wake modelling problem, where various machine learning models, including neural networks, are employed.

There are several reasons for scaling the data. Firstly, scaling helps prevent features with a larger magnitude from dominating the learning process. Many machine learning algorithms, particularly those relying on distance metrics or gradient-based optimization, may perform poorly if features have different scales. Scaling ensures that all features contribute equally to the learning process, leading to better convergence and performance.

Secondly, neural networks, as well as other machine learning models, are sensitive to the scale of input data. Larger input values can result in exploding gradients, making it challenging for the models to converge effectively.

Moreover, scaling can lead to faster training times and better generalization performance. Scaled data often improves the model's ability to generalize well to unseen data.

For all the experiments carried out in this project, all the input variables were scaled before feeding them into the models. Scaling involves transforming the data to a standardized range with a mean of 0 and a standard deviation of 1. In this case, it is particularly simple and convenient since the ranges of each variable are known in advance (refer to Table 4.1). Therefore, it is straightforward to apply a *min-max scaler* for each input variable v:

$$\overline{v} = \left(\frac{v - \min_v}{\max_v - \min_v} \right) \times (\overline{\max} - \overline{\min}) + \overline{\min} \tag{4.5}$$

where $\overline{v}$ is the scaled value, $\min_v$ and $\max_v$ are the minimum and maximum values of variable v, respectively, and $\overline{\min}$ and $\overline{\max}$ are the desired minimum and maximum values of the scaled range ($\overline{\min} = 0$ and $\overline{\max} = 1$). This process ensures that all input variables are within the desired standardized range, facilitating the training and performance of the machine learning models.

4.4 Modelling

To develop a data-driven surrogate model, a range of machine learning models were explored and assessed. Among them, the following models were identified as particularly relevant within the scope of this study:

- Univariate Regression Decision Tree (RDT);
- Univariate Multilayer Perceptron (MLP);
- Multivariate MLP;
- Univariate Neural Radiance Field (NeRF).

The decision tree was chosen to assess the performance of a straightforward approach in this domain, while the MLP models were selected for their capability to capture complex patterns and handle large datasets effectively. The motivations behind the choice of the NeRF model will be further elaborated in Section 5.1.2.

Additionally, a variety of other models have been tested and their analysis is provided in A, where a comprehensive examination of their performance and suitability in the wind turbine wake modelling context is presented.

4.4.1 Architecture and Hyperparameters

In this section, we present an overview of the architectures and hyperparameters used for the most relevant machine learning models. Each model brought its unique characteristics and capabilities to the table, paving the way for diverse experimentation with hyperparameter settings.

Several experiments were meticulously carried out to maximize the performance of each model, testing different combinations of hyperparameters to identify the most optimal configurations. Notably, the hyperparameters and architectures presented in the following paragraphs are those that emerged as the most promising and effective in our experimentation process.

However, it is important to note that the comprehensive analysis of hyperparameter tuning will not be presented in this report as the most compelling and enlightening insights that emerged from this research were derived from other types of experiments, which we will delve into in later sections.

Regression Tree For the Decision Tree, the default parameters provided satisfactory results for our task. Namely, the mean squared error served as the criterion (4.8) for splitting. Moreover, no maximum depth constraint was imposed on the Decision Tree, allowing it to grow freely. This decision was informed by comprehensive testing of various maximum depth values, the results of which can be found in Appendix B.

Now, let us explore the architectures and hyperparameters of more intricate models, such as the Multi-Layer Perceptron (MLP) and the Neural Radiance Field (NeRF), which exhibit greater complexity and potential for capturing intricate patterns in our wind turbine wake modelling problem. It is worth mentioning that, for the univariate approach, the batch size is set as a multiple of the number of cells $|\mathcal{M}| \times |\mathcal{N}| = 7,168$ in a wake field to ensure that each simulation is not split across multiple batches. This approach guarantees that the entire wake field data for a particular simulation is processed together within a single batch, maintaining the integrity of the simulation data during training, in a similar way as done for the multivariate approach (where an instance contains the whole wake field).

Univariate MLP This model consists of two hidden layers with 50 and 250 units, respectively, activated by the Rectified Linear Unit (ReLU) function. The output layer is not activated. The batch size is set to $8 \times 7,168$ instances, and a learning rate of 0.01 is used.

Multivariate MLP The multivariate approach employs an MLP with two hidden layers containing 50 and 500 units, activated by ReLU. The batch size is set to 64 instances, and a learning rate of 0.01 is utilized.

NeRF The NeRF architecture is structured with a Fourier Layer with 256 nodes applied only on the coordinates and three hidden layers of 256 units, all activated by ReLU. This architecture mirrors that of [66], as empirical investigations indicated its superior performance across all conducted experiments. The batch size is set to $8 \times 7,168$ instances, and a

learning rate of 0.0001 is employed. The Fourier Layer applies the Gaussian RFF mapping γ from [66], as it was shown to perform better in all the experiments. Considering the two parts of the input vector $\mathbf{u} = \begin{bmatrix}\mathbf{u}_{[1:3]} & \mathbf{u}_{[4:5]}\end{bmatrix}$, where the latter contains the coordinates x and y and the former all the other input variables, the feature mapping γ is defined as follows:

$$\gamma(\mathbf{u}_{[4:5]}) = \begin{bmatrix}\cos(2\pi\mathbf{B}\mathbf{u}_{[4:5]}) & \sin(2\pi\mathbf{B}\mathbf{u}_{[4:5]})\end{bmatrix}^\intercal \tag{4.6}$$

where each entry in $\mathbf{B} \in \mathbb{R}^{m\times d}$ is sampled from $\mathcal{N}(\mu, \sigma^2)$ (with $\mu = 0$ and $\sigma = 1$). The elements of the matrix $\mathbf{B}$ are the Fourier basis frequencies used to approximate the kernel. d is the number of coordinates (2, in this case) and m is the mapping size, i.e. the number of Fourier features, which is set to 64. The resulting mapping is concatenated to the other features and forwarded through the other fully-connected layers f with weights θ to obtain the final prediction $\hat{\mathbf{v}}$:

$$\hat{\mathbf{v}} = f\left(\begin{bmatrix}\mathbf{u}_{[1:3]} & \gamma(\mathbf{u}_{[4:5]})\end{bmatrix}; \theta\right) \tag{4.7}$$

A representation of this process is shown in Figure 4.5.

In all neural network-based models, the output layer is not activated. Additionally, we use the Adam optimizer with the specified initial learning rate and no weight decay. The training process has been conducted for a maximum of 500 epochs (with an early stopping strategy), consistently selecting the model with the smallest validation loss as the final trained model.

Loss Function The focus of this project is on predicting the wind deficit within a wind farm accurately. Converting it to an optimization problem, we want to minimize the prediction error of the overall field, this is why it is important to define a measure for the performance. For the surrogate models based on neural networks but also as a criterion for the decision tree, the Mean Squared Error (MSE) loss function is utilised:

$$\min : \text{MSE}\left(\hat{\mathbf{D}}^{(i)}, \mathbf{D}^{(i)}\right) = \frac{1}{|\mathcal{M}| \times |\mathcal{N}|} \sum_{j\in\mathcal{M}} \sum_{k\in\mathcal{N}} \left(\hat{\mathbf{D}}^{(i)}_{j,k} - \mathbf{D}^{(i)}_{j,k}\right)^2 \tag{4.8}$$

where $\hat{\mathbf{D}}^{(i)}$ represents the predicted wind deficit field, and $\mathbf{D}^{(i)}$ represents the actual one.

The MSE loss function calculates the squared difference between each predicted value $\hat{\mathbf{D}}^{(i)}_{j,k}$ and its corresponding ground truth value $\mathbf{D}^{(i)}_{j,k}$ for all cell locations. These squared differences are then averaged over the total number of cells to compute the overall MSE loss, and this process is applied for each simulation i. Minimizing the MSE loss during the training process helps optimize the model's parameters and improve the accuracy of predictions for the entire matrix of wind deficit values.

4.4.2 Performance Metrics

To assess and compare the performance of the data-driven surrogate models developed in this study, in addition to the previously mentioned MSE, we employ other common quantitative measures: **Mean Absolute Error** (MAE) and **R-squared** (R^2) score.

The MAE calculates the average absolute difference between the predicted and actual values. It provides a measure of the overall accuracy of the models, regardless of the direction of the prediction errors.

$$\mathrm{MAE}\left(\hat{\mathbf{D}}^{(i)}, \mathbf{D}^{(i)}\right) = \frac{1}{|\mathcal{M}| \times |\mathcal{N}|} \sum_{j \in \mathcal{M}} \sum_{k \in \mathcal{N}} \left| \hat{\mathbf{D}}^{(i)}_{j,k} - \mathbf{D}^{(i)}_{j,k} \right| \tag{4.9}$$

where $\hat{\mathbf{D}}^{(i)}$ represents the predicted wind deficit field, and $\mathbf{D}^{(i)}$ represents the actual one.

The R^2 score, also known as the coefficient of determination, quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables. R^2 score ranges from 0 to 1, with 1 indicating that the model perfectly predicts the target values and 0 indicating that the model provides no predictive value.

$$\mathrm{R}^2\left(\hat{\mathbf{D}}^{(i)}, \mathbf{D}^{(i)}\right) = 1 - \frac{\sum_{j \in \mathcal{M}} \sum_{k \in \mathcal{N}} \left(\mathbf{D}^{(i)}_{j,k} - \hat{\mathbf{D}}^{(i)}_{j,k} \right)^2}{\sum_{j \in \mathcal{M}} \sum_{k \in \mathcal{N}} \left(\mathbf{D}^{(i)}_{j,k} - \bar{\mathbf{D}} \right)^2} \tag{4.10}$$

where $\bar{\mathbf{D}}$ is the average of the actual wind deficits.

By employing these evaluation metrics on both training and testing sets, we can gain insights into how well the data-driven surrogate models perform in predicting the wind turbine wake effects and compare their performance with each other. These metrics will guide the evaluation process and contribute to understanding the effectiveness and accuracy of the proposed approach in the specific context of the current study.

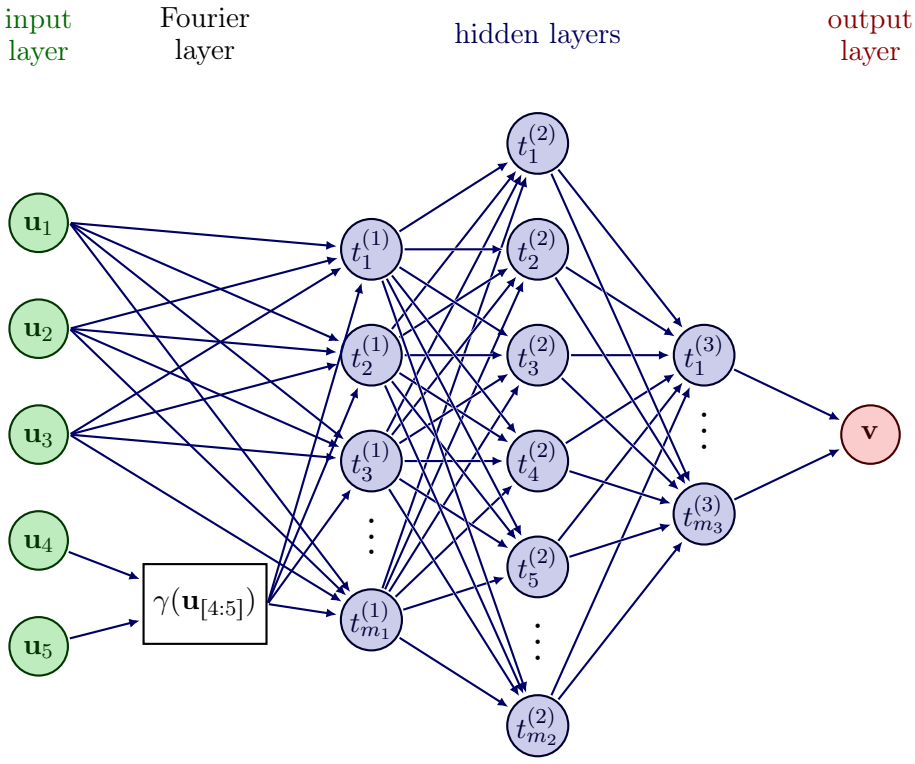


Figure 4.5: Neural Network Architecture with Fourier Layer: The diagram shows a standard neural network architecture, with an additional Fourier layer taking the coordinates x and y as input. The result, combined with the other input features, is passed through the fully-connected layers f to produce the final prediction.

It is important to emphasize that this comparison is solely within the context of the models developed in this study, and not with the related papers as explained in the next section.

4.4.3 Comparison with Related Work

In this section, we aim to highlight the key differences between the data-driven surrogate models developed in this study and the state-of-the-art works mentioned in Section 3.1, particularly from a qualitative perspective. A direct quantitative comparison becomes practically impossible due to the significant disparities observed among the datasets, methodologies, and evaluation metrics employed in different papers.

One of the main differences is the dataset's origin. To the best of our knowledge, there are no papers that use the Ainslie method to generate synthetic datasets in this domain. This implies significant differences in terms of quantity, quality and characteristics of the generated data.

Table 3.3 shows that the biggest number of simulations for generated datasets is 180, while the dataset used for this project, without considering wind speed, has a remarkable number $|\mathcal{S}'| = 8600$ of simulations. Moreover the resolution of the wake field is greater than most of the papers having a 2D field, contributing to a higher level of detail and accuracy in representing the wake effect.

Another pivotal divergence lies in the representation of data within the wake field. While the Ainslie implementation produces static 2D wake fields, certain other studies rely on dynamic simulations encompassing various time points to predict wake fields or 3D fields. However, even when considering only works using static 2D wake fields, comparisons would not be fair. This is not just a matter of data quantity; rather, it underscores the inherent differences in what the data generation methods are modelling. The underlying mathematical or physical principles behind these methods significantly differ, leading to contrasting outcomes.

These differences in the data and problem characteristics preclude also the application of some machine learning models in this work. For instance, the usage of LSTM models, as seen in [79], is not suitable for our specific task as there is no time dimension involved. However, other works have employed architectures similar to the ones used for this work. For example, tree-based models have been explored in [45] and [53], and Multi-Layer Perceptrons (MLPs) have been utilized in [75] and [69].

It is important to note that the NeRF structure achieving the best performance is the same as the one used in [66], even though the domains are entirely different. Surprisingly, this symmetric NeRF architecture consistently delivered superior performance in all conducted experiments. Larger or smaller asymmetric architectures exhibited significantly worse performance, including a much higher tendency to overfit (especially in the extrapolation setting). This finding underscores the critical role of the chosen NeRF structure and its impact on achieving successful results in our wind turbine wake modelling task.

Overall, the qualitative comparison provides valuable insights into the uniqueness of the approach used in this study, setting it apart from the existing literature in the field of data-driven surrogate models for wind turbine wake modelling. In the following chapter, these peculiarities will be tested at the effectiveness level.

Chapter 5

Experiment Analysis and Discussion

In this chapter, we present a comprehensive analysis and discussion of the experiments conducted to evaluate the performance and effectiveness of the data-driven surrogate models developed in this study. The primary objective of the experiments was to assess how well the most-promising surrogate models predict wind turbine wake effects and compare their performance with each other. To achieve this, a series of experiments with various model configurations and data splitting strategies were conducted to gain a thorough understanding of the models' capabilities and limitations.

Before discussing the detailed experimental results, we first address some preliminary discoveries in Section 5.1. This section includes insights on the relevance of input features in predicting the wake field. Additionally, intriguing artefacts of MLP-based models are explored. Understanding these aspects is essential to justify other parts of the experimental settings, such as excluding wind speed as an input variable and considering the use of NeRF.

The other two sections correspond to the two key perspectives that have been used to provide insights into the performance and the behaviour of different models under various settings and data reduction scenarios: **quantitative and qualitative performance**, respectively detailed in Sections 5.2 and 5.3. The former provides numerical measures of how well the models predict the wind turbine wake effect and the latter shows a visual evaluation of the shape and appearance of the wake fields predicted by the models. By addressing both the quantitative and qualitative aspects of the models' performance, our evaluation provides a comprehensive understanding of their capabilities and visualization effectiveness for wind turbine wake modelling.

5.1 Preliminary Findings and Exploratory Analysis

This section presents preliminary findings and a comprehensive data analysis to gain insights into the data-driven surrogate models used for wind turbine wake predictions. It explores the relevance of input features, particularly wind speed, in wake field predictions through correlation analysis and feature importance tests. Additionally, intriguing artefacts in MLP-based models are explored. Understanding these aspects is essential to justify other parts of the experimental settings, such as excluding wind speed as an input variable and considering the use of NeRF. These preliminary insights lay the groundwork for a comprehensive evaluation of the models' performance in the following sections, encompassing both qualitative and quantitative aspects of the experiments.

5.1.1 Assessing the Irrelevance of Wind Speed as Input Feature

In data analysis, understanding the relevance of input features is crucial for building accurate and efficient models. One essential input feature often considered in the context of wind turbine wake prediction is wind speed, which is used as a representative input for predicting the wake field. However, considering the Ainslie model and the experimental setup described before, it may not be the case for this project. Even if it is challenging to definitively prove that an input feature does not affect the data, we can provide evidence and perform analyses to suggest that the feature has little to no impact on the data or the model's performance.

To investigate the relevance of wind speed as an input feature in predicting the wake field, we analyzed the dataset $\mathbb{A}$ that has been generated, consisting of a large number of simulations, each representing a different combination of inflow parameters. In this analysis, we focused on the correlation matrix to identify the linear relationships between the input features and the wake field's wind deficit.

Correlation analysis helps identify the degree of linear relationship between variables, and in this case, between wind speed and the wind deficit. By calculating the correlation coefficients, we can assess the strength and direction of this relationship. A high correlation coefficient close to 1 would indicate a strong linear relationship, whereas a coefficient close to 0 would imply a weak or no linear relationship.

The correlation matrix in Figure 5.1 shows the correlation factor among the variables in $\mathbb{A}$. All the input variables are not related to each other, and the only coefficients different than 0 are between the input variables and the wind deficit ws. The matrix revealed that the correlation coefficient between wind speed ws and the wind deficit wd is zero, suggesting a lack of relationship between these two variables. This observation indicates that wind speed should not have an impact on the wake field's wind deficit in the context of the generated dataset.

In some cases, a low correlation coefficient may not necessarily indicate the feature's irrelevance but could be due to interactions with other features or non-linear relationships. While the former possibility can be excluded as the correlation matrix does not show any

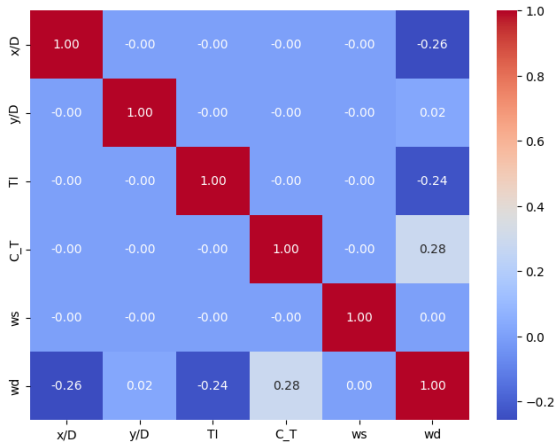


Figure 5.1: Correlation matrix of the variables of dataset $\mathbb{A}$.

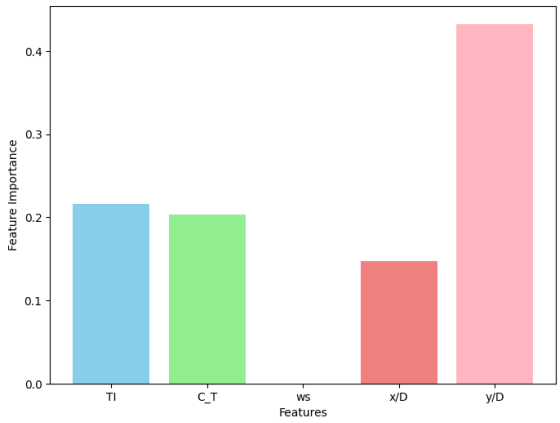


Figure 5.2: Feature importance computed through a Regression Tree.

correlation among the input variables, the latter requires more attention and it is essential to consider other factors when determining the relevance of an input feature.

As an example, the correlation matrix also shows that the y coordinate's correlation factor with the wind deficit wd is close to 0. However, this is not enough to consider it an irrelevant input variable because, as shown below, it is the most important one for the Regression Tree. Moreover, the y coordinate is significant in this project for the study of the wake field's symmetry (see Section 5.3.2).

In order to further study the impact of wind speed, an additional test has been carried out to study the importance of the wind speed feature. A Regression Tree has been fitted on the whole dataset $\mathbb{A}$ with different splitting approaches to study the features' importance. The results are shown in Figure 5.2, and they demonstrate the total irrelevance of ws.

A possible explanation for the irrelevance of wind speed as an input feature is that the wind deficit is not affected by the wind speed, which is mainly modelled by the thrust coefficient (as mentioned in Section 2.1.2.2). Thus, the wind deficit remains constant among different values of the wind speed keeping C_T constant. To validate this hypothesis,

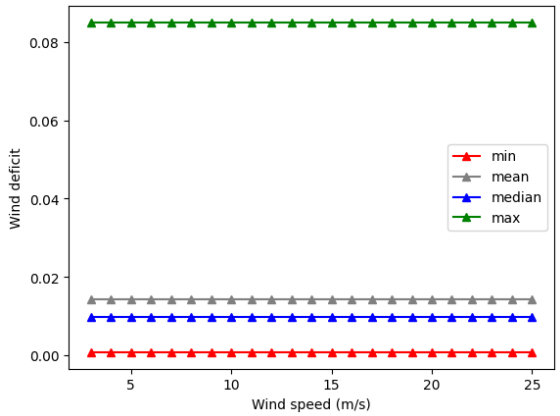


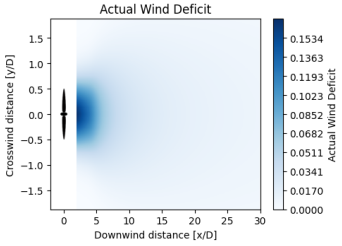
Figure 5.3: Statistics of the wind deficit over wind speed for a random TI-C_T combination.

the trend of the wind deficit has been studied through different wind speed values, and the results are shown in Figure 5.3. The chart is obtained by aggregating the wind deficit values of a wake field $\mathbf{D}^{(i)}$ for a combination of input parameters $TI^{(i)}$ and $C_T^{(i)}$ for different values of ws. The trend is the same among different TI-C_T combinations and also in case the same aggregations are carried out on all the wake fields for each ws value.

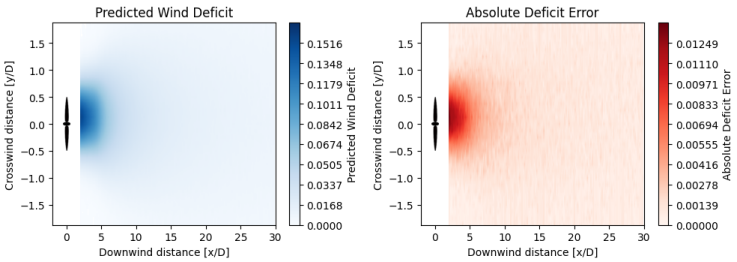
The wind speed's minimum, maximum, median, and average values remain constant across different ws values in the dataset. This observation strongly suggests that wind speed has no relevance as an input feature in predicting the wake field, given the characteristics of the generated dataset. Consequently, for all the experiments detailed in this report, wind speed will not be considered, and all the mentioned experiments rely on the dataset $\mathbb{A}'$.

5.1.2 Streaking and Fuzziness Artefacts in Neural Networks

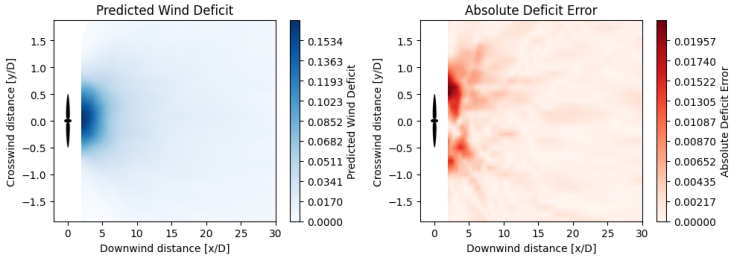
During the qualitative study of data-driven surrogate models based on MLPs, an intriguing artefact was observed in both the univariate and multivariate settings, manifesting in the appearance and smoothness of the wind deficit predictions when employed to predict the entire wake field. This phenomenon has been referred to as "Streaking" in the multivariate setting and "Fuzziness" in the univariate case. In the former scenario, the Streaking artefact presents as noticeable vertical patterns or streaks in the predicted wind deficit maps (see Figure 5.4b), while in the latter, it is evident as an overall lack of smoothness



(a) Original field



(b) Streaking prediction (by Multivariate MLP)



(c) Fuzziness prediction (by Univariate MLP)

Figure 5.4: Illustration of artefacts in MLP-based models. The original wake field to be predicted is shown on the left. (b) presents the prediction and error map of a Multivariate MLP model, exhibiting a streaking effect, particularly noticeable in the error map. (c) displays the prediction and error map of an Univariate MLP model, showing a tendency towards a fuzziness effect.

and the presence of more irregular shapes in the predicted wake fields (see Figure 5.4c).

These artefacts are consistently observed across various simulations, even with different combinations of input parameters. In some cases, the effects become apparent directly in the predicted wake field, while in others, they become more evident when examining the error maps. Notably, hyperparameter tuning, including changes to activation functions, did not lead to significant improvements in mitigating these artefacts.

The presence of these undesired patterns significantly impacts the overall appearance and coherence of the wake field predictions, potentially affecting the accuracy and reliability of the models. As a result, considerable effort has been directed towards investigating the origin of these artefacts and finding potential solutions to mitigate their impact.

Based on these observations, it is plausible to infer that the presence of these artefacts is likely due to architectural limitations in the MLP-based models, rather than issues related to hyperparameter settings or generalization capabilities. The manifestation of these undesired patterns significantly impacts the overall appearance and coherence of the wake field predictions, potentially affecting the accuracy and reliability of the models. One plausible hypothesis that emerged was the influence of *spectral bias* [56, 12], a phenomenon observed in machine learning models where certain high-frequency patterns or noise dominate the predictions, leading to non-smooth and non-physical structures in the generated images.

This notion aligns with previous studies on spectral bias in machine learning models, particularly those related to image generation tasks, which showed that non-linear transformations of the Cartesian coordinates could effectively augment the input space and mitigate the impact of high-frequency noise. By drawing upon these insights, it is possible to identify potential solutions to reduce or eliminate the spectral bias in the wake field predictions, thereby enhancing the overall performance and usability of the data-driven surrogate models for wind turbine wake modelling.

In particular, the noise-like pattern in Univariate MLP resembles the bluriness that was faced by Tancik et al. [66] in the 2D image regression (see Figure 3.1). Therefore, the exploration of the NeRF model represents a step towards addressing this artefact issue. By allowing the network to automatically learn non-linear transformations of the Cartesian coordinates through a Gaussian Fourier Feature mapping, the NeRF model has the potential to capture intricate patterns and high-frequency variations in the wake field, thereby enhancing the overall accuracy and realism of the predictions. In support of this idea, also the work of Wilson, Wakes, and Mayo [75] employed non-linear transformations of the Cartesian coordinates, such as manual transformations like x^2 and $1/y^2$, to augment the input space and enhance the generation of intricate patterns and high-frequency variations in the wind turbine simulations. However, the proposed approach in this work aimed to make these transformations automatic through Gaussian Fourier feature mappings (refer to Section 4.4.1). By leveraging the concepts behind Gaussian Fourier feature mappings, the hope was to enable the resulting Neural Radiance Field (NeRF) to capture complex and high-frequency variations present in the wind turbine wake data.

With this perspective in mind, the NeRF model was selected as a promising candidate for wind turbine wake modelling, as it presents the opportunity to overcome the limitations observed in the MLP-based models. The application of the NeRF model is motivated by the desire to achieve more accurate and visually coherent wake field predictions, thus

improving the overall performance and usability of the data-driven surrogate models in practical wind farm applications.

On the other hand, the Streaking artefact appears to be more closely related to the structure of the multivariate approach, which requires predicting the entire wake field with thousands of different regression values using only a small number of input variables. We empirically observed that having such a large output space can lead to this effect in the predictions. While further investigations and attempts to mitigate this artefact could potentially improve overall performance, the univariate approaches have shown better and more promising results, as discussed in the following sections. As a result, additional efforts to address the Streaking artefact were not prioritized. Future work in this area could explore an hybrid approach, such as using one of the two coordinates as an input variable and predict the corresponding set of values for the other coordinate. Such an approach may strike a balance between the simplicity and performance of univariate models and the information richness of multivariate models, potentially improving the predictions and reducing the impact of the Streaking artefact.

5.2 Quantitative Results and Comparisons

As detailed in Section 4.3, the analysis of the results is carried out with a focus on both interpolation and extrapolation. By investigating the generalization capabilities of the surrogate models under varying conditions, we aim to gain insights into their robustness and applicability to different scenarios. This approach takes into account the specific requirements of the domain, including how data might be generated in similar research studies and the practical needs of the wind-farm industry for layout design and other operational considerations.

In the following sections, we conduct a comprehensive examination of the performance of the main surrogate models – those that demonstrated superior performance during various experiments and have been fine-tuned to maximize their effectiveness. Additionally, Appendix A presents other models that did not yield promising results and were consequently excluded from further consideration.

For a thorough evaluation, we explore the effects of different dataset splitting strategies, as outlined in Section 4.3.4. This enables us to assess the models' effectiveness in handling diverse data distributions and provides valuable insights into their adaptability to real-world scenarios beyond the scope of the training dataset.

5.2.1 Interpolation on Inflow Parameters

In order to assess the models' ability to interpolate on the inflow parameters TI and C_T, we employed Strategy 4.3 to partition the dataset $\mathbb{A}'$ into training, validation, and test sets.

The reduction factors r_{TI} and r_{C_T} determine the number of input values considered for the training set. For the standard experiment/setting, we have chosen the following values: $r_{TI} = 4$ and $r_{C_T} = 4$. Higher values of reduction factors are further explored in 5.2.2. Lower values were not found meaningful for results as the amount of training data would have been too high.

Model Name	Train		Test	
	R^2	**MAE**	R^2	**MAE**
Multivariate NN	0.9910	0.0004	0.9901	0.0004
Univariate RDT	**1**	**3.2e-11**	0.9936	0.0009
Univariate NN	0.9985	0.0009	0.9979	0.0010
Univariate NeRF	0.9997	0.0004	**0.9996**	**0.0004**

Table 5.1: Performance of the surrogate models on interpolation with reducing factors: $r_{TI} = 4$ and $r_{C_T} = 4$

Table 5.1 presents the performance of the surrogate models on both the training and testing sets. Notably, all models achieved a high R^2 score exceeding 0.99, indicating their strong predictive capability. Surprisingly, the RDT model performed remarkably well in this context, and this trend continued in subsequent experiments. As a result, the RDT will be consistently compared with the most performing neural networks in future analyses, showing however always a slightly worse quantitative performance. This observation aligns with the findings of Purohit, Ng, and Kabir [53], where a tree-based model (XGBoost in their case, which has also been tested but found to be inferior to RDT, see Appendix A) exhibited worse performance compared to the model based on Artificial Neural Network. These variations in performance can be attributed to the different complexities and structural characteristics of the models involved.

MAE values reveal significant differences between the models, especially noticeable in the case of the Univariate MLP on the test set, exhibiting an error more than two and a half times higher than that of the Univariate NeRF model. Although the Multivariate MLP achieves the lowest MAE, it is accompanied by a considerably lower R^2 score. This particular architecture appears to face challenges in producing stable results, as evidenced by the presence of the Streaking artefact and the drop in metrics observed in additional experiments using this model, contrasting with the performance of the other surrogate models. Therefore, in the subsequent experiments, the Multivariate MLP will not be considered.

Additionally, the results highlight the expected tendency of decision trees to overfit the training data, as evidenced by the perfect metrics achieved on the training set. On the test set, the best-performing model turned out to be the NeRF, boasting a notably low MAE of only 0.0004. However, in the context of this experiment, the issue of overfitting does not pose a significant obstacle to our objectives. A comprehensive examination of this aspect is provided in Section 5.2.4 and Appendix B, wherein detailed analyses of overfitting and its effect are presented.

In general, this demonstrates the models' good capacity to generalize on unseen data within the range of the known data. One key factor influencing this performance is the total amount of data used, as even 6.4% of $\mathbb{A}'$ (the training percentage) corresponds to 550 simulations.

Additionally, this observation may suggest that the wake fields among different simulations do not significantly differ, and the models can effectively learn the patterns to predict accurate wake fields from just a few initial simulations. These findings provide the

basis for further analyses, focusing on reducing the training data even more to observe changes in performance and investigating qualitative results to gain deeper insights into the behavior of the different models.

5.2.2 Study of the Reduction Factors

In this section, we delve into the impact of reducing the training data further by considering different reduction factors. As discussed in the previous section, our initial experiments achieved commendable generalization performance even with a relatively small training dataset, representing only 6.4% of the available simulations from $\mathbb{A}'$. This observation raises intriguing questions about the models' ability to learn and generalize from a limited amount of data.

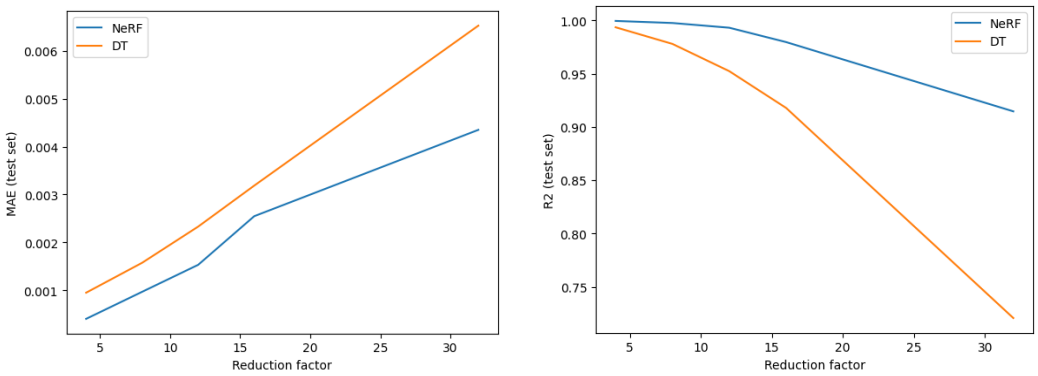
To explore this aspect, we conduct a systematic study with varying reduction factors, including 4, 8, 12, 16, and 32. The reduction factor refers to both TI and C_T, as the same value has been used for both ($r_{TI} = r_{C_T}$). The motivation behind this study lies in gaining deeper insights into the models' behavior and performance under data scarcity scenarios, which are extremely common in this study domain. Additionally, this investigation helps us understand how the models' generalization capabilities are influenced by the reduction in training data size, which could have significant implications in real-world applications where collecting extensive training data can be challenging or costly.

In Table 5.2, we present the performance metrics of the Univariate NeRF and Univariate RDT models under varying reducing factors for the training data. These models have been chosen among the most promising ones because of their peculiar characteristics and almost opposite behaviour. Univariate MLP, even if behaved competitively in the standard interpolation, did not give any reason for further explorations as the additional Fourier Layer seemed to improve its capabilities. As the reducing factor increases, the size of the training dataset gradually decreases, resulting in a data scarcity scenario. The metrics reported in the table pertain to the test data, providing an assessment of how well the models generalize to unseen data when trained on reduced datasets.

The Univariate NeRF models demonstrate remarkable performance even with significantly reduced training data, consistently achieving high R^2 scores and low MAE values. As shown, the model's performance gradually declines with higher reducing factors, in-

Red. Factor ($r_{TI} = r_{C_T}$)	**Training simulations**	**Univariate NeRF**		**Univariate RDT**	
		R^2	**MAE**	R^2	**MAE**
4	550 ($\approx 6.40\%$)	0.9996	0.0004	0.9936	0.0009
8	143 ($\approx 1.66\%$)	0.9976	0.0010	0.9780	0.0016
12	72 ($\approx 0.84\%$)	0.9932	0.0015	0.9524	0.0023
16	56 ($\approx 0.49\%$)	0.9797	0.0025	0.9179	0.0032
32	12 ($\approx 0.14\%$)	0.9147	0.0043	0.7207	0.0065

Table 5.2: Performance comparison on test set of Univariate NeRF and Univariate RDT models under different reduction factors for training data. The second column contains the number of simulations used for the training set and the percentage with respect to $\mathbb{A}'$.



(a) MAE (b) R^2 score

Figure 5.5: Comparison of performance between the NeRF and RDT models over different reduction factors.

dicating a slight decrease in generalization capabilities. However, even under a reducing factor of 32, the Univariate NeRF maintains impressive prediction accuracy.

On the other hand, the Univariate RDT models also exhibit reasonably good performance with low reducing factors. As the reducing factor increases, the RDT models experience a more noticeable decline in R^2 scores and a significant increase in MAE values. This observation suggests that decision trees are more susceptible to overfitting and have reduced generalization capabilities when confronted with limited training data.

Looking at Figure 5.5, the trends in performance become even more clear. The plots display the relationship between the reduction factor and the corresponding MAE and R^2 score for both Univariate RDT and Univariate NeRF models.

Regarding MAE (Figure 5.5a), we observe a linear increase for RDT as the reduction factor for both TI and C_T increases. Similarly, the R^2 score (Figure 5.5b) exhibits a decreasing trend, with a slightly parabolic shape. However, it is important to notice that a linear reduction in the reduction factor does not correspond to a linear reduction in the training data because of the relationship given by Equation 4.2.

Notably, the Univariate NeRF models consistently outperform the Univariate RDT models under varying data scarcity scenarios, both in terms of MAE and R^2 score. This observation suggests that the NeRF models are more robust to data scarcity and can better generalize from limited training data compared to the decision tree models.

Overall, this analysis provides valuable insights into the performance of the models under varying data availability scenarios. It highlights the trade-offs between performance and data availability and sheds light on the suitability of the models for wind turbine wake modelling tasks under different conditions. The findings underscore the importance of collecting larger quantities of data to improve model performance. As demonstrated by the linear trends in performance deterioration with increasing data reduction, a sufficient amount of high-quality training data is crucial for achieving superior predictive capabilities. Consequently, this study emphasizes the significance of data acquisition efforts in order to enhance the models' ability to generalize effectively and produce accurate wake field predictions in practical wind turbine wake modelling applications.

5.2.3 Interpolation including Coordinates

After thoroughly studying the interpolation performance on the inflow parameters TI and C_T, we aimed to expand our research by incorporating the spatial coordinates x and y into the analysis. While the primary focus remains on the models' capability to interpolate the data effectively, this extension introduces a physical basis into the investigation considering a regularly-space wake field. To achieve this, we introduced two new reduction factors, r_x and r_y, which control the reduction of cells seen by the surrogate model during training (refer to splitting strategy 4.4).

With the inclusion of spatial coordinates, the surrogate models now need to consider the spatial distribution of data points in addition to the inflow parameters. This approach provides valuable insights into how the models capture the spatial relationships within the wake field and how well they can generalize across different spatial locations.

However, it is important to note that this type of splitting strategy does not allow us to study multivariate approaches. In multivariate predictions, the model is trained to predict the entire wake field as a whole, utilizing only the inflow parameters. Therefore, no approach based on a multivariate prediction will be explored in this section.

Table 5.3 shows the metrics achieved by the main surrogate univariate models in this setting. Similarly to the previous interpolation experiment, the Decision Tree reaches the maximum metrics on the training set. However, when evaluating the models' generalization ability on unseen data, where the interpolation ability is of particular interest, NeRF outperforms other models and achieves the highest R^2 score. This outcome aligns with expectations, given NeRF's training on a regularly-spaced grid and subsequent testing on the remaining data, as was done in [66]. Interestingly, the Decision Tree achieves the lowest MAE score, which can be attributed to its capacity to capture certain spatial patterns effectively. On the other hand, the Univariate MLP struggles with interpolation, and its R^2 score does not reach 0.8 on the test set.

To further analyze the performance of Neural Network-based models, it is essential to compare their training and validation losses. The validation set is exclusively used for NN-based models and is entirely disjoint from the training and test sets. Figure 5.6 presents the training and validation loss values for the first 80 epochs for the standard interpolation ($r_{TI} = r_{C_T} = 4$), the interpolation involving coordinates ($r_{TI} = r_{C_T} = r_x = r_y = 4$) and the extrapolation (detailed in the next section) experiments. The Figure specifically refers to the NeRF architecture, but the behavior of the Univariate MLP is similar. In both interpolation experiments, the validation loss follows the trend of the training loss,

Model Name	**Train**		**Test**	
	R^2	**MAE**	R^2	**MAE**
Univariate RDT	**1**	**3.0e-11**	0.9446	**0.0029**
Univariate MLP	0.9923	0.0019	0.7902	0.0057
Univariate NeRF	0.9986	0.0007	**0.9715**	0.0034

Table 5.3: Performance of the surrogate models on interpolation on all the input variables with reducing factors: $r_{TI} = 4$, $r_{C_T} = 4$, $r_x = 4$, $r_y = 4$

and the models do not tend to overfit. However, when the model needs to interpolate on coordinates (in red and orange), the overall validation loss is consistently higher and, even if keeps decreasing, the gap with the training loss remains stable over time. This observation suggests that the spatial interpolation task poses additional challenges for the models, and it is reflected in the validation performance. A similar behavior, even more pronounced, is noted in the extrapolation experiment.

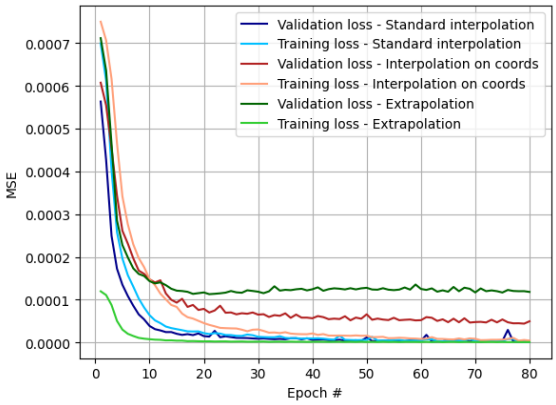


Figure 5.6: Training and validation loss of Univariate NeRF over the first 80 epochs among different experiments. The comparison is between the standard interpolation ($r_{TI} = 4, r_{C_T} = 4$, shown in dark and light blue), the interpolation on coordinates ($r_{TI} = 4, r_{C_T} = 4, r_x = 4, r_y = 4$, shown in red and orange) and the extrapolation ($I_{TI} = [0.15, 0.4]$ and $I_{C_T} = [0.3, 0.7]$, shown in dark and light green).

5.2.4 Extrapolation and Overfitting

After conducting several experiments on interpolation with different parameter settings, we now shift our focus to studying the extrapolation capability of the surrogate models. For this experiment, we utilize the data splitting strategy 4.5.

In the extrapolation experiment, two intervals, I_{TI} and I_{C_T}, are defined to represent the range of values of the inflow parameters considered for the training set. Specifically, we have chosen the values $I_{TI} = [0.15, 0.4]$ and $I_{C_T} = [0.3, 0.7]$, taking into account typical turbulence values and thrust curves. This deliberate choice ensures that the training and

Model Name	**Train**		**Test**	
	R^2	**MAE**	R^2	**MAE**
Univariate RDT	**1**	**2.7e-11**	0.6917	0.0088
Pruned Univariate RDT	0.9989	0.0004	0.7993	0.0064
Univariate NN	0.9987	0.0008	0.6893	0.0088
Univariate NeRF	0.9979	0.0008	**0.9013**	**0.0053**

Table 5.4: Performance of the surrogate models on extrapolation with reducing intervals: $I_{TI} = [0.15, 0.4]$ and $I_{C_T} = [0.3, 0.7]$

testing datasets contain realistic and typical values, while simultaneously reducing the variety of simulations available to the model during training.

In this section, we conduct a comprehensive analysis of the surrogate models' performance on the extrapolation task, which presents unique challenges compared to interpolation. Our primary focus is to assess their ability to make accurate predictions for inflow parameter combinations that lie outside the training data range. Additionally, we investigate how the reduction of training data in this experiment impacts the models' extrapolation capabilities and generalization performance.

The results of this experiment yield valuable insights into the suitability of the surrogate models for extrapolation tasks, shedding light on their characteristics and behavior when facing unseen data points. Furthermore, this investigation provides crucial knowledge for real-data applications, where data extrapolation is often necessary due to limited data collection time and the variations in atmospheric conditions that occur in different geographic locations. Understanding how the models handle extrapolation will enhance their reliability and applicability in practical scenarios, enabling better decision-making in wind farm operations and layout design.

Table 5.4 presents the performance metrics of the main surrogate models in this extrapolation setting, allowing us to assess how well the models perform when faced with the challenges of extrapolation. One evident observation from the results is the noticeable difference between the metrics on the training and test sets, indicating the presence of overfitting. This disparity indicates that the models struggle to generalize well to unseen data points beyond the training range, leading to higher errors and reduced performance in the test set. This phenomenon is further highlighted by Figure 5.6, where the extrapolation lines (depicted in dark green for validation and light green for training losses) underscore the learning process's struggle, particularly in contrast to the standard interpolation method (as well as the one on coordinates). The validation loss reaches convergence ahead of the training loss, albeit at a higher value, while the training loss continues to progressively decrease, converging towards zero.

This tendency does not seem to be influenced by the quantity of data used. For this experiment, approximately 11.6% of $\mathbb{A}'$ has been utilized, which is nearly twice the amount of data used in the standard interpolation setting discussed earlier (Section 5.2.1), where the models exhibited strong generalization capabilities. This observation leads us to believe that the quality of the data, particularly the diversity of wake fields, plays a more crucial role in causing the models' generalization struggles, rather than the sheer amount of data. It becomes evident that these surrogate models struggle to capture the complex patterns needed to accurately predict wake fields beyond the known data range.

In light of the potential issue of decision trees overfitting due to their inherent structure, we conducted new experiments with Extrapolation RDTs to ensure a fair comparison with the other models. In all the previously mentioned experiments, the regression trees were trained without any depth limitation, allowing them to naturally grow up to 40/50 levels with millions of nodes, which significantly increased the likelihood of overfitting. This decision was based on empirical findings, including the study of the impact of reducing the complexity of the tree on the resolution of the predicted wake field. Notably, reducing the depth of the tree had immediate consequences on the resolution, resulting in a less smooth and more 'blocky' shape in the predicted wake field (see Figure B.1).

Interestingly, in the different interpolation experiments, overfitting did not hinder the Decision Tree's ability to generalize well, as evidenced by highly competitive results on the test set. However, in the extrapolation experiment, a noticeable 30% deterioration in the R^2 score (and even more in the MAE) was observed between the training and test sets. To address this issue, we explored the use of a pruned version of the Decision Tree (Pruned Univariate RDT in Table 5.4). After several attempts, we limited the depth of the tree to 20 (from 47 when no limit was imposed), reducing the number of nodes from almost 15 million to less than 2 million. This pruning operation resulted in improved results on the test set. However, it is important to consider that this improvement comes at the cost of lower resolution in the wake field predictions.

Another solution for reducing the overfitting would be the ensemble methods, which have also been explored, as mentioned in Appendix A, but they did not result in noticeable improvements with respect to the standard simpler model.

In the end, we find that the NeRF model demonstrates again the best generalization capabilities on the test set, achieving the highest R^2 score and the lowest MAE. Despite this improvement, it is still essential to explore potential remedies to mitigate the overfitting issues (which make the R^2 score lose almost 30% in the test set), especially in cases where extrapolation is required.

The analysis of the extrapolation experiment provides valuable insights into the models' limitations and their suitability for handling unseen data points outside the known data range. By gaining a deeper understanding of their behavior under extrapolation, we can make informed decisions to enhance the models' reliability and robustness in practical wind farm planning and operations. Furthermore, this research contributes to advancing the field of wind energy research and development, where accurate predictions of wake fields beyond the observed data range are crucial for optimal wind farm design and performance.

5.2.5 Model Complexity

In this section, we delve into the study of computational efficiency, with a specific focus on two essential aspects: **training time** and **prediction time**. The training time refers to the duration required for the models to undergo the training process and obtain the final surrogate model. On the other hand, the prediction time represents the time taken

Model Name	Training Time (s)	Prediction Time (s)
Multivariate NN	10^2	10^{-5}
Univariate RDT	$\mathbf{10^1}$	$\mathbf{10^{-7}}$
Univariate NN	10^4	10^{-6}
Univariate NeRF	10^4	10^{-6}
Ainslie Model [51]	N/A	10^0

Table 5.5: Training and prediction times (orders of magnitude) for different models and PyWake's Ainslie implementation. The prediction time refers refers to the time to generate a simulation given the input parameters TI and C_T.

by the models to generate complete predictions, i.e., to forecast the entire wake field for a specific combination of inflow parameters.

The motivation behind exploring computational efficiency stems from the fact that it plays a pivotal role in the successful implementation of surrogate models, particularly considering the computational complexity of the currently used methods, as discussed in the initial chapters. Despite opting for a less complex model for data generation, it remains crucial to assess and compare the computational efficiency of the resulting surrogate models with respect to the Ainslie model, which serves as a direct point of comparison for this project (in particolar, the PyWake implementation [51] will be considered). Ensuring that the prediction time of the data-driven surrogate model is inferior to the model used for generating the data is paramount, as it determines the practical usability and viability of the surrogate model.

Moreover, the comparison of prediction times among different models not only sheds light on computational efficiency but also provides insights into the complexity of each model. A model that requires significantly less time for complete predictions might suggest a more efficient and streamlined approach to solving the wind turbine wake modelling problem.

Furthermore, the evaluation and comparison of training times add depth to the analysis. Training times are of utmost importance as they encompass the total duration needed to train the best model with specific parameters and settings. This duration spans from the initiation of training until the epoch where the model achieves the best MSE loss on the validation set. Considering the variations in the datasets and generation methods, there is a possibility that the models may require retraining in certain situations. Highlighting this possibility and providing reasons for its likelihood can help us comprehend the implications it holds for the practical application of these surrogate models in the wind turbine wake modelling domain.

Table 5.5 provides a summary of the training and prediction times in orders of magnitude for the different models, including the PyWake implementation of the Ainslie model. The prediction time specified in the table corresponds to the time required to generate a simulation given the input parameters TI and C_T, making the numbers perfectly comparable.

To present the training and prediction times in a concise and generic manner, we use the orders of magnitude notation, such as 10^x, where x denotes the exponent. This notation allows us to provide an overview of the relative time scales involved in the computations, making the comparison more intuitive and facilitating the identification of any significant differences in efficiency among the models. The actual training time naturally depends on various factors, including the amount of data, the model architecture, the complexity of the optimization process, and the computational resources available. As we perform experiments with different amounts of data and model complexities, it becomes impractical to provide precise training times for each specific experiment. Therefore, we leverage the orders of magnitude notation to present a general overview of the training times' trends across the different models and experiments.

It is worth mentioning that the training time reported does not include the time taken to load the data into memory and apply the corresponding splitting strategies and batching operation. This loading and splitting time is constant among the different models, and its

impact on the overall training time should not be equally distributed.

Overall, a significant difference is evident between the RDT-based and NN-based models in terms of training time. The NN-based univariate models require many hours to complete training, whereas the RDT-based models take just a few seconds. The Multivariate NN, on the other hand, falls in between, taking a few minutes for training and proving to be considerably faster than the other NN-based models. This difference in training time is attributed to the distinction in concepts of *instance* as mentioned in Section 4.3.2. The number of training instances for the Multivariate Neural Network is on a significantly lower order of magnitude compared to the other NN-based models, even though the overall dataset size is the same. This fundamental difference in the number of instances leads to a substantial disparity in the required training time, making the Multivariate NN the most time-efficient among the neural network models.

Regarding prediction time, we observe substantial differences between the surrogate models and the PyWake implementation of the Ainslie model, with the former obtaining at least 6 orders of magnitude faster predictions. While the Univariate Decision Tree exhibits faster prediction times than the neural network models, the overall difference is negligible compared to the significant gain achieved by employing the surrogate models instead of the Ainslie model. The surrogate models offer a more efficient and practical solution for wind turbine wake modelling, as they enable faster and accurate predictions without sacrificing the quality of results. This advantage is of paramount importance in real-world applications, where computational efficiency is crucial for timely decision-making and resource management, especially in scenarios where a large number of simulations need to be run.

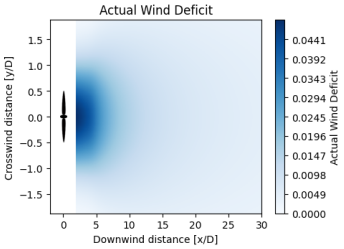
5.3 Qualitative Results and Visual Comparisons

In this section, we shift our focus to the qualitative evaluation of the surrogate models' performance. While quantitative metrics are essential for assessing the models' general predictive capabilities as they provide valuable insights into the models' predictive capabilities, they may not fully capture specific spatial variations, local anomalies, or subtle wake interactions that could be crucial in real-world scenarios. To complement the quantitative assessment, we now delve into the visual analysis of the predicted wake fields, which allows us to gain a deeper understanding of how well the models capture the intricate patterns, trends, and structures present in the wake field.

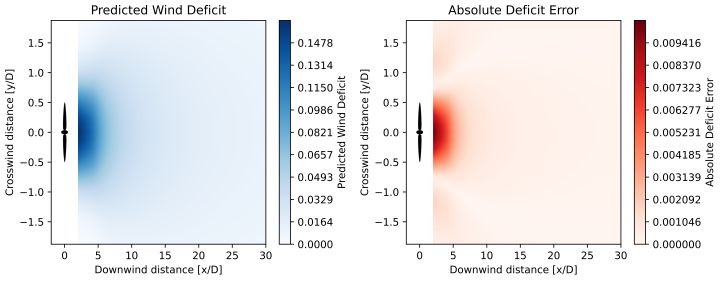
Through visualizations and graphical representations, we aim to assess the qualitative aspects of the surrogate models and their ability to visually represent the complex wake characteristics. This analysis provides a visual understanding of how the models' predictions compare to the actual wake fields, offering insights into the models' strengths and limitations from a different perspective. By visually inspecting individual wake fields in detail, we can identify potential areas of concern and assess the robustness of the models for practical wind turbine wake modelling applications. For instance, in wind farm layout design, engineers and stakeholders can gain a better understanding of how the wake interacts with surrounding turbines and how the overall wind farm performance might be impacted by visually inspecting the predicted wake fields.

Particular focus will be dedicated to the study of symmetry, which is of significant in-

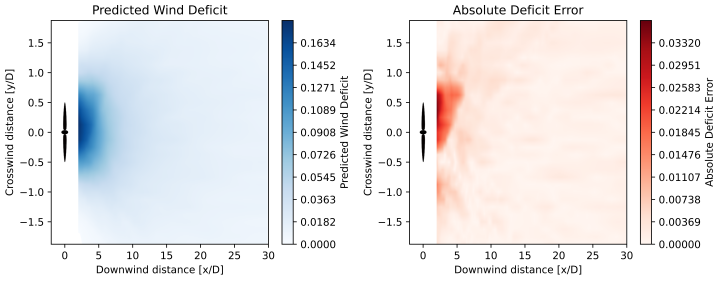
terest due to the radial symmetry of the generated data (as discussed in Section 2.1.1.1). While a quantitative approach could have been employed to measure the degree of sym-



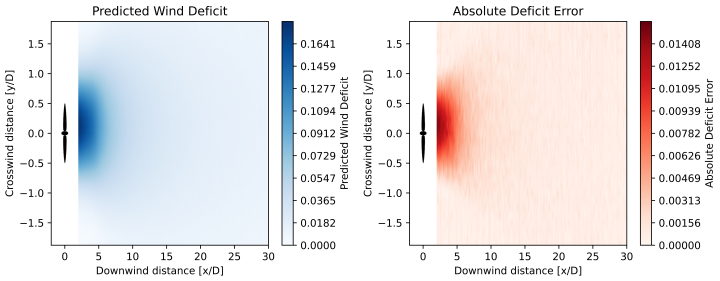
(a) Common original field



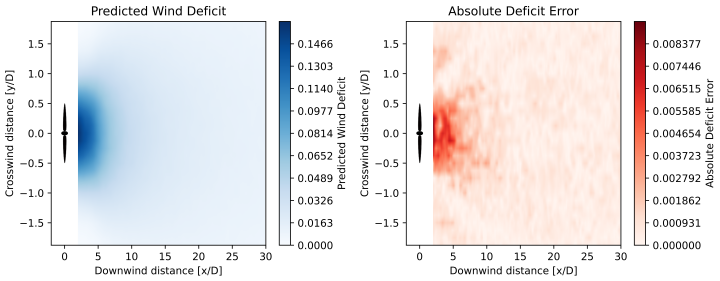
(b) Univariate RDT



(c) Univariate MLP



(d) Multivariate MLP



(e) Univariate NeRF

Figure 5.7: Predicted wake fields and corresponding error maps with respect to the common original wake field for the simulation $TI = 0.79, C_T = 0.37$. The models are trained in the standard interpolation setting ($r_{TI} = r_{C_T} = 4$).

metry in the predicted wake fields, we have chosen a visual analysis approach for several reasons. First, quantifying symmetry may be challenging due to the complexity and variability of wake patterns. Second, a visual inspection allows for a more intuitive and qualitative understanding of potential asymmetries and patterns. Additionally, given that symmetry is an important characteristic of the data, deviations from it can be easily observed through visualizations, making it a practical and effective approach for our analysis.

An important limitation of the qualitative evaluations lies in the sheer number of predicted simulations generated by the surrogate models. With potentially hundreds or even thousands of simulations, conducting an in-depth and rigorous visual analysis of every single prediction becomes infeasible and impractical. Consequently, we have selected a subset of representative images for an in-depth qualitative study, ensuring that they capture key variations and characteristics in the wake fields predicted by the different models. Due to space constraints, only a select few examples will be presented in this section. More images can be found in Appendix C. By carefully selecting representative images, we aim to provide insightful observations and valuable visual evidence of interesting patterns and potential deviations from the expected behaviour, while acknowledging that the conclusions drawn from this limited selection may not be fully representative of the entire prediction set.

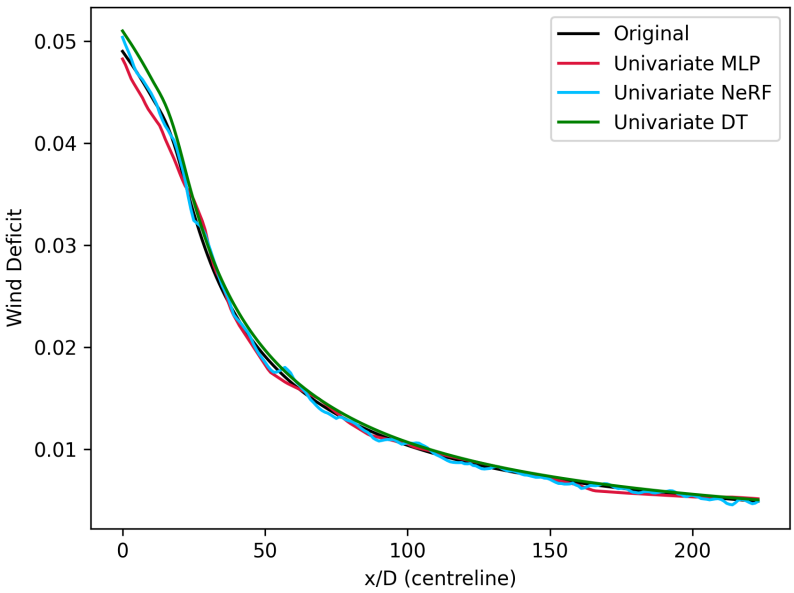
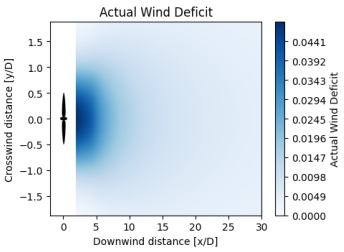


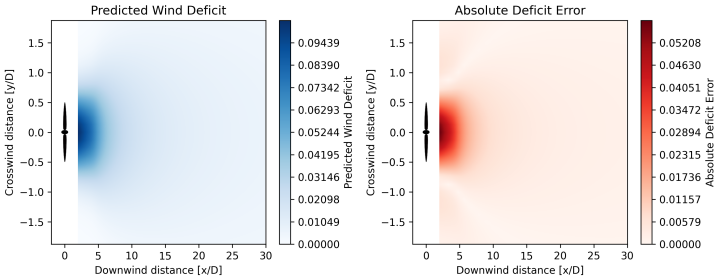
Figure 5.8: Visual comparison of the centreline predictions by different surrogate models in simulation $TI = 0.79$, $C_T = 0.37$.

5.3.1 Interpolation

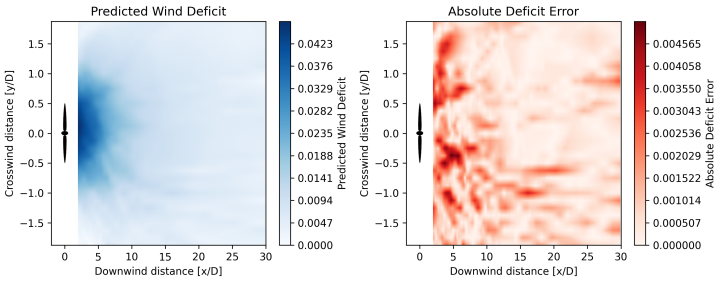
Figure 5.7 showcases the predicted wind turbine wake fields generated by the surrogate models mentioned in the previous section (Multivariate NN and Univariate RDT, NN, and NeRF). The image on the left represents the ground truth, i.e., the result of the Ainslie



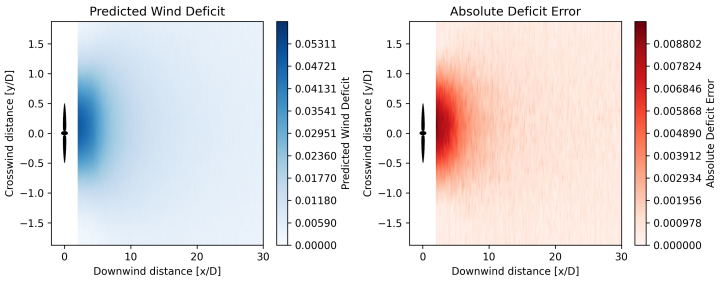
(a) Common original field



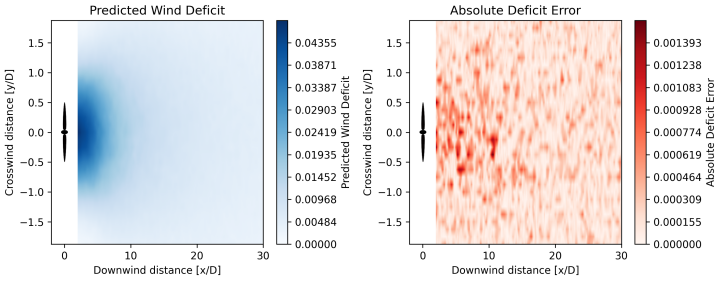
(b) Univariate RDT



(c) Univariate MLP



(d) Multivariate MLP



(e) Univariate NeRF

Figure 5.9: As for Figure 5.7, but the models are trained in the standard extrapolation setting ($I_{TI} = [0.15, 0.4]$ and $I_{C_T} = [0.3, 0.7]$).

model for the simulation ($TI = 0.79, C_T = 0.37$), which was not part of the training data. Each row on the right part displays the prediction and the corresponding error map for each model. The colour maps in the predictions and the original wake field depict the wind deficit values, with warmer colours indicating higher deficits and cooler colours indicating lower deficits. The error maps represent the absolute differences between the predicted and actual values, allowing for a better understanding of the shape and pattern of errors. This type of visualization serves the purpose of effectively comparing the predicted wake field with the original, thereby offering a clear view of the model's prediction performance, assisted also by the error map. Ultimately, this approach aids in assessing the visual implications of the model's predictions, which could be crucial when designing wind farm layouts.

Regarding the artefacts observed in the Univariate MLP and Multivariate models, as mentioned in Section 5.1.2, they are visible in Figures 5.7c and 5.7d. While the NeRF architecture (5.7e) partially reduces the noise in the error map compared to the Univariate MLP, some small irregularities are still present, albeit with significantly lower error values. It is worth noting that the NeRF architecture seems to reduce the fuzziness effect considerably, leading to a more coherent and smoother prediction in the wake field.

On the other hand, RDT exhibits the smoothest shape in the predicted field, with a coherent and evenly distributed error. In Appendix C, it is possible to find a similar figure for another simulation (see Figure C.3).

To provide a comprehensive visual comparison of all the models in a single image, Figure 5.8 presents a line chart displaying the centreline predictions of each model for the same simulation. The chart reveals the close proximity of the NeRF prediction (in cyan) to the original wake field (in black), indicating the model's ability to capture fine details and high-frequency variations. On the other hand, the line chart demonstrates the greater smoothness of the RDT predictions (in green), albeit at the cost of slightly worse performance. For a broader and more detailed overview of the models' behaviour and predictions, additional similar plots can be found in Figure C.1 with other simulations. Figure C.2e shows the same simulation obtained from models trained in the extrapolation setting, among other line charts in Figure C.2.

5.3.2 Symmetry

The radial symmetry characteristic implies that the 2D predicted wake fields, when compared to the horizontal centreline, are expected to exhibit symmetry between their upper and lower parts. Analyzing the predictions of the surrogate models, any lack of symmetry or the presence of asymmetrical details could be indicative of potential issues in the models, suggesting that the model has learned incorrect patterns or has limitations in its structure, such as bias or variance. Identifying such issues through visual analysis is essential for gaining deeper insights into the models' behavior and performance.

Looking at Figure 5.7, interesting differences in symmetry among the four models can be observed. RDT demonstrates the most pronounced symmetry, followed by the Multivariate MLP. NeRF also exhibits a relatively symmetrical shape in the predicted wake field, but the irregularities in the error map are not perfectly symmetrical. On the other hand, the Univariate MLP shows the least symmetry and coherence in the predicted

wake field and its error map.

5.3.3 Extrapolation

Figure 5.9 displays the predicted wind turbine wake fields generated by the surrogate models in the standard extrapolation setting, where the models are trained using data from the intervals $I_{TI} = [0.15, 0.4]$ and $I_{C_T} = [0.3, 0.7]$. As with Figure 5.7, each row represents the model's prediction and its corresponding error map. The ground truth wake field is depicted on the left side of each row.

Examining the figure, similar conclusions can be drawn as in the standard interpolation setting. The models' ability to predict the unseen simulations is visibly challenged, highlighting the complexities and limitations of extrapolation tasks. The figure emphasizes the observed patterns, differences among the models, and other observations noted in the previous experiment, reaffirming the importance of careful model selection and evaluation in the context of wind turbine wake predictions and the challenges of an extrapolation setting.

Comparing the interpolation and extrapolation settings using an additional simulation (refer to Figures C.3 and C.4), we encounter a case where the visual differences, both in terms of shape and patterns, may not be immediately apparent unless the colour scale is aligned. However, upon inspecting the respective colour bars, it becomes evident that errors tend to be higher among the different models in the extrapolation setting. This observation further accentuates the challenges posed by extrapolation tasks.

Chapter 6

Conclusion and Perspectives

In this final chapter, we present the conclusions and perspectives of our study on surrogate models for wind turbine wake predictions. Section 6.1 provides the final conclusions regarding the performance and limitations of the developed surrogate models, aiding in the selection of the most suitable model. In Section 6.2, we discuss the broader implications and applicability of surrogate models for wind turbine wake modelling, while also addressing their limitations. Finally, the last section (6.3) explores future perspectives and offers recommendations for advancing wind turbine wake modelling using surrogate models, highlighting additional ideas and potential research directions for practical applications.

6.1 Final Considerations on the Surrogate Models

In the previous chapter, we conducted a comprehensive analysis and discussion of the experiments performed on the promising surrogate models for wind turbine wake predictions. Among the different experiments, it was observed that the univariate approaches, specifically the Regression Tree (RDT) and Neural Radiance Fields (NeRF), exhibited particularly interesting and stable performance.

Both RDT and NeRF demonstrated exceptional applicability in the standard case, requiring only a small number of simulations, and they maintained their effectiveness even when the amount of data was drastically reduced. Notably, NeRF consistently outperformed RDT in terms of quantitative performance across all experiments, showing better trends in both MAE and R^2 score with decreasing reduction factors and extrapolation. NeRF displayed a remarkable capability of generalization, making it less likely to overfit the data compared to RDT. However, this enhanced predictive capability of NeRF came at the cost of a more complex model that required significantly longer training times compared to RDT. Nevertheless, the prediction times of both models remained comparable, and they both provide a significant reduction in the simulation time with respect to the original simulation method.

In terms of qualitative performance, visualizations of the predicted wake fields and error maps revealed slight differences between the two models. RDT generated smoother and more coherent shapes, while NeRF exhibited slightly more unstable and noisy patterns. The visualizations of the centreline prediction further reinforced this observation, with RDT presenting a precise and elegant curve, while NeRF's line appeared more vibrating and unstable, even if closer to the ground truth curve.

Despite the presence of the Fuzziness artefact in NeRF, it is important to emphasize that NeRF remains a valuable tool for predicting wind turbine wake effects. The impact of the artefact may vary depending on the specific requirements of the prediction task, and

in certain applications, it may not significantly affect the overall performance, especially considering the small quantitative error that it may achieve.

On the other hand, the Regressor Tree still behaves very well, especially considering its simplicity, and it could be a good choice in case the training time is a big concern (although considering that it's a one-time cost) or in case a small improvement in prediction time is important for a particular application due to a large number of simulations to run or other constraints.

The choice between these models depends ultimately on the specific needs and constraints of the wind turbine wake modelling task at hand. Nevertheless, both models represent highly viable alternatives to the original Ainslie model, boasting simulation times that are millions of times faster. This underscores the immense potential of surrogate modelling in wake predictions.

6.2 General Conclusions, Limitations and Applicability

The examination of reduction factors provides insights into the data requirements to achieve specific results. Tens or hundreds of simulations may be necessary based on the desired level of performance and error tolerance. Importantly, the study reveals that the quantity of data alone is not a reliable indicator of model generalization. Instead, models perform significantly better when a smaller quantity of data covers a wider range of potential values. This emphasizes the challenge of model generalization in real-data scenarios, where the available data often covers only a limited subset of possible atmospheric conditions and turbine types. On the other hand, in the case of synthetic generation through more complex models that can only be run with a few simulations, these models prove highly suitable. In such scenarios, it becomes crucial to select parameters that span a broader range, covering as much of the variability as possible and leveraging the interpolation capabilities of the model for effective wake predictions.

It is essential to acknowledge that these findings are limited to the Ainslie model, which exhibits less complex predictions compared to computational models. Thus, the models' behavior could differ when applied to scenarios with more irregular shapes or complex wake structures. This observation opens up many perspectives for future work and calls for further investigation into the models' performance in more challenging and realistic scenarios.

Regarding the applicability of this work, the developed surrogate models offer compelling solutions for single-wake scenarios in wind turbine wake modelling and wind farm layout design. These models offer a substantial edge over the original Ainslie model, boasting remarkable computational efficiency enhancements that result in simulation times millions of times faster. This leap forward renders them especially attractive for entities such as companies, researchers, and stakeholders who currently depend on the Ainslie model or other engineering models with similar or lower precision. The surrogate models enable these users to significantly curtail computational expenses while still attaining reasonably accurate wake predictions.

It is important to acknowledge that while these surrogate models present an attractive alternative to expensive methods like CFD simulations, they are trained on less precise ground truth data. Therefore, users should be mindful of their limitations in replicating

real-world wake effects with high precision. Despite this, the models' impressive predictive capability and efficient computation make them a viable and practical option for many wind turbine wake modelling scenarios.

Moreover, these surrogate models hold the potential to extend their application from single-turbine wake predictions to more complex multi-turbine layouts using superposition models. By combining the individual predictions of multiple single wakes, it becomes feasible to simulate the intricate interactions between multiple turbines. Although this multi-turbine simulation was beyond the scope of this project, readers interested in exploring this aspect further can refer to specialized literature in the field of wind turbine wake modelling and wake interaction studies.

6.3 Future Perspectives and Recommendations

While there is room for improvement of the surrogate models developed in this study, the changes are not expected to be revolutionary. The models achieved nearly perfect results with a decent amount of data, and improving the models in scenarios with limited data may not be meaningful due to the non-prohibitive cost of simulations using the Ainslie model. Therefore, extensive efforts to enhance the Ainslie models may not be the most compelling direction for future work.

Instead, an intriguing avenue for exploration lies in adapting Neural Radiance Fields to other data scenarios. Despite its promising characteristics, NeRF has not been previously investigated in the context of wake predictions. Its inherent suitability for handling 2-dimensional or 3-dimensional nature of wake fields makes it an appealing choice for further research. Exploring the performance of NeRF on 3D predictions would be particularly intriguing, as most of the existing research on NeRF primarily focuses on this type of data.

Furthermore, it is worth considering more complex data scenarios, such as simulations involving multiple turbines or more intricate wake interactions. Such scenarios can introduce a wider range of parameters and conditions, making the modelling problem more challenging and reflective of real-world complexities.

The necessity for substantial quantities of data has been confirmed through experiments on extreme values of interpolation and extrapolation, where performance significantly drops with data reduction. This highlights the importance of acquiring extensive data when employing other generation methods for future work. However, researchers should be mindful of the often prohibitive cost associated with these methods.

Another intriguing avenue for future work is to merge the Ainslie model data with computational simulations. While similar approaches have been explored in the context of analytical models combined with real data [45], this work could focus on generating a substantial amount of data using the Ainslie model, which has been proven feasible and not overly expensive. By combining this Ainslie-generated data with fewer simulations from complex and more reliable models, researchers could potentially achieve improved performance through transferred learning or fine-tuning techniques.

Finally, exploring the possibility of incorporating uncertainty estimation in the surrogate models would be valuable. Given the inherent uncertainties in wind turbine wake predictions, providing uncertainty estimates alongside the predictions can be useful for

decision-making and risk assessment in real-world applications, especially when designing wind farm layouts. While some attempts have been carried out in this regard for this work but without achieving the desidered results, more effort in this direction could yield valuable insights and improvements.

In conclusion, while the surrogate models presented in this thesis offer significant improvements in computational efficiency and prediction accuracy, there are several exciting avenues for future exploration and research. By advancing the state-of-the-art in wake prediction modelling, researchers can contribute to the advancement of wind energy technologies and their widespread adoption in sustainable energy applications.

Appendix A

Less Promising Models

In this appendix, we explore alternative models that were considered for predicting wind turbine wake effects but did not achieve promising results. While the primary focus of this thesis is on the development and evaluation of successful surrogate models, it is essential to acknowledge the other modeling attempts and the insights gained from their evaluation.

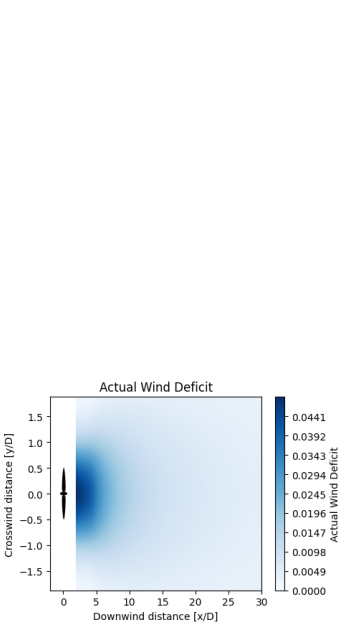
Among the less promising models, both univariate and multivariate **Linear Regression** (LR) are included. Although these models served as a baseline for comparison, they were not expected to be competitive due to their limitations in capturing the inherent non-linearity of the wind turbine wake data. The simple and quick nature of Linear Regression made it valuable for assessing the performance of more complex models, especially in the multivariate setting, where only the Neural Network-based model surpassed the initial exploration.

Additionally, other tree-based models were considered, namely **Random Forest** (RF) and **XGBoost** (XGB), motivated by the competitive results obtained with the basic Regression Tree (RDT). For Random Forest, 100 estimators have been used to leverage bagging and achieve robust predictions, no maximum depth was set and Mean Squared Error was chosen as the criterion for splitting. However, the tree-based models' performance, while interesting and competitive, was only comparable to or slightly worse than the RDT baseline (see Table 5.1). Consequently, there was little incentive to adopt more complex models.

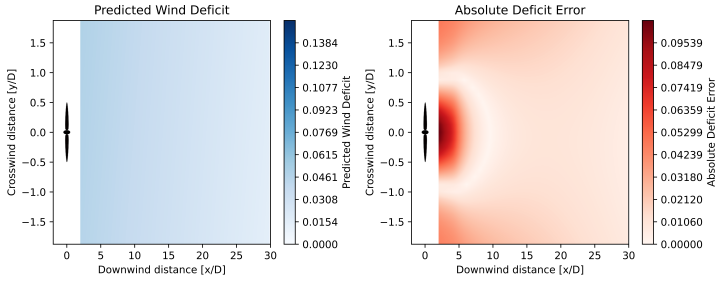
Gaussian Process (GP) modelling, with the added benefit of measuring uncertainty, was also considered due to its potential advantages for capturing complex patterns and providing predictive uncertainties, even though GP models are more suitable in scenarios with small data. Attempts were made to employ GP in both univariate and multivariate settings, but memory errors were encountered during the experiments. These errors were due to the vast amount of memory required for the covariance matrix multiplication [60],

Model Name	**Train**		**Test**	
	R^2	**MAE**	R^2	**MAE**
Multivariate LR	0.7404	0.0067	0.7518	0.0063
Univariate LR	0.2034	0.0154	0.2083	0.0147
Univariate RF	1.0000	3.3e-5	0.9936	0.0010
Univariate XGB	0.9955	0.0012	0.9888	0.0016
Univariate AGP	0.5750	0.0119	0.5859	0.0113

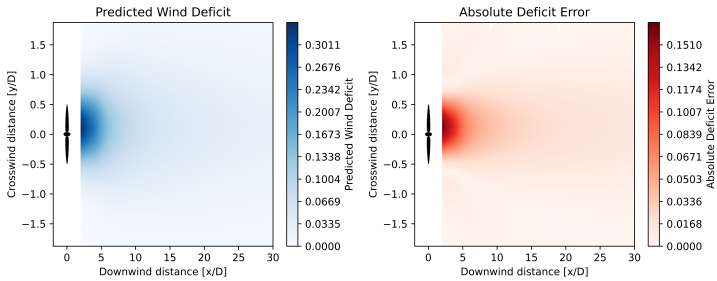
Table A.1: Best results of the excluded models in the interpolation with reducing factor of 4



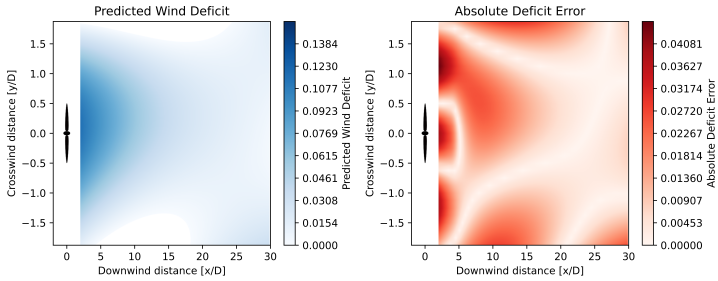
(a) Common original field



(b) Univariate LR



(c) Multivariate LR



(d) Univariate AGP

Figure A.1: As for Figure 5.7, but for the less promising models.

which became impractical for the abundant data in this study, and even more powerful computing resources could not circumvent the challenge. Although reducing the amount of data could have addressed this issue, empirical attempts indicated that the treatable data size would be too limited to provide meaningful results for this project (limiting the amount to one or two simulations at most).

To explore a computationally more efficient and feasible version of GP, the **Approximate Gaussian Process** based on Random Fourier Features (AGP) was also tested. This model employs an approximated radial basis function (RBF) kernel with 32 components using $1/(\text{num features} \times X.\text{var}())$ as the value of γ [58]. However, it did not exhibit promising performances either. While this approximation aimed to overcome the computational challenges faced with full GP, it still failed to achieve competitive results compared to other selected models.

Table A.1 summarizes the best results obtained by these excluded models in the standard interpolation setting with a reduction factor of 4. We observe that the tree-based

models (RF and XGB) demonstrated more competitive performances, but they did not surpass the direct competitor based on one tree (see Table 5.1). Furthermore, AGP showed less accurate predictions compared to the other mentioned models. This table also serves the purpose of establishing lower thresholds for acceptable R^2 scores and MAE errors within this problem, thereby providing insights into performance boundaries.

The inclusion of LR aimed to explore their performance in capturing the complex relationships within the wind turbine wake data. It was clear from the outset that Linear Regression might struggle to provide competitive results when dealing with highly nonlinear data. The results of the former totally confirmed this expectation. The model's inability to capture the intricate nonlinear patterns in the wake field led to relatively low performances, with a stable R^2 value of 0.20 both in training and test sets. Multivariate LR exhibited relatively better performance but still fell short in capturing the complexity of the wake effects. While the R^2 values for Multivariate LR were higher than the Univariate version, reaching 0.7404 on the training set and 0.7518 on the test set, the model's predictive power was still limited, resulting in insufficient predictive accuracy.

Figure A.1 visually highlights the limitations of the LR models and AGP, not only in terms of performance but also in faithfully replicating the shape of the original wake field. Additionally, in Figure A.2, a direct comparison of these models is presented within a single plot, specifically focusing on centreline predictions. This plot employs the same simulation and settings as shown in Figure 5.8. Furthermore, it is discernible that Multivariate LR is adept at capturing a less linear trend due to its numerous linear outputs, resulting in a shape closer to that of the centreline.

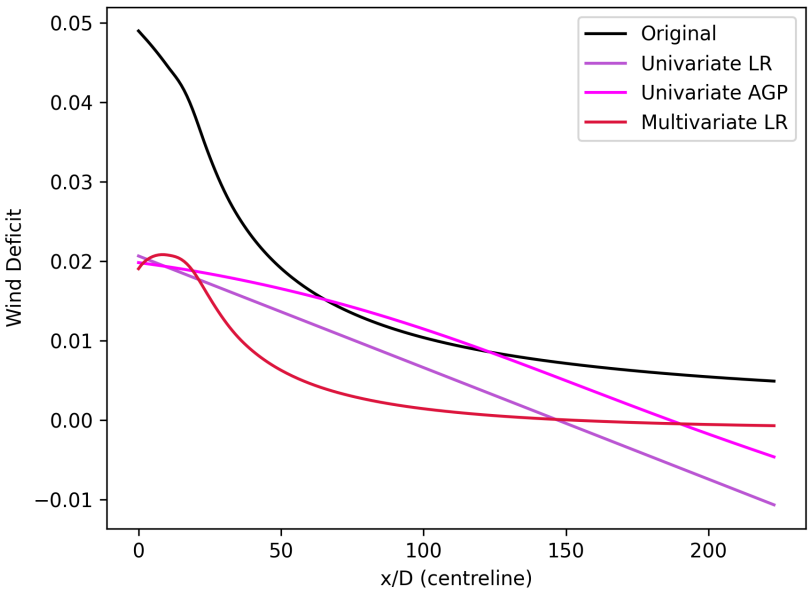


Figure A.2: As for Figure 5.8, but for the less promising models

In summation, the evaluation of these less promising models yielded no substantial enhancements or compelling reasons for further exploration. Consequently, our focus

remained centred on the Neural Radiance Fields and other successful techniques that exhibited a greater potential for accurately predicting wind turbine wake effects. The exclusion of these models from further discussions allows us to streamline our efforts and concentrate on the most promising methodologies.

Appendix B

Impact of Decision Tree Depth on Wake Field Resolution

In this appendix, we delve into the impact of decision tree depth on the resolution of predicted wake fields in the context of our wind turbine wake modelling study. Decision trees are known for their tendency to overfit the training data, and one of the factors contributing to overfitting is their depth. However, in the current research domain, pruning the tree to mitigate overfitting directly affects the resolution of the wake field, thus necessitating a thorough investigation to strike the right balance between model performance and prediction quality.

B.1 Motivation

Throughout the interpolation and extrapolation experiments, we have extensively discussed the performance of Regression Trees and noted their susceptibility to overfitting. While decision trees may achieve perfect metrics on the training set due to their ability to grow deeper, the true test of a model's capability lies in its generalization on unseen data, represented by the test set. We observed a notable difference in behaviour between the interpolation and extrapolation settings regarding decision tree depths.

Table B.1 presents a comparison of the metrics achieved in both experiments, considering different limits on the depth of the tree or no limits at all (last row of each experiment). In interpolation, allowing the tree to grow freely led to the best results. However, when evaluating the same depth values in the extrapolation setting, we observed a reduction in performance (both in terms of R^2 score and MAE) when no depth limit was imposed, and the optimal results were obtained by setting a maximum depth of 20 (as mentioned in Section 5.2.4).

To gain a deeper understanding of the relationship between decision tree depth and wake field resolution, we conduct a comparative analysis. Specifically, we investigate how different decision tree depths impact the shape and resolution of the predicted wake fields.

B.2 Visual Comparative Analysis

In the case of overfitting, such as the one observed in the interpolation experiment, a possible approach is to reduce the depth of the tree. While this may yield a quantitative improvement in terms of generalization performance, we need to carefully consider its impact on the qualitative aspect, particularly the resolution of the predictions.

Examining decision tree models with shallow depths reveals a distinct outcome. These models tend to produce coarse and 'blocky' predictions for the wake field. Due to the

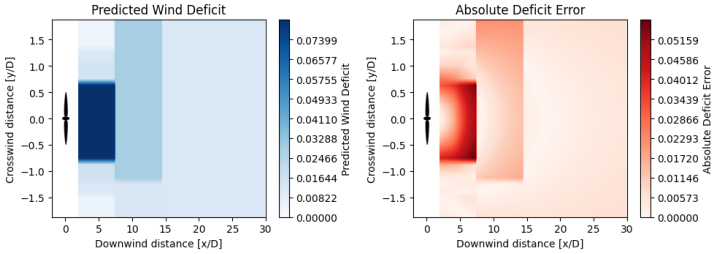
Experiment	Depth	# Nodes	Train		Test	
			R^2	MAE	R^2	MAE
Interpolation	5	63	0.6337	0.0088	0.6421	0.0084
	10	$\approx 2,000$	0.9323	0.0033	0.9257	0.0034
	15	$\approx 65,000$	0.9886	0.0012	0.9820	0.0015
	20	$\approx 1,350,000$	0.9990	0.0003	0.9929	0.0009
	38 (max)	$\approx 4,300,000$	**1**	**3.2e-11**	**0.9936**	**0.0009**
Extrapolation	5	63	0.7942	0.0064	0.3274	0.0130
	10	$\approx 2,000$	0.9599	0.0026	0.6758	0.0084
	15	$\approx 65,000$	0.9924	0.0011	0.7630	0.0070
	20	$\approx 1,900,000$	0.9989	0.0004	**0.7993**	**0.0064**
	47 (max)	$\approx 14,700,000$	**1**	**2.7e-11**	0.6917	0.0088

Table B.1: Comparison of model performance on the interpolation and extrapolation tasks with different decision tree depths. The table showcases the performance metrics (R^2 score and MAE) for each experiment, considering various depth limits or no limit (indicated as "max" in the Depth column). The number of nodes in each decision tree is also provided.

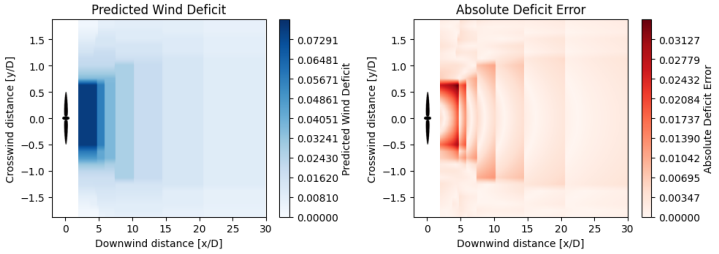
limited number of splits in the tree, large regions of the feature space are assigned the same predicted wind deficit value. Consequently, the wake field predictions lack fine-grained details and may fail to capture intricate spatial patterns present in the data. This reduction in resolution can be concerning, especially when precise predictions are essential for practical applications, such as wind farm layout design.

To visually compare the wake field resolution, we present a series of predictions in Figure B.1. The leftmost image (B.1a) displays the original wake field, which is common to all experiments and obtained using the inflow parameters $TI = 0.5$ and $C_T = 0.53$. Notably, this simulation has not been included in the training set. The images on the right (from B.1b to B.1f) depict wake fields predicted by decision tree models trained with different depths. Accompanying each prediction, we include absolute error maps that highlight patterns in the resolution of the predicted wake fields.

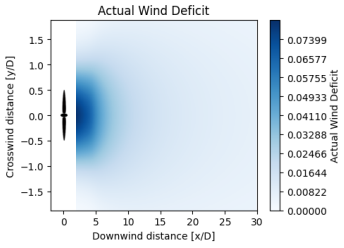
The results from these visualizations clearly illustrate a direct and proportional relationship between decision tree depth and wake field resolution. Therefore, when considering tree pruning to improve generalization while preserving resolution, it is crucial to find the optimal depth limit that ensures sufficient resolution in the predicted wake fields. This understanding can aid in identifying potential areas of concern and gaining deeper insights into the wake interactions with surrounding turbines, ultimately enhancing the practical applicability of the models in real-world scenarios.



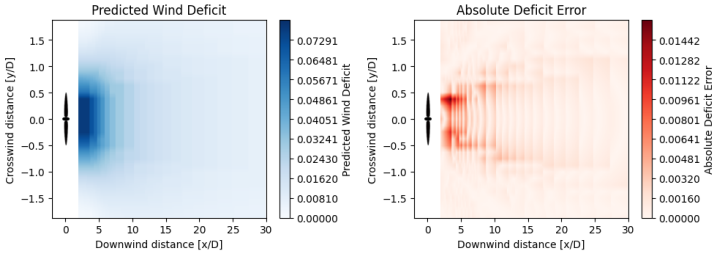
(b) Max depth = 5, Number of nodes = 63.



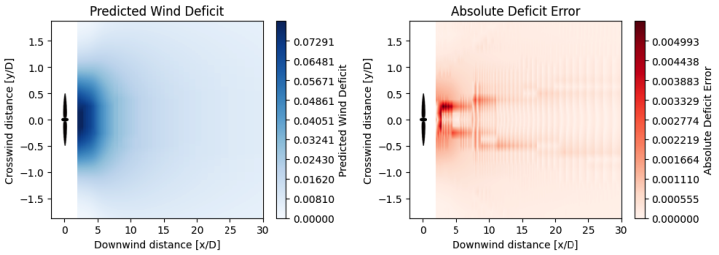
(c) Max depth = 10, Number of nodes $\approx 2,000$.



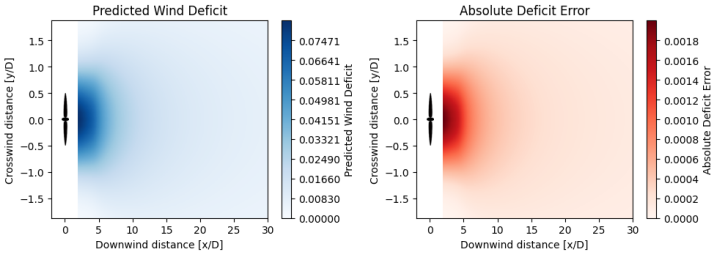
(a) Common original field



(d) Max depth = 15, Number of nodes $\approx 65,000$.



(e) Max depth = 20, Number of nodes $\approx 1,350,000$.

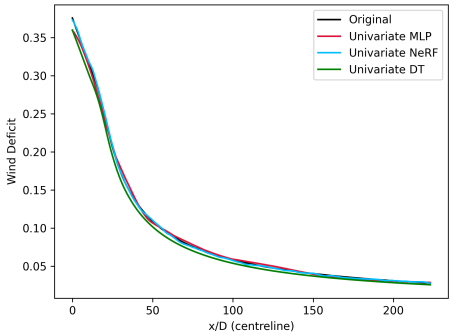


(f) No limits: Max depth = 38, Number of nodes $\approx 4,300,000$.

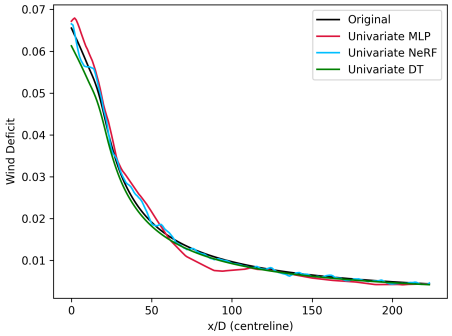
Figure B.1: Comparison of wake fields predicted by RDT models trained with varying maximum depths on the test-set simulation $TI = 0.5$, $C_T = 0.53$. (a) displays the original wake field, common to all experiments. The images on the right (from (b) to (f)) show the predictions of the models trained with different depths. Corresponding absolute error maps accompany each prediction to highlight patterns in the resolution.

Appendix C

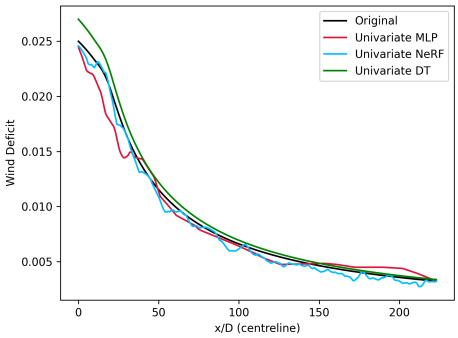
Additional Results Figures



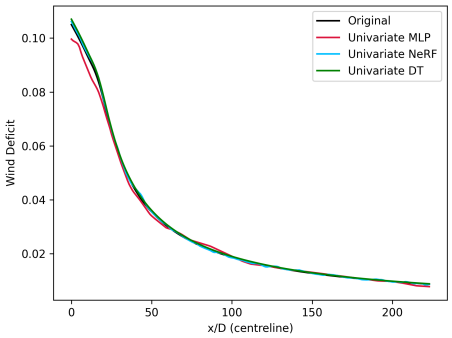
(a) Simulation: $TI = 0.11$, $C_T = 0.51$



(b) Simulation: $TI = 0.36$, $C_T = 0.22$

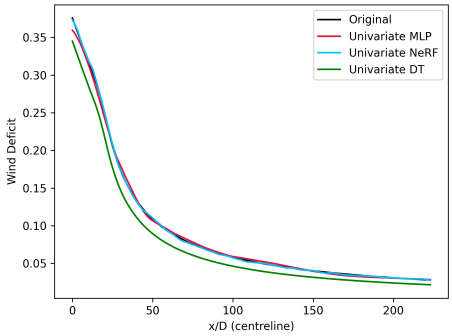


(c) Simulation: $TI = 0.68$, $C_T = 0.24$

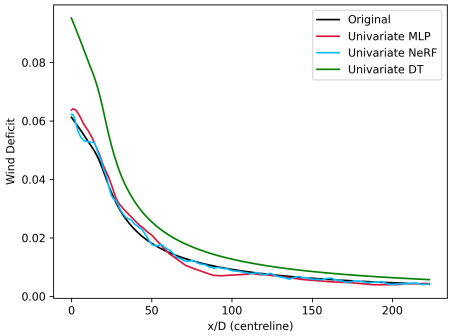


(d) Simulation: $TI = 0.85$, $C_T = 0.64$

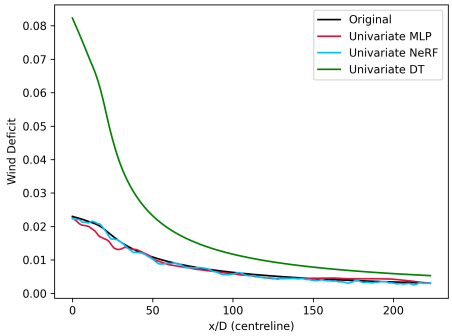
Figure C.1: Visual comparison of the centreline predictions by different surrogate models trained in standard interpolation and tested on four different simulation.



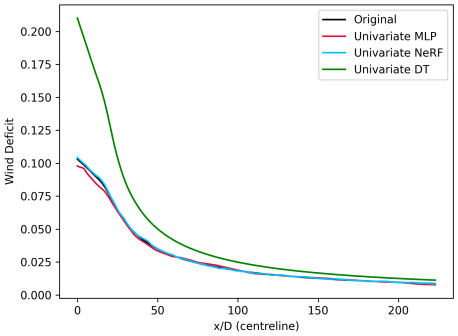
(a) Simulation: $TI = 0.11$, $C_T = 0.51$



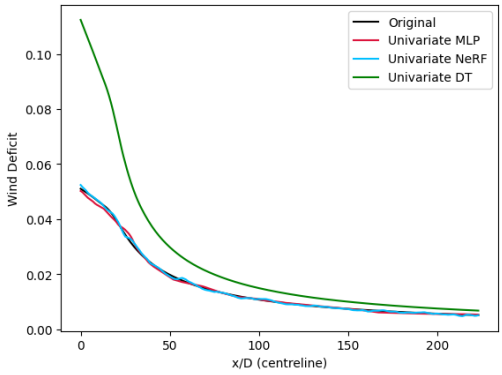
(b) Simulation: $TI = 0.36$, $C_T = 0.22$



(c) Simulation: $TI = 0.68$, $C_T = 0.24$

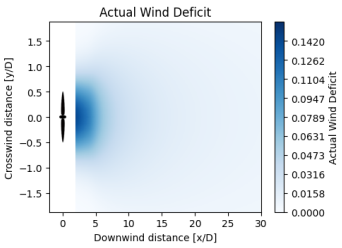


(d) Simulation: $TI = 0.85$, $C_T = 0.64$

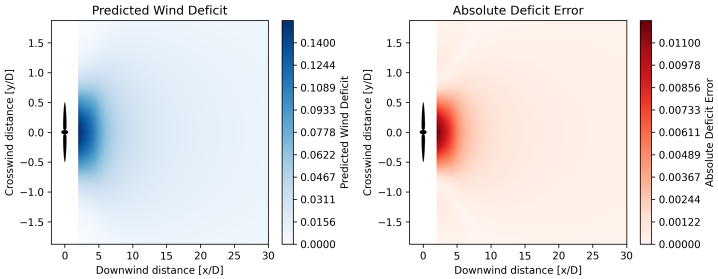


(e) Simulation: $TI = 0.79$, $C_T = 0.37$

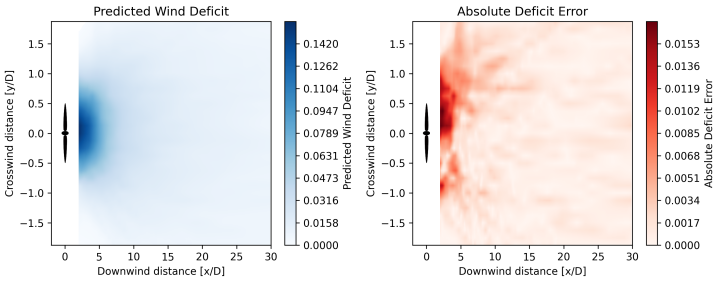
Figure C.2: As for Figure C.1, but the models are trained in the extrapolation setting. Moreover, the simulation (e) is added to be compared with Figure 5.8.



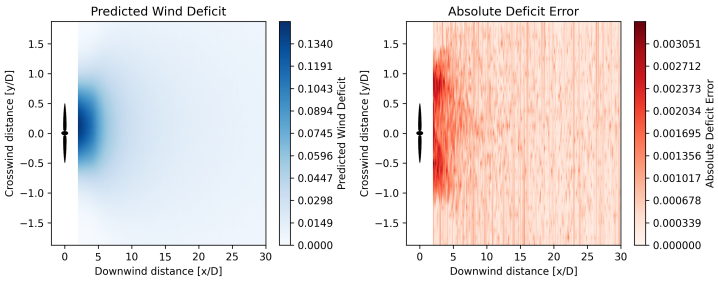
(a) Common original field



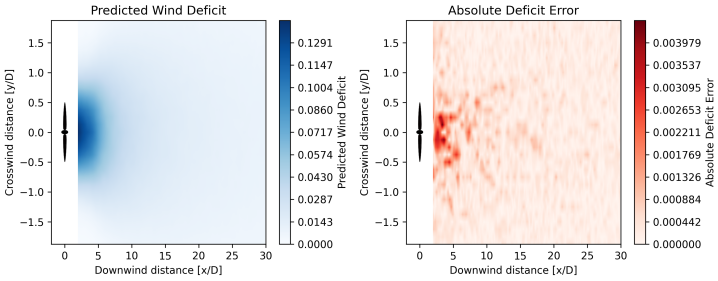
(b) Univariate RDT



(c) Univariate MLP

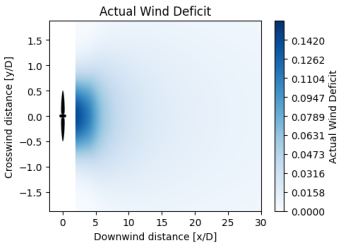


(d) Multivariate MLP

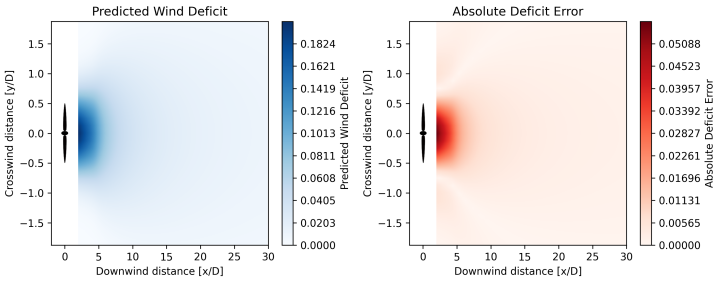


(e) Univariate NeRF

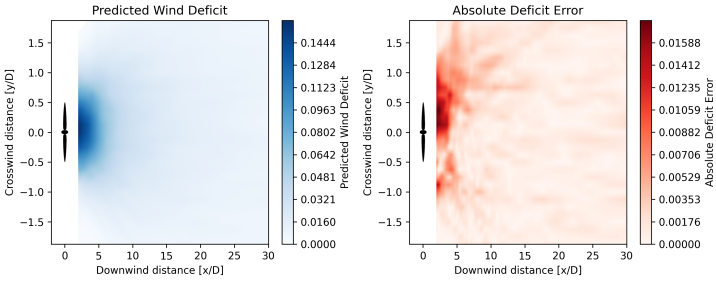
Figure C.3: As for Figure 5.7 but for the simulation with $TI = 0.45, C_T = 0.62$. The models are trained in the standard interpolation setting ($r_{TI} = r_{C_T} = 4$).



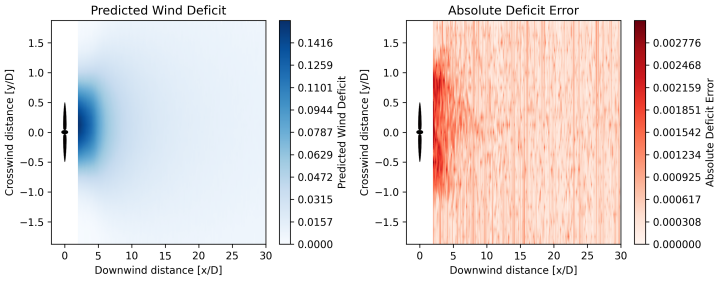
(a) Common original field



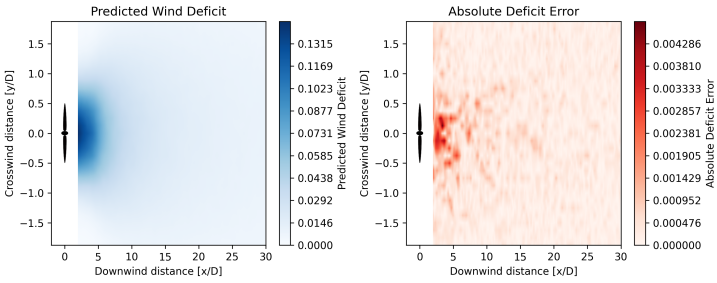
(b) Univariate RDT



(c) Univariate MLP



(d) Multivariate MLP



(e) Univariate NeRF

Figure C.4: As for Figure 5.9 but for the simulation with $TI = 0.45, C_T = 0.62$. The models are trained in the standard extrapolation setting ($I_{TI} = [0.15, 0.4]$ and $I_{C_T} = [0.3, 0.7]$).

Bibliography

[1] Global Wind Energy Council (GWEC). Global Wind Report 2022. Brussels, Belgium: GWEC, 2022. URL: https://gwec.net/global-wind-report-2022/.

[2] REN21 (Renewable Energy Policy Network for the 21st Century). Renewables 2021 Global Status Report. 2021. URL: https://www.ren21.net/gsr-2021.

[3] IEA (International Energy Agency). World Energy Outlook 2022. https://www.iea.org/reports/world-energy-outlook-2022. License: CC BY 4.0 (report); CC BY NC SA 4.0 (Annex A). Paris, 2022.

[4] JF Ainslie. "Development of an eddy viscosity model for wind turbine wakes". In: 7th BWEA Wind Energy Conference, Oxford, UK. 1985, pp. 61–66.

[5] John F Ainslie. "Calculating the flowfield in the wake of wind turbines". In: Journal of wind engineering and Industrial Aerodynamics 27.1-3 (1988), pp. 213–224.

[6] Giancarlo Alfonsi. "Reynolds-averaged Navier–Stokes equations for turbulence modeling". In: Applied Mechanics Reviews 62.4 (2009).

[7] Mike Anderson. Simplified solution to the eddy-viscosity wake model. Tech. rep. 01327-000202. Renewable Energy Systems Ltd., July 10, 2009.

[8] Cristina L Archer et al. "Review and evaluation of wake loss models for wind energy applications". In: Applied Energy 226 (2018), pp. 1187–1207.

[9] S Ashwin Renganathan et al. "Data-driven wind turbine wake modeling via probabilistic machine learning". In: Neural Computing and Applications (2022), pp. 1–16.

[10] R. J. Barthelmie et al. "Quantifying the Impact of Wind Turbine Wakes on Power Output at Offshore Wind Farms". In: Journal of Atmospheric and Oceanic Technology 27.8 (2010), pp. 1302–1317. DOI: https://doi.org/10.1175/2010JTECHA1398.1. URL: https://journals.ametsoc.org/view/journals/atot/27/8/2010jtecha1398_1.xml.

[11] R.J. Barthelmie et al. "Modelling and measuring flow and wind turbine wakes in large wind farms offshore". In: Wind Energy 12 (2009), pp. 431–444. DOI: 10.1002/we.348.

[12] Ronen Basri et al. "Frequency bias in neural networks for input of non-uniform density". In: International Conference on Machine Learning. PMLR. 2020, pp. 685–694.

[13] Majid Bastankhah and Fernando Porté-Agel. "A new analytical model for wind-turbine wakes". In: Renewable Energy 70 (2014). Special issue on aerodynamics of offshore wind energy systems and wakes, pp. 116–123. ISSN: 0960-1481. DOI: https://doi.org/10.1016/j.renene.2014.01.002. URL: https://www.sciencedirect.com/science/article/pii/S0960148114000317.

[14] Saakaar Bhatnagar et al. "Prediction of aerodynamic flow fields using convolutional neural networks". In: Computational Mechanics 64 (2019), pp. 525–545.

[15] Atharv Bhosekar and Marianthi Ierapetritou. "Advances in surrogate based modeling, feasibility analysis, and optimization: A review". In: Computers & Chemical Engineering 108 (2018), pp. 250–267.

[16] Laurens Bliek et al. "Online optimization with costly and noisy measurements using random Fourier expansions". In: IEEE transactions on neural networks and learning systems 29.1 (2016), pp. 167–182.

[17] David Bock and XSEDE. Visualization of Flow in Wind-Farms. Visualization created as part of XSEDE's Extended Collaborative Support Services. Image source: National Center for Supercomputing Applications.

[18] Leo Breiman. "Random forests". In: Machine learning 45 (2001), pp. 5–32.

[19] Leo Breiman et al. Classification and regression trees. CRC press, 1984.

[20] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. ACM, Aug. 2016. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145%2F2939672.2939785.

[21] D. Conti, N. Dimitrov, and A. Peña. "Aeroelastic load validation in wake conditions using nacelle-mounted lidar measurements". In: Wind Energy Science 5.3 (2020), pp. 1129–1154. DOI: 10.5194/wes-5-1129-2020. URL: https://wes.copernicus.org/articles/5/1129/2020/.

[22] TJ Craft, BE Launder, and K Suga. "Development and application of a cubic eddy-viscosity model of turbulence". In: International Journal of Heat and Fluid Flow 17.2 (1996), pp. 108–115.

[23] Kimberly E Diamond and Ellen J Crivella. "Wind turbine wakes, wake effect impacts, and wind leases: using solar access laws as the model for capitalizing on wind rights during the evolution of wind policy standards". In: Duke Envtl. L. & Pol'y F. 22 (2011), p. 195.

[24] A Duckworth and RJ Barthelmie. "Investigation and validation of wind turbine wake models". In: Wind Engineering 32.5 (2008), pp. 459–475.

[25] Ju Feng and Wen Zhong Shen. "Solving the wind farm layout optimization problem using random search algorithm". In: Renewable Energy 78 (2015), pp. 182–192. ISSN: 0960-1481. DOI: https://doi.org/10.1016/j.renene.2015.01.005. URL: https://www.sciencedirect.com/science/article/pii/S0960148115000129.

[26] Sten Frandsen et al. "Analytical modelling of wind speed deficit in large wind farms". In: Wind Energy 9 (Apr. 2006), pp. 39–53. DOI: 10.1002/we.189.

[27] David A Freedman. Statistical models: theory and practice. Cambridge University Press, 2009.

[28] Xiaoxia Gao, Hongxing Yang, and Lin Lu. "Optimization of wind turbine layout position in a wind farm using a newly-developed two-dimensional wake model". In: Applied Energy 174 (2016), pp. 192–200. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2016.04.098. URL: https://www.sciencedirect.com/science/article/pii/S0306261916305633.

[29] Pierre Geurts, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees". In: Machine learning 63 (2006), pp. 3–42.

[30] Harry Hochstadt. Differential equations: A modern approach. Courier Corporation, 2014.

[31] IPCC. Global Warming of 1.5°C. Tech. rep. Geneva, Switzerland: IPCC, 2018.

[32] Tomas Jansson, Larsgunnar Nilsson, and Marcus Redhe. "Using surrogate models and response surfaces in structural optimization–with application to crashworthiness design and sheet metal forming". In: Structural and Multidisciplinary Optimization 25 (2003), pp. 129–140.

[33] Farah Japar et al. "Estimating the wake losses in large wind farms: A machine learning approach". In: ISGT 2014. IEEE. 2014, pp. 1–5.

[34] N.O. Jensen. A note on wind generator interaction. English. Risø-M 2411. Risø National Laboratory, 1983. ISBN: 87-550-0971-9.

[35] A Jimenez et al. "Advances in large-eddy simulation of a wind turbine wake". In: Journal of Physics: Conference Series 75.1 (July 2007), p. 012041. DOI: 10.1088/1742-6596/75/1/012041. URL: https://dx.doi.org/10.1088/1742-6596/75/1/012041.

[36] Yaochu Jin. "Surrogate-assisted evolutionary computation: Recent advances and future challenges". In: Swarm and Evolutionary Computation 1.2 (2011), pp. 61–70.

[37] I.F.S.A. Kabir and E.Y.K. Ng. "Effect of different atmospheric boundary layers on the wake characteristics of NREL phase VI wind turbine". In: Energy 130 (2019), pp. 1185–1197. DOI: 10.1016/j.renene.2018.08.083.

[38] I. Katic, J. Højstrup, and N.O. Jensen. "A Simple Model for Cluster Efficiency". English. In: EWEC'86. Proceedings. Vol. 1. Ed. by W. Palz and E. Sesto. European Wind Energy Association Conference and Exhibition, EWEC '86 ; Conference date: 06-10-1986 Through 08-10-1986. A. Raguzzi, 1987, pp. 407–410.

[39] Adam R Kosiorek et al. "NeRF-VAE: A Geometry Aware 3D Scene Generative Model". In: Proceedings of the 38th International Conference on Machine Learning. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 5742–5752. URL: https://proceedings.mlr.press/v139/kosiorek21a.html.

[40] Jakub Kudela and Radomil Matousek. "Recent advances and applications of surrogate models for finite element method computations: A review". In: Soft Computing 26.24 (2022), pp. 13709–13733.

[41] James F. Manwell, Jon G. McGowan, and Anthony L. Rogers. Wind Energy Explained: Theory, Design and Application. 2nd. John Wiley & Sons, 2009, p. 704. ISBN: 978-0-470-01500-1.

[42] Ricardo Martin-Brualla et al. "NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections". In: (2021). arXiv: 2008.02268 [cs.CV].

[43] Ben Mildenhall et al. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". In: (2020).

[44] Patrick J Moriarty and A Craig Hansen. AeroDyn theory manual. Tech. rep. National Renewable Energy Lab., Golden, CO (US), 2005.

[45] Guo Nai-Zhi, Zhang Ming-Ming, and Li Bo. "A data-driven analytical model for wind turbine wakes using machine learning method". In: Energy Conversion and Management 252 (2022), p. 115130.

[46] Amin Niayifar and Fernando Porté-Agel. "Analytical modeling of wind farms: A new approach for power prediction". In: Energies 9.9 (2016), p. 741.

[47] NREL. "OpenFAST: Open-Source Wind Turbine Simulation Code". In: NREL Software and Data Catalog (2021). Accessed: June 2021.

[48] Jonathan J Oliver and David J Hand. "On pruning and averaging decision trees". In: Machine Learning Proceedings 1995. Elsevier, 1995, pp. 430–437.

[49] OpenCFD Limited. OpenFOAM: The Open Source CFD Toolbox. https://www.openfoam.com. Accessed: June 2023. ESI Group, 2021.

[50] Leandro Parada et al. "Wind farm layout optimization using a Gaussian-based wake model". In: Renewable Energy 107.C (2017), pp. 531–541. DOI: 10.1016/j.renene.2017.02.. URL: https://ideas.repec.org/a/eee/renene/v107y2017icp531-541.html.

[51] Mads M. Pedersen et al. "PyWake 2.5.0: An open-source wind farm simulation tool". In: (Feb. 2023). URL: https://gitlab.windenergy.dtu.dk/TOPFARM/PyWake.

[52] Fernando Porté-Agel et al. "Large-eddy simulation of atmospheric boundary layer flow through wind turbines and wind farms". In: Journal of Wind Engineering and Industrial Aerodynamics 99.4 (2011). The Fifth International Symposium on Computational Wind Engineering, pp. 154–168. ISSN: 0167-6105. DOI: https://doi.org/10.1016/j.jweia.2011.01.011. URL: https://www.sciencedirect.com/science/article/pii/S0167610511000134.

[53] Shantanu Purohit, Eddie Yin Kwee Ng, and Ijaz Fazil Syed Ahmed Kabir. "Evaluation of three potential machine learning algorithms for predicting the velocity and turbulence intensity of a wind turbine wake". In: Renewable Energy 184 (2022), pp. 405–420.

[54] Narendra Kurnia Putra et al. "Multiobjective design optimization of stent geometry with wall deformation for triangular and rectangular struts". In: Medical & Biological Engineering & Computing 57 (2019), pp. 15–26.

[55] Joaquin Quinonero-Candela and Carl Edward Rasmussen. "A unifying view of sparse approximate Gaussian process regression". In: The Journal of Machine Learning Research 6 (2005), pp. 1939–1959.

[56] Nasim Rahaman et al. "On the spectral bias of neural networks". In: International Conference on Machine Learning. PMLR. 2019, pp. 5301–5310.

[57] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". In: Advances in neural information processing systems 20 (2007).

[58] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". In: Advances in neural information processing systems 20 (2007).

[59] Ali Rahimi and Benjamin Recht. "Uniform approximation of functions with random bases". In: 2008 46th annual allerton conference on communication, control, and computing. IEEE. 2008, pp. 555–561.

[60] Carl Edward Rasmussen, Christopher KI Williams, et al. Gaussian processes for machine learning. Vol. 1. Springer, 2006.

[61] Vaishali Sohoni, SC Gupta, and RK Nema. "A critical review on wind turbine power curve modelling techniques and their applications in wind based energy systems". In: Journal of Energy 2016 (2016).

[62] Dan Steinberg and Phillip Colla. "CART: classification and regression trees". In: The top ten algorithms in data mining 9 (2009), p. 179.

[63] N Stergiannis et al. "CFD modelling approaches against single wind turbine wake measurements using RANS". In: Journal of Physics: Conference Series 753.3 (Sept. 2016), p. 032062. DOI: 10.1088/1742-6596/753/3/032062. URL: https://dx.doi.org/10.1088/1742-6596/753/3/032062.

[64] Richard J.A.M. Stevens and Charles Meneveau. "Flow Structure and Turbulence in Wind Farms". In: Annual Review of Fluid Mechanics 49.1 (2017), pp. 311–339. DOI: 10.1146/annurev-fluid-010816-060206. URL: https://doi.org/10.1146/annurev-fluid-010816-060206.

[65] Chaoli Sun et al. "A fitness approximation assisted competitive swarm optimizer for large scale expensive optimization problems". In: Memetic Computing 10 (2018), pp. 123–134.

[66] Matthew Tancik et al. "Fourier features let networks learn high frequency functions in low dimensional domains". In: Advances in Neural Information Processing Systems 33 (2020), pp. 7537–7547.

[67] Roger Temam. Navier-Stokes equations: theory and numerical analysis. Vol. 343. American Mathematical Soc., 2001.

[68] K. Thomsen and P. Sørensen. "Fatigue loads for wind turbines operating in wakes". In: Journal of Wind Engineering and Industrial Aerodynamics 80 (1999), pp. 121–136.

[69] Zilong Ti, Xiao Wei Deng, and Hongxing Yang. "Wake modeling of wind turbines using machine learning". In: Applied Energy 257 (2020), p. 114025.

[70] Zilong Ti, Xiao Wei Deng, and Mingming Zhang. "Artificial Neural Networks based wake model for power prediction of wind farm". In: Renewable Energy 172 (2021), pp. 618–631.

[71] Linlin Tian et al. "Development and validation of a new two-dimensional wake model for wind turbine wakes". In: Journal of Wind Engineering and Industrial Aerodynamics 137 (2015), pp. 90–99. ISSN: 0167-6105. DOI: https://doi.org/10.1016/j.jweia.2014.12.001. URL: https://www.sciencedirect.com/science/article/pii/S0167610514002505.

[72] S.D.O. Turner et al. "A new mathematical programming approach to optimize wind farm layouts". In: Renewable Energy 63 (2014), pp. 674–680. ISSN: 0960-1481. DOI: https://doi.org/10.1016/j.renene.2013.10.023. URL: https://www.sciencedirect.com/science/article/pii/S0960148113005545.

[73] L.J. Vermeer, J.N. Sørensen, and A. Crespo. "Wind turbine wake aerodynamics". In: Progress in Aerospace Sciences 39 (2003), pp. 467–510. DOI: 10.1016/S0376-0421(03)00078-2.

[74] Paul Westermann and Ralph Evins. "Surrogate modelling for sustainable building design–A review". In: Energy and Buildings 198 (2019), pp. 170–186.

[75] Brett Wilson, Sarah Wakes, and Michael Mayo. "Surrogate modeling a computational fluid dynamics-based wind turbine wake simulation using machine learning". In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE. 2017, pp. 1–8.

[76] Yu-Ting Wu and Fernando Porté-Agel. "Atmospheric Turbulence Effects on Wind-Turbine Wakes: An LES Study". In: Energies 5.12 (2012), pp. 5340–5362. ISSN: 1996-1073. URL: https://www.mdpi.com/1996-1073/5/12/5340.

[77] Cheng Yan et al. "Surrogate-based optimization with improved support vector regression for non-circular vent hole on aero-engine turbine disk". In: Aerospace Science and Technology 96 (2020), p. 105332.

[78] Aston Zhang et al. Dive into deep learning. 2021.

[79] Jincheng Zhang and Xiaowei Zhao. "A novel dynamic wind farm wake model based on deep learning". In: Applied Energy 277 (2020), p. 115552.

[80] Jincheng Zhang and Xiaowei Zhao. "Wind farm wake modeling based on deep convolutional conditional generative adversarial network". In: Energy 238 (2022), p. 121747.