Detection of cup anemometer anomalies using machine learning

Y. C. Heijne

TU Delft

Cover page: Snapshot of the Test Centre for Large Wind Turbines at Østerild
Credit: DTU Wind Energy
URL: `https://www.youtube.com/watch?v=g7NZ_COJ4uE&feature=youtu.be` - *Accessed on 29-07-2020*

Detection of cup anemometer anomalies using machine learning

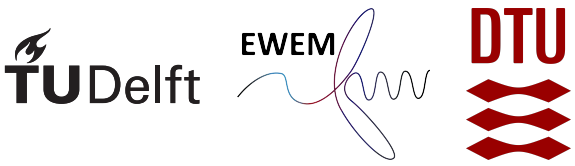
by

Y. C. Heijne

to obtain the degree of:
Master of Science in Aerospace Engineering at Delft University of Technology (TUD)
Master of Science in Engineering (European Wind Energy) at Technical University of Denmark (DTU)

Student number:	4389476 (TUD); s183039 (DTU)		
Project duration:	November 1, 2019 – July 31, 2020		
Supervisors:	Simon Watson	Professor	Delft University of Technology
	Michael Courtney	Senior Researcher	Technical University of Denmark
	Elliot Simon	Senior R&D Engineer	Technical University of Denmark
	John Thomas Lyons	Development Engineer	Technical University of Denmark

An electronic version of this thesis is available at `http://repository.tudelft.nl/` or `http://findit.dtu.dk`.



d

Preface

It is safe to say that these past nine months have been quite an adventure, considering I lived and worked on this thesis in three different countries. I would not have been able to complete this project on my own, as many people have helped me along the way.

First and foremost, I want to thank my project supervisors Michael Courtney and Simon Watson, from the DTU and TU Delft respectively. I am thankful for the knowledge and experience you shared with me during our meetings, as well as the useful feedback and encouragements that allowed me to truly make this project my own.

I also want to thank my daily supervisors Elliot Simon and John Thomas Lyons from the Test and Measurements (TEM) and Test and Calibration (TAC) sections of the DTU Wind Energy department. You both have provided me with the right help when I needed it and I enjoyed our interesting and fun conversations both related and unrelated to the project.

I want to extend my gratitude to the employees of TEM and TAC as well. You were all very helpful and kind in case I needed something and I wish I would've been able to enjoy more hygge Friday morning breakfasts.

A journey is nothing without great company, which is why I want to thank my fellow EWEM students and friends from all over Europe for the great experiences we've had. I want to thank my family, both in the Netherlands and in the United States, for providing me with unconditional support and all the monitors I need while working from home. Last but not least, I want to thank my beloved fiance Elea, without whose support this would not have been possible.

These are of course weird times and it's hard to close this chapter of my life without being able to properly celebrate this in person. Nevertheless I hope to see and talk to you all again soon to properly express my gratitude for helping me achieve this big milestone in life.

Wormer, the Netherlands, July 29, 2020

Y.C. Heijne

f

Summary

Past research has successfully increased the accuracy and quality of horizontal wind speed measurements made by cup anemometers. Errors are introduced in a variety of ways: when operational conditions deviate from calibration, due to flow distortion and as a result of the mechanical properties of an instrument changing during its lifetime. Specific investigations on the latter source of errors have resulted in means to detect such damages, e.g. icing and bearing degradation. These methods, however, often require impractical measurement set-ups. A data-driven approach using expert knowledge of cup anemometer performance that uses two operational anemometers at the same altitude is suggested as a novel means of anomaly detection. The proposed anomaly detection model uses regression to predict the next residual between both instruments. Evaluation between measurements and predictions through confidence intervals reveal whether or not a flag is placed on the measurements.

Data is obtained at the Light Mast North (LMN) in Østerild National Test Centre for Large Wind Turbines, Denmark. Wind speed measurements are taken at 10Hz from 2016-02-10 15:00 (UTC+2) until 2020-01-29 00:00 (UTC+2). The training set, used to develop the regression models, is comprised of normal data. The collocated cups show high correlation after designated filters, at both a 10-min and 10Hz scale, have removed the errors described above. Measurements of wind direction are required as well, since the residual between both instruments is dependent on where the wind is coming from. Directions where either cup is downstream of the mast are removed as well, since correlation becomes lower in these regions.

Three regression models of increasing complexity are trained on a subset of data. A comparison is made between a Stochastic Gradient Descent (SGD), ARIMA and MultiLayer Perceptron (MLP) model. The MLP performed best in terms of false positive rate and mean squared error on a test set due to its adaptive confidence intervals and ability to model non-linearities. The trained MLP model is employed on regular time series and validation sets that contain periods of icing and structural damage. The former reveals that the number of flags is sensitive and proportional to the wind speed. At high wind speeds, the limitation of the model become clear, as it reflected a large rise in supposed anomalies. Large prediction uncertainties made it harder to detect failure at lower wind speeds, but sharp increases in the number of flags right before and after cup anemometers froze over were observed. Periods where structural damage affected the measurements stood out from normal behaviour, in terms of density of anomalies and extremeness. The promising capabilities of the MLP model are evident in that faulty events were labelled correctly, but its dependency on the magnitude of wind speed under normal operation indicates that sound user judgement is still required to interpret whether or not certain data can be deemed anomalous.

h

Contents

CHAPTER 1

Introduction

Accurate estimations of the wind resource at potential sites are of great importance. The wind energy market is bigger than ever, and it is expected to continue to grow in the coming years (GWEC, 2019), meaning that wind energy production is increasing rapidly. The Annual Energy Production (AEP), which results in revenue, yielded by a wind turbine is directly related to the available wind resource. Highly accurate wind speed and direction measurements are therefore both valuable and desirable for siting engineers. Recent developments within metrology have provided new techniques, such as ground- or nacelle-based lidar and spinner anemometers, to measure wind speed and direction for power performance purposes. The most popular and common measurement instrument, however, remains the cup anemometer. According to Pindado et al. (2011), the Measuring Network of Wind Energy Institutes (MEASNET) was responsible for several thousands of anemometer calibrations between 1998 and 2007 due to an increasing demand from the industry.

1.1 Problem statement

The cup anemometer still plays a vital role in wind speed measurements and is used extensively on met masts in the field. Cup anemometers are a popular choice because they are relatively inexpensive and require little to no power. Improving the quality of cup measurements is therefore of great importance within both the academic world and the wind energy industry. Cup anemometers do not measure wind speed directly, but rather use their rotational velocity to quantify the incoming wind. Although the aerodynamics involved are rather complex, the relationship between the wind speed and rotational velocity of the cups turns out to be linear in the domain of interest (Kristensen, 1998). This means that cups can be calibrated in a wind tunnel, where two constants describing this linear relationship (a gain and an offset) are determined.

However, once a cup anemometer is operational in the field, these determined calibration coefficients may give incorrect measurements. The calibration coefficients are only valid for the exact conditions in the wind tunnel during the calibration process, which is rarely the case during an anemometer's operation. Even exact conditions in different wind tunnels can lead to different calibration results. Deviations from wind tunnel conditions in anemometer measurements can be grouped into three distinct types of errors: operational errors that arise due to this change of conditions, mounting errors that arise due to flow distortion from the masts and booms that the instruments are mounted on and mechanical errors, which originate from a change in the physical properties of the anemometer. The first two error sources are accounted for by means of a quantified uncertainty. At the moment, manual filtering is still standard procedure

for removing faulty data points originating from the third error source. Even experts, however, can not remove every data point that corresponds to a measurement error. An automated process for filtering anomalous data would result in a higher quality of wind speed measurements. Furthermore, current data filtering practices (Bailey et al., 1997) are often very wasteful. For example, data entries in which a temperature of 2 degrees Celsius or lower is encountered are discarded such that potential icing events, which change the aerodynamic properties of the cups, are excluded. Because of the increases in knowledge of machine learning applications, methods to train models that can evaluate and flag invalid data persistently without any human interference and without having to throw away large amounts of correct measurements are now possible.

1.2 State-of-the-art techniques

The literature review as presented in the following section aims to define the state-of-the-art regarding quantification of cup anemometer performance and anomaly detection on cup anemometer measurements. Research regarding this subject either aims to reduce the prescribed uncertainty or create automated models that filter cup data more accurately, thus increasing the quality of measurements as well as the amount of valid data. Setting the state-of-the-art reveals how this work could add to the body of scientific knowledge.

One of the first applications of automated anomaly detection on wind speed measurements from anemometers was introduced by Hill and Minsker (2010). They trained four regression models, a naive predictor, nearest cluster predictor, single-layer linear network and an Artificial Neural Network (ANN), on the wind speed data stream generated at a frequency of 1Hz from a propeller anemometer. The last 30 measurements in a time series were used to predict the next point. New measurements were compared to this prediction and flagged when the measurement fell outside of a statistically based confidence interval centered around the predicted value. The linear and neural network proved to be most successful, where the latter was able to detect anomalous data with only 1% of false positives (incorrectly labelled as anomalous) and 2% of false negatives (not labelled as anomalous when they actually were) when a confidence interval of 95% was used. ANNs have also been found to be successful in reducing introduced errors from turbulence, and thus operational uncertainty (Bégin-Drolet et al., 2013). The developed ANN was able to reduce errors on wind speed measurements with 0.18% to 0.53% on test data, which had an overall error of 1%, using a small subsequence of the time series around the point of interest.

More models were developed as part of the PHM 2011 Data Challenge Competition, which revolved around detecting cup anemometer degradation in test sets containing five days worth of 10-minute averages of paired cup anemometer data located at either the same or different altitudes as well as wind direction and temperature data. One of the recommendations of Hill and Minsker (2010) included the use of additional sensors to improve the accuracy of anomaly detection models. Siegel and Lee (2011) won the competition with their model that used K-means clustering on the residual between the paired anemometers. Their model was able to detect clusters of negative differences

between the two measured wind speeds, meaning that one of the anemometers was in a degraded state. The runner-ups attempted to detect deterioration through pattern identification in the residual as a function of wind speed (Sun et al., 2012).

Paired cup data can also be used to assess and correct ageing effects. Azorin-Molina et al. (2018) showed that anemometer degradation resulted in biases that were dependent on both the wind direction and time-scale, conforming with the results from the the PHM Data competition. Two regression models were developed by Azorin-Molina et al. (2018). The first regression model is based on predictions using correlated measurements in time and calm/strong wind conditions, whereas the second model uses harmonic terms related to different time scales and wind direction. Both these models were able to mimic the ageing effects on this specific type of cup anemometer and reduced these biases. Icing events can also be detected using residuals between two anemometers. Wadham-Gagnon et al. (2015) and Swytink-Binnema et al. (2019) showed that these events could be flagged as measured wind speed differences between heated and unheated cups dropped below a calculated limit curve.

Most of the aforementioned models work inside the time domain. Anomaly detection methods have also been reported in the frequency domain (Vega et al., 2014). Time series of the rotational frequency can be decomposed using Fourier decomposition. Pindado et al. (2013) showed that the third harmonic can be used for anemometer performance analysis due to its high correlation with the rotational frequency. Damage to the cup anemometer might not always be obvious from the time series, as they can still show a perfectly linear calibration curve, but changes in aerodynamic behavior produce lower Fourier coefficients (Vega et al., 2014), making Fourier decomposition a suitable method for anomaly detection. Aerodynamic perturbations mostly present themselves in the first harmonic term, which has also been shown to be potentially relevant for detecting ageing effects in cup anemometers (Pindado et al., 2015).

1.3 Project scope

It can be concluded that effective methods to improve cup anemometer measurements have been developed, but that there is still much to be gained in this endeavour. A general method to flag anomalous cup anemometer measurements that does not focus on a specific type of error and uses multiple sensors does not yet exist on a 10Hz scale. The scope of this project revolves around the proof-of-concept of such a novel method that can detect these faulty measurements and anomalous events.

Machine learning techniques offer a useful approach to such a problem, as they are concerned with creating automated and generic processes with minimal to no contribution of human intervention (Herlau et al., 2019). Anomaly detection is a specific branch of machine learning applications that can be defined as the finding of data points or patterns that do not conform to normal or expected behaviour, according to Chandola et al. (2010). Two distinct types of machine learning algorithms can be considered for the anomaly detection context: supervised and unsupervised learning. The former uses a given output, be it a numerical value or label, whereas the latter aims to find patterns in

the data.

Supervised learning algorithms are concerned with learning a function $f(x)$ to generate output y. These algorithms learn from a training set x, which consists of a large number of objects that have one or more attributes together with the corresponding target values y. Supervised learning algorithms are able to solve two different types of problems: regression and classification. Regression problems revolve around the prediction of a continuous variable based on a set of attributes of x. A very simple example of a regression problem is creating a least-squares solution $f(x)$ based on a set of points x to predict y as accurately as possible. The second problem for which supervised learning can be used is classification. Here, the machine learning model assigns categorical class labels to data points. The main difference with regression problems is that the vector y can only take on distinct variables. Unsupervised learning algorithms do not need a solution y to train. They can thus be used to perform different tasks, such as clustering, density estimations and association mining. All these tasks revolve around characterizing patterns in the data set. Clustering algorithms aim to create groups of data objects with common attributes. Density estimations try to establish how likely future observations are, given the past observations and association mining techniques use data sets to establish prevalent association rules that hold within the given data set.

Classification techniques would be perfect for anomaly detection, but these models can only be trained if target labels are given for each object in the provided data set. Since labelled data is rarely available, classification is generally not suitable for anomaly detection. However, supervised learning remains a possibility. Predictions of data points can be made using regression techniques, which can be used to assign labels through set confidence intervals. Given the success regression techniques have in predicting wind speed signals in previous work (Azorin-Molina et al., 2018; Hill & Minsker, 2010), they are used as a starting point of the development of an anomaly detection model in this work. Gupta et al. (2013) provides a detailed survey of algorithms from all categories applied to anomaly detection problems. From a review of the works related to regression, a general outline of a suitable detection model can be created. This outline is shown in Figure 1.1. The flowchart shows the workings of Normal Behaviour Modelling (NBM). The data used for training heavily influences the predictions made by a regression model (Herlau et al., 2019). As this training data describes purely normal behaviour, the model should make more accurate predictions on unseen normal data when compared to unseen abnormal data. This discrepancy in accuracy can be used to label measurements based on comparison with the provided predictions.

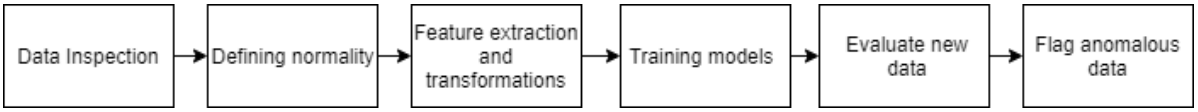


Figure 1.1: Flowchart describing the general anomaly detection model.

Within the project context, the outlined steps in Figure 1.1 are elaborated upon such that a proof-of-concept on an algorithm that can flag faulty cup anemometer measurements, particularly those that are operational on met masts in the field. The

raw data undergoes data inspection first, after which it will be filtered so that all remaining data conforms to the prescribed definition of normalcy. The definition of normalcy is created through a detailed literature study on cup anemometer dynamics and known measurement errors related to cup anemometers, further discussed in Chapter 2. Chapter 3 gives a description of the measurement site and used instruments, from which the measurements are subjected to filtering procedures to create a data set that conforms to the set definition of normalcy. First, errors present on a 10-minute scale are removed in Chapter 4 as means of top level inspection. Every data set inherently contains outliers, correct but unusual measurements, which are removed along with faulty data at this step. In Chapter 5, all remaining 10Hz data is loaded and this set's completeness is verified. This verification step is required due to the large size of the dataset, which contains multiple years' worth of data. This set is subjected to extra filters to ensure it adheres to the required definitions of normalcy. Other than normal data sets, validation data sets containing faulty data and means of validation are identified.

The regression models are elaborated upon in Chapter 6. First, dataset transformations will be applied to make the normal data suitable for the specific machine learning algorithms. Multiple prediction models of increasing complexity are introduced and tuned to create accurate predictions. Confidence limits are set on these predictions such that the models can be used as means to anomaly detection. A subset of the normal data is used for training of these models to limit the computational costs. All trained models are compared in Chapter 7 on a given set of criteria that describe the effectiveness and accuracy of the models, after which the best performing model is trained using all available data. The separated faulty sets are then used to evaluate the model performance in Chapter 8, which aims to create a deeper understanding of the proposed detection model, as well as conclusions on the proof-of-concept of NBM on cup anemometer measurements.

CHAPTER 2
The cup anemometer

The invention of the cup anemometer dates back to the 19th century, where T.R. Robinson used four rotating cups attached to a vertical axis to measure wind speed (Robinson, 1850). This remained the standard until Patterson (1926) proposed the concept most commonly used today: an anemometer using three cups. The major advantage of three cups compared to four is that the torque on the shaft becomes more uniform during a revolution of the anemometer. A comprehensive overview of early research on cup anemometers is given by Wyngaard (1981). Wyngaard provides a detailed summary of the many theories and models developed to describe cup anemometer dynamics as well as the results of experimentally determined lift and drag forces, which are responsible for making the cups rotate. The working principles of cup anemometers are described first in this section. This knowledge is essential to later identify the different mechanisms that can lead to measurement errors. Proposed uncertainties aim to quantify the effects of a number of these sources on measurements, allowing for a definition of anomalies that may be detected.

2.1 Cup anemometer working principles

Modern cup anemometers measure the horizontal wind speed indirectly through the rotational frequency of the anemometer. Their main component is the horizontal rotor, consisting of three cups spaced in an axisymmetric manner, i.e. 120° apart. The cups rotate around a vertical shaft that drives a signal generator. In a general Cartesian coordinate system, incoming wind with speed $\boldsymbol{U}_t(t) = \left[U_x(t), U_y(t), U_z(t)\right]^T$ induces a resultant aerodynamic force $\boldsymbol{F}$ in the plane of rotation. These aerodynamic forces are proportional to the the square of the horizontal velocity U, where $U(t) = \sqrt{U_x^2(t) + U_y^2(t)}$. For horizontal, unsteady wind speeds, the vector then becomes $\boldsymbol{U}_t(t) = \left[\overline{U} + u(t), v(t), w(t)\right]^T$. The lower case wind speed symbols u, v, w describe the three fluctuating, orthogonal components of wind speed, with u aligned with the mean flow direction. Ideally, cup anemometers should have a cosine response $\left(U(t) = |\boldsymbol{U}_t(t)| \cos\left(\frac{w(t)}{\overline{U}+u(t)}\right)\right)$ to non-zero values of w, since the cup anemometer only measures horizontal wind speed and thus the horizontal projection of the velocity. However, this is rarely the case, as different responses are encountered depending on the design of the anemometer (Papadopoulos et al., 2001). Responses of cup anemometers to flow inclination are further discussed in Section 2.2.2. The resultant force on each cup as exerted by the wind collectively results in an aerodynamic moment Q_A, causing the vertical shaft to accelerate. The net moment is slightly lower due to a small, counteracting moment caused by friction.

The cups accelerate the shaft until the net moment is zero, which occurs at a certain angular velocity ω. A schematic overview of these forces and resultant torque is provided in Figure 2.1.

The rotating shaft drives the output system, which converts the rotational velocity into an electric signal of pulses. The time between two consecutive pulses Δt_{pulses} or counting the number pulses N_{pulses} over a period of time T with N_0 pulses per revolution allows one to thus measure the rotational frequency of the cup, as shown in (2.1). The pulses can be generated with a magnet- or optoelectrics-based system, which produces 1-3 pulses or 6-44 pulses per rotation respectively (Pindado et al., 2011). Recent research also investigates the possibility of calculating the output frequency through a Fast Fourier Transform (FFT), but this was shown to perform worse than the aforementioned Counting Pulses (CP) (Ramos-Cenzano, Ogueta-Gutierrez, et al., 2019).

$$f(t) = \frac{N_{pulses}(t)}{N_0 T} = \frac{1}{\Delta t_{pulses}} = 2\pi\omega(t) \tag{2.1}$$

2.1.1 Calibration

The remaining necessary element for obtaining measurements of horizontal wind speed is the relation between wind speed and measured angular velocity. The equation of motion for a cup is given by (2.2), where Q_f is the frictional torque introduced by the bearings of the system. The frictional torque can be assumed to be neglected for wind speeds higher than 1m/s (Solov'Ev et al., 2004). The equation of motion can be rewritten in a generic differential form, as done in (2.3) (Kristensen, 1998), where $\Omega(t)$ represents the peripheral speed and is simply described by $\omega(t) r_{cup}$.

$$I\frac{\mathrm{d}\omega(t)}{\mathrm{d}t} = Q_A(t) + Q_f \tag{2.2}$$

$$\frac{\mathrm{d}\Omega(t)}{\mathrm{d}t} = F\left(\Omega(t), U(t), w(t)\right) \tag{2.3}$$

Assuming a steady state, where $U(t)$ and $\Omega(t)$ are constant, and purely horizontal wind ($w = 0$), equation (2.3) can be solved. Kristensen (1998) denotes that the result is a steady-state response curve that can be considered linear. His well-known result is shown in (2.4), where U_0 is the so-called offset speed and l_0 the calibration length. The offset speed is always positive and is indicative of when the rotor starts turning, meaning that Q_A is large enough to overcome Q_f. The calibration length can be seen as the length of the column of air that needs to pass for the rotor to turn 1rad and is indicative of the response of the rotational velocity to a change in horizontal wind speed. Equation (2.4) can be rewritten as a generic linear function, as shown in (2.5), such that the relationship between the rotational frequency and horizontal wind speed is established.

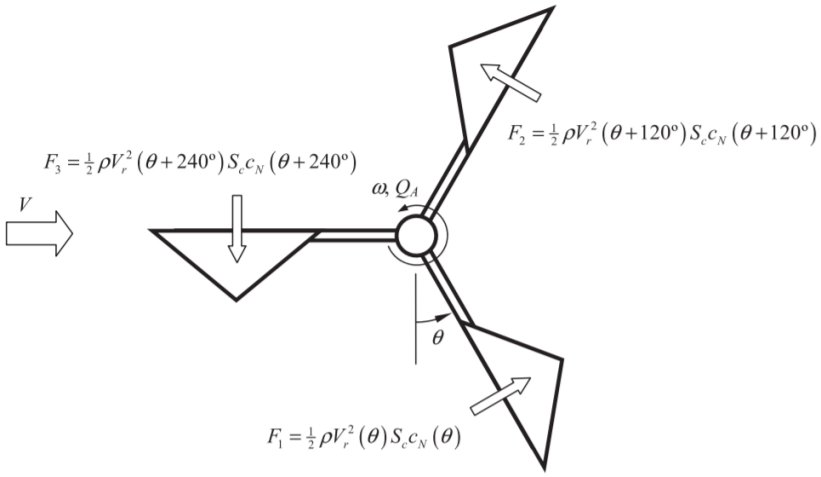


Figure 2.1: Schematic overview of a modern cup anemometer, where frictional torque is neglected (Pindado et al., 2015).

$$\Omega = \frac{U - U_0}{l} \tag{2.4}$$

$$U = 2\pi r_{cup} l f + U_0 = A \cdot f + B \tag{2.5}$$

In order to arrive at the resulting (2.5), the assumption of a steady state wind and no flow inclination was made, meaning that wind tunnels are suitable to determine gain A and offset B to perform calibration. Such assumptions can be made in a controlled environment. Accredited tunnels and calibrated pitot tubes, which allow the cup anemometer to be traceable, can be used to create calibration curves, where the known horizontal wind speed can be plotted as a function of measured rotational frequencies. A simple linear fit will provide A and B and make the cup anemometer suitable for use in the field. Note that only the scalar value of the horizontal wind speed can be obtained since the rotational frequency is a scalar value as well. The rotor is adirectional due to the equal spacing of the three cups.

As mentioned, this linearity only holds under the assumption that the contribution of Q_f to the resultant torque is negligible, which can be assumed for wind speeds above 1m/s. For smaller wind speeds, the magnitude of the frictional torque is not small in comparison to the aerodynamic torque and, as such, can no longer be neglected. The resulting loss of linearity is shown in Figure 2.2. Intuitively, it can be understood that significant effects of friction will result in a lower rotational velocity compared to the linear response. Calibration procedures, as provided by MEASNET (2009), have a minimum wind speed of 4m/s, where the loss of linearity is not an issue.

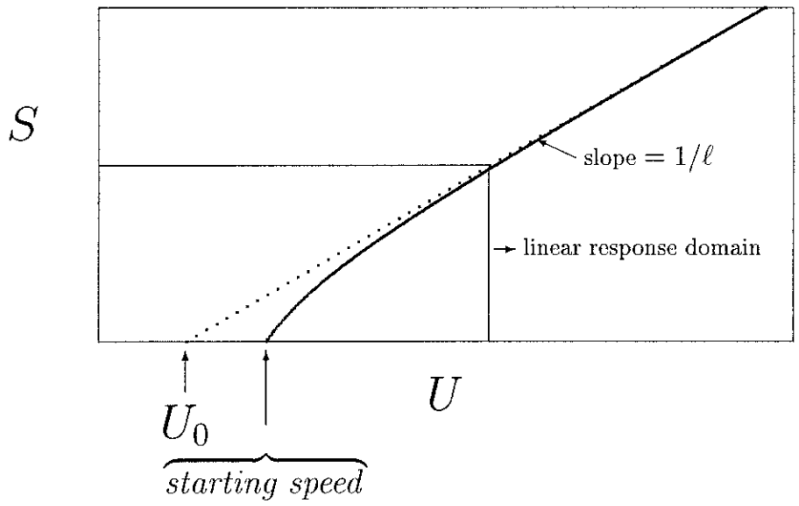


Figure 2.2: Actual response of rotational velocity, here denoted as S, compared to the linear response curve (Kristensen, 1998).

2.2 Measurement errors

The need for accurate measurements has led researchers to reveal the origins and quantify the effects of different error sources. A measurement error is said to exist when the measured horizontal wind speed is not representative of the true wind speed. The linear calibration curve was created under the assumption of steady-state horizontal flow in a wind tunnel with one set of ambient conditions. Conversely, operational cups on met masts will be subjected to turbulent wind fields, so this assumption no longer stands to reason. Aside from the turbulence-induced biases, the effects of flow inclination and ambient conditions can create more operational errors. Another type of error results from the mounting of a cup on a met mast. This error source does not originate from the cup anemometer itself, but comes from flow distortion induced by the intrusive met mast in the flow. The last type of error comes from changes in the physical properties of the cup anemometer. Icing and mechanical damage on the anemometer change its moment of inertia and aerodynamic properties, whereas ageing of the bearings results in higher frictional torque that changes the gain and offset compared to the initial calibration constants.

2.2.1 Turbulence

Several of the first and most researched measurement errors for cup anemometers are related to turbulent wind speeds. Cup anemometers have an asymmetric response to changes in wind speed, where the anemometer responds faster to increases in wind speeds compared to decreases (Busch & Kristensen, 1976). This means that a positive bias in mean wind speeds is introduced. Maccready (1966) assigns four different types of errors related to turbulent wind fields: the u-error, v-error, w-error and DP-error (Data

Processing).

The u-error, also referred to as overspeeding, is related to speed fluctuations in the mean direction of the incoming wind. Assuming no fluctuations in w (error related to fluctuations in this direction are discussed below), and noting that r_{cup} is constant, the generic equation of motion in (2.3) can be reduced to (2.6). Kristensen (1998) proposed a semi-empirical relationship for F as shown in (2.7) in a simple form. In (2.7), the non-negative constant β is introduced, as well as the distance constant l_0. This distance constant is an important parameter for the dynamic characteristics of a cup anemometer, as it describes the length of fluid that needs to pass the rotor for it to respond to 63.2% of a step change in wind speed (Manwell et al., 2002). (2.7) can be rewritten using (2.5), resulting in (2.8) (Bégin-Drolet et al., 2013). From this form, it can be determined how an increase in $U(t)$ leads to a faster response in U_{meas} compared to a decrease of the same magnitude. It can thus be concluded that rotors with low inertia have lower u-errors and that cup anemometers can be designed for low distance constants based on (2.8) and (2.2).

$$\frac{d\omega(t)}{dt} = F\left(\omega(t), u(t)\right) \tag{2.6}$$

$$F\left(\omega(t), U(t)\right) = \frac{\left[U(t) - U_0 - l\omega(t)\right]\left[U(t) - U_0 + \beta l\omega(t)\right]}{l_0 l(1+\beta)} \tag{2.7}$$

$$\frac{dU_{meas}(t)}{dt} = \frac{\left[U(t) - U_{meas}(t)\right]^2}{l_0(1+\beta)} + \frac{\left[U(t) - U_{meas}(t)\right]\left[U_{meas}(t) - U_0\right]}{l_0} \tag{2.8}$$

The v-error arises from fluctuations in wind directions. It should be noted that the v-error is not even a bias at all if one considers the scalar mean wind speed instead of the magnitude of the average wind velocity vector (Kristensen, 1998). In fact, Maccready (1966) defines Kristensen's v-error as the DP-error. The DP-error occurs because a scalar average is always higher than a vector average. An overestimation of speed arises when averaged cup measurements are directly compared to, for example, vector-averaged sonic anemometer measurements.

The w-error refers to fluctuations in the vertical direction, orthogonal to the horizontal plane. The w-error is manifested similarly to the DP-error described above. Instantaneous non-zero values of U_w cause an inclination angle ϕ of the incoming wind. A positive bias is introduced due to the fact that the scalar mean wind speed is taken over a scalar mean direction. Kristensen (1998) assigns one more error related to the w component: the stress bias. This error arises because cups do not have a cosine response. For this reason, the cup never measures the true horizontal wind speed. The size and polarity of the error depends on the exact nature of the deviation of the actual angular response from an ideal cosine. No stress bias would be introduced if an anemometer had a perfect cosine response.

2.2.2 Flow inclination

The effects of flow inclination can degrade measurement accuracy. As mentioned, cup anemometers do not have a cosine response and cause an increase in the w-error and stress bias. Follrichs and Glocker (1997) tested seven MEASNET accredited cup anemometers in a wind tunnel for different inclination errors. From their measured inclination responses, it became clear that cup anemometers have an asymmetric response that is generally more sensitive to inclinations compared to the standard cosine response. Flow inclinations can be present in the field due to complex terrain or improper installment of an anemometer. Papadopoulos et al. (2001) performed a similar test where measurements for cup anemometers from five different manufacturers were taken at different inclination angles. Their measurements were compared to a reference horizontal cup anemometer. An example of the resulting response curve is shown in Figure 2.3. The research concludes that errors related to flow inclination are significant and that the wind tunnel result can be indicative of the response, but operational anemometers will have smoother response curves.

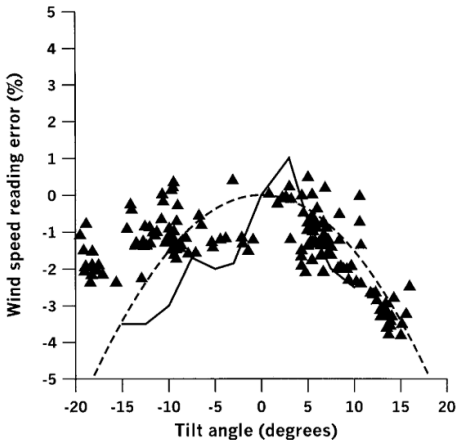


Figure 2.3: Relative percentage error of the Vector 100AK anemometer over the effective tilt angle. The solid line represents data from the wind tunnel. The dotted line represents the cosine response. Data from all wind directions and from two wind speed bins centered around 6 and 8m/s have been used (Papadopoulos et al., 2001).

2.2.3 Mounting errors

Cup anemometry is an intrusive measurement technique, meaning that cup anemometers need to be placed at the exact location where the wind speed measurement is required. This means that met masts are generally tall structures that allow the anemometer to measure at the desired altitudes. The tower and boom distort the flow, which affects the incoming direction and magnitude of the wind speed. This means that the measured wind speed can be higher or lower than the true free wind speed, but not as a result of the cups' poor performance. Measurement errors up to $\pm 2\%$ can be observed when the wind direction is in line with the anemometer and the tower (Lindelöw et al., 2010).

Conscientious placement and orientation of the mast and its booms can reduce this error significantly, as the mounting error depends on the direction of the incoming wind (see Figure 2.4), meaning that these errors are not always prevalent. Similar results were found by Baseer et al. (2016), where the ratio of the median from cup anemometers at the same altitude but opposite sides of a met mast were plotted as a function of wind speed. These studies indeed confirmed degradation in accuracy of similar magnitude in directions where either one of the instruments is downstream of the mast.

2.2.4 Ambient conditions

Ambient conditions alter the measurements as well. Changes in density also change the aerodynamic moment Q_A. Therefore, cups may rotate faster or slower compared to the calibrated conditions at the same free wind speed, resulting in measurement errors. Pindado, Sanz, et al. (2012) showed that the effects of density on wind speed measurement can not be neglected. A similar argument can be made for the effects of ambient temperature on the measurements. Decreases in temperature cause an increase in bearing friction Q_f due to increasing viscosity of bearing grease. Although the bearing friction is of less importance around normal calibration temperatures (15-20°) and higher wind speeds (> 5m/s), offset values for calibration can change significantly below freezing temperatures (Dahlberg et al., 2001).

2.2.5 Icing

Freezing temperatures may cause icing to occur on an anemometer, which can result in large measurement errors. The accumulation of ice causes the profile of the cups to change, causing an increase in inertia as well as changed aerodynamic coefficients and a different Q_A (Kimura et al., 2001). As a result, ice formation on cup anemometers can produce underestimated wind speeds. Cup anemometers can also completely freeze over, therefore registering the offset B as wind speed. Instruments equipped with heating devices have been developed to prevent icing, but the addition of heaters increases the rotor inertia as well, making the anemometer respond slower to changes in wind speed and thus perform worse for turbulent measurements.

2.2.6 Ageing of bearings

As cups are operational, bearings wear down over time and as a result increase the friction torque Q_f. Pindado, Barrero-Gil, et al. (2012) showed that anemometers reach stable performance after a transition period following the first calibration and that cup anemometers age differently over time regardless of their maintenance schedule or usage. Significant degradation in accuracy of the NRG #40 type anemometer was detected using a set of paired anemometers (Clark et al., 2009), but no distinct pattern for ageing can be defined. The accumulation of damage over time must also be considered. During a cup anemometer's lifetime, it will be subjected to not only normal wear, but also electrostatic events or collision with objects (e.g. falling ice) such that its inertia changes

(Pindado et al., 2014). These events may cause damage immediately or can accumulate over time and introduce a bias to the reported measurements. Similar errors may arise from dirt accumulation on the cups, which changes their aerodynamic profile and thus the aerodynamic properties of the instrument.

2.2.7 Output system

Lastly, errors can occur within the output system of the cup anemometer. Opto-electrical systems are said to create 6-44 pulses per rotation. Improper maintenance or contamination from the environment may cause one of these pulses to be missed. It was shown that a contaminated Thies Clima First-Class Advanced cup anemometer, equipped with a 37 pulse opto-electrical system, in which one slot was contaminated with an oil drop, experienced a significantly increased error when compared with the same but uncontaminated instrument (Ramos-Cenzano, Ogueta-Gutiérrez, et al., 2019). According to the CP method, the velocity residual between the two anemometers increased from $10^{-3.5}$ m/s to $10^{-1.5}$ m/s after contamination, regardless of the sampling rate.

2.3 Measurement uncertainties

Taking into account this broad range of errors, much work has been done to quantify the accuracy of cup anemometer measurements. It is important to know which error sources are already considered for uncertainties, such that the uncertainty can be reduced, or which unquantifiable effects can be addressed in the future to further increase cup anemometer performance. The relevant standard that is to be referred to for cup anemometer measurements is the IEC 61400-12-1:2017 (2017). The IEC 61400-12-1:2017 is used for "Power performance measurement of electricity producing wind turbines" and includes requirements for mast mounted anemometers. These requirements include general requirements for the calibration procedure as well as means to determine uncertainty. This uncertainty includes three components, namely contributions from the calibration procedure, mounting and its operational environment.

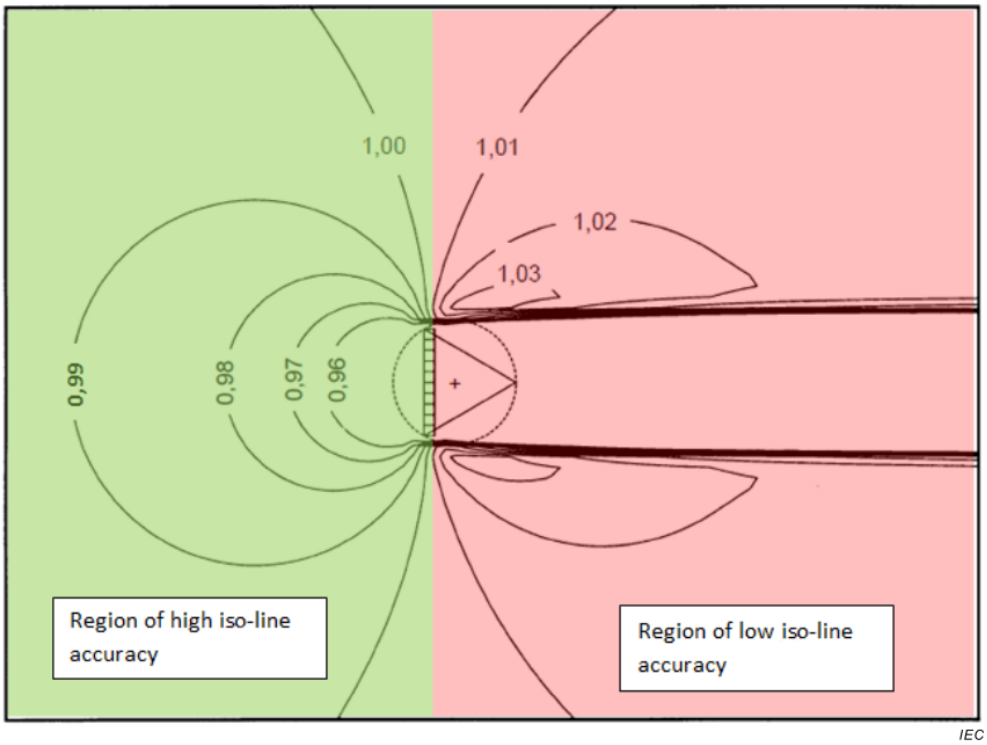
Annex F (IEC 61400-12-1:2017, 2017) describes the required calibration procedure for cup anemometers. It is necessary to calibrate the instruments in an accredited wind tunnel equipped with traceable instruments. Calibration is to be performed several times for wind speeds between 4 to 16m/s with 1m/s increments such that repeatability can be ensured. Guidelines have been established for calibrating cups, such as those done by MEASNET (2009), to improve the quality of calibration procedures in accordance to IEC 61400-12-1:2005 (the preceding version of the standard established in 2005 that did not include remote sensing devices). Annex F also describes how to determine the calibration uncertainty $u_{cal,1}$ with contributions from the wind tunnel environment and reference measurement devices, such as pitot tubes. The commonly used MEASNET guidelines also include one more component, $u_{cal,2}$, which accounts for the fact that wind tunnels do not perfectly agree with each other. Objects that are placed within a cross-section suffer from blockage effects, resulting from flow distortions due to the intrusive instrument and locally reduced cross sections. Potential flow models can be used to approximate

these effects, but no direct or analytical solution currently exists to correct for blockage. Round-robin calibrations of the same cup anemometer in different wind tunnels reveals results that differ by ±1%, which is acceptable within the MEASNET system. Assuming a uniform distribution, a standard uncertainty of $1\%/\sqrt{3}$ is commonly used to represent this spread.

Mounting can introduce large errors, as mentioned before. Flow distortion from the mast and its booms can cause the anemometer to measure wind speeds that are different from the true wind speed. Annex G (IEC 61400-12-1:2017, 2017) describes separation distances for top-mounted and boom-mounted anemometers that should keep these biases to a minimum. For instance, boom-mounted cups should be placed at least 20 boom diameters above the boom to keep the influence of the boom below 0.5%. Wind direction is also of great influence on the tower induction, as shown in Figure 2.4. It can easily be seen how the booms also need to separate the tower and the cup far away such that this induction is minimal. The standard uncertainty due to mounting u_{mount} is often taken as 1% for boom-mounted cups and 0.5% for a top-mounted cup anemometer, if all spacing guidelines in annex G are followed accurately.

Lastly, operational uncertainty u_{ope} needs to be quantified. Wind tunnels have a very controlled environment and do not account for variations in ambient conditions when calibration is performed to ensure repeatability. It has been shown that varying values of turbulence intensity, temperature, air density and flow inclination can lead to differing measurements of the same true wind speed. This final contribution of uncertainty will take into account the fact that these parameters affect the measured wind speeds. Annex I (IEC 61400-12-1:2017, 2017) deals with classification for cup anemometers, from which operational uncertainty can be obtained. Four different classes (A,B,C and D) are given, depending on the future applications for a type of cup anemometer. Class A is applicable for moderate climates with flat terrain, whereas class B is used for a similar climate but complex terrain. Class C also applies to flat terrain, but should be used for cold climates. Finally, cup anemometers that are classified for class D can be used for cup anemometers that will be used in complex terrain and in a cold climate.

Together with the class for which the cup anemometer can be used, a class number k is assigned to quantify u_{ope} through (2.9). This class number describes the systematic deviation from calibration due to influences from the environment. This systematic deviation requires many different calibration curves under many different ambient conditions, making modelling a more suitable solution. The model used in IEC 61400-12-1:2017 is the ACCUWIND model (Dahlberg et al., 2006). This model uses calibration data, e.g. torque curves and angular responses, to simulate cup anemometer responses under a set of given input parameters. Different 10 minute simulations covering the entire range of variables for the desired class, as shown in Figure 2.5, are performed and compared to the the obtained calibration curve. The maximum absolute deviation as a percentage of concurring wind speed will then be the class number k. This process is done for five different anemometers of the same type and these are combined to determine the overall classification of a certain type of cup anemometer. The total cup uncertainty is obtained by assuming all contributions are uncorrelated, meaning they can be combined using (2.10).



NOTE The speed is normalised by free-field wind speed (from the left); analysis by 2 dimensional Navier-Stokes computation and actuator disc theory for which the green shaded region on the left is considered accurate and the red shaded region on the right inaccurate due to inaccuracies modelling flow in and behind the separation zone.

Figure 2.4: Isolines of wind speed around a triangular lattice met mast with a C_T equal to 0.5 (IEC 61400-12-1:2017, 2017)

$$u_{ope} = \frac{k}{100}\frac{1}{\sqrt{3}}(0.5U + 0.5) \tag{2.9}$$

$$u_{tot} = \sqrt{u_{cal,1}^2 + u_{cal,2}^2 + u_{mount}^2 + u_{ope}^2} \tag{2.10}$$

$$\tag{2.11}$$

Uncertainty thus takes into account the effects of turbulence, flow inclination, mounting errors and ambient conditions. Reducing this kind of uncertainty deals with cup anemometer error estimation of undamaged cups, but does not take into account errors introduced due to a change of the physical properties of the instrument, e.g. icing, ageing and damage. Pindado et al. (2014) draws a similar conclusion that three issues that deteriorate cup anemometer measurement: 1. ageing of the bearings and mass addition on the cups, e.g. dirt or ice accumulation, changing the inertia of the rotor; 2. damage to the anemometer due to lightning, storms or hail; and 3. failure in the output signal system. Manual filtering is commonly applied to remove data entries where icing may be an issue, e.g. remove all entries where $T < 2°C$. Frequent maintenance, recalibration and

	Class A **Terrain meets requirements in Annex B**	**Class B** **Terrain does not meet requirements in Annex B**	**Class C** **Terrain meets requirements in Annex B**	**Class D** **Terrain does not meet requirements in Annex B**	**Class S**[34] **Special class with user defined ranges**
	Range	Range	Range	Range	Range
Wind speed V (m/s)	4 to 16	4 to 16	4 to 16	4 to 16	4 to 16
Turbulence intensity	0,03 to 0,12 + 0,48/V	0,03 to 0,12 + 0,96/V	0,03 to 0,12 + 0,48/V	0,03 to 0,12 + 0,96/V	User defined
Turbulence [35] structure $\sigma_u/\sigma_v/\sigma_w$	1/0,8/0,5*	1/0,8/0,5*	1/0,8/0,5*	1/0,8/0,5*	User defined or 1/0,8/0,5*
Air temperature (°C)	0 to 40	−10 to 40	–20 to 40	−20 to 40	User defined
Air density (kg/m^3)	0,9 to 1,35	0,9 to 1,35	0,9 to 1,35	0,9 to 1,35	User defined
Average upflow angle (°)	−3 to 3	−15 to 15	−3 to 3	−15 to 15	User defined
Wind direction (°)[36]	Cups and sonics: 0° to 360°	Cups and sonics: 0° to 360°	Cups and sonics: 0° to 360°	Cups and sonics: 0° to 360°	Cups: 0° to 360° Sonics: user defined
* A non-isotropic Kaimal turbulence spectrum with turbulence length scale 350 m.					

Figure 2.5: Ranges of influence parameters for each class (IEC 61400-12-1:2017, 2017).

preemptively replacing anemometers are applied to ensure ageing does not deteriorate the measurements. However, these filtering practices lead to an enormous loss of valid measurements, and Operation & Maintenance (O&M) procedures cost money and time. Anomaly detection methods, meaning methods for identifying those measurements that do not represent the real wind speed due to factors that are not yet accounted for within the prescribed measurement uncertainty, can make filtering procedures less harsh.

CHAPTER 3
Measurement Sites and Equipment

This section provides an accurate description of the measurement campaign and related instruments from which the data for this work is taken. All data belongs to and is obtained by DTU Wind Energy, who take continuous measurements at the Risø campus as well as at both of their test centres for large wind turbines, Høvsøre and Østerild. Site investigation will reveal information about the surroundings of the used masts that will be relevant during data inspection. Furthermore, it is verified that all instruments have been correctly calibrated according to the field standard (IEC 61400-12-1:2017, 2017) during their operative lifetimes on the masts.

3.1 Østerild measuring site

The Østerild National Test Centre for Large Wind Turbines is located in Northern Jylland, Denmark. The region is known for its high wind speeds due to its coastal location and flat terrain. The test centre currently has seven test stands operational where prototypical turbines are installed for tests and measurements. These test stands are aligned along the north-south direction and are accompanied by a 250m met mast on either side. These masts are referred to as Light Mast North (LMN) and Light Mast South (LMS). Recent expansions of the test centre allow for turbines to measure up to 330m from ground to blade tip in upright position on three of the existing stands, compared to the 250m before the expansion. Two new stands have been added as well, but are not currently in use. The height and distributed instrumentation of the LMN and LMS allow for a set of measurements of wind speed, direction and ambient conditions across the full rotor plane of installed turbines. An overview of the immediate surroundings of both the LMN and LMS are shown in Figure 3.1 and 3.2 respectively. It can be seen that both towers, especially the LMS, are surrounded by vegetated regions. It is also seen that the LMN is located in the wake of the nearby turbine when the wind is coming from approximately 150°, with North being 0°. A similar situation occurs for LMS when the wind is coming from approximately 35°.

The design of the LMN and LMS is identical, and their schematic overview can be found in Figure A.1. The masts are three-legged lattice towers with an equilateral triangular cross-section, where each side has a length of 1.2m. Booms are connected to the western side of the tower and are pointing towards $0° \pm 1°$ (North boom) or $180° \pm 1°$ (South boom). Note that the instrumentation on the left side in Figure A.1 points south for the LMN and north for the LMS.

Figure 3.1: Surroundings of the LMN, pinned at 57°5’13.43"N 8°52’50.35"E.

Figure 3.2: Surroundings of the LMS, pinned at 57°2’55.99"N 8°52’50.79"E.

Two pairs of P2546A OPR cup anemometers (Wind Sensor, no date) are identified at 106m and 178m altitude, which are selected for further analysis. All four of the cup anemometers are spaced at least 4.8m from the lattice tower, in accordance with IEC 61400-12-1:2017 (2017) annex G. All but one of the cup anemometers have a vertical separation of 1m with their respective booms, which also conforms to the same annex G. The southern cup anemometer for LMN (northern for LMS) has an elevation of 0.75m, due to the boom being attached 0.25m, meaning that both cups still measure at a 106m altitude. The 178m pair will be used for model development, as both anemometers are installed in an identical manner. In addition to the cup anemometer pairs, a Metek P2901 USA-1 sonic anemometer (Metek, no date) and Risø P2426B temperature sensor are located three meters below each pair. This type of sonic anemometer provides the three dimensional wind vector from which the wind direction can be measured, which is necessary to include due to the dependency of direction on the wind speed residual/ratio. The temperature measurements are used for filtering out potential icing events. A radiation shield is also present on the temperature sensor, which prevents the instrument from reading a higher temperature due to incoming solar radiation.

All measurements are logged by what is referred to as a Data Acquisition Unit (DAU). The anemometer pair at 178m and the related sonic anemometer and temperature sensor are connected to the DAU, which is located in an enclosure at 170m. The pair located at 106m and all relevant instruments surrounding these two instruments are also connected to a DAU in an enclosure at 90m. These DAUs are connected to an ethernet interface panel, which is used to upload all incoming measurements to the online database. Lastly, a power supply is present in the enclosure, providing a 24V DC signal to the system, allowing all measurements from the mast to be transmitted to the online database.

3.2 Light Mast North

During the measurement campaign of the LMN, many irregularities and scheduled maintenance have occurred. Most of the instruments that were initially mounted on the masts have been replaced since their construction. Furthermore, site-specific events, changes in set-up and irregular behaviour of the DAU have all taken place. These instances are reported in a weblog, from which the most significant events have been displayed in Figure 3.3.

Other than the scheduled replacement of instruments and temporary outages, it can be seen that several lightning strikes have occurred. Most noticeably, one of the lightning strikes caused serious damage to the northern cup anemometers at 106m and 178m, as one of the arms broke off. Even though this event took place at the end of January 2019, the broken instruments were not replaced until April 2019. It is for this reason that the data from the LMN will be used for this report, as all invalid data during this period can serve as validation. Outside of this event, several other periods with invalid measurements are identified due to communication or power issues.

Considering multiple wind speed sensors have been operative on the mast, it is important to note that all instruments have been accredited and calibrated such that it

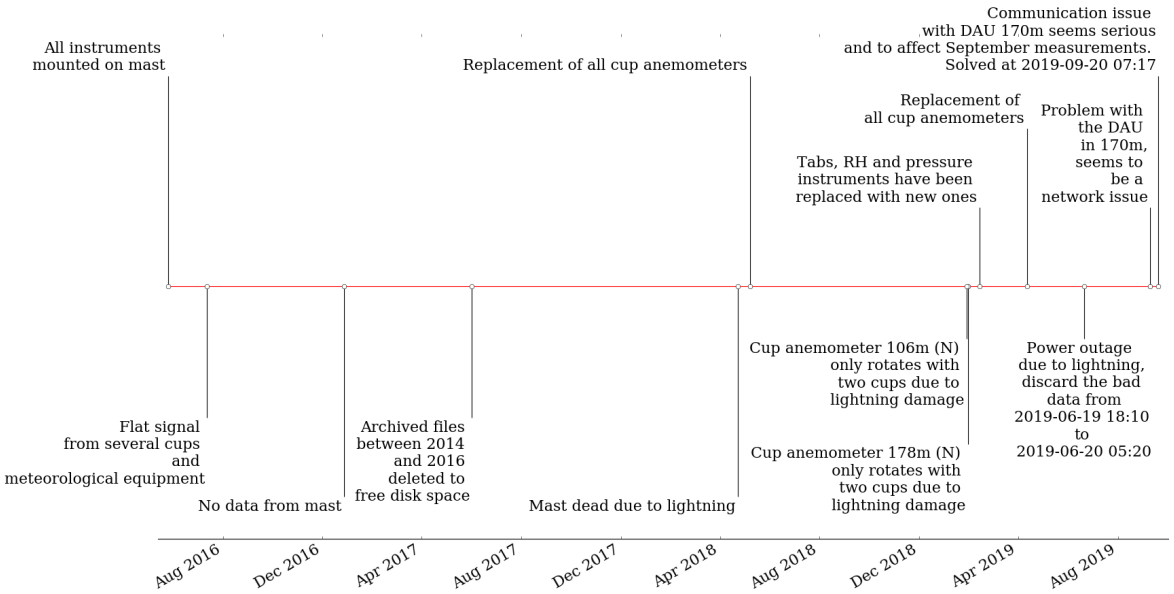


Figure 3.3: Major events related to the LMN as denoted in the weblog. Note that all lines indicate the start of each event.

can be assumed that the horizontal wind speed under normal conditions is accurately captured. Furthermore, damage may occur and cup anemometers can be subject to ageing, meaning that their accuracy might deteriorate during the end of their operation. In order to verify that the anemometers have been calibrated and installed correctly, the calibration certificates of the pre- and post-calibration are inspected. It is verified that each instrument is traceable and the calibration procedure is in accordance with IEC 61400-12-1:2017 (2017). It is also verified that the obtained calibration coefficients have been correctly established on the DAU after installment through the weblog. The result of the calibrations of all cup anemometers that have been installed on the LMN are shown in Table 3.1.

It can be seen that the gain and offset between the pre- and post calibration and instruments with similar orientation are varying with some of these parameters changing with 10%. A different gain can, in conjunction with a different offset, give a similar wind speed. The resulting wind speeds for common frequencies are displayed for all calibration results in Table 3.2. Here, it can be seen that the residuals of wind speed between the pre- and post-calibration coefficients are insignificant. It is confirmed that all instruments have been replaced before the calibration certificates expired. Furthermore, it can also be seen that no real pattern in ageing is present, where the first set of instruments all saw minor increases in their measured wind speeds, whereas the second set resulted in minor decreases. It can thus be concluded that all instruments are valid and that the wind speeds during the operation can be used. Another important conclusion is that it can not be verified whether or not the final model will be able to detect ageing effects, since these are either not present or quantifiable during the relatively short operation of the used anemometers.

Table 3.1: Pre- and post-calibration results of the different cup anemometers that have been installed at the LMN. Note that all calibrations had a fit with R^2>0.99999.

Time of operation	**Orientation**	**Date of calibration**	**Offset** $[m/s]$	**Gain** $[(m/s)/Hz]$	**Date of post calibration**	**Offset** $[m/s]$	**Gain** $[(m/s)/Hz]$
09-05-2017 to 07-05-2018	106m N	04-10-2016	0.2106	0.61588	09-08-2018	0.232	0.61608
	106m S	04-10-2016	0.218	0.61478	09-08-2018	0.2363	0.61618
	178m N	23-12-2016	0.2214	0.61553	09-08-2018	0.2345	0.616282
	178m S	23-12-2016	0.2085	0.61603	09-08-2018	0.2248	0.61663
07-05-2018 to 10-04-2019	106m N	13-04-2018	0.2442	0.61623	*Damaged*	-	-
	106m S	13-04-2018	0.2364	0.61748	22-08-2019	0.2357	0.61615
	178m N	13-04-2018	0.2379	0.61712	*Damaged*	-	-
	178m S	13-04-2018	0.2466	0.61683	22-08-2019	0.2327	0.61664
10-04-2019 to (10-04-2020)	106m N	21-01-2019	0.2448	0.61712	*Operational*	-	-
	106m S	21-01-2019	0.2232	0.6182	*Operational*	-	-
	178m N	21-01-2019	0.2393	0.61674	*Operational*	-	-
	178m S	18-01-2019	0.2327	0.61659	*Operational*	-	-

Table 3.2: Pre- and post-calibration results of wind speeds for common rotational frequencies. All units are in m/s unless specified otherwise.

	10Hz			**15Hz**			**20Hz**		
	Pre	*Post*	*Residual*	*Pre*	*Post*	*Residual*	*Pre*	*Post*	*Residual*
106m N	6.3694	6.3928	0.0234	9.4488	9.4732	0.0244	12.5282	12.5536	0.0254
106m S	6.3658	6.3981	0.0323	9.4397	9.479	0.0393	12.5136	12.5599	0.0463
178m N	6.3767	6.39732	0.02062	9.45435	9.47873	0.02438	12.532	12.56014	0.02814
178m S	6.3688	6.3911	0.0223	9.44895	9.47425	0.0253	12.5291	12.5574	0.0283
106m N	6.4065	0		9.48765	0		12.5688	0	
106m S	6.4112	6.3972	-0.014	9.4986	9.47795	-0.02065	12.586	12.5587	-0.0273
178m N	6.4091	0		9.4947	0		12.5803	0	
178m S	6.4149	6.3991	-0.0158	9.49905	9.4823	-0.01675	12.5832	12.5655	-0.0177
106m N	6.416	0		9.5016	0		12.5872	0	
106m S	6.4052	0		9.4962	0		12.5872	0	
178m N	6.4067	0		9.4904	0		12.5741	0	
178m S	6.3986	0		9.48155	0		12.5645	0	

CHAPTER 4
10-minute average data inspection

All raw data is stored in an online database, which is accessible through Structured Query Language (SQL). The relevant wind speed signals are sampled at 10Hz from the cup anemometers; the wind speed direction as determined from the sonic anemometers is sampled at 20Hz. Temperature measurements are taken at 1Hz. The database for the Østerild masts is organized as follows:

- `caldata_yyyy_mm_10Hz`, where all the 10Hz data is stored from year `yyyy` and month `mm`.
- `caldata_yyyy_mm_20Hz`, where all the 20Hz data is stored from year `yyyy` and month `mm`. Note that from 2017 onwards `caldatacorr_yyyy_mm_20Hz` is used.
- `caldata_yyyy_mm_01Hz`, where all the 1Hz data is stored from year `yyyy` and month `mm`.
- `calmeans`, where all the 10-minute averaged data is stored from the aforementioned tables.
- `calstdvs`, where the observed standard deviations are stored during the same 10-minute periods.

Continuous measurements have been taken and are logged in their relevant tables. The data collection and inspection is carried out with a top-down approach, where the 10-minute averages are loaded and analyzed first. Afterwards, a similar procedure is followed for the fast data. This approach is justified due to the sheer size of the data. The first data entry is logged at 2016-02-10 15:00 (UTC+2) and ranges until 2020-01-29 00:00 (UTC+2), a period spanning 259,096 10-minute intervals and corresponding to over 1.5 billion 10Hz measurements. The 10-minute averages will inherently contain unwanted data objects, such as obviously faulty measurements. Filtering out these unwanted entries will significantly prune the data in an effective and organized manner. It should be noted that all measurements are taken in the same time zone, which will not be reported in this work for the sake of clarity.

All data from the `calmeans` and `calstdvs` tables are loaded and stored locally in a Hierarchical Data Format (HDF) file to obtain the 10-minute averaged data from the aforementioned time period. The data entries are ordered through a unique code that depicts the time period over which the average is taken. For example, a the code '201602101500' indicates that the average is taken over 2016-02-10 15:00.00 to 2016-02-10 15:09.59. These strings are stored in the `name` column. For each of these time stamps,

the relevant wind speed, wind direction and temperature around the 178m LMN pair are selected using the following queries:

```
SELECT name, wsp_178m_LMN, wsp_178m_N_LMN, Sdir_175m_LMN,
       Tabs_175m_LMN;
FROM oesterild_light_masts.calmeans;
ORDER BY name;

SELECT name, wsp_106m_LMN, wsp_106m_N_LMN, Sdir_103m_LMN,
       Tabs_103m_LMN;
FROM oesterild_light_masts.calstdvs;
ORDER BY name;
```

The pymysql module (Matsubara & contributors, 2018) is used to create a connection to the database and run the queries. The resulting tables that are obtained are stored in a pandas dataframe. Pandas (McKinney, 2010) is a python module that is commonly used for one dimensional data analysis. A dataframe can be seen as a multidimensional table, where data objects each have a unique index and their attributes are stored in a columnar manner. In this case, the timestamps are used as indices and eight columns are defined that indicate the average and standard deviation of all four attributes. The averages and standard deviations are merged when they possess identical time stamps to align the statistics in time. Pandas also allows for easy storage and retrieval of dataframes to and from HDF files.

4.1 Raw data filters

The resulting time series from the raw data from the 10-minute statistics of the wind speed, direction, and temperature are shown in Figure 4.1 and 4.2 respectively.

From the plot, it becomes clear that filtering is necessary. Average wind speeds above 10^6 m/s were reported, which are obviously faulty. In addition, empty measurements are observed, which are denoted with Not a Number (NaN) values. Data points with standard deviations equal to 0m/s or 0° are also observed, indicating failure of the instruments or the data logger. Similar conclusions are drawn from the temperature plots, where abnormal temperatures below -50°C are observed. A set of NaN values and zero standard deviations are found in this set as well. Filtering out all these obviously faulty data points produces the time series shown in Figure 4.3 and 4.4.

The effects of the filters are obvious. Negative or unrealistic magnitudes of wind speed are no longer present. The time series of the wind direction is clearly bounded between 0° and 360°. The remaining temperatures show logical values and trends. Seasonal variations stand out most in this time series, as well as daily variations between night and day. One clear outlier remains around February 2019, which will be filtered out by the set of filters discussed in Section 4.3.

This initial data set after these first filters remains far from clean, as observed from the residual and ratio of both cup anemometers when plotted over the corresponding

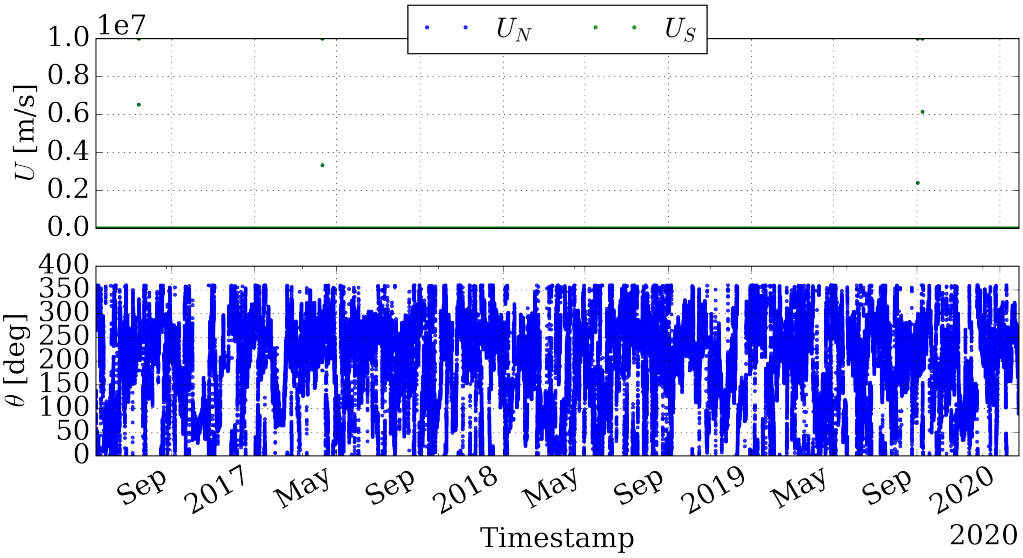


Figure 4.1: Raw 10-minute averages of the wind speed, measured by the north and south cup anemometers at 178m, and wind direction, measured at 175m.

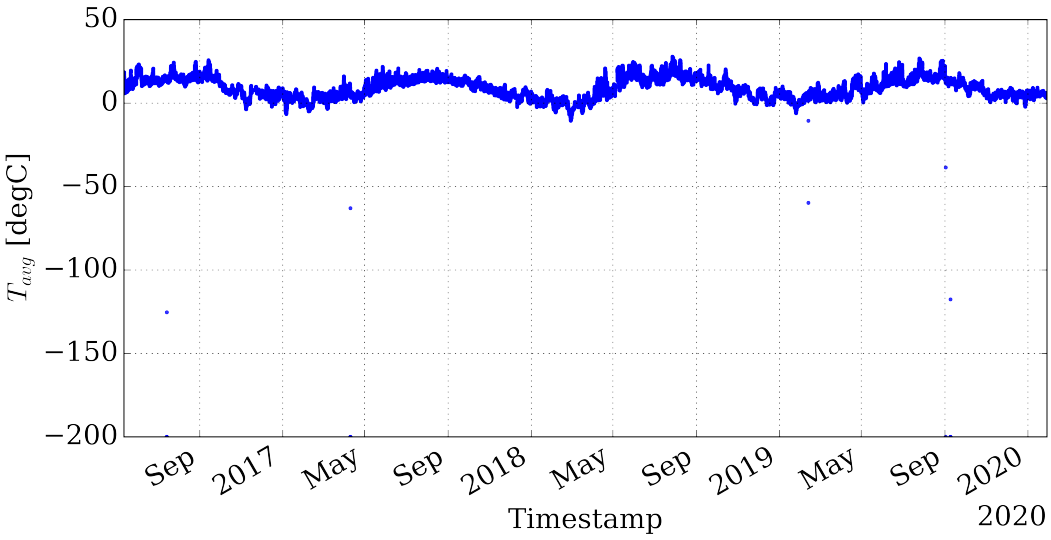


Figure 4.2: Raw 10-minute averages of the temperature, measured at 175m.

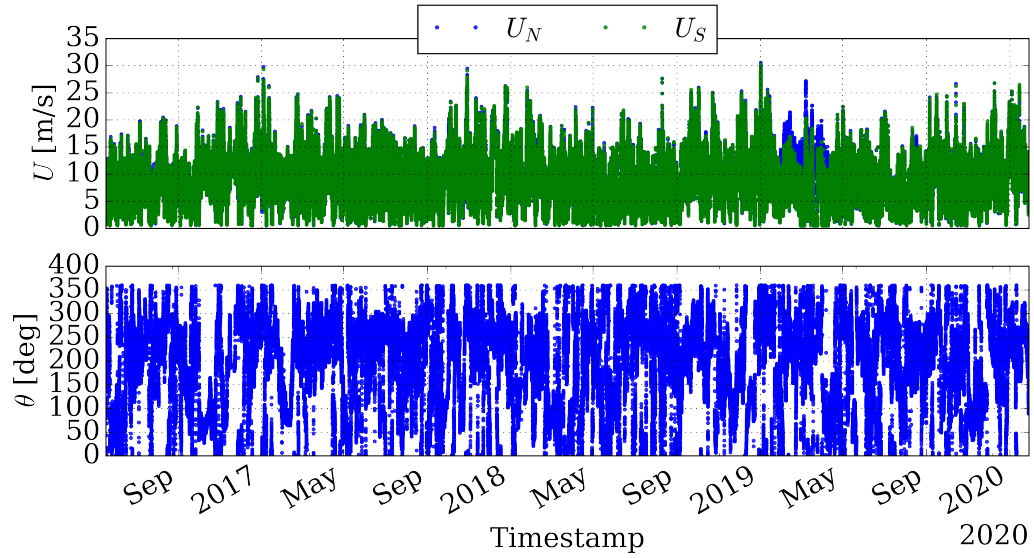


Figure 4.3: Filtered 10-minute averages of the wind speed, measured by the north and south cup anemometers at 178m and wind direction measured at 175m.

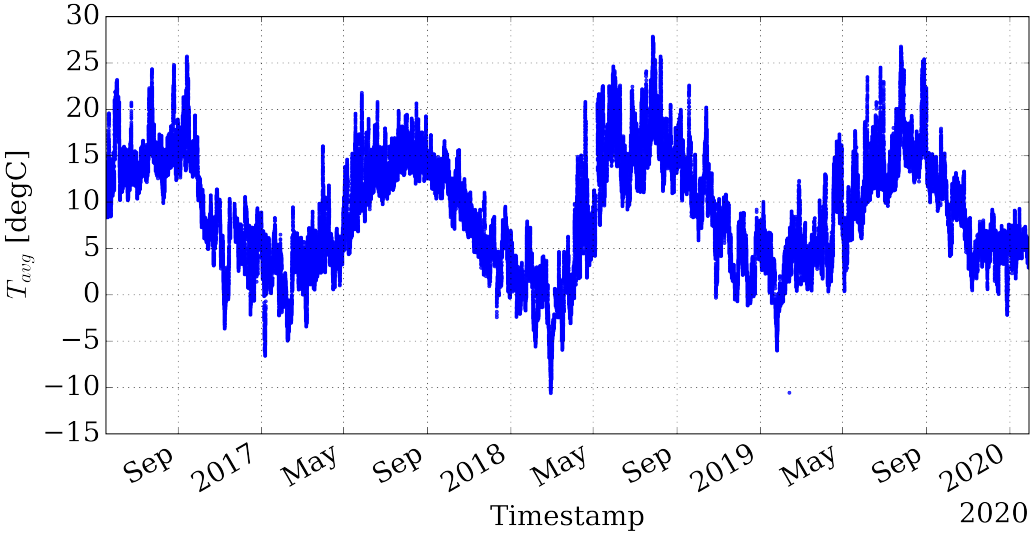


Figure 4.4: Filtered 10-minute averages of the temperature, measured at 175m.

wind direction, as shown in Figure 4.5. One expects a clean set of data to show a clear correlation, as discussed in Section 2.2.3. Clear groups of outliers can be seen below the expected curves, as well many single points that are scattered from this curve.

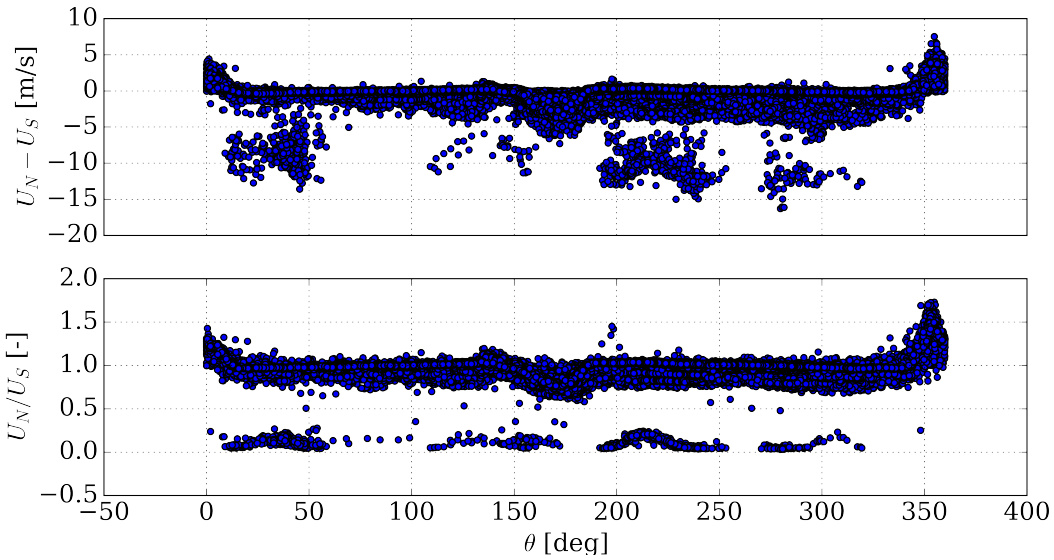


Figure 4.5: Wind speed residual and ratio of the two cup anemometers at 178m as a function of wind direction at 175m, measured at LMN.

4.2 Weblog Filters

The timeline of events at LMN in Chapter 3 gives clear indications that incorrect measurements have been collected. Periods during which no data is gathered do not affect the measurements directly, as these simply result in gaps in the time series. The most striking event is characterized between 2019-01-29 12:20 and 2019-04-11 15:40, where a cup broke off the northern anemometer as a result of lightning damage. Note that the cup anemometer at 106m was struck at 2019-01-27 15:40. Wind speed measurements from the damaged cup anemometers clearly do not conform to the desired normal behaviour and are thus filtered out. Furthermore, issues with the DAU at 170m were present for the first half of September 2019. The first reported issues were documented at 2019-08-29 06:10, which were not fixed until 2019-09-19 07:20. The weblog explicitly mentions that measurements between this period have been affected and these are thus also removed from the data set. The weblog also asks for the faulty averages due to the power outage on 2019-06-19/20 to be discarded. However, these points are already filtered out, as standard deviations for all variables were below the set threshold in the section above. The impacts of these filters on the directional plots are displayed in Figure 4.6 and 4.7, where the deleted points and remaining points are plotted respectively.

The damaged cup clearly slowed down the anemometer, as the large clusters of extremely low differences were generated when the damaged cup was in operation after

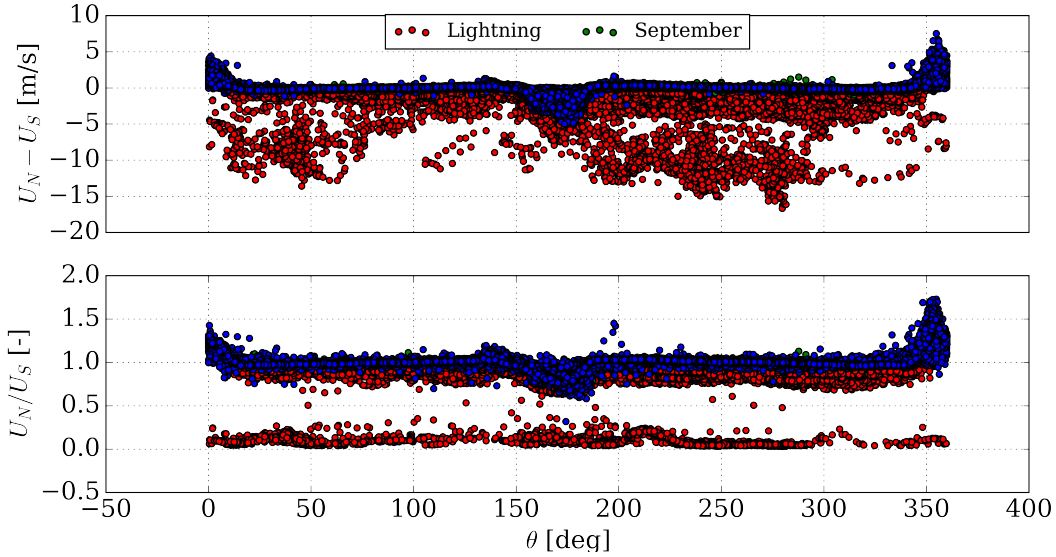


Figure 4.6: Filtered out wind speed residuals and ratios of the two cup anemometers at 178m as a function of wind direction at 175m, measured at LMN, within the identified abnormal periods from the weblog.

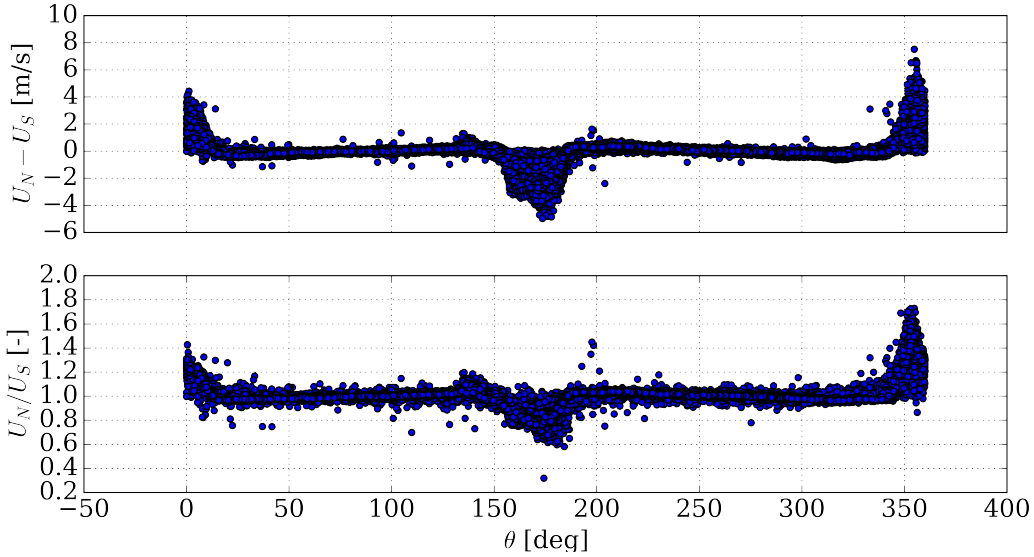


Figure 4.7: Wind speed residual and ratio of the two cup anemometers at 178m as a function of wind direction at 175m, measured at LMN, after filtering out abnormal periods as identified from the weblog.

the lightning strike. The effects on the time series are further discussed in Chapter 5. The points that correspond to the corrupted data during September 2019 have fewer outliers when compared to the remaining data, but some anomalies around 275 degrees are removed where a positive bias is found. The remaining points already show major improvement, as a better correlation can be observed.

4.3 General filters

Large scatter is still observed, indicating that further filtering is required. Due to mast shadow effects, where one cup anemometer is directly downstream of the mast, one of the instruments may start or stop before its counterpart does. These effects around 0 and 180° are clearly visible in Figure 4.6. This results in outlying points that do not conform to the expected behavior and thus must be filtered. Low wind speeds (<4m/s) in general are not relevant to wind energy applications due to the cut-in speeds of wind turbines. It is for this reason that the standard suggests that cup anemometers are to be calibrated within the range of 4 to 16m/s (IEC 61400-12-1:2017, 2017). Furthermore, icing events may be present in the data set. Periods where instruments are completely frozen for longer than ten minutes have been filtered out, as these will have no standard deviations. However, measurements taken by partially frozen instruments are most likely still present in the current data set. These temperature-induced errors are removed from by filtering out all data points below 2°C. Lastly, periods with relatively large standard deviations of wind direction are present. Large uncertainties in the wind direction averages can cause tower shadow effects to be misrepresented. Inspection of the outlying points show that these are all associated with a standard deviation larger than 12°. The filtered out points and remaining data are plotted in Figure 4.8 and 4.9 respectively.

A majority of the points are removed by the low wind speed filter. This becomes predominantly visible when inspecting the ratios, as these values are amplified around low wind speeds. More outlying objects are removed by the temperature filter, indicating potential icing events. The filter related to the standard deviation of the wind direction mainly removed outliers around the peaks that result from the mast shadow, which is to be expected.

4.4 Local Outlier Factor filter

Some general outliers can still be observed in the data set. Further inspection of these points does not reveal a distinctive error source, but rather indicates that they originate from noise. Every dataset inherently has outliers. In order to remove them, and other similar outlying points that may have slipped through the set filters, the Local Outlier Factor method is applied.

Local Outlier Factor (LOF), proposed by Breunig et al. (2000), is a density-based approach for outlier detection. LOF does not attribute a binary property to potential outliers, i.e. a candidate is either an outlier or not, but rather assigns an outlier factor to outlying objects. LOF also uses the local neighborhood of data points to identify

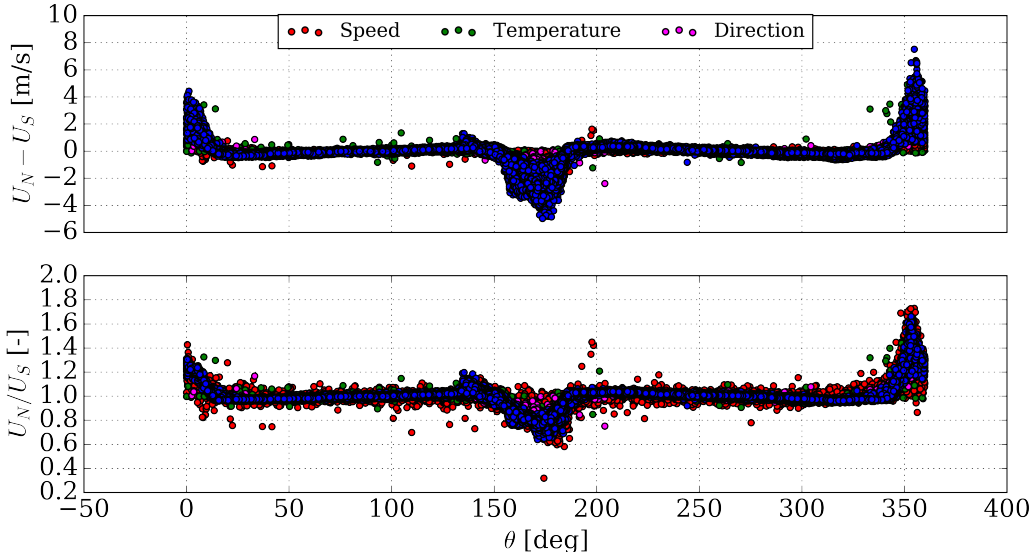


Figure 4.8: Filtered out wind speed residuals and ratios of the two cup anemometers at 178m as a function of wind direction at 175m, measured at LMN, based on remaining errors.

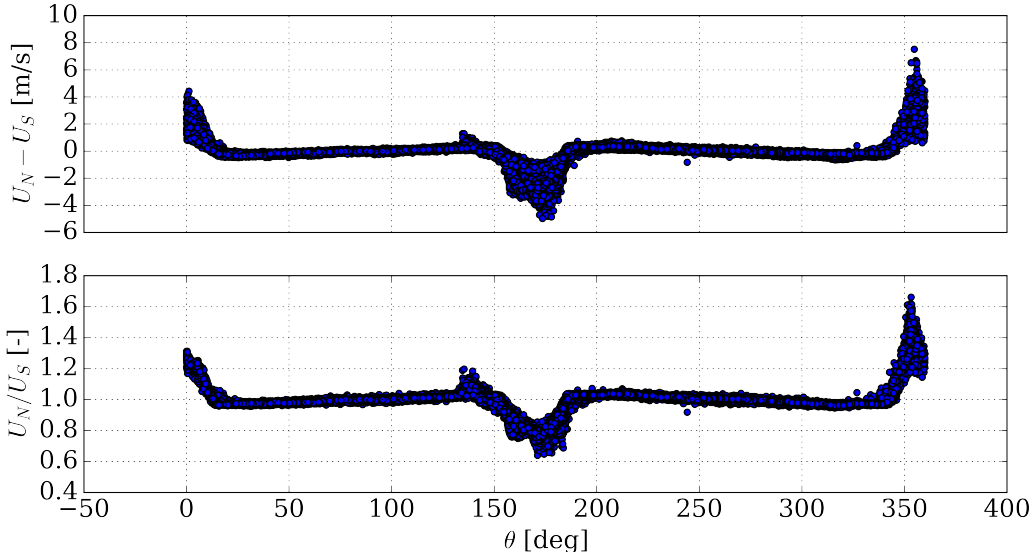


Figure 4.9: Wind speed residual and ratio of the two cup anemometers at 178m as a function of wind direction at 175m, measured at LMN, after filtering out abnormal periods from remaining error sources.

outliers, instead of using the entire data set to determine global outliers, making it a suitable technique for removing the desired data points.

LOF makes use of the K-Nearest Neighborhood (KNN) of an object p, denoted by $kNN(p)$ to assign the outlier score. It does so by first defining a so-called reachability distance, as shown in (4.1). Note that the *k-distance* simply denotes the euclidean distance between the furthest point in the $kNN(o)$ to point o itself. Whenever objects p and o are relatively far away from each other, the reachability distance is equal to their euclidean distance, but if p belongs to the KNN of o, the reachability distance becomes *k-distance*. This reachability distance is used to calculate the local reachability density, as shown in (4.2). Relatively large values of $lrd\,(p)$ imply that p does not belong the kNN of the objects o in its own kNN, thus providing a measure of relative isolation of p. Finally, the LOF is determined using (4.3). The LOF thus defines the average ratio of isolation between a test object p and objects o in its KNN. If objects o have high values of lrd, meaning their nearest neighbors are close-by, but $lrd_k\,(p)$ is low, meaning that p is isolated, a relatively high value of $LOF_k\,(p)$ is assigned.

$$\textit{reach-dist}_k\,(p,o) = \max\left(\textit{k-distance}(o), d(p,o)\right) \tag{4.1}$$

$$lrd_k\,(p) = \left[\frac{\sum_{o\in kNN(p)} \textit{reach-dist}_k\,(p,o)}{|kNN(p)|}\right]^{-1} \tag{4.2}$$

$$LOF_k\,(p) = \frac{1}{|kNN(p)|}\sum_{o\in kNN(p)}\frac{lrd_k\,(o)}{lrd_k\,(p)} \tag{4.3}$$

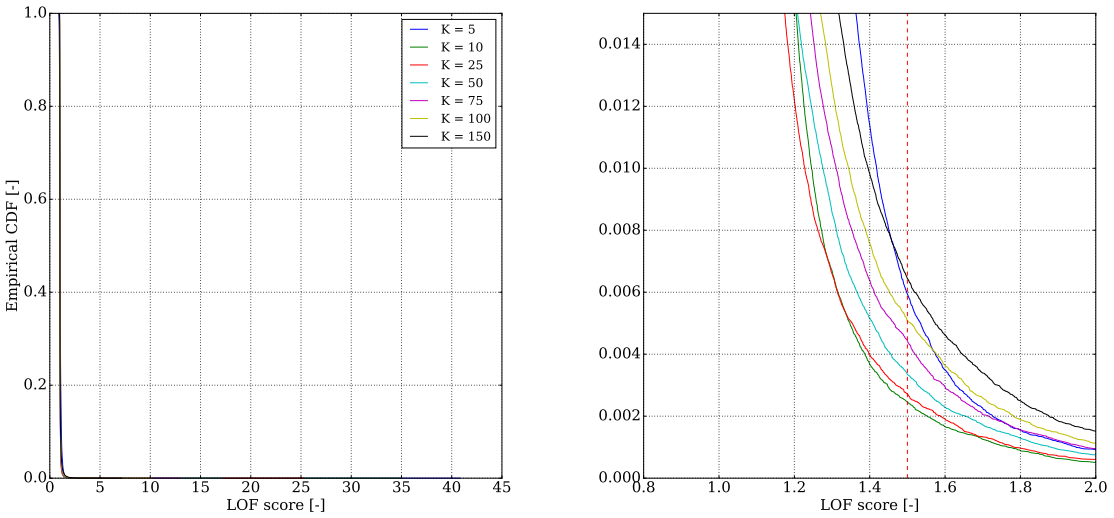
The drawbacks of this method include the sensitivity of K and the expensive nature of the computations. Breunig et al. (2000) provides direction to obtain a range of values for K, but tuning is required. Furthermore, the LOF algorithm is computationally expensive, as every object has to be visited with the addition of its k neighbors.

Parameter tuning

The data is subjected to the LOF technique with use of the scikit-learn module (Pedregosa et al., 2011) in python. The scikit-learn implementation of LOF requires two input parameters that are to be tuned: a contamination value and K. The contamination value defines the percentage of outliers in the dataset, and the aforementioned K describes how many of the closest points make up $kNN(p)$.

As mentioned, the LOF value does not directly indicate if a point is an outlier or not. Rather, its value is relative to the scores of the other points and defines the 'outlierness' of a point. The contamination parameter in the scikit-learn module solves this problem. For example, a contamination of 0.05 means that the 5% of the most negative scores are defined as outliers. Breunig et al. (2000) proves that inliers have LOF scores close to 1.0 and deducted from a series of case studies that outlying values have scores above 1.5.

The LOF scores for a set of different K are plotted for the filtered 178m LMN paired data set in Figure 4.10.



(a) Original

(b) Zoomed

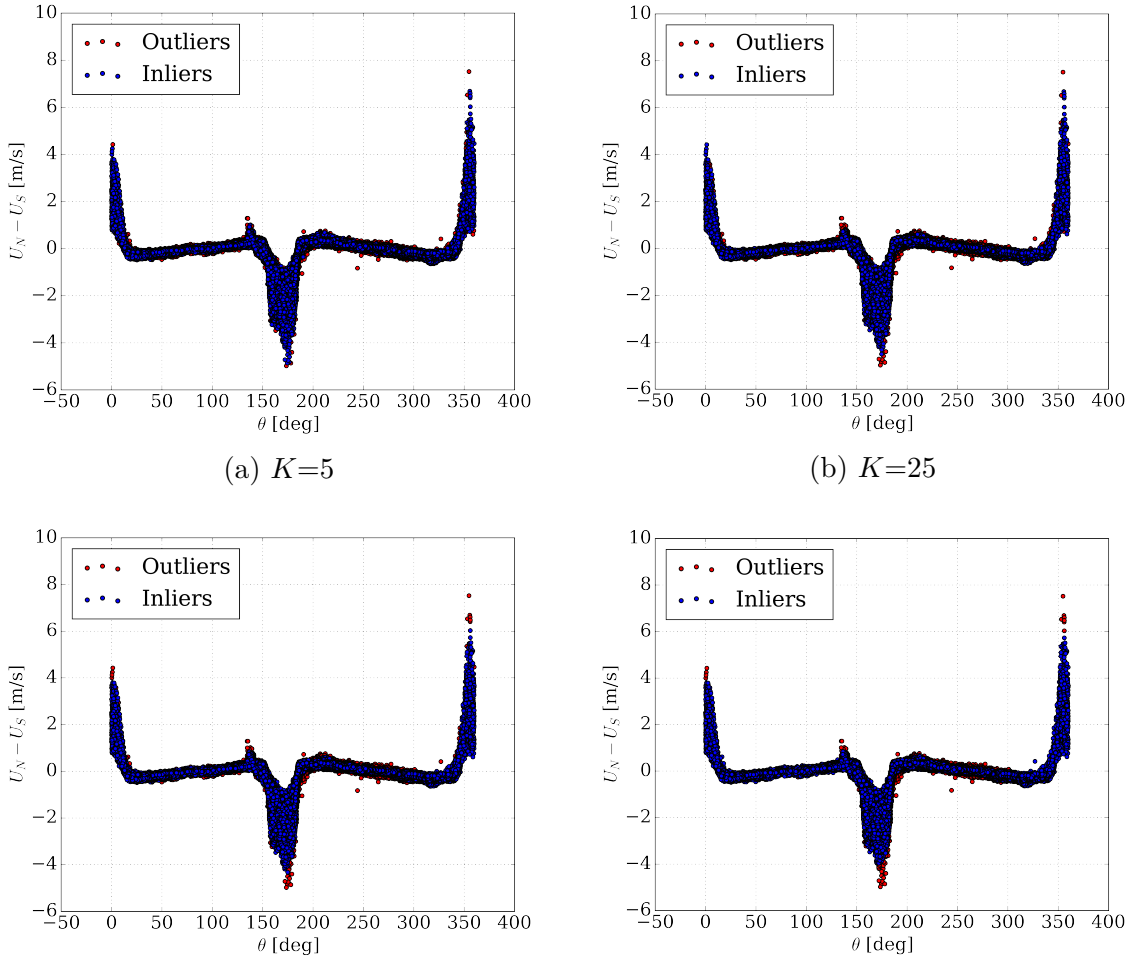
Figure 4.10: Empirical CDF of the calculated LOF scores for the LMN 178m pair, for several different values of K

It can be seen that the vast majority of the points have a score close to 1.0, regardless of K. This is a logical outcome, as the data set has already been thoroughly filtered. It can also be concluded that the contamination at the cut-off point of 1.5 is dependent on the K and does not converge within a range of sensible neighbors. No exact method can be applied to tune this data set, and due to the lack of convergence of the statistics, e.g. the standard deviation in the set of outliers, collecting empirical observations for different K is the leading method in ensuring the desired effect is reached. The results for $K = 5, 25, 50$ and 100 are shown in Figure 4.11.

From these figures, several observations can be made. Using $K = 5$ is clearly not feasible, as the neighborhoods contain too few points to capture the outlying point around the peaks at 0° and 10°. Taking K=100, however, seems to overestimate the number of outliers. For instance, one of the clear outliers around 300° is labelled as an inlier. Through empirical tuning, it can be seen that K=50 strikes a good middle ground and can be used to remove these statistical outliers.

4.5 Filtering results

Finally, all remaining data points are displayed in Figure 4.12. Compared to the raw data, it can be seen that a pattern is present and that both anemometers agree for most directions, i.e. are centered around a ratio of 1.0 or difference of 0.0m/s. The regions where the averages do not seem to agree are not due to faulty behaviour, but are a result



(a) K=5 (b) K=25

(c) K=50 (d) K=100

Figure 4.11: In- and outliers separated at a score of 1.5, as determined by the LOF algorithm for $K = 5, 25, 50, 100$ for the LMN 178m pair.

of mast shadow effects, as explained in the next chapter. On a 10-minute average level, the data set is said to conform to normalcy. All error sources as described in Chapter 2 that are to be flagged as anomalous have been rid of on this time scale. An overview of all applied filters is provided in Table 4.1.

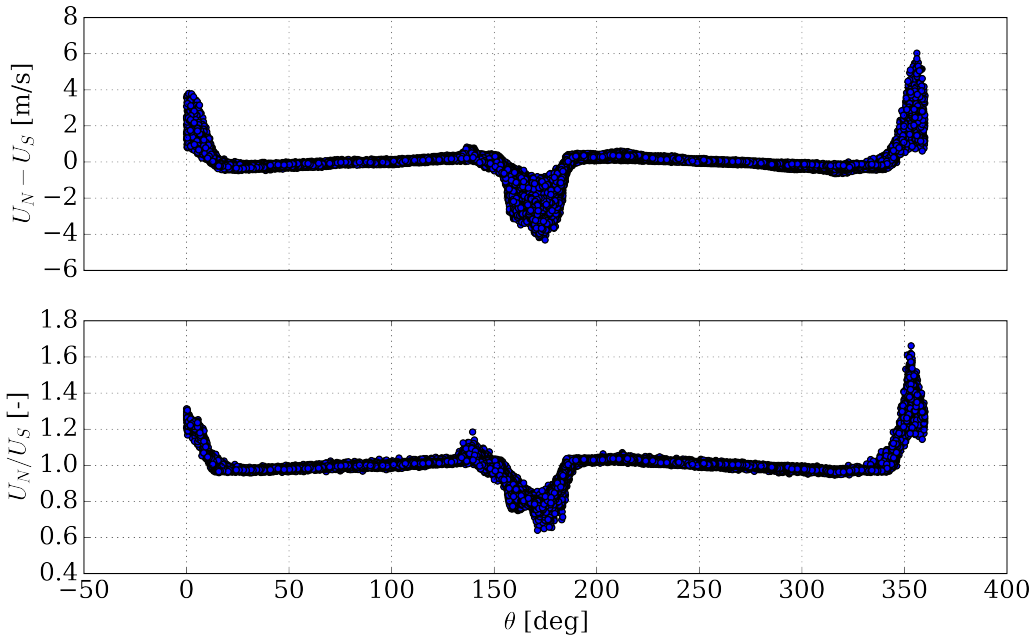


Figure 4.12: Clean plots of wind speed residual and ratio of the two cup anemometers at 178m as a function of wind direction at 175m, measured at LMN.

Table 4.1: Summary of applied filters to the 10-minute averages, combined with the total recovery rate after the listed filters are applied.

Category	**Condition**	**Recovery**
Faulty	$U_{avg} > 100$m/s $U_{std} < 0.001$ m/s $\theta_{std} < 0.001°$ $-40°$C$> T_{avg} > 50°$C $T_{std} < 0.001°$C Any row with a NaN	97.2%
Weblog	2019-01-29 12:20.00 $\leq$ Time $\leq$ 2019-04-11 09:30.00 2019-08-29 06:10.00 $\leq$ Time $\leq$ 2019-09-19 07:20.00	91.0%
General	$U_{avg} < 4$ m/s $\theta_{std} > 12°$ $2°$C$\leq T_{avg}$	70.8%
LOF	score > 1.5	69.5%

CHAPTER 5
10Hz dataset formation

The clean data set allows for the formation of the 10Hz data set that will be used for training and testing of the anomaly detection models. The regression techniques will use the wind speed and direction as input to predict the next residual between both anemometers. With this setup, a univariate model can be used which includes information of both signals whilst keeping the complexity and dimensionality low. The residual is taken since small differences are amplified at low wind speeds, which is undesirable when both anemometers agree with each other. Special consideration for the mast shadow removal is given first, after which the loading of the 10Hz data is discussed including with extra filters required. The 10Hz data is distributed over several batches, which are loaded in separately. This allows for out-of-memory techniques to be used, as loading in all data at once is not possible.

5.1 Mast shadow

Although the filtering procedure, described in the previous chapter, defines normal behaviour, some limitations are still evident. A clear definition of normalcy is sought after such that abnormal behaviour stand out. This clear definition does not seem to be present in regions where the effects of mast shadow seem to dominate, which can be observed in Figure 5.1, where the binned averaged directional plot with 0.5° bins together with the standard deviation inside each bin.

It can be seen that in regions where one of the cups is not downwind of the lattice tower, relatively low standard deviations are present with slight variations in the mean due to flow distortion. In the northern and southern direction, however, the standard deviation is several orders of magnitude larger. There also appears to be a slight offset on the sonic anemometer, but this does not need to be taken into account since only the relative relationship between the residual and direction is important.

The effects of mast shadow on the correlation between both anemometers can be seen in Figure 5.2, where the wind speed of concurrent measurements are plotted against each other. The measurements that were taken whilst the wind was coming from directions outside the regions of increased variance are shown separately.

The value of filtering out the measurements that are affected by the mast shadow effects is immediately evident. The large scatter where measurements show significant differences are filtered out by removing the tower shadow effects. Using Figure 5.2, it is concluded that no clear definition can be established when measurements affected by mast shadow are present. Therefore, a separate flag will be attached to such measurements. It can also be concluded that normalcy in the 10-minute averages is achieved, since all

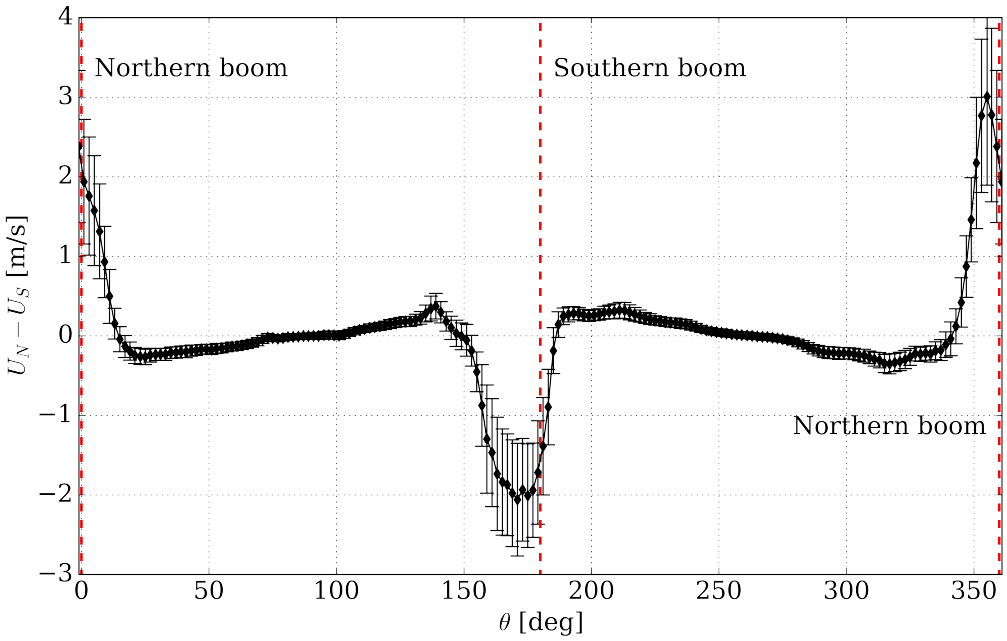


Figure 5.1: Bin-averaged wind speed residual of the two cup anemometers at 178m as a function of wind direction at 175m, measured at LMN.

concurrent averages from both cup anemometers agree with each other with few scatters, as expected from collocated instruments.

5.2 Batch splits

The 10Hz data set can now be created, as all valid 10-minute conform to the created definition of normalcy. The fast wind speed and direction data are loaded and saved in similar manner as the 10-minute averages. In order to check for completeness of the results, sequences of connecting 10-minute averages are built, e.g. the first two sequences for the 178m LMN pair range from 08:20 to 09:10 and 09:30 to 13:00. The shortest sequences are only 10 minutes long, whereas the longest uninterrupted sequence spans over a day. If these sequences are complete, one can expect 6000 10Hz measurements per 10 minutes. The following queries have been applied for the same cup pair to the databases and the resulting dataframes are stored in HDF5 files. Note that the wind direction data is resampled to 10Hz, as the measurements are taken at 20Hz.

```
SELECT name, scan_id, wsp_178m_LMN, wsp_178m_N_LMN;
FROM oesterild_light_masts.caldata_yyyy_mm_10Hz;
WHERE name BETWEEN yyyymmddHHMM AND yyyymmddHHMM;
ORDER by name;

SELECT name, scan_id, Sdir_175m_LMN;
```

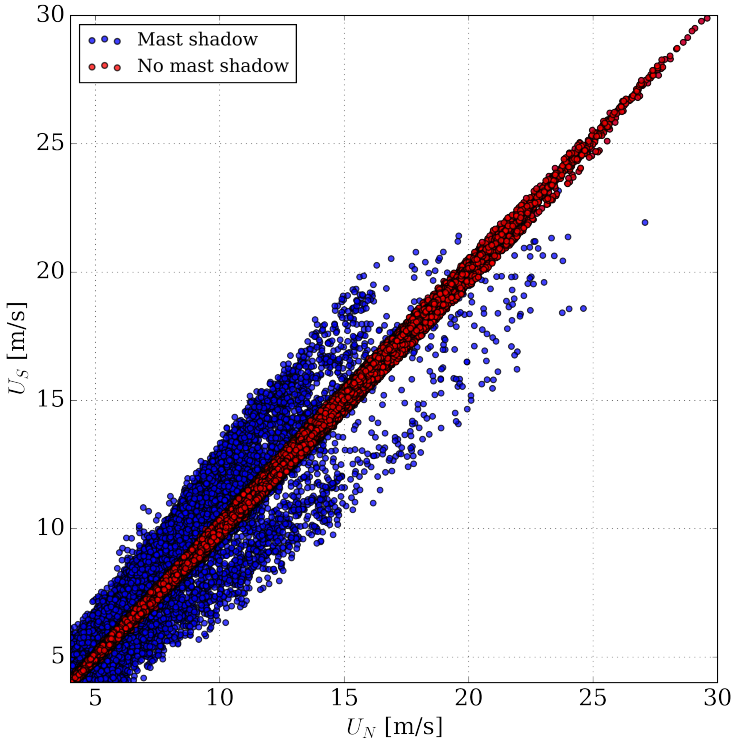


Figure 5.2: Relationship between the measured wind speed from the LMN 178m cup anemometer pair with and without measurements taken when either cup is affected by mast shadow effects.

```
FROM oesterild_light_masts.caldata_yyyy_mm_20Hz
WHERE MOD(scan_id,2) !=0;
AND name BETWEEN yyyymmddHHMM AND yyyymmddHHMM;
ORDER by name;
```

The new column `scan_id` is used to create new time stamps with a resolution of 0.1ms, as all the `name` tags have a resolution of 10 minutes. The start and end date of a sequence is given by `yyyy-mm-dd HH:MM`. In case the start and end date are from different months, two queries are run from both months, as the data is stored in tables that correspond to one month only.

Two sequences with missing data were identified for the 178m LMN pair. One of these sequences was missing two minutes worth of data. This could be explained using the weblog, where it was recorded that maintenance had just been performed and the system rebooted slightly after a 10-minute period. The other incomplete sequence corresponded to a period for which 10-minute averages were available, but without any corresponding 10Hz data. These unrecorded measurements take place between 30-09-2017 22:50 until 31-09-2017 02:20. As such, this sequence is split in two sequences of connected 10-minute

time steps, meaning that all sequences are verified as complete.

Each sequence can now be split up into batches and be subjected to further filtering. Even though the data conforms to the defined standard of normalcy at the 10-minute scale, some unwanted events may be averaged out. For instance, on the first connecting sequence as shown in Figure 5.3, in the start of first ten minute period, wind speeds of 0m/s are logged. Because the measurement campaign did not start at exactly the beginning of a ten minute interval, these measurements are averaged out over such an interval. The low wind speeds, however, fall below the set threshold for 10-minute periods. A similar observation is made for the wind direction; it can be seen that the wind is coming from regions where the effects of the tower shadow are significant. In order to filter these periods out, each sequence is split up in periods of one minute. If any measurement within this one minute batch has wind directions outside of the valid regions, or goes below wind speed of 2m/s, a batch is considered invalid. Furthermore, an extra filter for batches containing NaNs values was applied as well, to ensure all batches contained continuous measurements throughout the periods.

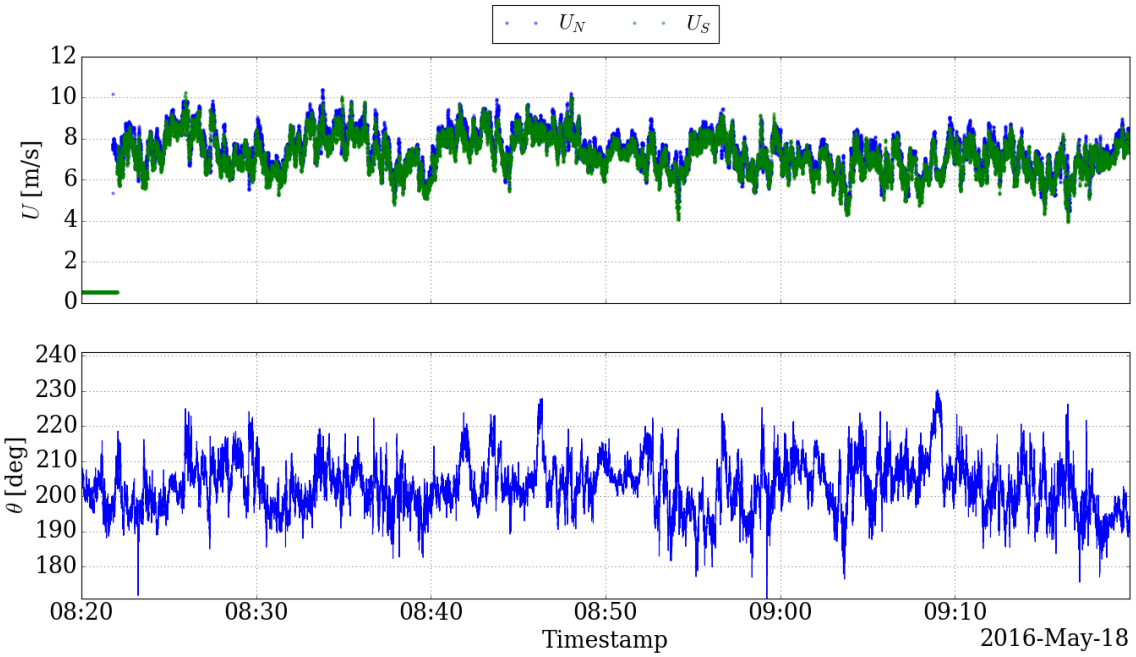


Figure 5.3: Sequence of wind speed and direction measurements ranging from 08:20 until 9:20 at May 8th 2016.

From inspection of different sequences, it was seen that some sequences contained one minute batches with wind speed differences that were exhibiting relatively extreme behaviour. Although this is not considered abnormal behaviour, these batches are considered inefficient for training purposes because they are outliers. In order to save time and make training more efficient, these batches are not used for training. An example of such a time series with labelled batches, with all filters applied as described, is given in Figure 5.4.

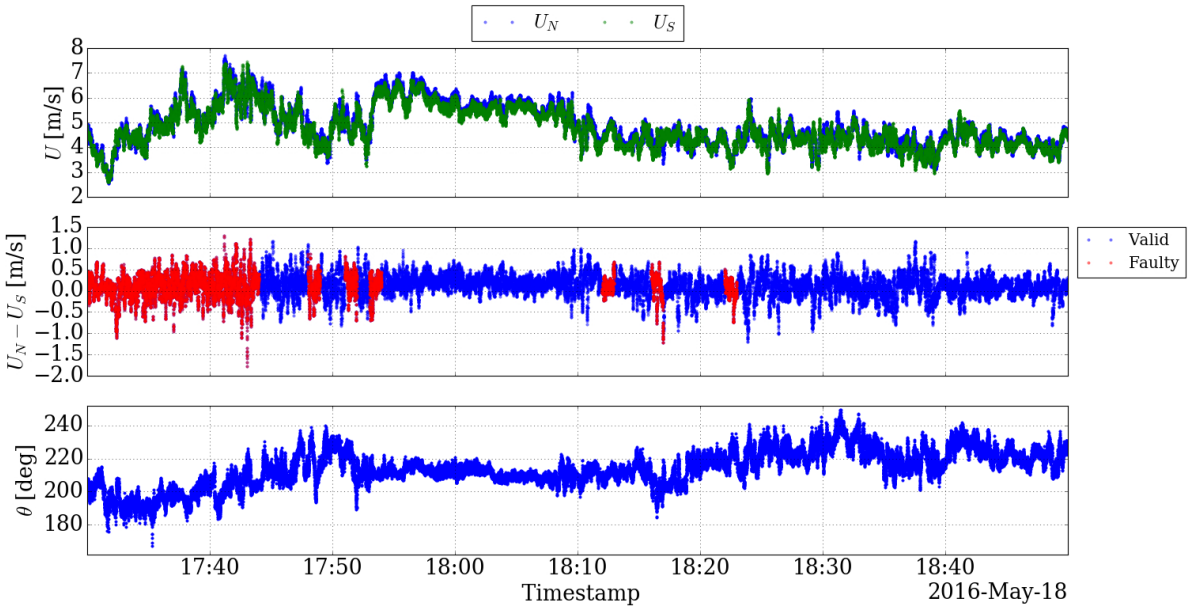


Figure 5.4: Sequence of wind speed and direction measurements ranging from 17:30 until 18:50 at May 18th 2016, where all valid and invalid minute windows are denoted according to the set filters.

A combination of all these filters can be found in Table 5.1. From the remaining valid batches, larger batches can be formed by finding consecutive one minute batches, such that each batch will always contain data without any gaps.

Table 5.1: Summary of applied filters to the one minute windows containing the 10Hz wind speed and direction measurements.

Category	**Condition**
Wind direction	$\theta_{min} \leq 25°$ $\theta_{max} \geq 330°$ $130° \leq \theta_{max} \leq 195°$ $130° \leq \theta_{min} \leq 195°$
Wind speed	$U_{N,min} < 2\text{m/s}$ $U_{S,min} < 2\text{m/s}$
Faulty	Any row containing NaNs
Outlier	$\sum_{i=1}^{600} \left(\|(U_N - U_S)_i - \mu_{(U_N - U_S)}\| > 3\sigma_{(U_N - U_S)} \right) > 9$

5.3 Training and test splits

As mentioned, not all data can be loaded in at once. The models in Chapter 6 are trained using batches. This means that, for example, training is performed on equally sized batches of 6000 data points (10 minutes worth of data) until every batch has been seen by the model. The size of these batches can affect the outcome, and the result after all data has been seen will generally not be the same. As such, batches are formed on three different time scales: 1, 5 and 10 minutes. For these sizes, 901,678, 134,379 and 48,057 valid batches are found, respectively. Due to the increasing batch sizes and the requirement that batches contain no gaps, relatively fewer valid batches are found.

Each batch is categorized through a unique identifier based on the sequence and position of the batch within the sequence, which are stored and saved in a list. This list will be used for the rest of the work in order to make train and test splits whenever necessary. The start and end date of each sequence is also stored along with this list. Subsets of the valid batches can be made by filtering out sequences according to date. For example, a subset of all data coming from 2017 can be drawn by selecting the keys that correspond to sequences that were measured within this year.

5.4 Validation data

The ultimate goal of the model is for it to be unable to accurately predict abnormal behaviour, which has different underlying mechanisms than the defined set of normal data. To ensure that the model reaches this goal, validation is to be performed. Since no heated anemometers are present, no reference wind speed can be used for detection of icing, and because no camera images are available for visual inspection, the data itself is needed to provide qualitative validation. The raw data have been filtered on the 10-minute scale such that no abnormalities are present in the clean set. These filtered points are then further inspected to create validation sets where known anomalous behaviour is present. The weblog is also used as reference, as the cup anemometer pair has had logged events of known abnormalities.

5.4.1 Lightning damage

During the data inspection, periods of data were found where one of the three cups of the northern anemometer broke off during the period between 2019-01-29 12:20 and 2019-04-11 15:40. According to the weblog, this was caused by a lightning strike that permanently damaged this cup anemometer. This incident can be used for validation purposes, as the broken anemometer provided known faulty measurements until it was replaced. The 10-minute averaged time series during this period for both wind speed and direction are shown in Figure 5.5.

According to Figure 5.5, faulty data is generated during this period. The northern cup anemometer stops spinning on several occasions during March and April. Furthermore, at the end of January, the northern anemometer starts to rotate noticeably slower. Further

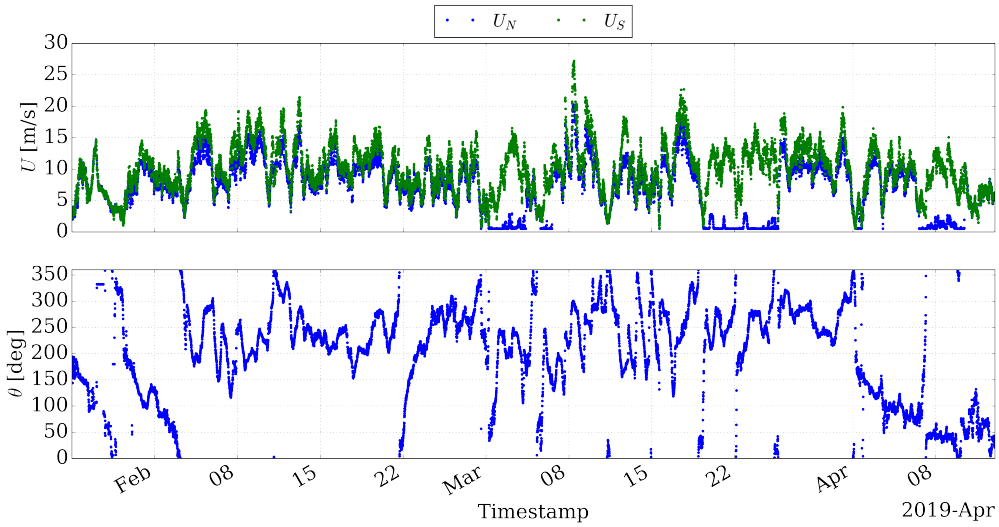


Figure 5.5: Time series of the lightning event, with the 10-minute averaged wind speed measured from the cup anemometer pair at 178M and temperature at 175m at LMN.

inspection is taken on the 10Hz scale at the moment the lightning strike supposedly hit. These time series, logged on January 29th, 2019, are displayed in Figure 5.6.

The moment of the lightning strike, around 12:30, is visible due to a single, sudden jump in the wind speed. It can be seen that the signal from the northern cup anemometer behaves erroneously during several periods after the incident. These periods can also said to be anomalous and, as such, can be used for validation of the developed anomaly detection models.

5.4.2 Icing events

Some of the most common issues with anemometry are icing events. As mentioned, conservative temperature filters are applied in order to remove potential icing data. However, this practice does not mean that all data below 2°C corresponds to faulty measurements. In order to assess whether or not icing conditions were present and impacting the data, a detection method based on ambient conditions proposed by Wadham-Gagnon et al. (2015) is used. Relative humidity and temperature data are available on the LMN, meaning that the Magnus-Tetens equations for dew point temperature, as shown in (5.2), can be used to assess whether atmospheric conditions allow for the accumulation of water or ice (Heldman & Moraru, 2010). If an object has a surface temperature below the dew point temperature, vapor inside the air can condensate on the object. The collection of water on the surface then might freeze if the ambient temperature is sufficiently low and result in icing. Wadham-Gagnon et al. (2015) has tested a variety of their suggested icing detection methods for wind speed measurements and concluded from comparison that this technique has a False Positive Rate (FPR) of 75%, meaning that many of the

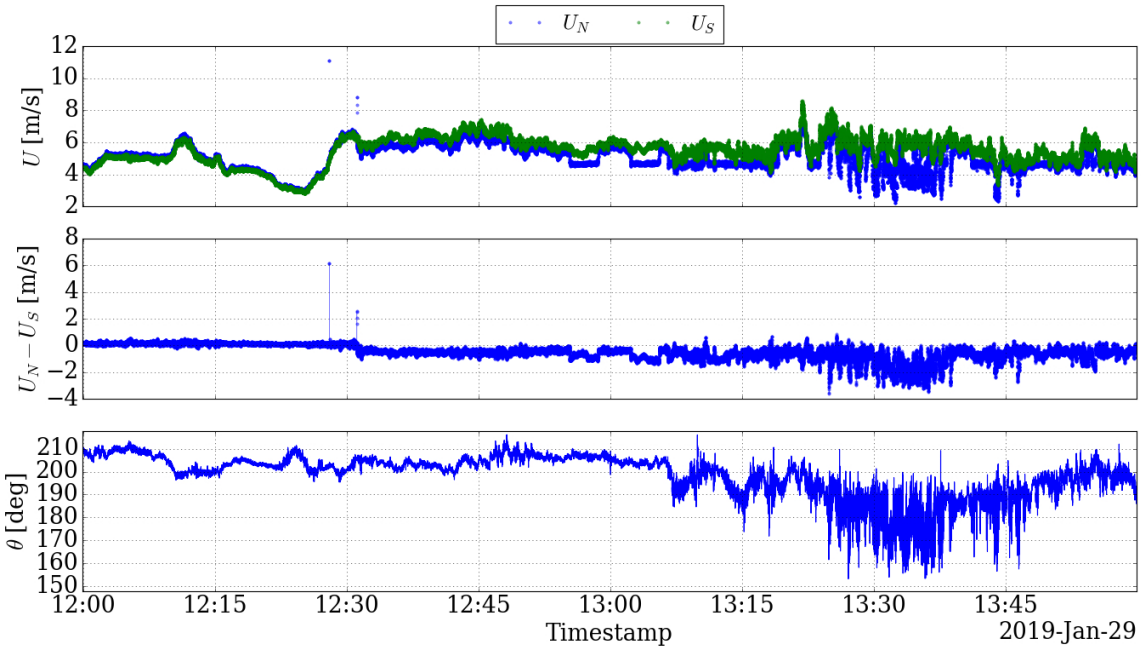


Figure 5.6: Time series of the start of the lightning event, with the wind speed measured from the cup anemometer pair at 178M and direction at 175m at LMN, sampled at 10Hz.

indicated periods did not coincide with icing events. This means that the atmospheric data can give an indication that the atmospheric conditions were appropriate and that further inspection of the times series may reveal icing events, which is the present desired outcome.

$$p_v = 0.615RH \exp \frac{17.27T}{237.7 + T} \tag{5.1}$$

$$T_d = \frac{\log\left(\frac{p_v}{0.6105} b\right)}{a - \log\left(\frac{p_v}{0.6105}\right)} \tag{5.2}$$

Inspection of the mast diagram shows that no instruments that measure relative humidity are present at 178m, but are located at 103m and 241m. Relative humidity is slightly higher at 241m than at 103m, since the air is less dense, resulting in more conservative calculations. It is important to note that this humidity is only used for the qualitative indication of an icing event. The timestamps of potential icing events are identified from the unfiltered time series, shown in Figure 5.7, where all red markers indicate such conditions.

Marked timestamps where icing may occur can be inspected to reveal periods where the cup anemometers may have been partially frozen. An example of such a time series is shown in Figure 5.8, where the ten minute averages of wind speed, wind direction and temperature are displayed in regions that were marked in Figure 5.7. It can be seen that the both anemometers slow down during the night of March 12th, when the ambient

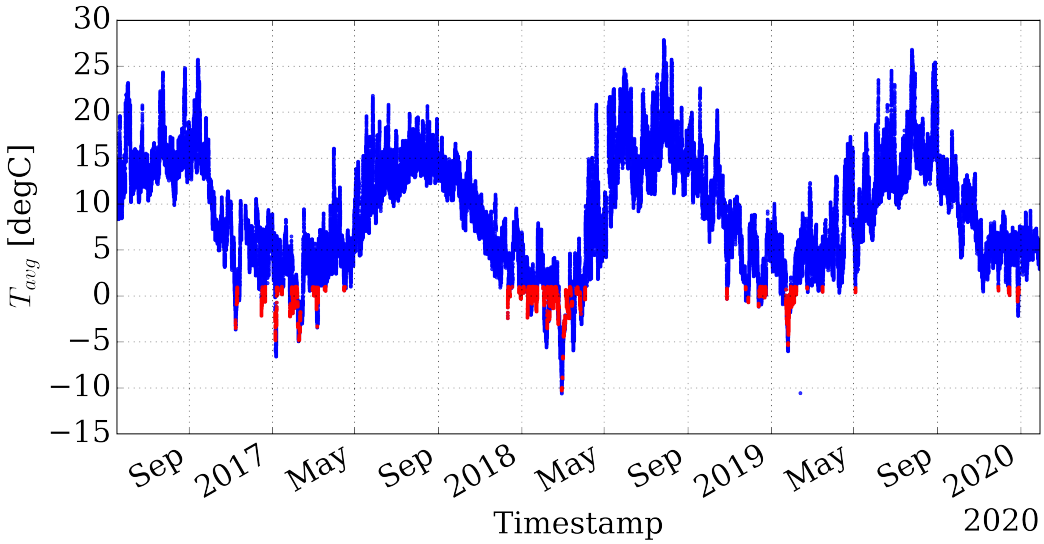


Figure 5.7: Unfiltered time series of temperature at 175m, where potential icing events according to the Magnus-Tetens equations (Heldman & Moraru, 2010) are indicated in red.

temperature is below freezing, and both anemometers completely stop right before 09:00. As the temperature rises, the northern anemometer seemingly starts unfreezing as the signal picks backs up around 18:00. After three hours, the southern anemometer also seemingly unfreezes and starts rotating with a similar frequency as the northern cup anemometer.

The event shown above can be classified as anomalous, most likely due to icing. Since anomaly detection is performed at a 10Hz scale, a closer inspection of the start of the event is conducted, when the southern cup freezes right before the northern cup does as well. These time series are shown in Figure 5.9 and further indicate that icing may be present, as the wind direction signal that is retrieved from the sonic anemometer right below the collocated instruments also gives a flat signal. The event can also be identified from the residual between the cups. More icing events have been identified and performance on such series are analyzed for validation purposes in Chapter 8.

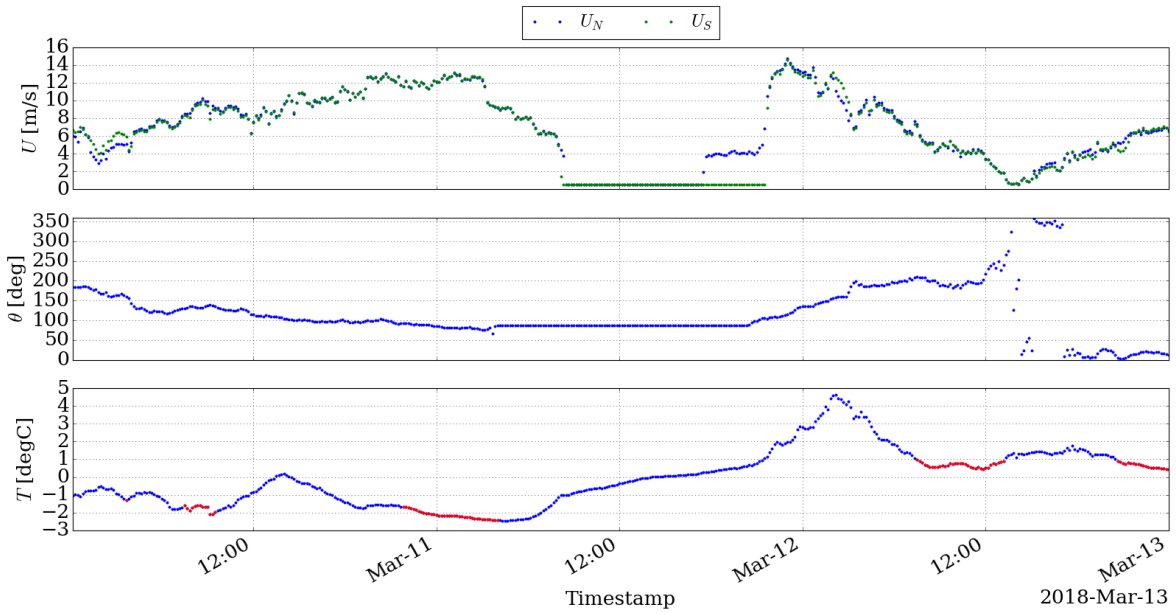


Figure 5.8: Time series of a supposed icing event, with the 10-minute averaged wind speed measured from the cup anemometer pair at 178M and temperature at 175m at LMN.

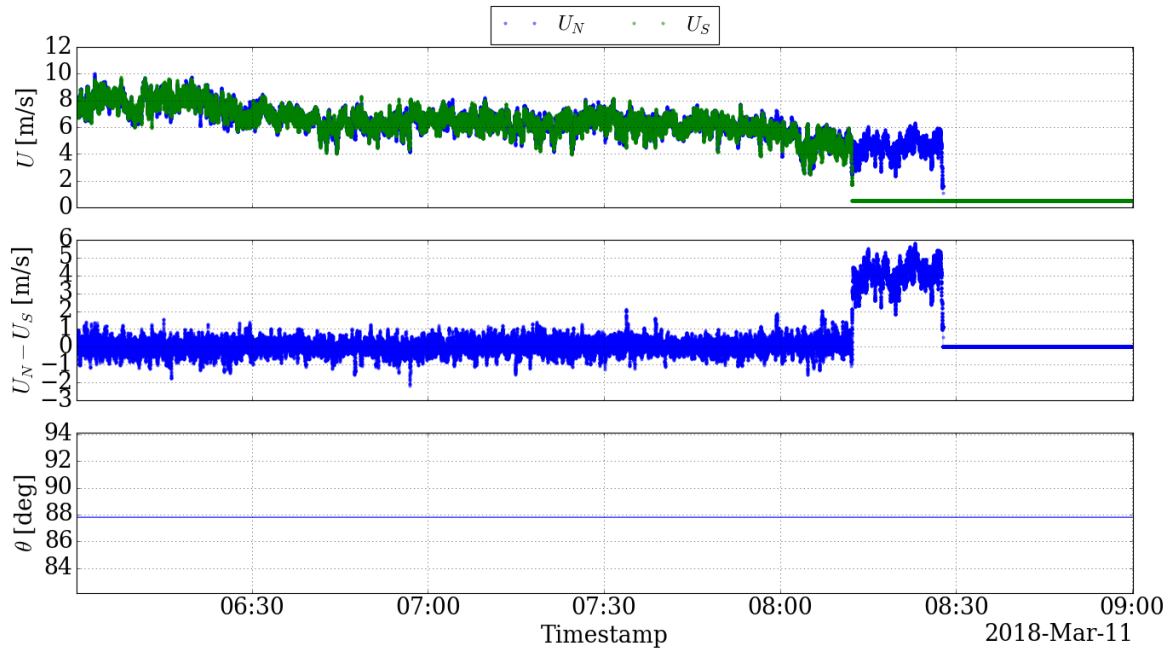


Figure 5.9: Time series of the start of an alleged icing event, with the wind speed measured from the cup anemometer pair at 178M and direction at 175m at LMN, sampled at 10Hz.

CHAPTER 6
Prediction Models

Prediction models that apply machine learning can be used as means for anomaly detection. The available batched data as presented in Chapter 5 define normal behaviour of the cup anemometer pair. The following models use these measurements to learn how to predict wind speeds under such normal conditions. A window of past wind speed measurements can be used to make predictions for the next time step, which can be compared to the actual measurements as they come in to judge whether or not the new, actual measurements are anomalous. The wind speed difference signal can be used for this one-step-ahead prediction, which keeps the dimensionality low and the prediction univariate whilst maintaining information from both anemometers. The wind direction is also included as input, due to the directional bias of the wind speed difference. Considering the fact that anomalous sources change the dynamic response of the cup, and therefore the wind speed measurements, it is argued that accurate estimation of normal data leads to inaccurate predictions of abnormal data. Due to the normalcy in the dataset, the models only learn how to predict under normal circumstances.

This chapter focuses on three different regression models of increasing complexity, preceded with a section about the preprocessing of the input data that will be fed into these models. The background and development of each model is elaborated upon together with proposed definitions of confidence intervals, providing quantitative measures of whether or not certain deviations from the models are abnormal or not. The following models use a subset of all available data, containing random samples from 2017.

6.1 Preprocessing

Preprocessing of the data is required to ensure that the regression models use the data in the right format to make predictions. Taking the northern and southern wind speed and wind direction as features involves two steps: transformation of the wind direction and normalization of all features. The former is required to ensure the latter process is meaningful. The transformation of the wind direction takes into account the periodic boundary that exists between 0-360°. Averages of wind direction that are taken with values close to the periodic boundary can lead to inaccurate statistical moments. The average of 10° and 350° should be 0 or 360°, whereas half of their sum equals 180°. This periodicity of the wind direction is taken into account by using the sine and cosine of the wind direction as features, which can be used to decompose any wind direction into two unit vectors. The statistics of the sine and cosine projections can be used without worrying about the periodic nature.

Normalizing input data is common practice for regression problems, especially for data with different features that are not of the same order of magnitude (Herlau et al.,

2019). Wind directions, for example, can reach up to 360°, whereas wind speeds above 40m/s are extremely rare. A 1m/s change in wind speed is thus much more significant than a 1° change in wind direction, which needs to be taken into account. The features are scaled by subtracting the mean and dividing the remainder by the standard deviation, as shown in (6.1), where X describes a feature and $\hat{X}$ the normalized feature. This feature scaling results in a data set with zero mean and unit variance, regardless of the original order of magnitude.

Feature scaling should always be applied over all data in the training set. The training data have been divided up into batches since not all data stored locally at once. The mean and standard deviation of both wind speed and direction features need to be calculated incrementally. This is done through the scikit-learn implementation of the `partial_fit` method that belongs to the StandardScaler class (Pedregosa et al., 2011). This implementation can transform any given dataset $\boldsymbol{X}$ exactly as shown in (6.1). The mean and standard deviation can incrementally be updated through the `partial_fit` method, which makes use of (6.2) and (6.3) (Chan et al., 1983). In these equations, an existing mean and variation μ_X and σ_X^2, based on a sample size of N points, can be combined with an extension of M samples belonging to the set $\boldsymbol{Y}$. The scaling can be applied to any given input that uses the northern and southern wind speed, as well as the sine and cosine projection of the wind direction, once every batch has been passed along such that the total mean and variance of the training set features are known.

$$\hat{\boldsymbol{X}} = \frac{\boldsymbol{X} - \mu_X}{\sigma_X} \tag{6.1}$$

$$\mu_{X,Y} = \frac{1}{N+M}\left(N\mu_X + \sum_{i=1}^{M} y_i\right) = \frac{1}{N+M}\left(N\mu_X + M\mu_Y\right) \tag{6.2}$$

$$\sigma_{X,Y}^2 = \sigma_X^2 + \sum_{i=1}^{M}\left(y_i - \mu_Y\right)^2 + \frac{M}{N(N+M)}\left(\frac{N}{M}N\mu_X + \sum_{i=1}^{M} y_i\right)^2 \tag{6.3}$$

6.2 Stochastic Gradient Descent

Linear predictors have worked reasonably well on a single wind speed data streams, as shown by Hill and Minsker (2010). Linear models uses a set of weights to each input variable to linearly combine these variables into a single or multiple predictions. These weights are obtained through training. The general linear prediction model that will be used for the posed prediction problem is shown in (6.4), where c, a_i, $a_{\cos}$ and $a_{\sin}$ are the weights of the model that are to be trained, stored in column vector $\boldsymbol{a}$, and $\overline{\theta}$ is the average wind direction over the window w. The regression model makes use of the preprocessed, normalized features.

$$x^{(t+1)} = c + \sum_{i=1}^{w} a_i x^{(t-i+1)} + a_{\cos} \cos\left(\theta^{(t-i+1)}\right) + a_{\sin} \sin\left(\theta^{(t-i+1)}\right)$$
$$= \begin{bmatrix} 1 & \boldsymbol{x}^{(t)} & \cos\left(\boldsymbol{\theta}^{(t)}\right) & \sin\left(\boldsymbol{\theta}^{(t)}\right) \end{bmatrix} \boldsymbol{a} \quad (6.4)$$

From the above expressions, a loss function that quantifies the misfit of the model is to be chosen and minimized. This results in an optimal set of weights $\boldsymbol{a}^*$ for the posed optimization problem. The most common loss function for linear regression is the squared error. A regularization term λ is introduced to penalize the complexity of the model. This, for example, can be accomplished through inclusion of the so-called $L1$- or $L2$-norm. The former tends to prefer sparse solutions, where features are eliminated for predictions, whereas the latter provides solutions where certain features may have weights that are approximately zero (Herlau et al., 2019). Small changes may result in completely different weight selections for the $L1$-norm, and therefore the $L2$-norm will be used. The total loss function can then be described by (6.5). Here, $\tilde{\boldsymbol{X}}$ is the normalized matrix, containing all possible windows of normal behavior subtracted by their column mean and divided by the standard deviation of the column, and $\tilde{\boldsymbol{y}}$ the column vector of every corresponding x_{i+1} subtracted by the mean. Note that normalization has to be applied as the weights are relative to the order of magnitude of the features. The optimum set up weights can then simply be found through (6.6). The analytical solution can easily be constructed from this by taking the derivative of the loss function with respect to the weights, as shown in (6.7). The resulting solution is depicted in (6.8).

$$E(\boldsymbol{a}, \lambda) = \left|\left|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{a}\right|\right|^2 + \lambda \boldsymbol{a}^T \boldsymbol{a} \quad (6.5)$$
$$\boldsymbol{a}^* = \arg\min_{\boldsymbol{a}} E\left(\boldsymbol{a}, \lambda\right) \quad (6.6)$$
$$\frac{\partial E}{\partial \boldsymbol{a}} = -\tilde{\boldsymbol{X}}^T \left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{a}^*\right) + \lambda \boldsymbol{a}^* = 0 \quad (6.7)$$
$$\boldsymbol{a}^* = \left(\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I}\right)^{-1} \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{y}} \quad (6.8)$$

The need for an optimization model to generate a solution for large datasets, i.e. large $\tilde{\boldsymbol{X}}$, is evident, as obtaining the inverse of such a matrix is computationally expensive. Instead, the optimization problem posed in (6.6) can be solved using Stochastic Gradient Descent (SGD). SGD is an extension to regular Gradient Descent (GD). GD is a method that aims to reduce the loss function by iteratively updating the weights. It does this by taking proportional steps along the direction of the most negative gradient. The weights are then updated according to (6.9) (Bottou, 2012), where this proportionality is expressed through the learning rate η. The learning rate decays throughout the iterations, meaning that relatively large steps are taken in the beginning to speed up the algorithms, but smaller steps are taken close to convergence to avoid overshooting. SGD follows from GD by picking a random data object from the training set to compute $E(\boldsymbol{a}, \lambda)$, compared to using the full set of objects. For every batch that is presented, the SGD model will

update the weights according to (6.9) by calculating the mean gradients of the batch and update the weights accordingly. The learning rate decays with each iteration to avoid oscillations in the loss function and a sub-optimal set of weights. The GD method is said to have found a minimum when the loss function does not decrease more than a threshold `tol` of five iterations in a row.

$$\boldsymbol{a}^{(i+1)} = \boldsymbol{a}^{(i)} - \eta \left(\frac{\partial E(\boldsymbol{a}, \lambda)}{\partial \boldsymbol{a}} \right) \tag{6.9}$$

6.2.1 Hyperparameter tuning

The SGD implementation for python is given by the scikit-learn module (Pedregosa et al., 2011). From the equations presented above, one can recognize several hyperparameters that require tuning, such that the most optimal solution can be found. These hyperparameters are as follows: window size w, regularization parameter λ, initial learning rate η_0 and the learning rate decay parameter t. Furthermore, the tolerance must be tuned such that an optimal solution is found within minimal computational expense. These parameters are tuned from two cross-validations (CVs). K-fold CV splits the available data up in K approximately equal sizes of chunk, e.g. 5-fold CV creates five splits with 20% of the data. The SGD models are therefore trained five times on 80% of the data and tested on the remaining 20%. As such, more general conclusions can be made as the results of a single fold have a dependency on the training set.

Two seperate 5-fold CVs are used to determine a useful set of hyperparameters for the linear model. Since the number of possible combinations per CV are reduced, this strategy ultimately reduces the number of times the model has to be trained. This choice can be further justified in that it is not of interest within the context of this work to find the perfect combination of hyperparameters that leads to the most optimal solution, but rather to find a combination leading to an effective tool for anomaly detection. It is for that same reason that effective orders of magnitude are sought after instead of a fine grid that calculates the most optimal value of each hyperparameter.

The variation of hyperparameters in both CVs are displayed in Table 6.1. The first CV seeks an optimal set of window sizes and a sufficient tolerance for several levels of regularization. The second CV aims to find the optimal learning parameters from a large range of regularization parameters and an order of magnitude above and below the default values of scikit-learn for both the initial learning rate and the decay parameter.

All parameters are tuned on a smaller subset of the data. SGD coefficients have empirically been shown to converge after approximately one million samples (Pedregosa et al., 2011). CVs are therefore conducted with 1500 10-minute batches, randomly sampled from the valid batches from 2017. This random sampling will remove any dependence on seasonal effects or atmospheric conditions. The performance of each model is measured through the mean squared error (MSE) between the predictions on the test set and the actual measured difference between the cups. The results for the first CV are shown in Figure 6.1.

Table 6.1: Model parameters that are used and varied for the two 5-fold CV runs.

Parameter	**CV 1**	**CV 2**
Window size w	[1, 3, 5]	w_{opt}
Tolerance `tol`	[10^{-1}, 10^{-3}, 10^{-5}]	$\texttt{tol}_{opt}$
Regularization α	[10^{-2}, 10^{-4}, 10^{-6}]	[10^{-1},..., 10^{-6}]
Learning rate η_0	0.01	[0.1, 0.01, 0.001]
Decay rate x	0.25	[1.0, 0.25, 0.05]

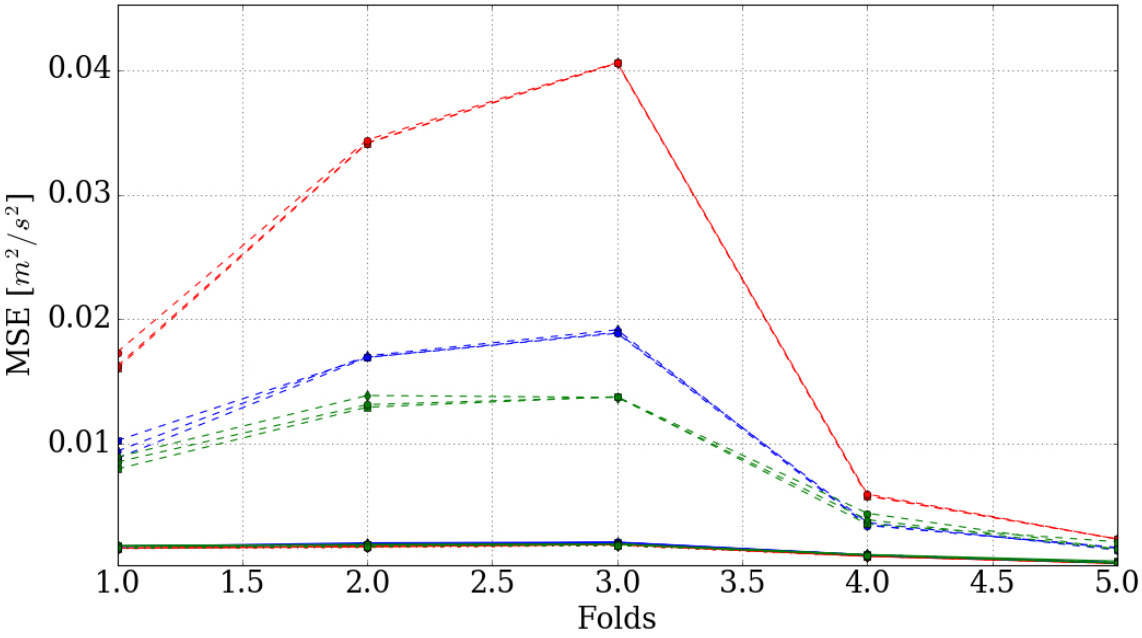


Figure 6.1: 5-fold cross validation of the SGD regressor, trained on 1500 10-minute batches that are randomly sampled from the available 2017 data. The models show variation in window size (1 = red, 3 = blue, 5 = green), regularization (10^{-2} = squares, 10^{-4} = diamonds, 10^{-6} = circles) and tolerance (10^{-2} = dashed, 10^{-3} = solid, 10^{-5} = alternated).

From Figure 6.1, it is evident that a tolerance of 10^{-2} is clearly insufficient, as their MSEs over all folds are significantly higher than both other options. It is also noted that no real improvements are found by decreasing the tolerance from 10^{-3} to 10^{-5}, and thus it is concluded that a tolerance of 10^{-3} is both sufficient and computationally efficient. Outside of models with insufficient tolerance values, window sizes of 1 are most accurate, and adding more points seems to increase the MSE. Including more information through larger window sizes deteriorates the effectiveness over all folds. These results are consistent over the variety of regularization parameters.

With the window size and tolerance fixed, the optimal order of learning parameters is sought after with the second CV. A similar procedure is taken using the parameters in

the right column of Table 6.1. From the resulting MSE over all five folds, it becomes clear that the standard parameters worked well for the problem at hand. Furthermore, models with almost no regularization seem to be favored. The coefficients of the optimal model are tracked throughout the iterations and presented in Figure 6.2, in which it can be seen that the model weights clearly converged. Interestingly, the influence of the wind direction is negligible. The northern and southern wind speeds have a positive and negative coefficient of similar magnitude respectively. This is as expected, as the wind speed difference is defined as $U_N - U_S$. The final results of the parameter tuning for SGD can be found in Table 6.2.

6.2.2 Persistence model

An important observation can be made from the coefficients of the most optimal model. The weights that are corresponding to the northern and southern wind speed measurements are the inverse of the applied scaling, meaning that this linear model creates a naive prediction by copying the last observed difference. This type of model is called a persistence model and is typically used as a baseline for comparison to a more complicated model.

In order to use the persistence model as means for anomaly detection, a degree of abnormality has to be established. For a persistence model, thresholds will be established for the predicted wind speed difference, meaning that an observed difference between the cup anemometers can not vary more than a set amount between 0.1s or 10Hz. These thresholds are set by ensuring that most of the data points are captured within these limits. The relative number of occurrences of residuals between the predicted and measured wind speed differences are displayed in Figure 6.3.

Most noticeably, a residual of 0.0m/s occurs far more often than any other residual. The histogram displays no skewness and larger magnitudes of residuals between predictions and measurements are less common than residuals closer to zero. An explanation for repeated wind speed difference can be found in the signal generator of the cup anemometer. The P2546A cup anemometer has a magnetic switch that opens and closes twice per revolution, and the wind speed is converted through timing of one rotation. The mean gain and offset of this type of anemometer is $0.620(m/s)/Hz$ and $0.21m/s$ respectively (Wind Sensor, no date). If an anemometer is rotating at 10Hz, it corresponds to approximately 6.4m/s. Any measurement below this wind speed will cause the anemometer to rotate slower than its sampling rate, and thus might cause it to repeat its last measurement. This also gives some insight as to why the SGD converged to a persistence model at this sampling rate. In order to create quantile-based thresholds, the relative frequencies are presented in their empirical cumulative density function (cdf), shown in Figure 6.4.

In Figure 6.4, the 2.5% and 97.5% percentiles are denoted by the green lines and the 0.25% and 99.75% percentiles are denoted by the red lines. The repeated measurements, resulting in 0.0m/s residuals, account for less than 10% of the data. 95% of all data is captured within ± 0.185m/s and 99.5% is captured within ±0.37m/s of the predictions. A 95% quantile for normality is used to attempt to find the middle ground between minimizing the FPR and minimizing the number of anomalies that are not flagged, or

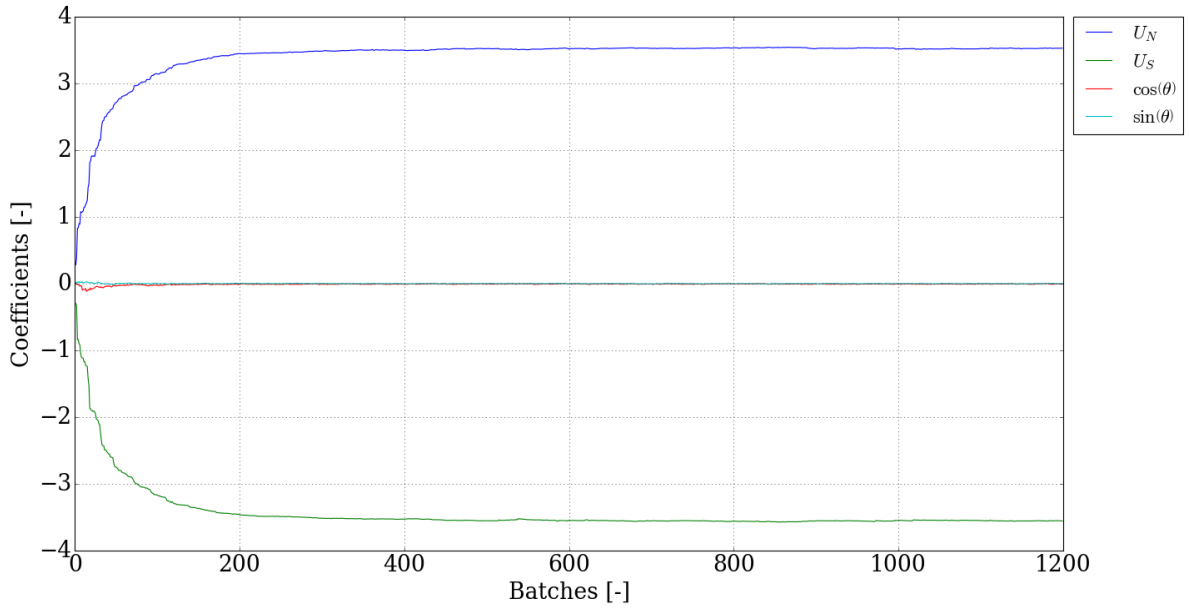


Figure 6.2: Trained coefficients on the first fold for the most optimal SGD parameters as determined across all folds during the second CV.

Table 6.2: Final parameters and settings used for SGD.

Parameters	**Value**
Window size w	1
Tolerance `tol`	10^{-3}
Regularization α	10^{-6}
Learning rate η_0	0.01
Decay rate x	0.25

false negative rate (FNR). It is important to note that this is based on a quantative approach, since no true negatives are known.

6.3 Stochastic Linear Processes

One of the properties of time series data is that points that are close in time are correlated. Similar to the spatial correlation of two cup anemometers that are used for anomaly detection, measurements that are taken close in time have temporal correlations. Implementations of stochastic linear processes aim to model the time series such that the residuals between observed and predicted time series are white noise (uncorrelated with zero mean).

The simplest of these models follows a naive prediction in which the model simply assumes the value of the next time step equals its preceding value. This corresponds to the persistence model. The error ϵ_{t+1} associated with this prediction $\hat{Y}_{t+1}$ can be determined

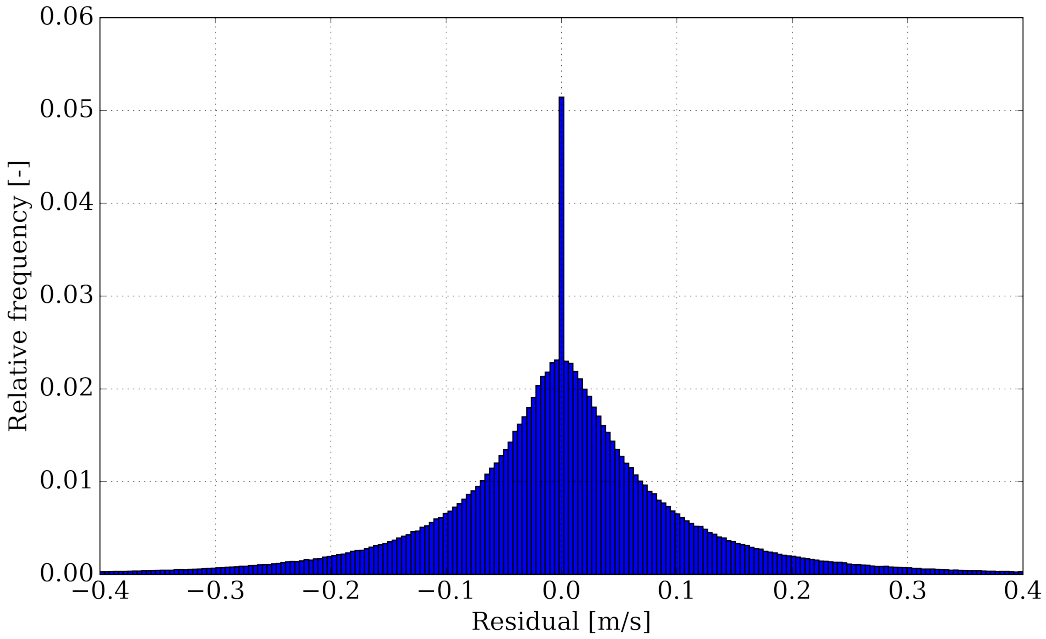


Figure 6.3: Relative frequency of the observed residuals between the predictions of the persistence model and the actual measurements from 1500 sets of 10-minute batches.

once the actual value Y_{t+1} is known, where the time series $\{\epsilon_t\}$ are uncorrelated variables with zero mean (white noise). A combination of the naive prediction and white noise creates the Random Walk (RW), as indicated in (6.10) (Madsen, 2007). Another way of writing the RW model is through an integration of noise terms, as shown on the right hand side.

$$Y_{t+1} = Y_t + \epsilon_t = \epsilon_t + \epsilon_{t-1} + ... \tag{6.10}$$

The RW model can grow unbounded, as all noise terms are random uncorrelated samples with zero mean. Therefore, RW describes a non-stationary process. Other commonly used linear processes are the Auto Regressive (AR(p)) process and the Moving Average (MA(q)) process (Madsen, 2007). An AR process, seen in (6.11), predicts the new value based on p previous observations. A MA process, seen in (6.12), uses the q previous error terms to predict the next value Y_t. A combination of the two can also be made, which is referred to as the ARMA(p, q) process, as depicted in (6.13). These models can all be used to describe stationary processes.

$$Y_t + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} = \epsilon_t \tag{6.11}$$

$$Y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} \tag{6.12}$$

$$Y_t + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} \tag{6.13}$$

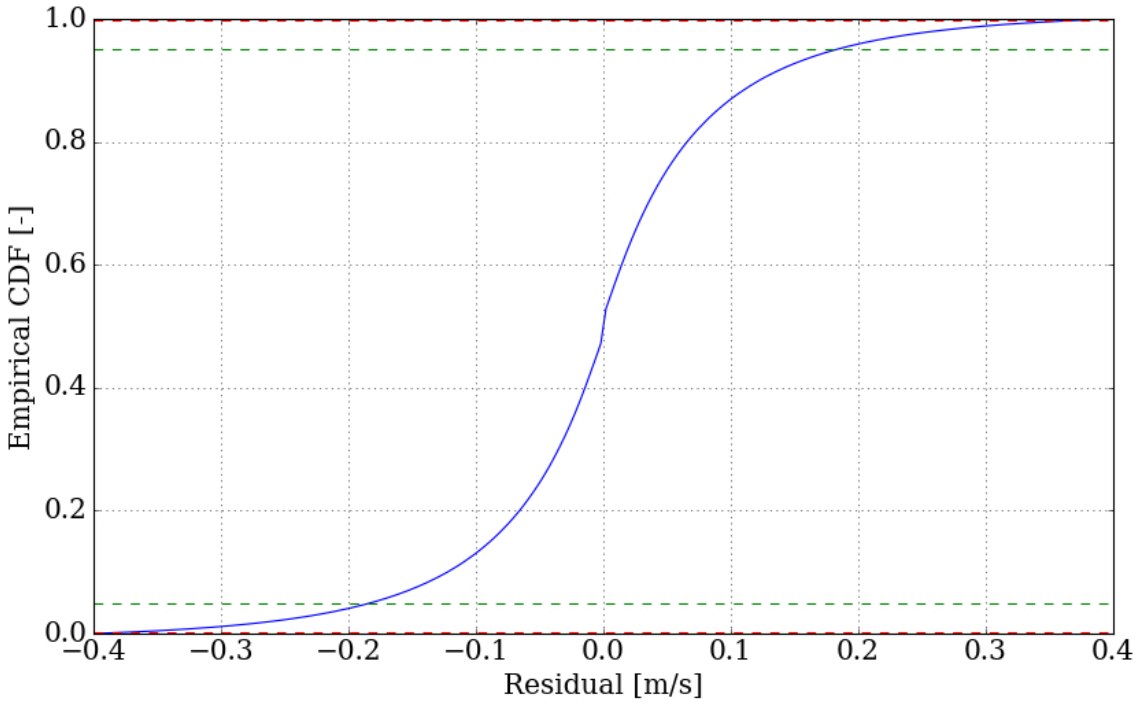


Figure 6.4: Empirical CDF of the observed residuals between the predictions of the persistence model and the actual measurements from 1500 sets of 10-minute batches. The green lines denote a 95% interval centered around 50%, the red lines for 99.5%.

Time varying behaviour is often characterized by non-stationary processes. In order to make the ARMA model suitable for these processes, the Auto Regressive Integrated Moving Average (ARIMA(p, d, q)) model is introduced. The ARIMA model differences the time series first, which removes trends from the data. As a result, a stationary signal is obtained where the ARMA model works. Most methods use either first order difference ($y_t = Y_t - Y_{t-1}$) or second order difference ($y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$). From the ARIMA model, the other models can be obtained, such as RW (ARIMA(0,1,0)), AR (ARIMA(1,0,0)), MA (ARIMA(0,0,1)), and ARMA (ARIMA(1,0,1)).

6.3.1 Identification ARIMA inputs

Identification of the ARIMA inputs, p, d and q, is done through autocorrelation and partial autocorrelation analysis. Autocorrelation describes the correlation between a signal and a lagged, or shifted, copy of its own and is defined as (6.14). Here, the covariance the lag k describes the lag between the signal y_t with length n and its copy. The sample mean of the signal is denoted by $\overline{y}$. The definition is interpreted as the ratio between the sample covariance between the two signals and their respective sample variances.

Autocorrelations contain influences from previous k points. The partial autocorrelations at lag k , ϕ_{kk}, can be determined recursively through the Yule-Walker equations, as

shown in (6.15) (Madsen, 2007). Lastly, confidence bands are established for the (partial) autocorrelation as in (6.16), based on a 95% band for a Gaussian distribution.

$$\rho(k) = \frac{\frac{1}{n-k}\sum_{t=k+1}^{n}(y_t - \overline{y})(y_{t-k} - \overline{y})}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \overline{y})^2}\sqrt{\frac{1}{n-k}\sum_{t=k+1}^{n}(y_{t-k} - \overline{y})^2}} \tag{6.14}$$

$$\begin{bmatrix} \rho(0) & \rho(1) & \dots & \rho(k-1) \\ \rho(1) & \rho(0) & \dots & \rho(k-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(k-1) & \rho(k-2) & \dots & \rho(0) \end{bmatrix} \begin{bmatrix} \phi_{k1} \\ \phi_{k1} \\ \vdots \\ \phi_{kk} \end{bmatrix} = \begin{bmatrix} \rho_0 \\ \rho_1 \\ \vdots \\ \rho_k \end{bmatrix} \tag{6.15}$$

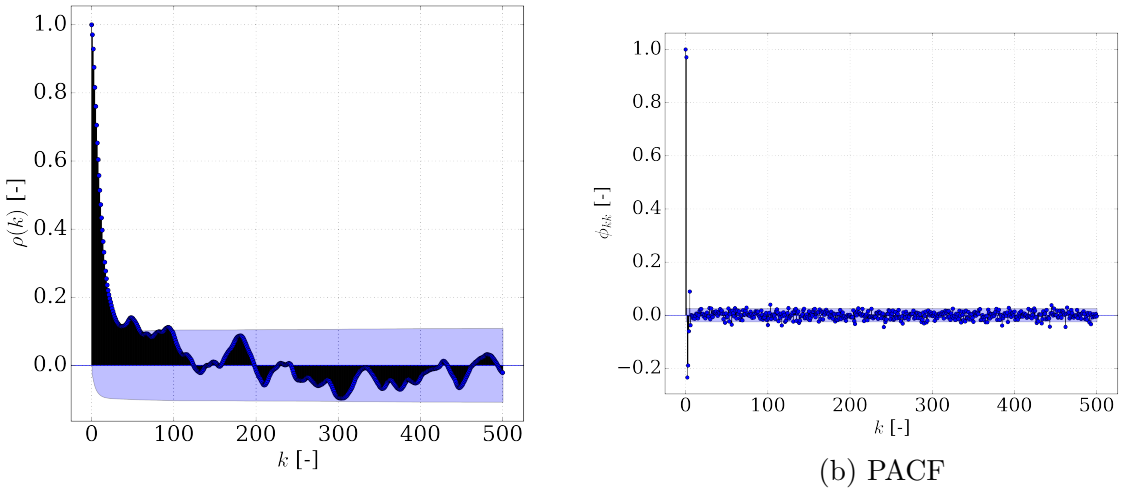
$$band = \pm z_{97.5\%}\sqrt{\frac{1}{n}\left(1 + 2\sum_{i=1}^{k}\rho(i)^2\right)} \tag{6.16}$$

The degree of differencing d has to be determined first. According to Box et al. (2015), non-stationary processes can be identified through a very slowly decaying autocorrelation function (ACF). This slow decay becomes apparent when inspecting the ACF from the difference signal of a sample 10-minute batch, shown in Figure 6.5. Here, a sharp cut-off is observed with the partial autocorrelation function (PACF), whereas the ACF decreases slowly for increasing lags. This behaviour is consistent across all batches for $d = 0$. Increasing the order of differencing, setting $d = 1$, results in the ACF and PACF shown in Figure 6.6. A clear cut-off is now apparent in both graphs, with some correlation in the first couple of lags with no significant contributions from future lags. It is thus concluded that a $d = 1$ is a sufficient order of differencing, making the difference between the northern and southern wind speed stationary.

From the ACF and PACF plots, indications for an appropriate selection of the number of AR and MA terms, p and q, can be obtained. The required number of parameters depends on the cut-off in the ACF and PACF. For example, Figure 6.6a indicates that at the least the first three lags share significant correlation and therefore p may be chosen as three. The PACF, in Figure 6.6b, shows a relatively significant correlation up until the fourth lag. However, these results differ across the batches. All batches are subjected to a variety of ambient and flow conditions, which influence the underlying mechanism of the time series. In order to make an estimation of the required parameters, a grid search will be used with $d = 1$ and p and q ranging from one to six, which seems to suffice based on inspection of random samples.

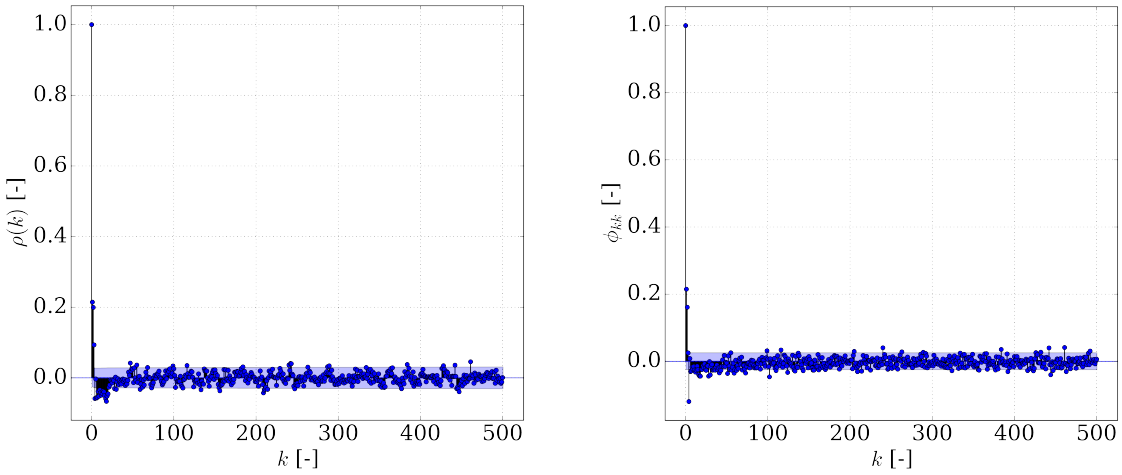
6.3.2 Estimation ARIMA parameters

When the ARIMA parameters (p, d, and q) are known, their coefficients require estimation such that the model can be used for predictions. In order to set up the ARIMA model and estimate the required parameters, the python module `statsmodels` (Seabold & Perktold, 2010) is used. The `statsmodels` provides a straightforward establishment of the ARIMA model, as well as means to estimation of the coefficients. It does so by optimizing the Maximum Likelihood $\hat{L}$, which gives a measure of the loss between the



(a) ACF (b) PACF

Figure 6.5: Autocorrelation and partial autocorrelation for the first 500 lags of the residual between the measured wind speed of the northern and southern wind speed, logged between 2016-05-18 17:54-18:04.



(a) ACF (b) PACF

Figure 6.6: Autocorrelation and partial autocorrelation for the first 500 lags of the differenced residual between the measured wind speed of the northern and southern wind speed, logged between 2016-05-18 17:54-18:04.

actual data and the predictions. For every estimation, information about the quality of the fit is provided as well. The relative quality of each model is evaluated through the Akaike information criterion (AIC) (Akaike, 1974). AIC is defined as in (6.17), where k is defined as the number of estimated parameters.

$$AIC = 2k - 2\ln\left(\hat{L}\right) \tag{6.17}$$

Following (6.17), a relatively lower AIC means that a better set of coefficients is found. The AIC is used to perform the grid search for the optimal combination of p and q. The same batches that were used for CV on the SGD model are used to train 1500 ARIMA models for every combination of p and q. For each batch, the minimum AIC for the set of parameters is logged, giving rise to the frequencies displayed in Figure 6.7.

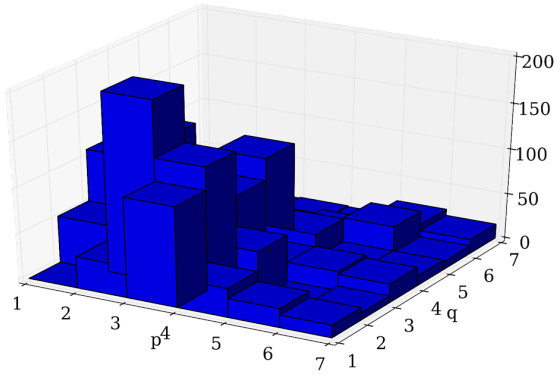
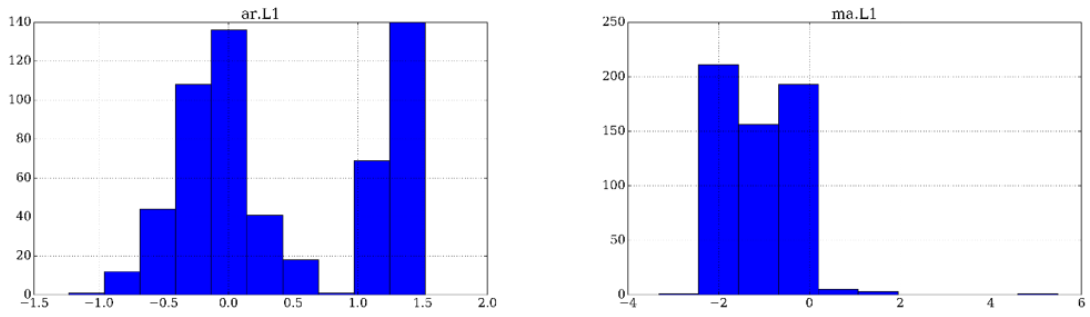


Figure 6.7: Optimal combination of p and q of ARIMA(p,1,q), trained on 1500 10-minute batches randomly sampled throughout 2017.

The combination (2,1,2) parameters most often result in the lowest AIC, but there are many batches that have their optimal parameters ranging between 2 and 4. The first fitted AR and MA coefficients from the (2,1,2) model are also logged. These results are denoted in Figure 6.8. The MA coefficient centers around a distinct mean, but the AR coefficient displays different behaviour, meaning that the batches can not be used for training due to their varying behaviour.

Instead, the longest sequence is used for training, since the batch approach does not lead to a set of conclusive coefficients. It can be argued that this time series experiences the largest variety of ambient and flow conditions. Novel approaches, such as a nested ARIMA approach (Sim et al., 2019), are not applicable in this case. Too many sizeable gaps between the sequences are present, leading to inaccurate and unpredictable time series. The longest sequence is observed from 18-10-2016 15:00 until 24-10-2016 19:50, meaning that six days worth of data is captured. A similar grid search reveals that ARIMA(2,1,3) fits best. The optimal coefficients for all terms are displayed in Table 6.3.



(a) (b)

Figure 6.8: Histogram of the first AR and MA coefficients of the 1500 trained ARIMA(2,1,2) models.

Note that among those coeffients, the variance σ^2 of the white noise is included as well. It is verified that this residual is also uncorrelated.

Table 6.3: Estimated ARIMA(2,1,3) parameters based on the uninterrupted sequence of measurements from 18-10-2016 15:00 until 24-10-2016 19:50

Parameter	**Value**
AR1	1.21
AR2	-0.27
MA1	-0.81
MA2	-0.15
MA3	-0.037
σ^2	$0.0152 m^2/s^2$

6.3.3 Model evaluation

The ARIMA(2,1,3) parameters are fitted based on one sequence, together with an estimation of the variance of the residuals. This variance can be used to construct confidence intervals, as the residuals of a well-designed ARIMA model are normally distributed. Since much of the data is not used for training, the capabilities of the model trained on one sequence can be tested on the remaining sequences and batches. First, an assessment can be made on the predictive capabilities. In Table 6.4, the MSE is given of the predictions on the folds that were also used for testing the SGD results. The ARIMA(2,1,3) has not been trained on this data, since it considers 1500 10-minute batches that were randomly sampled from 2017.

It can be seen that the model performs reasonably well, with lower MSE than the persistence model. Furthermore, the MSEs over all folds, with the exception of the last fold, lie very close to the expected variance. Closer inspection of certain batches

Table 6.4: Mean Squared Error of the ARIMA(2,1,3) model, evaluated on 1500 10-minute batches that are randomly sampled from all available data in 2017. All units are in m^2/s^2.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
ARIMA(2,1,3)	0.0154	0.0147	0.0155	0.0155	0.0164

results in three scenarios: one where the residuals closely resemble white noise, one where the kurtosis becomes too large and one where the distribution becomes uniform. The first case indicates that the built-in confidence intervals are accurate. These confidence intervals are similar to the thresholds introduced for the persistence model. The fitted variance of the residual of the ARIMA model is $0.015\text{m}^2/\text{s}^2$. Since it is assumed that the residual comes from white noise, an interval of four standard deviations centered around the prediction describes a 95% confidence interval. This corresponds to an interval of 0.24m/s, which is smaller than the persistence model, which has a two-sided interval of 0.37m/s.

6.4 Multilayer perceptron model

The methods described above both can both be considered linear models. In order to model non-linearties, a Multilayer Perceptron (MLP) is implemented. A MLP is identical to a feed forward Artifical Neural Network (ANN). ANNs have been shown to perform well in predicting non-linear behavior using non-linear activation functions, including corrections on overspeeding (Azorin-Molina et al., 2018). Similar to the linear prediction method, an ANN can be used to make a prediction about the wind speed difference with inputs from both wind speed signals and wind direction. According to Bishop (2005), a single layer feed forward ANN is sufficient to model non-linear behaviour. Feed forward networks have only one direction of information (input layer to hidden layers to output layer) and thus no feedback loops between the layers.

An ANN is built up by sets of neurons and is used within the context of supervised learning. These neurons are simple computing nodes that have weighed connections with each other at each layer. The most simple ANN has three layers: an input layer, a hidden layer and an output layer. Multiple hidden layers can be added to the architecture of an ANN model as well, allowing for the system to potentially learn non-linear patterns better, albeit with greater computational expense. For a feed forward ANN, all neurons are connected, as depicted in Figure 6.9.

ANNs pass information forward through "activation" on nodes. The nodes on the input layer are initialized by assigning the ith input value x_i to each input neuron i. These inputs are linearly combined on node j on the next layer k using the weights $\boldsymbol{w}_j^{(k)}$, which are to be learned. A constant can be added to this linear operation in case a bias term w_0 is used. This linear combination is said to be the activity a_j^k that is given to the respective neuron.

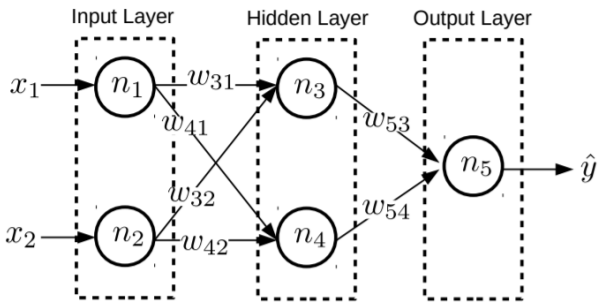


Figure 6.9: Simple ANN with three layers and two hidden nodes in the hidden layer (Herlau et al., 2019).

ANNs are able to model non-linearities through their activation functions, which are used to transform given activities. These transformed activities are then passed along further down the network. A popular choice as activation function is the Rectified Linear Unit (ReLU), shown in (6.18) (Goodfellow et al., 2016). Advantages of the ReLU function are its computational simplicity and its sparsity due to the possibility that activations are true zero values. The information that is passed forward from neuron j in layer k, $z_j^{(k)}$ can be expressed by (6.19). Note that, for the input layer, $\boldsymbol{z}^{(0)} = \boldsymbol{x}$. This process is repeated for as many layers are present until the output layer is reached. The activation $\boldsymbol{z}^{(k+1)}$ then becomes the output of the ANN.

$$h(x) = \max(x, 0) \tag{6.18}$$

$$z_j^{(k+1)} = h^{(k+1)}\left(w_{0j} + \left[\boldsymbol{z}^{(k)}\right]^T \boldsymbol{w}_j^{(k+1)}\right) = h^{(k+1)}\left(w_0 + \sum_{i=1}^{N} z_i^{(k)} w_{ji}^{(k+1)}\right) \tag{6.19}$$

The weights that are used to connect the neurons in the network are to be trained to give accurate predictions. The process to train these weights is similar to the SGD method. A cost function is to be defined, such as the MSE, which is minimized through gradient descent. It should be noted that the number of weights that are to be minimized is significantly less. Whereas the SGD model was trained with four features and thus four weights, an ANN with an identical feature space, four hidden neurons and one output neuron consists of 20 weights. Computing the gradients for this many weights, however, is too expensive. Gradients are therefore computed per layer through the chain rule by moving backwards through the network. This is also referred to as backpropagation (Goodfellow et al., 2016). The same out-of-memory approach can be used as well, meaning that the weights are updated after evaluation of a batch.

6.4.1 Hyperparameter tuning

The MLP model is implemented in python using pytorch (Paszke et al., 2019). The parameters of the MLP, including the number of layers, batch size, window size, input layer and neurons for each layer require tuning. Tuning all these parameters in one

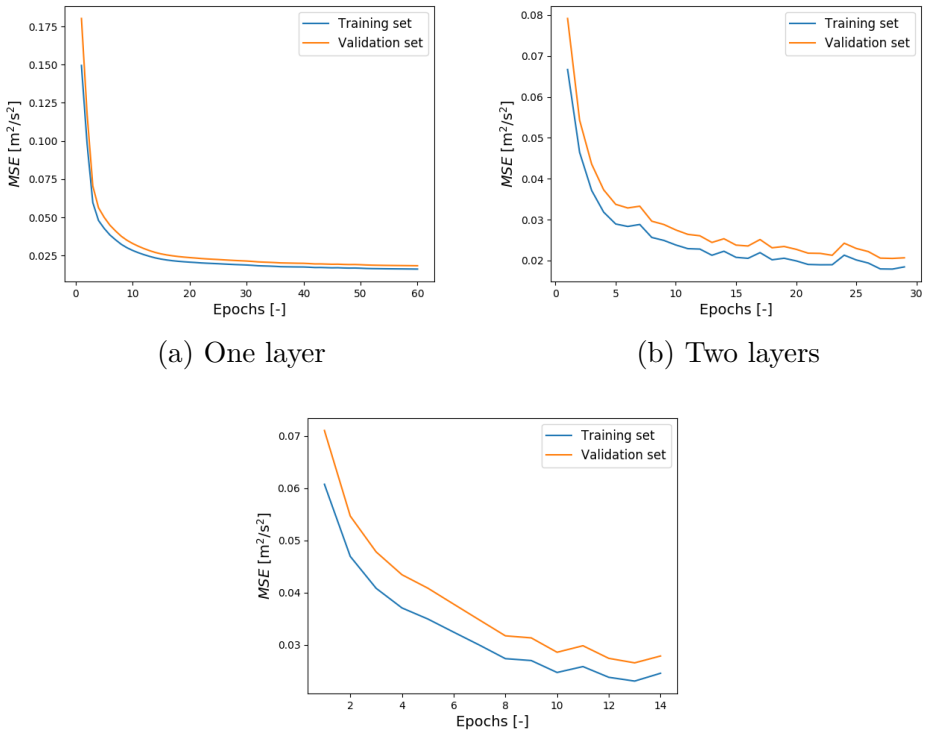
cross validation requires many models to be trained multiple times, which is extremely computationally expensive. For this reason, higher level parameters are tuned using learning curves, such as the number of layers and batch size. Once these are set, a CV is done to determine the number of neurons on each layer. Similar reasoning can be used for this approach as done with SGD, where parameters defining an efficient model are more important than a thorough search leading to the perfect solution. Therefore, searching grids are coarse and a 5-fold CV is justified for parameter tuning.

Learning curves are established by tracking the MSE on the training set and validation set as a function of epochs. The validation set is created by leaving out 10% of the keys of the training set after creating a training and test set. 10% is chosen since fewer batches indicates that the set will possess insufficient variation, giving a biased estimate, whereas a higher percentage is detrimental to the accuracy of the model due to an unnecessarily small training set. This validation set is used to ensure that the data is not overfitting. Overfitting means that the model learns how to predict on the training set, but fails to perform well on unseen data. A smoothly decreasing MSE is expected over multiple epochs on the training set, which is not guaranteed on the validation set. These validation curves can also be used to determine if convergence is achieved; if a significant decrease in MSE is observed over several epochs, the model is said to be optimized under its current settings. Validation curves of MSE are tracked for all trained MLP models and ideally are slightly higher when compared to the training set, meaning that the model performs slightly worse on unseen data but is able to capture the underlying mechanisms of the data well (Herlau et al., 2019).

First iterations of the model were trained with one, two and three layers to determine the optimal number of layers. Training is performed five times on the same 1500 random samples from 2017 as the previous models. For each layer, 75, 100, and 125 neurons were used with a window size of 5. The training curves' multiple layers with 100 neurons each are shown in Figure 6.10. It can be seen that multiple layers result in unwanted fluctuations in both the training and validation MSE. Furthermore, it is observed that training errors are significantly higher for an increasing number of layers. This result is consistent among different numbers of neurons and with comparable work in the literature (Hill & Minsker, 2010).

A similar analysis is performed on the effect of batch size. For 75, 100, and 125 neurons, different batch sizes are used. This includes 1-, 5-, and 10-minute batches. Their training curves are depicted in Figure 6.11. Here, it becomes apparent that the batch size does not affect the desired smooth training curves. 10-minute batches therefore become the most optimal set, as fewer batches have to be loaded in, making training more computationally efficient. It can be confirmed that the validation curve also shows desirable characteristics, having a slightly higher MSE per epoch than the training curve with a smooth converging trend.

Having set the number of layers and batch size, a 5-fold CV is performed to find the number of optimal number of neurons in the input layer (window size) and number of neurons in the hidden layer. The output layer always has one output node, as the difference is to be predicted. For the CV, the test set is used to construct the MSE. Training curves are kept to ensure that training goes smoothly and no unexpected



(a) One layer

(b) Two layers

(c) Three layers

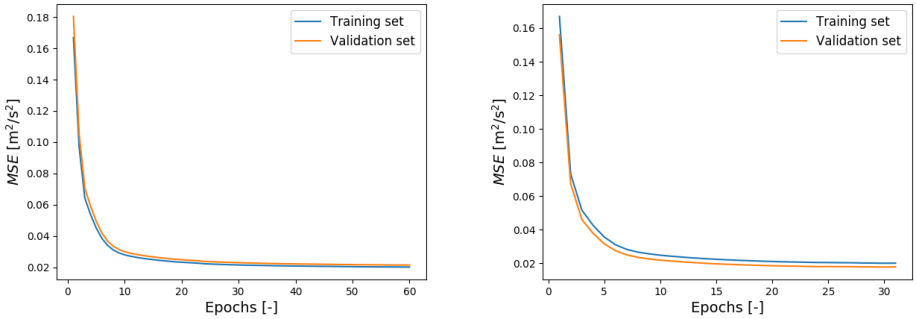
Figure 6.10: MSE on the training and validation set for the MLP with 100 neurons/layer, for different number of layers, trained on 1080 10-minute batches sampled from 2017.

behaviour is observed. From the previous iterations, it was observed that smaller window sizes and neurons produce more accurate predictions. The cross-validated parameters therefore are done for window sizes of 1, 2, 5, and 10 and for 25, 50, 75, and 100 neurons. The results of this CV can be seen in Figure 6.12.

From Figure 6.12, it becomes clear that a window size of two (and thus eight nodes in the input layer) gives better results with a lower MSE than the other window sizes. Having the last two points as inputs clearly provides more necessary information than one, but adding more information results in a higher MSE. It also becomes clear from Figure 6.12 that using 50 hidden neurons is necessary to make more accurate predictions. Using 100 neurons, accompanied with a window size of two, produces the best results and will thus be chosen for the final MLP model. An overview of all tuned hyperparameters can be seen in Table 6.5.

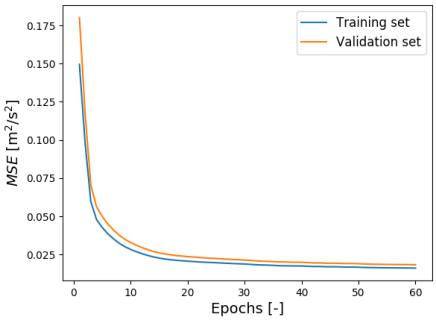
6.4.2 Confidence intervals

Confidence intervals for the MLP are created by introducing dropout to the model. A dropout layer can be described as nodes that either pass forward their given information or discard it. The dropout nodes, equal to the number of nodes in the preceding layer,



(a) One minute

(b) Five minutes



(c) Ten minutes

Figure 6.11: MSE on the training and validation set for the MLP with 100 in the single hidden layer, trained on 2160, 2160 and 1080 1-, 5- and 10-minute batches sampled from 2017 respectively.

each have an independent probability p to drop out. This also means that, on average, a fraction of p nodes are not used in the linear combinations on the neurons in the next layer. Dropout layers can be used to prevent overfitting (Srivastava et al., 2014), but have also been used to established confidence intervals (Gal & Ghahramani, 2014).

A major advantage of using dropout is they enable the establishment of dynamic confidence limits. The persistence and ARIMA model work with fixed limits. Dropout networks be used to calculate statistics of a prediction by making multiple predictions with the same input. Due to the independent dropout probabilities, these multiple passes vary and can be used to address the uncertainty in the made prediction. Cortes-Ciriano and Bender (2019) use this variation to calculate non-conformity scores of the ANN in the validation set and then use the standard deviation of multiple predictions of the same objects.

A similar approach can be taken to create a dynamic confidence interval through tuning of the dropout rate. By having a higher dropout rate, less information is passed forward, which results in a wider range of predictions for every pass through the network. This means that the dropout rate can be empirically tuned to give accurate confidence limits, i.e. a dropout rate can be selected such that, for a given confidence interval,

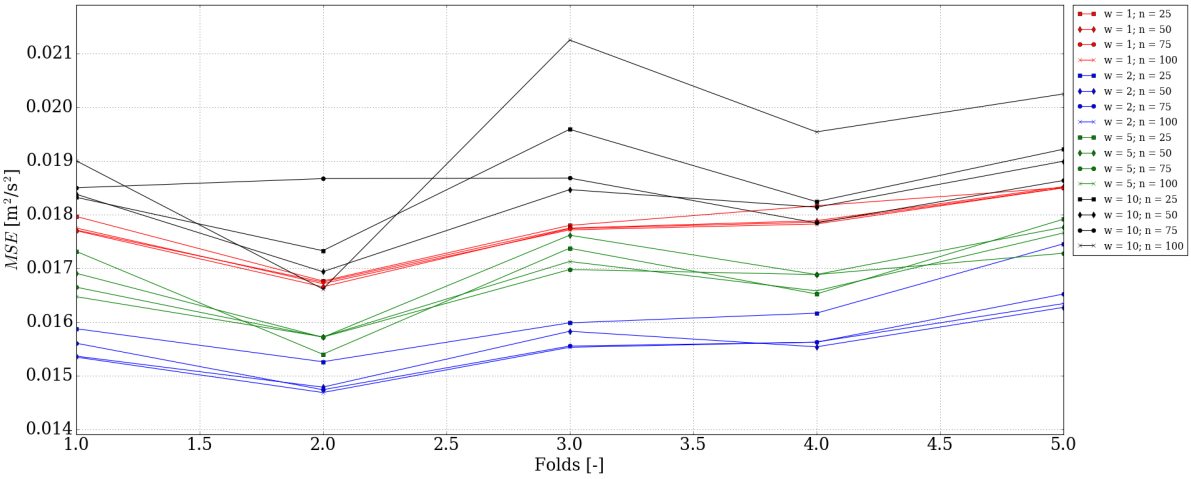


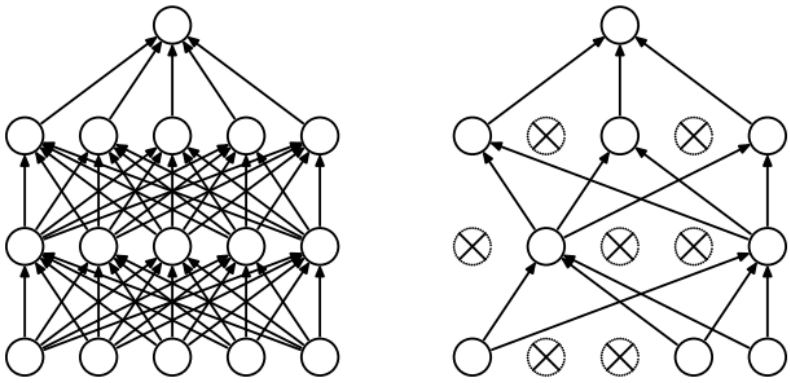
Figure 6.12: Cross validation of the MLP model for different input and hidden layer structures, based on 1500 randomly sampled 10-minute batches from 2017.

Table 6.5: Final parameters and settings used for MLP.

Parameters	**Value**
Window size	2
Batch size	10 minutes
Input nodes	8
Hidden nodes	100
Output nodes	1
Activation function	ReLU
Learning rate	10^{-4}

an identical percentage of measurements are inside the confidence intervals around the prediction. For example, when a 95% CI is chosen, 95 out of 100 measurements are within the the confidence limits and will thus not be flagged. This ideal dropout rate is found through empirical tuning with the following process:

1. For a given sequence with N_{obs} observations, $N_{obs} - 2$ predictions are made with no dropout, giving the most accurate predictions $\hat{\boldsymbol{x}}$.
2. A dropout layer with rate p is introduced between the hidden layer and the output layer, after which N_{passes} predictions are made for each observation, resulting in the matrix $\hat{X}$ with shape N_{passes} x $N_{obs} - 2$.
3. The absolute deviations with the most accurate predictions $\left\|\hat{\boldsymbol{x}} - \hat{X}\right\|$ are sorted from low to high in each column such that the observed quantiles can be found in each row. For example, the 95th row is said to empirically describe the 95% confidence interval for each of the $N_{obs} - 2$ predictions in case $N_{passes} = 100$.

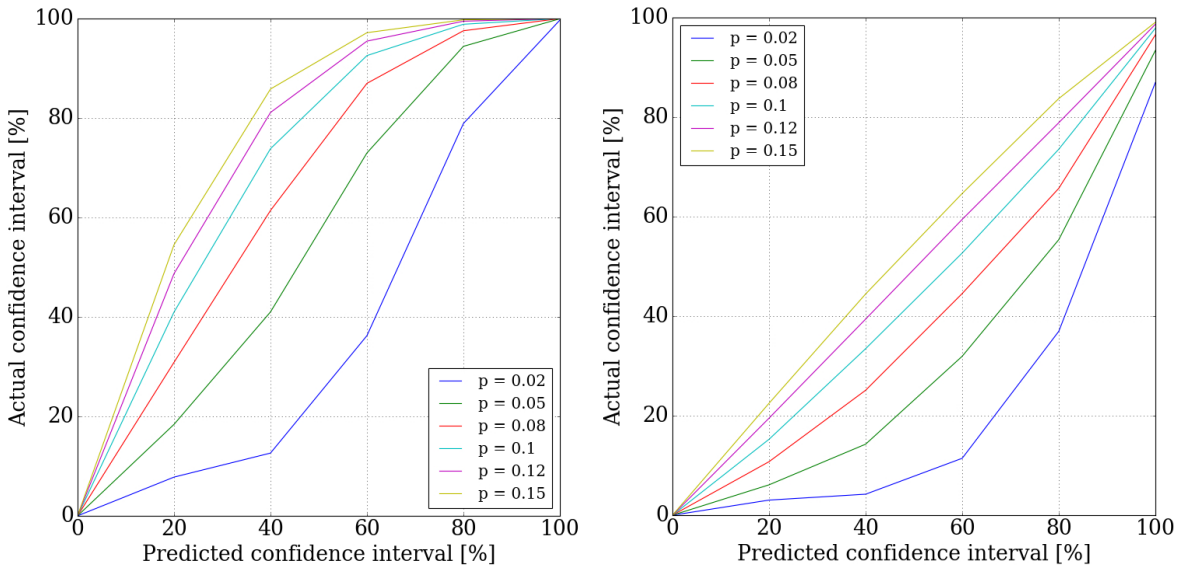


(a) Standard Neural Net (b) After applying dropout.

Figure 6.13: Schematic overview of a fully connected network on the left and a dropout network on the right (Srivastava et al., 2014).

4. From the $N_{obs} - 2$ obtained confidence intervals, the percentage of observed differences that fall inside the confidence limits of all predictions give the actual confidence interval.

The procedure applied above can be used with multiple dropout rates for which several predicted quantiles can be compared to their actual quantiles. The dropout rate for which the predicted confidence interval matches the actual confidence interval best is said to be the ideal dropout rate. The process described above is visualized for two sequences in Figure 6.14.



(a) Sequence from 01-01-2017 01:20 to 16:20. (b) Sequence from 23-12-2017 20:40 to 21:00.

Figure 6.14: Predicted and observed confidence intervals for several dropout rates for two different sequences.

Two distinct patterns can be observed in Figure 6.14. Figure 6.14a shows that the lowest dropout rate underestimates the lower range of observed confidence intervals, whereas it overestimates the highest confidence intervals. All higher dropout rates consistently overestimate the confidence intervals. Figure 6.14b shows that, for the same dropout rates, $p < 0.1$ results in systematic underestimations and $p > 0.1$ results in slight over estimations. A dropout rate of 0.1 seems to be ideal in this case, since it seems to predict the actual confidence limits well. Both plots do show that increasing the dropout rate results in a wider spread of predictions and thus larger confidence limits.

Closer investigation of both of sequences reveals that their major difference lies in the magnitude of the measured wind speeds. The sequence ranging from 01-01-2017 01:20 to 16:20 has an average horizontal wind speed of 11.1m/s with a maximum of 15.5m/s, whereas the other sequence has an average wind speed of 25.5m/s with a 35.7m/s maximum. The relationship between the ideal dropout rate and average wind speed is quantified by determining the former for all full sequences in 2017 and plotting them against the average wind speed as measured by the northern cup anemometer. Measurements that were affected by mast shadow were not used to make predictions with the model. The result is shown in Figure 6.15.

As expected, the optimal dropout increases as the average wind speed increases. It can thus also be concluded that the difference between both anemometers fluctuates more for higher wind speeds since more dropout is required to capture these fluctuations within the confidence limits. Low wind speeds therefore require very little dropout, as observed. These smaller fluctuations are in accordance with earlier observations, where measurements within this region of lower wind speeds may be repeated. These repeating measurements also explain why almost all dropout rates in Figure 6.14a overestimate the confidence intervals. High wind speeds require more dropout, but occur less frequently. It is important to select one dropout rate, as a fixed definition of anomalous data is required for the static model that is being made. Overall, a dropout rate of p=0.05 seems a suitable compromise; it is most optimal for the widest variety of wind speeds. Wind speeds below 10m/s will overestimate the confidence intervals for normal data, meaning that more false positives will not be flagged in this range. High wind speeds (above 22 m/s) will be underestimated, but these do not occur often enough to justify raising the dropout rate.

Using $p = 0.05$, these confidence intervals can be visualized for any time series. For the same sequences in Figure 6.14, a minute's worth of data is visualized in Figure 6.16. Here, it can be observed that the confidence intervals adapt and have different orders of magnitude that scale with the input horizontal wind speed.

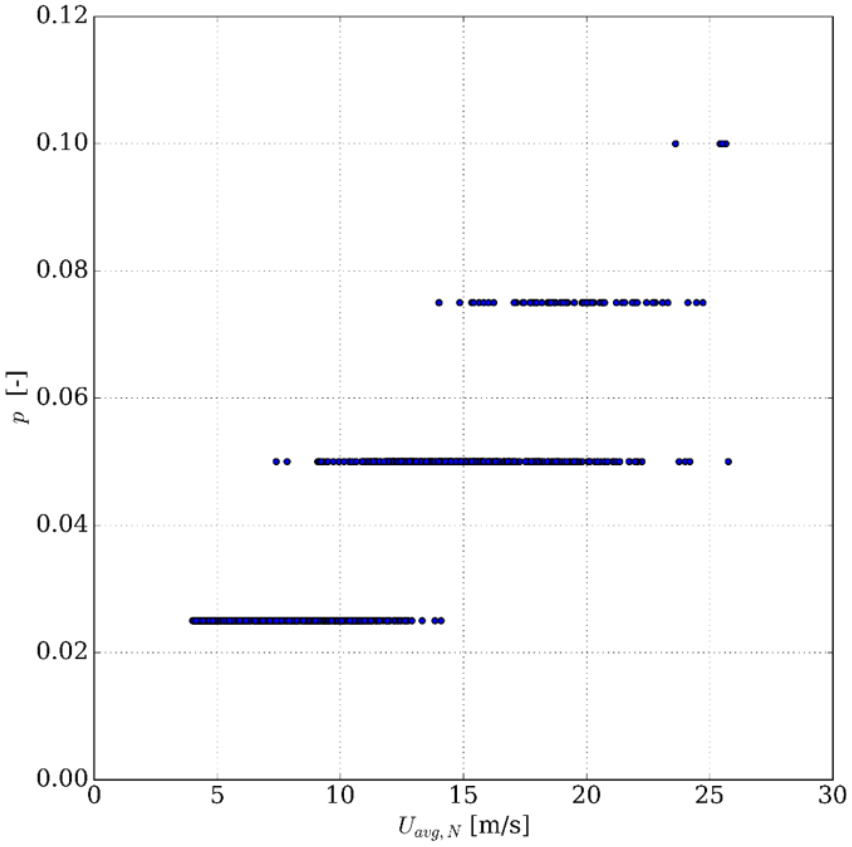
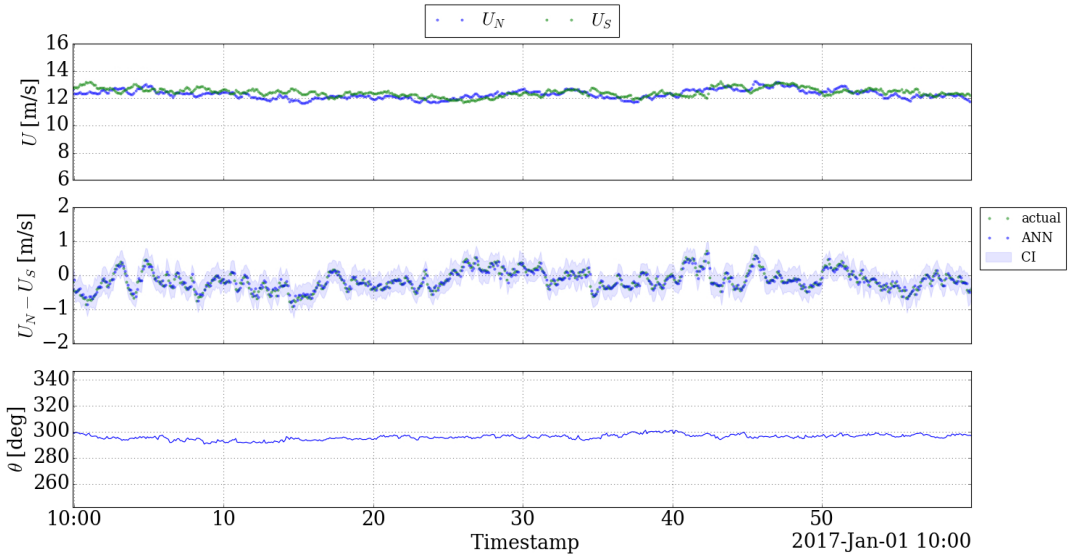
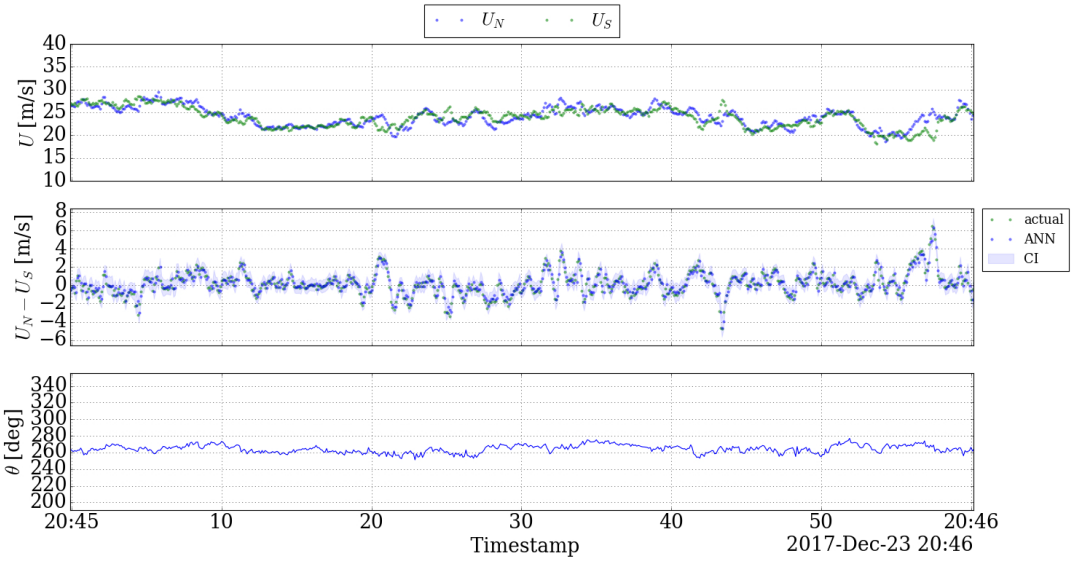


Figure 6.15: Ideal dropout rates, for which the predicted and observed confidence intervals align best, as a function of average wind speed for all uninterrupted sequences of normal data in 2017.



(a)



(b)

Figure 6.16: MLP prediction model with adaptive confidence intervals applied to two time series with different magnitudes in wind speed.

CHAPTER 7
Model comparison & selection

The persistence, ARIMA and MLP anomaly detection models described in the previous chapter were trained using a sample set of valid batches from 2017. These models can now be compared, and the best performing model can be trained on all data and used for a thorough analysis on the entire data set. In order to determine which model performs best, a set of criteria are used that incorporates both the accuracy of the predictions and the effectiveness of the confidence intervals. Since the models are trained on a data sample from 2017, the conducted analyses regarding the anomaly detection in this chapter incorporate all 10Hz sequences corresponding to the remaining ten minute averages in 2017.

7.1 Model comparison

A model performs well when it is able to not only predict the next wind speed residual accurately, but also capture the confidence intervals. Therefore, the comparison is made on based on two separate aspects: the MSE and the FPR for 95% intervals as determined by all models.

Cross-validation was applied during training for the selection of appropriate hyperparameters. One layer CV, as done in Chapter 6, has its limitation in prediction of the generalization error (the error a model has when prediction unseen data). Results are directly dependent on the training set, and as such give a biased error. This technique is therefore only suitable for parameter tuning and model comparison, since all MSE share that same bias when they are trained on identical sets. Two layer CV can be used for more statically significant predictions of the generalization error (Herlau et al., 2019), but such predictions are not relevant to the current work as the generalization error does not contribute to the definition of anomalies. The persistence and MLP models were trained on identical splits, and thus the MSE across the test splits can readily be compared. The ARIMA model has been trained on the longest available sequence, that beginning in October 2016, and thus can be tested on the same test folds. The MSE for each prediction model are noted in Table 7.1.

As illustrated, the ARIMA model outperforms the persistence model, which is to be expected considering that the ARIMA equations remove correlation structures that are present in the prediction errors. The underlying non-linearities that the MLP is able to model are also present in the MSE, as the MLP predicts the differences more accurately compared to the persistence model. Between the ARIMA(2,1,3) and MLP model, almost identical results are obtained. When a more accurate prediction is required during this time period, the ARIMA model may be more suitable due to its lower complexity.

Table 7.1: Mean Squared Error of the prediction models, evaluated on 1500 10-minute batches that are randomly sampled from all available data in 2017 and divided into five training and test splits. All units are in m^2/s^2.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Persistence	0.0181	0.0171	0.0182	0.0183	0.0190
ARIMA(2,1,3)	0.0154	0.0147	0.0155	0.0155	0.0164
MLP	0.0156	0.0147	0.0155	0.0156	0.0163

Since the predictions are used to assess whether or not points are anomalous through the established confidence intervals, an assessment of the FPR can be made on all sequences from 2017. Considering the sequences are filtered such that they define normalcy, any labelled data point can be considered a false positive. All confidence intervals are set to 95% to create fair comparisons, meaning that FPRs around 5% are to be expected. The persistence and ARIMA models use a fixed threshold of 0.19m/s and 0.12m/s respectively to label measurements, whereas the MLP model uses a dropout rate of 0.05 after making a prediction to create an adaptive confidence interval. All wind speeds that have concurrent wind direction measurements in the excluded sectors, as explained in Chapter 5, are not taken into account. The FPRs across all sequences in 2017 are shown in Table 7.2.

Table 7.2: False positive rates of the three different detection models, across all normal sequences from 2017, together with the maximum FPR on a sequence.

Model	**CI**	**FPR**	**Max FPR**
Persistence	0.19m/s	15.0%	75.7%
ARIMA(2,1,3)	0.12m/s	8.30%	69.7%
MLP	*adapt*	1.50%	15.5%

The MLP outperforms both the ARIMA and persistence model significantly in that it has a lower FPR and the latter two have higher FPRs than predicted. The persistence and ARIMA model show a higher FPR than expected for a 95% CI. Higher wind speeds result in higher residuals and are as such more prone to fall out of the set bounds for higher speeds that occur less frequently. The persistence model underestimates CIs due to its dependency on the training data. If the sampled 1500 batches have a lower mean wind speed, the same threshold to capture certain quantiles will be underestimated. Since the ARIMA model was trained on the longest sequence, the expected variation in the residuals is directly linked to the magnitude of the wind speeds during that sequence. An acute disadvantage of the ARIMA model is that it can not be trained incrementally. The effects of higher wind speeds can be seen in the maximum percentage of anomalies encountered in a sequence. In total, 1432 sequences of varying lengths of normal data are present in 2017, from which the percentage of false positives within each sequence

can be determined. In Table 7.2, it can be seen that this maximum is much lower for the MLP model than the ARIMA and persistence models. All of these maxima come from the same sequence with the highest average wind speed of 25.5m/s.

Overall, it can be concluded that the MLP model works best and will be a suitable method to train on all data. The MLP is able to make the most accurate predictions, though it is the most complex of the models. Its ability to introduce dropout and create adaptive confidence intervals means that a static parameter, the dropout rate, can be set to create a defined limit under which a said percentage of normal measurements fall, which holds for higher wind speeds as well.

7.2 Final model

The MLP model that will be used to test the effectiveness, advantages and limitations of the developed technique will be trained on all normal data, available from May 2016 up until January 2020. The large number of measurements (288,342,000) verifies that the training procedure goes as expected with use of the validation curve. Furthermore, a different dropout rate may be required in order to accurately represent confidence limits. These matters are also taken into account.

A validation curve is generated over the course of the training process, tracking the MSE over a portion of the training data that is held out. Ten percent of the data is allocated for this before the training starts to ensure training goes properly and the most optimal model can be selected. No train and test splits are made, since no estimation of the generalization error is required. Since the hyperparameters are already tuned, one layer CV would be sufficient to make an estimation. However, the most accurate prediction is not relevant for the purposes at hand, and the generalization error does not give an indication of potential anomalies. Doing CV for such a large dataset is very computationally expensive and therefore not conducted. The validation curve can be seen in Figure 7.1.

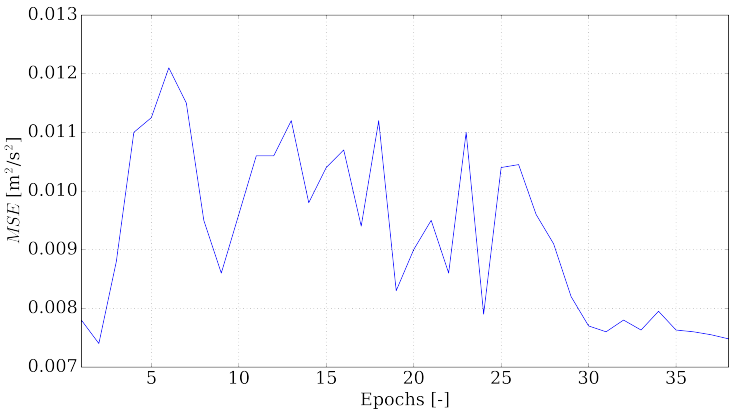


Figure 7.1: MSE on the validation set from all available normal data.

It is evident that the validation error decreases on the second epoch, but then increases steeply. Fluctuation with a downward trend around 0.011 m^2/s^2 seems to be present, after which the curve seemingly converges on the MSE found at the start. It is also noted that the magnitude of the MSE is lower on every epoch when compared to the general validation errors encounter during model development in Section 6.4, which is to be expected for a larger training set. Overall, it can be concluded that the downward trend and decreased fluctuations mean that training is performed well and that the model can be used to determine the ideal dropout rate that will accurately describe the confidence intervals. The ideal dropout rate for each sequence is plotted as a function of average wind speed and illustrated in Figure 7.2.

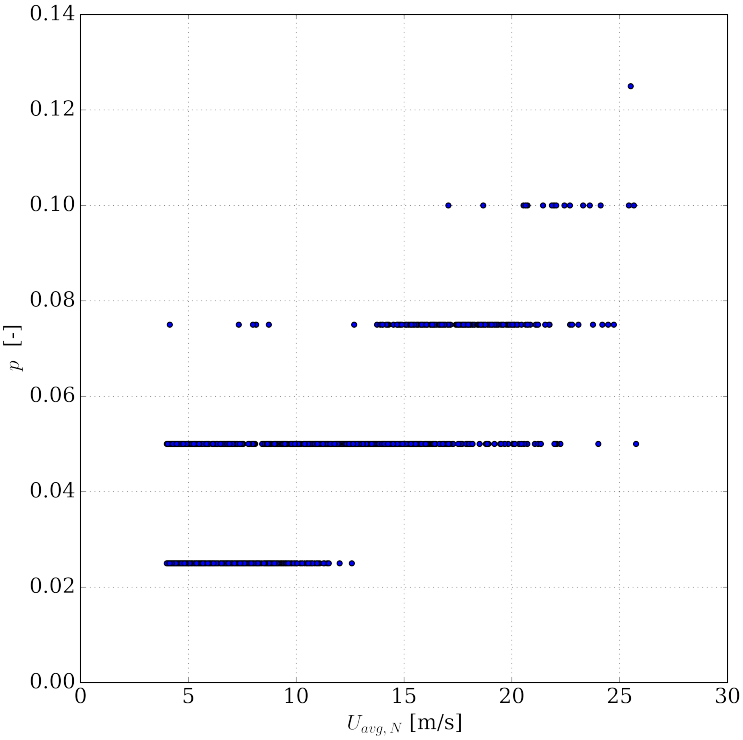


Figure 7.2: Ideal dropout rates, for which the predicted and observed confidence intervals line up best, as a function of average wind speed for all uninterrupted sequences of normal data in 2017 for the fully trained MLP model.

Similar conclusions can be drawn as for the model that was trained on a subset of the data. A dropout rate of 0.05 still covers most of of the data points and even seems to be optimal for lower wind speeds when compared to the model used in comparison. A subset of all sequences in 2017 is still used for this plot because of its realistic range of wind speeds from which relevant conclusions can be drawn. It can be concluded that the fully trained MLP model can be used for a detailed analysis of anomalous structures present in all available data.

CHAPTER 8
Performance evaluation

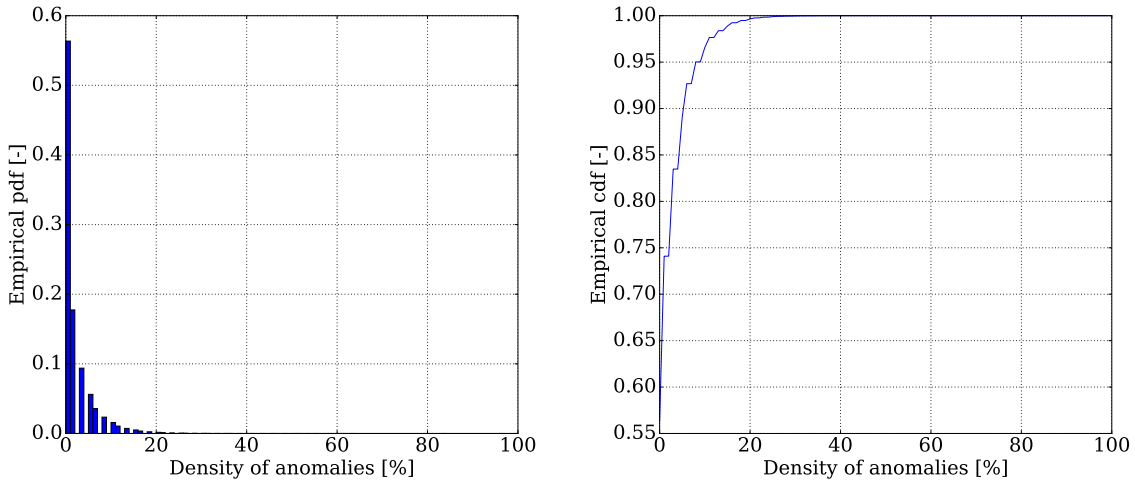
With the MLP model in place and trained, it can be utilized on the sequences it was trained on to establish limits of normality and performance measures. These are then tested on unseen normal data, measured in 2020 after the training set was created, to see if similar results are produced. Its performance is also evaluated on the validation sets of periods with icing events and during the period of the damaged operational cup anemometer, as described in Chapter 5. These analyses result in validation of and conclusions about the limits of the model performance.

8.1 Normal sequences

Assessment of anomalous time series relies on the distinction with the results compared to a normal data set. The model is therefore employed on all normal training sequences to see how it behaves under the created definition of normality. Although the FPR can be estimated by counting the number of anomalies in this set, the estimated FPR will not give insights on the details of model behaviour on these sets. More information can be gathered by looking, for example, at the density of anomalous labels and at the extremeness of anomalies, i.e. how many confidence intervals a flagged measurement is from its prediction. The estimated FPR is observed to be 2.06%. It should be noted that this estimation only serves to provide an estimation of the true FPR, which must be calculated through CV. Due to the large computational time that is required, more emphasis is placed on the proof of concept through analysis of qualitative performance than the quantitative evaluation of the model effectiveness.

Observing the density of anomalous points provides unique insight. It is reasoned that abnormal behaviour rarely occurs instantaneously, but rather affects the measurements for a longer period of time. For this reason, abnormal sequences should contain a distinctly larger number of flagged measurements compared to sequences of normal behaviour with several false positives. It is therefore important to look at the performance on the training data to see which densities are common and which may indicate anomalous sequences. The densities that are referred to in this section are defined as the percentage of anomalous points in a six-second window, or 60 measurements. The length of this window is carefully chosen, as a smaller window allows for little variability of densities, whereas a larger window may lead to anomalous sequences being averaged out. The empirical pdf and cdf of these densities across all normal sequences are displayed in Figure 8.1.

As expected, the most common densities are relatively low. More than 56% of the six-second windows contain no flags at all, and 83.5% contain two anomalies or less. Upon closer inspection, it is observed that more than 99.5% of all windows have 20% of their



(a) (b)

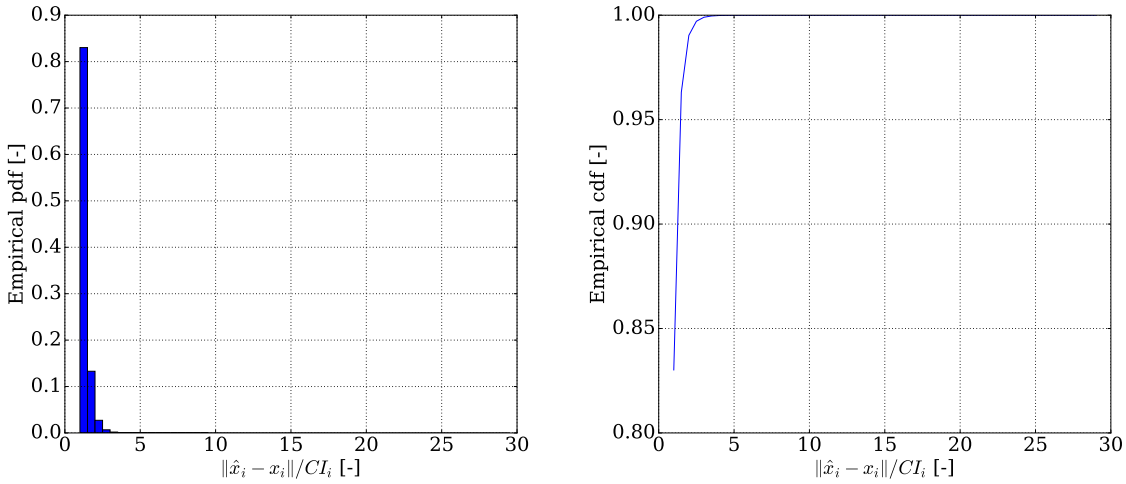
Figure 8.1: Empirical pdf (left) and cdf (right) with 1% bins of the density of flagged measurements over a window of 60 measurements across all normal sequences.

points flagged, or 12 out of 60 measurements. It can thus be concluded that densities that exceed these percentages with a similar window size require further inspection and may be anomalous.

A similar argument can be made for the extremeness of an anomaly. False positives on the training data set mostly fall just outside of the confidence limits. A level of extremeness can be established based on the ratio of the prediction residual $\|\hat{x}_i - x_i\|$ and confidence interval CI_i. For flagged anomalies, this measure is always greater than one. For the observed extremeness on the false positives among the normal sequences, the empirical pdf and cdf are determined. The results are displayed in Figure 8.2.

The expectation that most anomalies in the normal sequences lie close to the confidence limits is confirmed since 99% of all anomalies lie within two confidence intervals. The most extreme values are encountered at $12.5CI$ & $18.5CI$. In Chapter 5, the averaged-out startup of the first sequence was used to display the need for additional filters for the 10Hz sequences. The model results on the start of this sequence are shown in Figure 8.3.

Extremes above $10.5CI$ are flagged in Figure 8.3 as such. It can be seen that the first two measurements, after the northern cup anemometer starts spinning, are denoted as extremes. The first and second extreme points both fall above 18.5 and 12.5 their respective CI limits. It is desirable to flag these points as the southern cup anemometer is not operative and, as such, these extreme values are removed from the set. It can also be seen that a high density of flags exist during the period where the southern cup is not measuring the wind speed, showing that the discussion about the density of anomalous sequences holds here as well. By excluding the two extremes shown above, the maximum in extreme that is encountered has a measurement that is between 10 to 10.5 times its CI. In the following analyses in this chapter, extreme labels are denoted as observed residuals of larger than $10.5CI$.



(a) (b)

Figure 8.2: Empirical pdf (left) and cdf (right) with 0.5 bins of the extremeness of flagged measurements across all normal sequences.

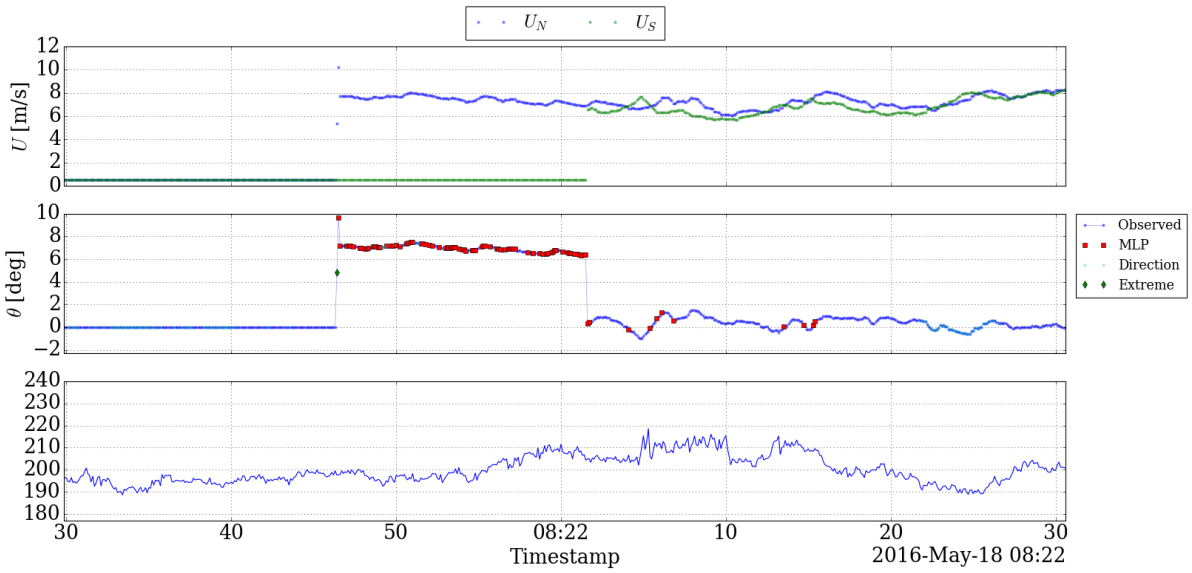


Figure 8.3: Measurements taken during the start of the operation of the LMN, where extreme values as determined by the MLP model are encountered.

The number of flags per 10 minutes can also be used as a measure to evaluate sequences. The majority of filtering takes place on a 10-minute scale, which significantly reduces the number of data points. A smaller number of data points also enables analyses over larger time periods, e.g. spanning over multiple days. Furthermore, as shown in Chapter 5, anomalous behaviour becomes apparent on a 10-minute level when looking for validation sets. In order to evaluate the detection model results on larger time scales, the 10-minute averaged time series for wind speed and direction can be plotted together with the number of flags during each of the 10-minute periods. The results of the longest normal sequence are shown in Figure 8.4.

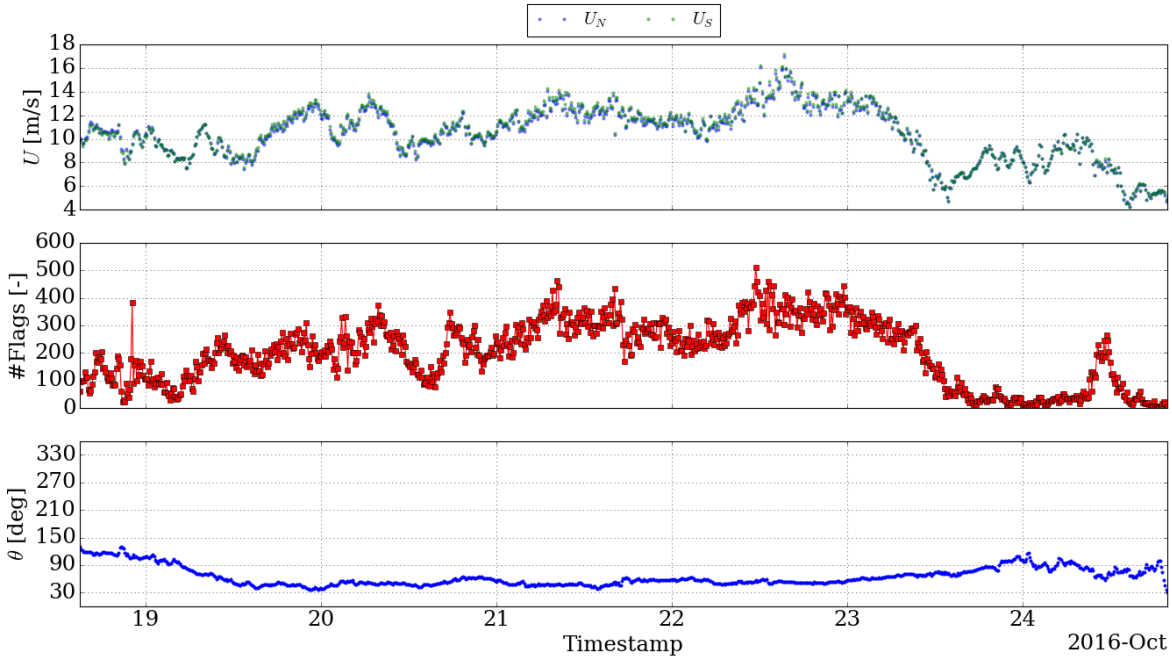


Figure 8.4: 10-minute averaged time series of the longest sequence conforming to normalcy. For each 10-minute period, the number of flags that are assigned on the 10Hz data is given as well.

An important result is observed in Figure 8.4. When comparing the trend of both wind speed signals and the number flags, a very strong correlation becomes apparent. An increase in wind speed results in an increase in the number of flags. It was already concluded that higher wind speeds require a higher dropout rate, since more variability is necessary to create larger confidence intervals, and the correlation here further supports that conclusion. A static definition of normalcy is required and, as such, an increase in false positives under normal conditions is to be expected. For most wind periods, at moderate and low wind speeds, the number of observed flags is under 300 or less than 5% of the measurements, corresponding to the expected results.

The relation between the wind speed and number of flags becomes even more apparent when the number of flags is normalized with either the northern or southern wind speed, as shown in Figure 8.5. It can be seen there that the number of flags contain significantly

less trends that are similar to the wind speed, especially between the 21st and 23rd of October, when compared to Figure 8.4. In other regions, such as around 00:00 on 24-10, both anemometers seem to agree well with each other, which results in almost no flags or patterns. The dependency on the wind speed is a significant limitation of this model that needs to be taken into account when interpreting results.

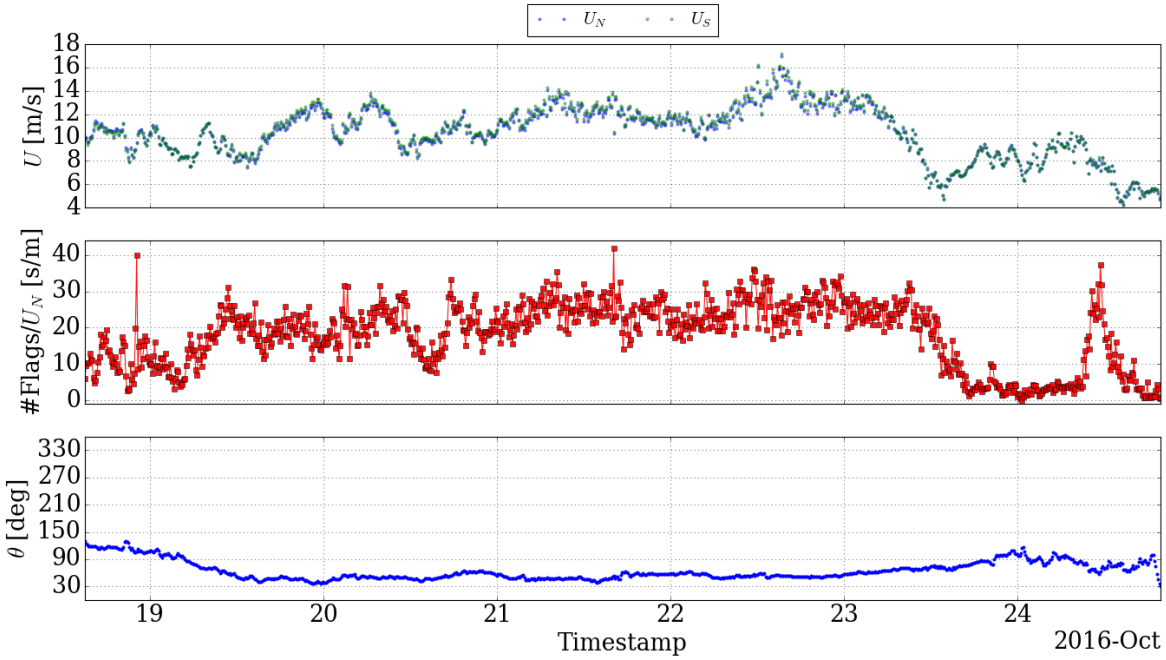


Figure 8.5: Number of flags per 10 minutes of the longest sequence conforming to normalcy, normalized by the wind speed of the northern cup anemometer.

8.2 Untrained time series

Since the analysis above is done on the training data, performance on time series that were not used during training must distinguish whether the model is able to perform well on unseen data and whether it correctly recognizes normal behaviour. 10% of the data was withheld during training in order to create the validation curve, meaning that the model can be employed on the previously withheld data to assess its detection capabilities with unseen data. Furthermore, the dataset was created in January 2020 and as such does not contain more recent data than that which was available at that time. Since then, new measurements have been taken that are filtered according to the procedures presented in Chapter 4 and evaluated as more unseen data.

8.2.1 Validation set batches

Across all batches in the validation set, no extremes are found and only 436 six-second windows have more than 20% anomalous points total. Considering that each batch

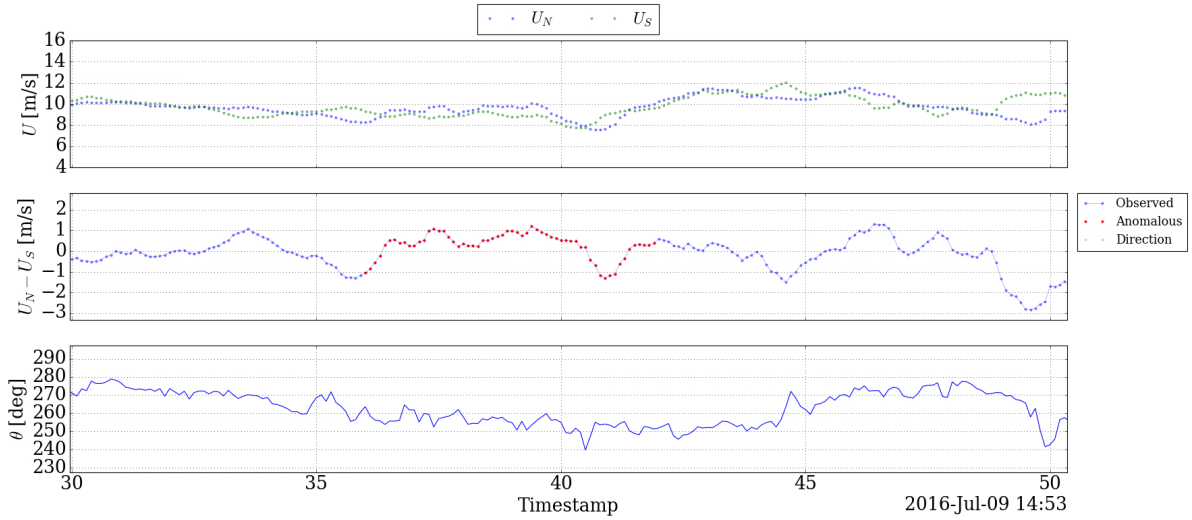
contains 1000 windows, and 4805 batches are present in the validation set, only 0.009% of the windows are faulty. This is below the estimated interval of 0.5% on the training set and therefore complies with the established definition of normalcy. Overall, it can be said that the model performance on the batches in the validation set does not have a significantly higher FPR compared to the performance on the training set and thus conforms to the established definition of normalcy.

Of the batch with the highest detection rate, one of the windows with an uncommon density of anomalies is highlighted in Figure 8.6. From the time series, engineering judgement can be used to conclude that the displayed signals are not faulty according to the given definition and thus should not be seen as anomalous. Rather, this behaviour most likely comes from a gust or a turbulent structure that passes by. Around 14:53.40, the northern cup anemometer lags behind the southern cup by about a second. This corresponds roughly to an eddy travelling at 10m/s over 10.8m, the separation between both cups. From the predictions and confidence intervals in Figure 8.6c, it can be seen that this lag causes inaccurate predictions of the difference between both anemometers. Specifically, when both anemometers have a similar downward trend after 14:53.40, the confidence limits become smaller. When the southern cup first starts logging a faster wind speed, flags occur. A similar result is observed when the northern cup then starts measuring a higher wind speed about a second later. It is important to note that an uncommonly dense collection of flagged measurements do not directly indicate that the sequence is indeed faulty, but also that the evaluation of the batches indicates that the performance indicators see these batches as conforming to normality. Specifically when both anemometers exhibit a similar trend, the confidence intervals seem to shrink, resulting in an increase of flags when this local trend changes.

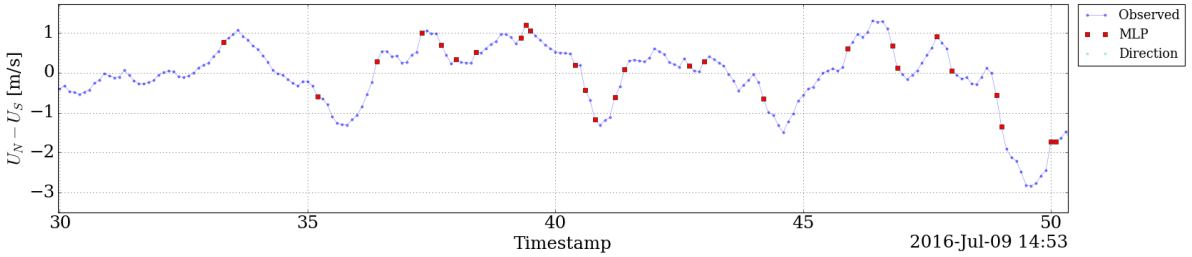
8.2.2 2020 sequences

The training and validation set used for the model training contains data obtained in May 2016 to January 2020. Since the LMN was operational after January 2020, more unseen data is available that can be used for evaluation and testing. As mentioned, it is important to validate the model's ability to both flag faulty data as well as recognize normal data. Since the raw data may contain both types of data, and faulty sequences are available from the filtered out objects, the new raw data is filtered using the same procedure explained in Chapter 4 to create a set of normal data between February 2020 and May 2020. The directional bias of the difference between 10-minute averages of both cup anemometers is shown in Figure 8.7. From here, an identical pattern is visible when compared to the results of Chapter 4, meaning that the resulting sequences can thus be used for additional validation of the model's performance on normal time series.

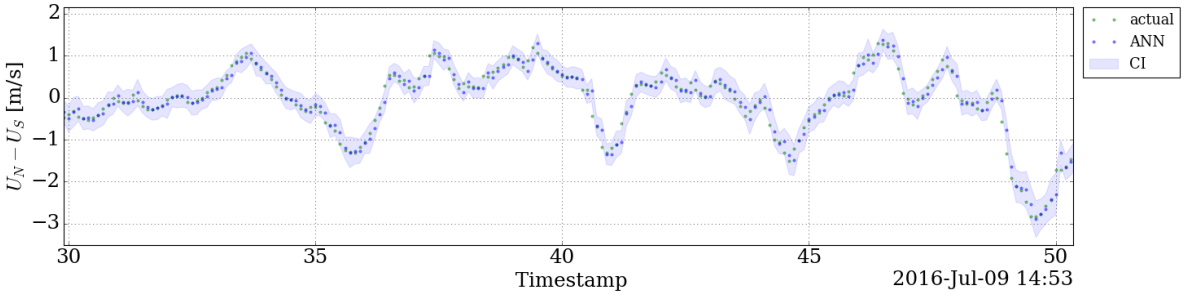
For these unseen sequences, a similar analysis to that of the batches in the validation set is conducted. In total, 2.99% of the measurements are flagged in these sequences. A slight increase in detection rate is observed when compared to the FPR on the training set, which equalled 2.06%. This is an important result and indicates the model is not under- or overfitted. An underfitted model has no significant increase when the results on a training and test set are evaluated, whereas an overfitted model preforms significantly



(a) Anomalous densities



(b) Flagged measurements



(c) Predictions

Figure 8.6: Model performance on part of the 10-minute batch from the validation set with the highest detection rate.

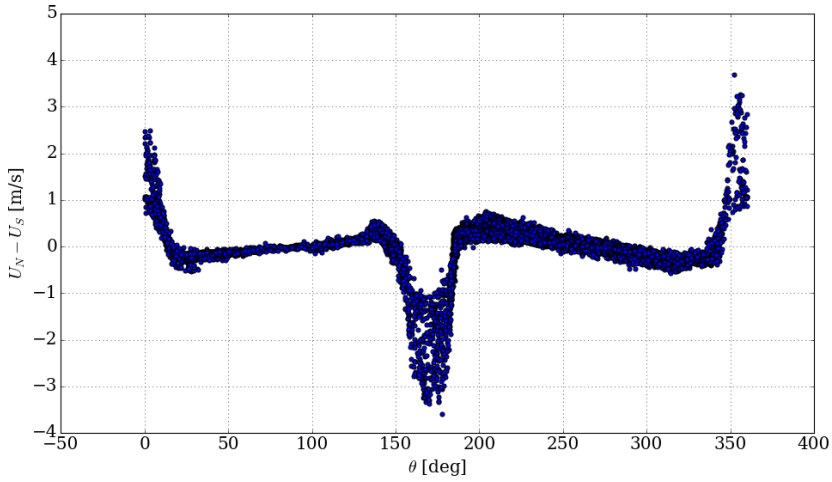


Figure 8.7: Wind speed residual of the two cup anemometers at 178m as a function of wind direction at 175m, measured at LMN, after filtering for normalcy between February and May 2020.

worse on a test set compared to a training set (Herlau et al., 2019). A well-trained model strikes the middle ground, as does the model in question.

One extreme is encountered during the period of February 2020 to May 2020 and originates from a series where the mast shadow affects all the surrounding measurements. The extreme is assigned when the instantaneous wind direction is slightly below the set limits. As such, this point does not conform to a faulty measurement that slipped between the filters. From all six-second windows, 0.9% of these had more than 20% of the measurements flagged. This is slightly higher when compared to the training set, which is to be expected, as the model has not seen this data before. These three indicators lead to a similar conclusion that the performance on unseen data is less effective when compared to training data, but still can be said to correctly recognize the unseen sequences as normal behaviour.

Inspection of the 10-minute averages collected at the beginning of February reveals that on the 9th of February, exceptionally high wind speeds were observed with local maxima above 30m/s. These time series are shown with the detection model results in Figure 8.8. The effects of the storm on the detection model are observed through the large rise in flags, which come in much higher quantity when compared to performance on the training data. On February 7th, under calm conditions, the model performs well and does not see any consistent abnormalities. The relatively large discrepancy on February 8th can be explained by looking at the wind direction, as the the measurements are affected by the mast shadow during these periods. This means that no predictions can be made and that a low number of flags are observed. However, when the wind moves away from this direction, a large increase in flags can be observed with some 10-minute periods containing more than 15% flagged measurements. This percentage is similar to the observed maximum during the model comparison in Chapter 7, which corresponded to a sequence with an average wind speed of 25.5m/s. After this peak, the

number of flags seem to follow the trend of the wind speed as time progresses, similar to the observations on the training data set. The number of flags in each period, as well as the magnitude of the wind speed, are significantly higher.

However, when looking at the number of flags normalized by the northern wind speed in Figure 8.9, a similar result is observed when compared to the training data. The number of normalized flags is of a similar order of magnitude and seems to be de-trended when compared to the windspeed. However, on the morning of February 11th, the number of flags are amplified, which occurs simultaneously with a sharp increase in wind speed.

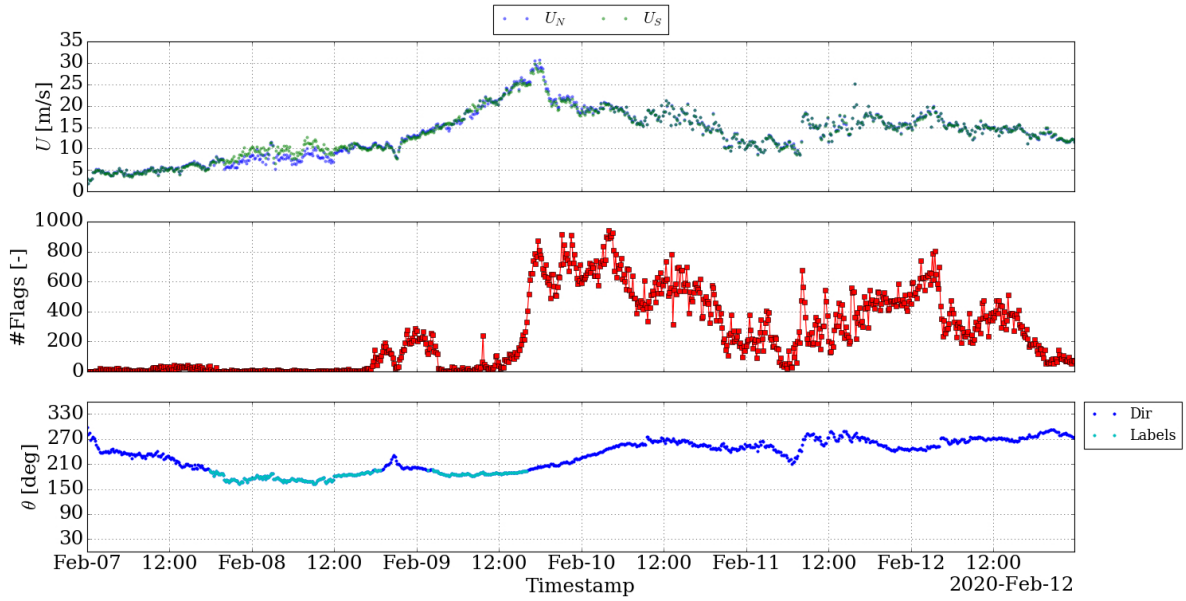


Figure 8.8: 10-minute averaged time series of the longest sequence conforming to normalcy. For each 10-minute period, the number of flags that are assigned on the 10Hz data is given as well.

Focusing these regions with a high number of flags, for example that shown in Figure 8.10, reveals that anomalous densities are also observed relatively frequently. This can be explained by the fact that high wind speeds give rise to significantly more flags, meaning that more dense regions start to exist. Higher wind speeds prove to be a clear limitation of the model, where it becomes harder to distinguish anomalous from normal behaviour. The model performs well on average, since wind speed of relatively high magnitudes occur somewhat infrequently, but the increase of detection rate when these wind speeds do occur impose a limitation of the reliability of the flags under these conditions. These results again confirm that expert judgement is required to interpret and discard these results, as the time series reveal no clear anomalous behaviour even though a significant increase in flags is observed. Due to the lack of clear anomalous events under high wind speed conditions, it can not be verified if these events are indeed anomalous or belong to a set of measurements which are taken by cup anemometers under normal operation.

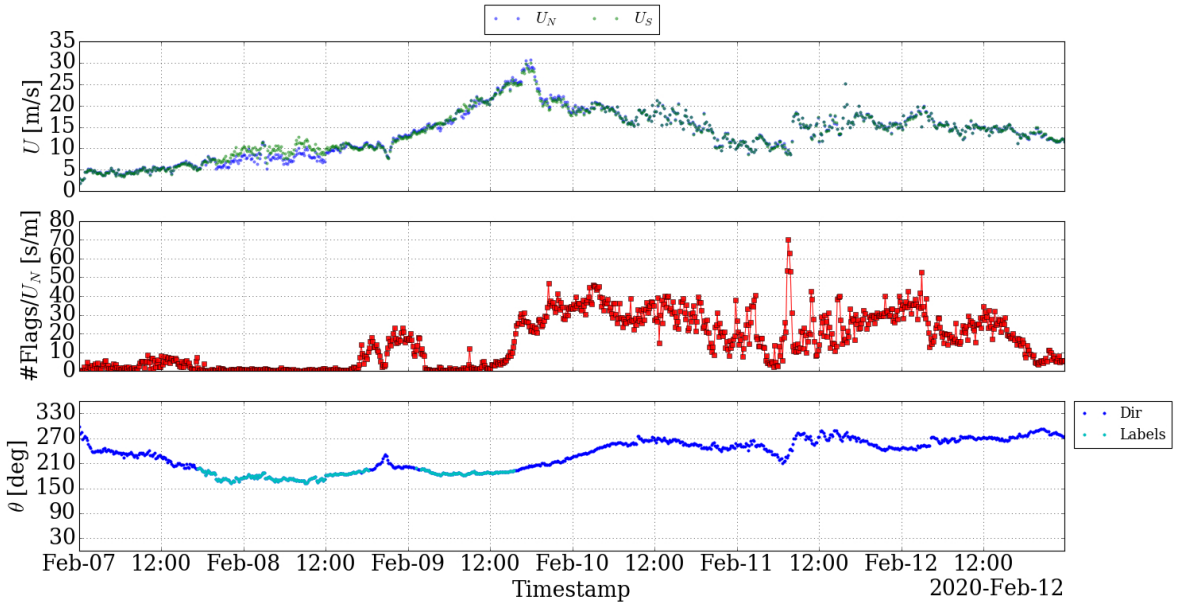


Figure 8.9: 10-minute averaged time series of the longest sequence conforming to normalcy. For each 10-minute period, the number of flags that are assigned on the 10Hz data is given as well which are normalized by the northern wind speed.

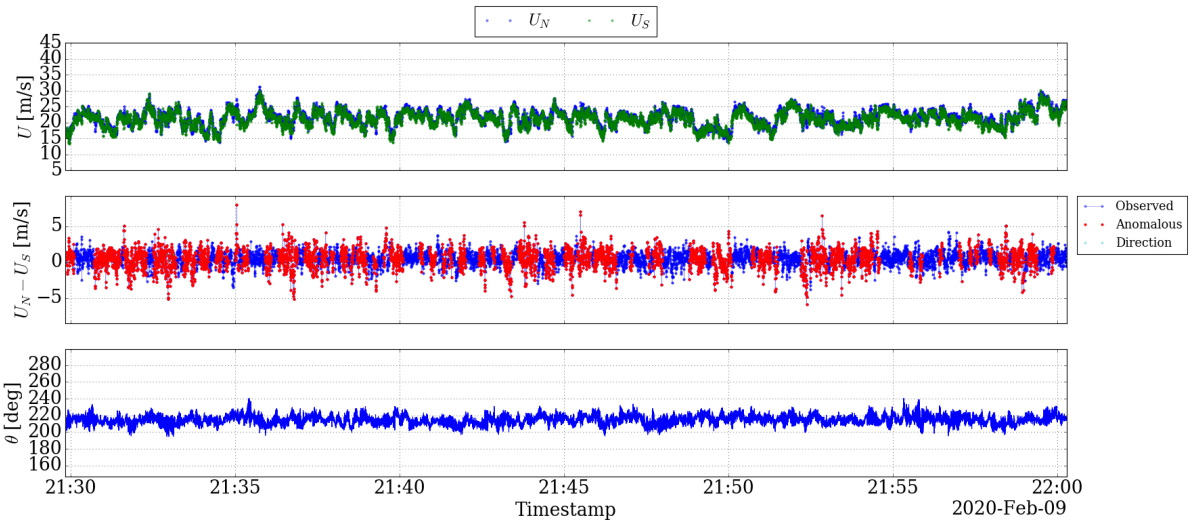


Figure 8.10: Time series with the model predictions during a period of high wind speed. The marked sequences comprise of at least 20% anomalous flags.

8.3 Lighting time series

As mentioned in Chapter 4, during the period between January 29th 2019 and April 10th 2019, the northern cup anemometer was struck by lightning. The arm of one of the cups broke as a result, giving inaccurate measurements during this period of time. Several sequences of interest occurring during this period are evaluated in detail to further validate the model and assess its anomaly detection capabilities.

An overview on the 10-minute scale of this event is shown in Figure 8.11. It is observed that the northern cup anemometer logs faulty measurements between January 29th and 31st, as the wind speed signal shows clearly erroneous patterns. The model recognizes the erroneous patterns, and an increase in flags can be observed after this period. Although the rise in wind speed also increases the number of flags, the observed number of flags is noticeably higher compared to the number of flags on normal sequences at similar wind speed magnitudes (<400 flags). On February 5th, this becomes even more apparent, with periods that consistently contain more than 1000 flags per 10-minute window. When looking at the normalized number of flags, as seen in Figure 8.12, the large number of flags remain clearly visible. The number of normalized flags is significantly higher as well when compared to the normal data, reaching consistently above $50 s/m$. On a 10-minute scale, the model clearly recognizes the present anomalies. A closer analysis is performed on the 10Hz time series to observe how these errors are treated on a smaller time scale.

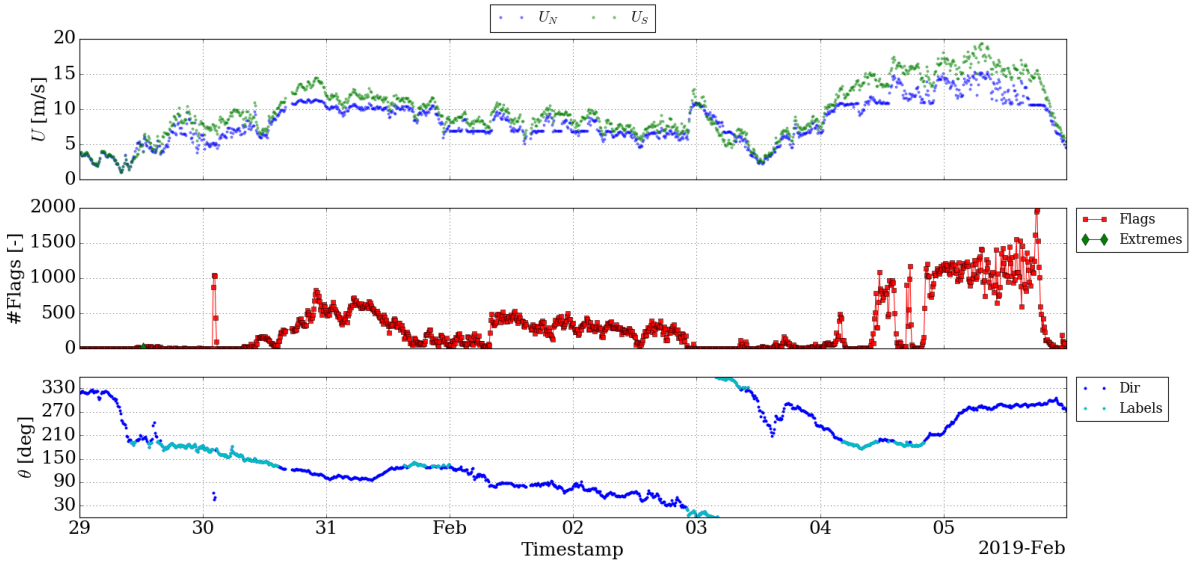


Figure 8.11: 10-minute averaged time series of the period during a lightning strike that results in anemometer damage. For each 10-minute period, the number of flags that are assigned on the 10Hz data is given as well.

First, a closer taken at the point in time where the lightning supposedly struck. The event is logged at 12:20 29-01-2019 in the weblog. The resulting time series and the model results are shown in Figure 8.13. Right after 12:28, it can be seen that an instantaneous jump in the northern wind speed signal is observed. The detection model flags this as

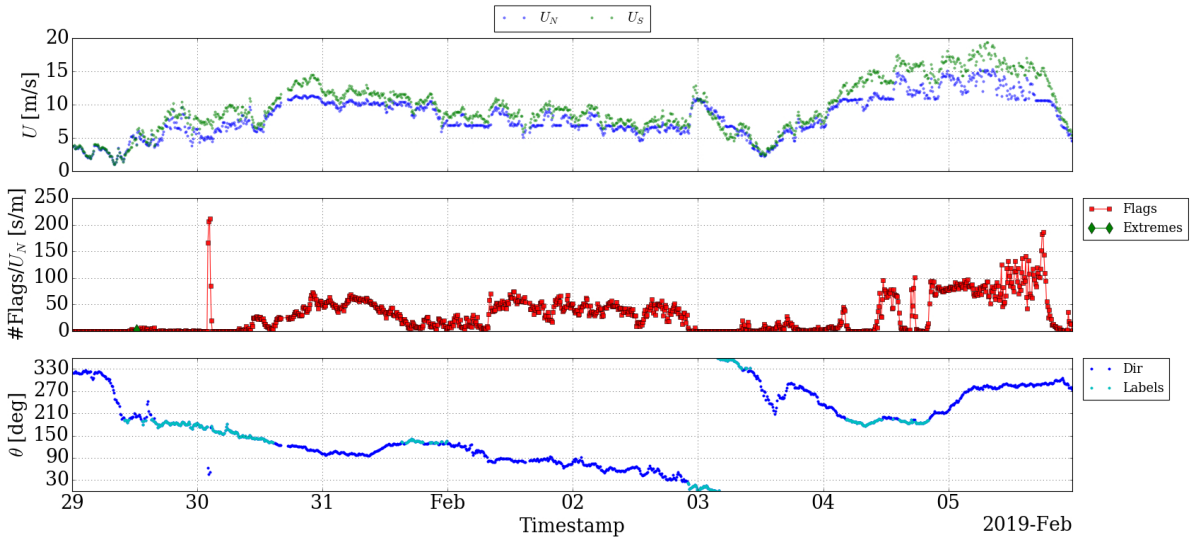


Figure 8.12: 10-minute averaged time series of the period during a lightning strike that results in anemometer damage. For each 10-minute period, the number of flags that are assigned on the 10Hz data is given as well which are normalized by the northern wind speed.

an extreme point, which is to be expected from the relatively large sudden jump in the signal. After 12:31, a second jump is observed and flagged. It can be seen that, after this second jump, the northern wind speed signal starts to deviate from the southern cup anemometer's measurements. The model picks up on the sudden changes and recognizes the lightning strike as an extreme event through the assigned labels on the difference signal around 12:27 and 12:32.

The series in Figure 8.14 shows typical behaviour of both anemometers after the lightning strike. Instead of looking at the individual labels, six-second windows with more than 20% of flagged points are marked as anomalous. The effects of the broken cups are clearly visible in Figure 8.14a; the northern cup reports a lower wind speed and responds differently to changes in wind speed compared to the southern cup anemometer. These time series also illustrate that a broken anemometer does not necessarily lead to obviously faulty averages, as the magnitude of the wind speed stays reasonable. 10-minute averages from the northern anemometer are lower than those of its southern counterpart, but not obviously wrong.

Only 0.5% of all six-second windows in the training set contained more than 20% of flags. The number of identically sized windows containing more than 20% of anomalies is much higher than 0.5% in Figure 8.14a, meaning that the model is able to classify most of the time series after the cup arm broke off as anomalous, even when the magnitude of the faulty measurements is not clearly wrong.

It should be noted, however, that other periods that are faulty are not flagged. For example, between 05:31-05:32, a significant bias is present, yet no significant number of flags is placed in this period. The anomaly detection model seems to be able to

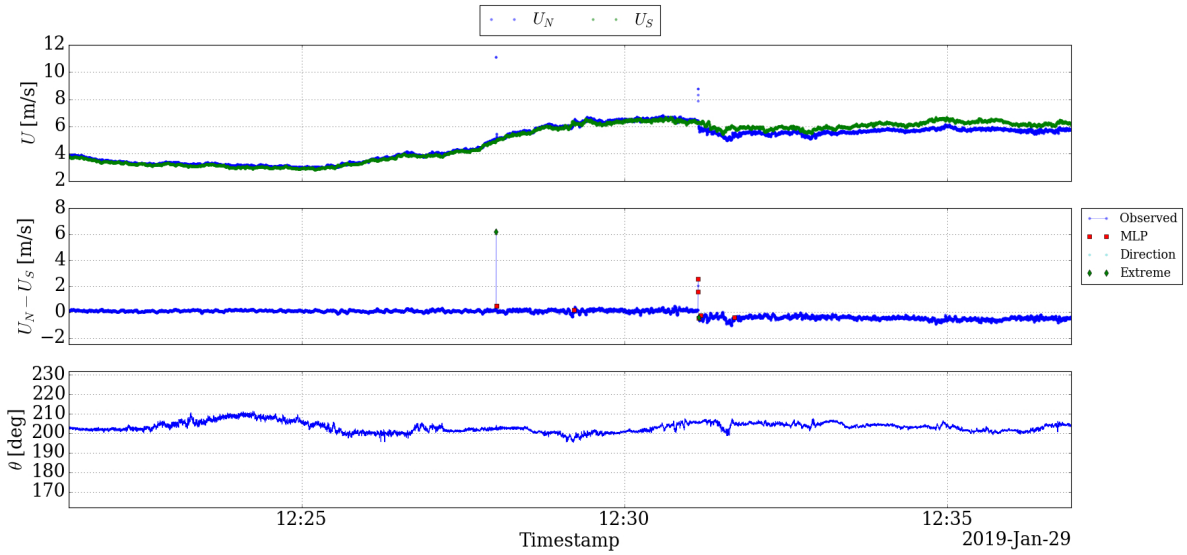


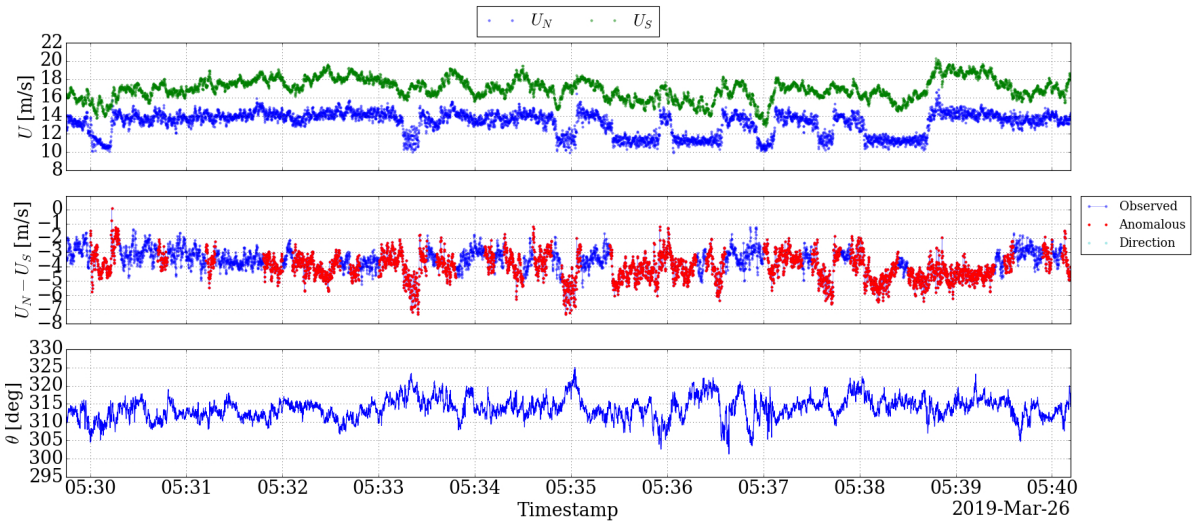
Figure 8.13: Time series with the model flags as the northern cup anemometer is struck by lightning, thus breaking the northern cup anemometer.

determine inaccurate data based on the faulty responses of the northern cup, which are not as apparent during that specific calmer period. The predictions the MLP model made are shown, with confidence intervals, in Figure 8.14b. It is observed that the model indeed tries to predict the bias that is present due to the broken cup, but fails to do so accurately when compared to normal behaviour.

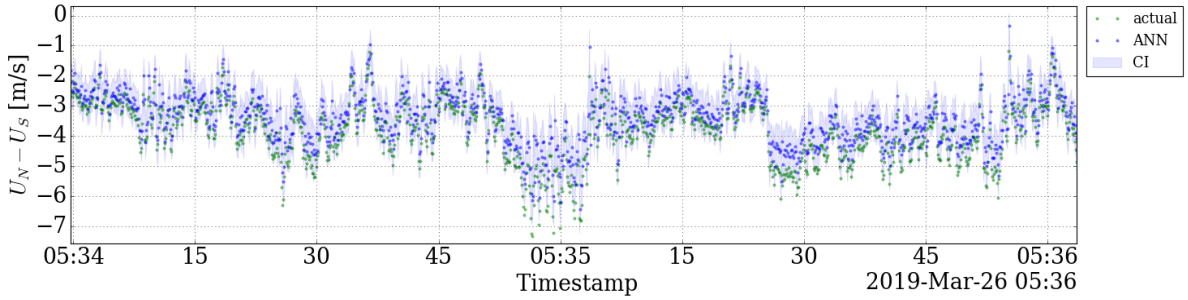
This hypothesis is also confirmed when looking at time series when the northern cup anemometer comes close to a standstill. In Figure 8.15, it can be seen that the damage inflicted on the instrument causes the rotor to stop spinning whilst the southern cup anemometer continues logging normal measurements. As a result, almost all windows of the wind speed residual are marked as anomalous. Although the southern cup anemometer works well, the MLP model learned to predict the difference on normal data in which both cups work normally. Since one of the cups is severely malfunctioning, the detection model is able to flag all sequences during these periods.

Once the broken anemometer is evidently replaced, the model stops labelling sequences as anomalous data. In Figure 8.16, it is observed that two gaps in the time series exist due to scheduled maintenance. A large period of anomalous data is observed between 6:50 and 14:50 with five recorded extremes. One of these is not shown for the sake of visualization of the rest of the time series, because it corresponds to a wind speed of 120m/s. It can also be seen that, between 10:00 and 11:00, the northern cup records measurements that are correlated better with the southern cup, and over time, fewer and fewer anomalous sequences are recorded.

It can be concluded that the anomaly detection model is able to identify and flag a basic anomaly when one of the cups fails completely. Although it does not recognize the large bias, it tries to predict the next wind speed difference based on the dynamics of two working anemometers and therefore fails to do so accurately due to the normalcy of the



(a) Labels



(b) Zoomed-in predictions

Figure 8.14: Time series with the model predictions after one of the arms of the northern cup anemometers broke off. The marked sequences comprise of at least 20% anomalous flags.

training data. In the time series where one of the cups fails partially, a higher density of anomalous sequences appears, but not all periods are labelled. Extreme points are also observed during the lightning strike and during maintenance. These indications validates that extremes, in this case, would have revealed the lighting event. In combination with the following increase in both the number of flags and number of anomalous densities, the model is able to correctly recognize this data set as anomalous.

8.4 Icing time series

In Chapter 5, the Magnus-Tetens equations were used to obtain indications of potential icing events. Inspection of the 10-min averages of wind speed and temperature time series confirmed the suspicion that ice may have formed on the cup anemometers. These time series are analyzed more closely and in order to provide further validation of the model and address the model's performance during another type of anomalous event:

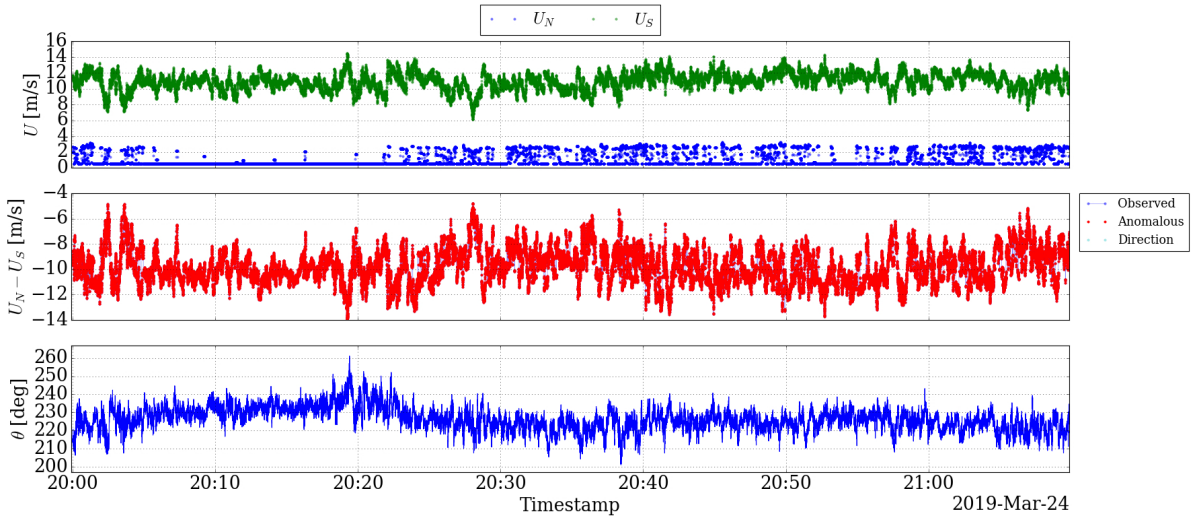


Figure 8.15: Time series with the model predictions when the northern cup anemometer fails as a result of lightning damage. At least 20% of the measurements were labelled in each of the marked six-second windows.

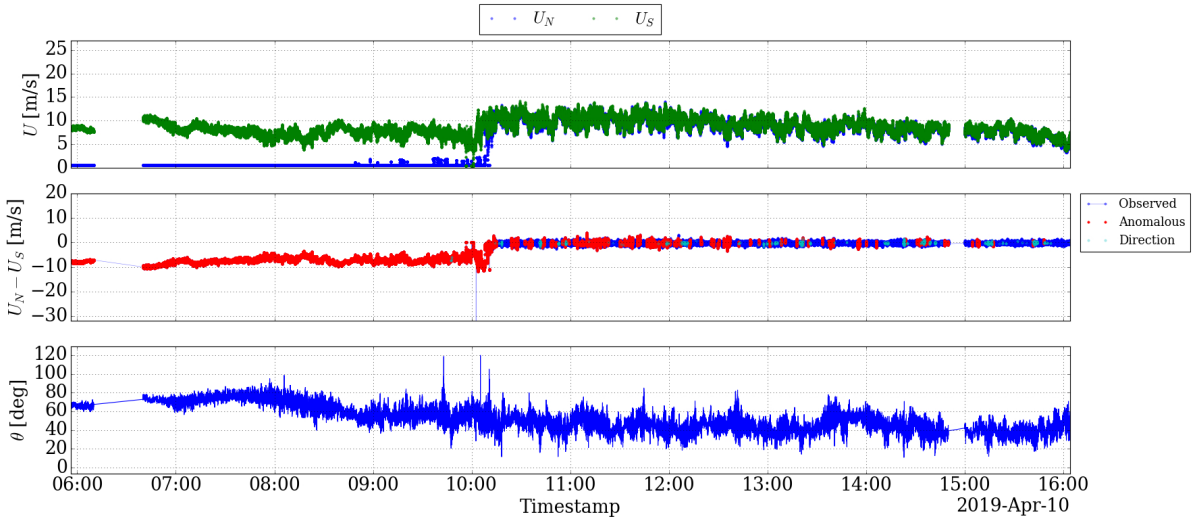


Figure 8.16: Time series with the model predictions during the period when the northern anemometer is replaced. The marked sequences comprise of at least 20% anomalous flags. One point in the time series right after 10:00 is not shown for purposes of visualization; the recorded speed of 120m/s is clearly faulty.

icing.

In Chapter 5, several time series containing a potential icing event were already highlighted, such as that occurring on the 12th of March. The 10-minute averaged time series, together with the number of flags in each period, are displayed in Figure 8.17. Here, it becomes evident that an icing event takes place around the 11th of March on noon, where both anemometers are standing still and there are no apparent fluctuations of wind direction. It can be seen that the model starts to flags more measurements preceding the icing event around 12:00, indicating that anomalous behaviour might be occurring. It can not be said whether this increase and subsequent drop in flags is due to some gradual ice formation or to an increase in wind speed. The number of flags are similar as that observed on normal sequences, meaning that these results from the potential icing even on the 12th of March are inconclusive. Right before the freezing event, a significant spike is observed that can be linked to the freezing of the cups. After first the northern and then later the southern cup anemometer unfreeze, two spikes in flags can be observed, meaning that these events are distinguishable.

A similar conclusion can be drawn when looking at the normalized flags, as shown in Figure 8.18. Right before both anemometers come to a standstill, the number of normalized flags is significantly higher compared to normal data. Similarly, as the northern cup is seemingly partially frozen, the same observation is made.

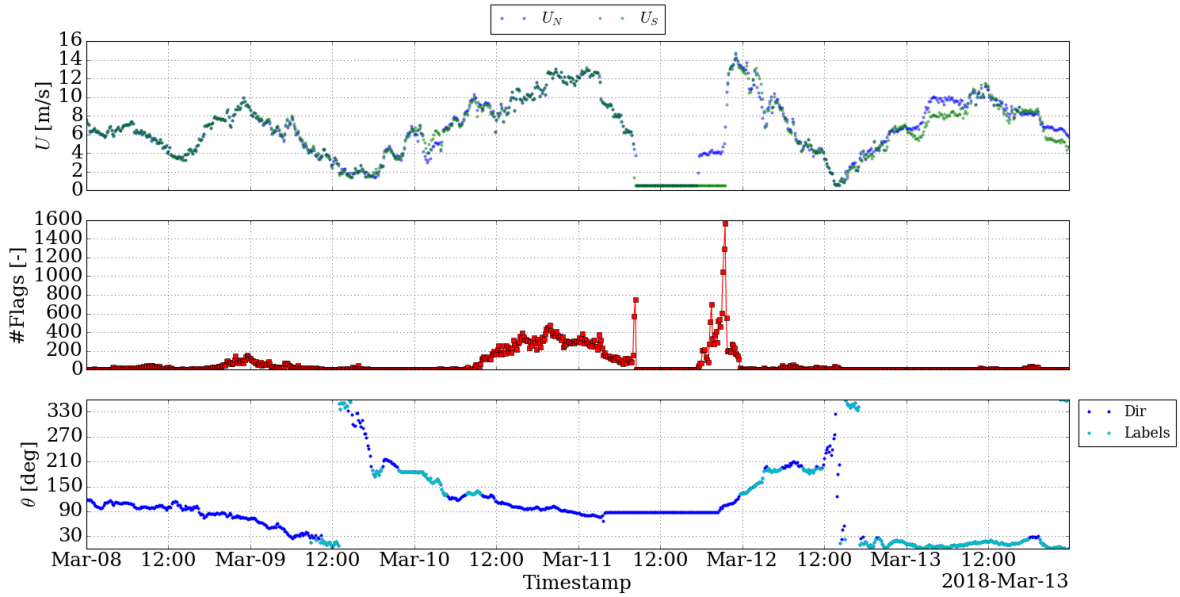


Figure 8.17: 10-minute averaged time series of the wind speed and direction during which all instruments freeze, together with the total number of flags per 10-minute period.

The time before and after the icing event is observed closer on a 10Hz scale. These time series right before both anemometers freeze are displayed in Figure 8.19. During these periods, it can be seen that the sonic anemometer is indeed already frozen, and that the northern cup stops about 15 minutes before the southern cup. During the periods in which the northern cup is stopped, some anomalous windows are observed between 08:10

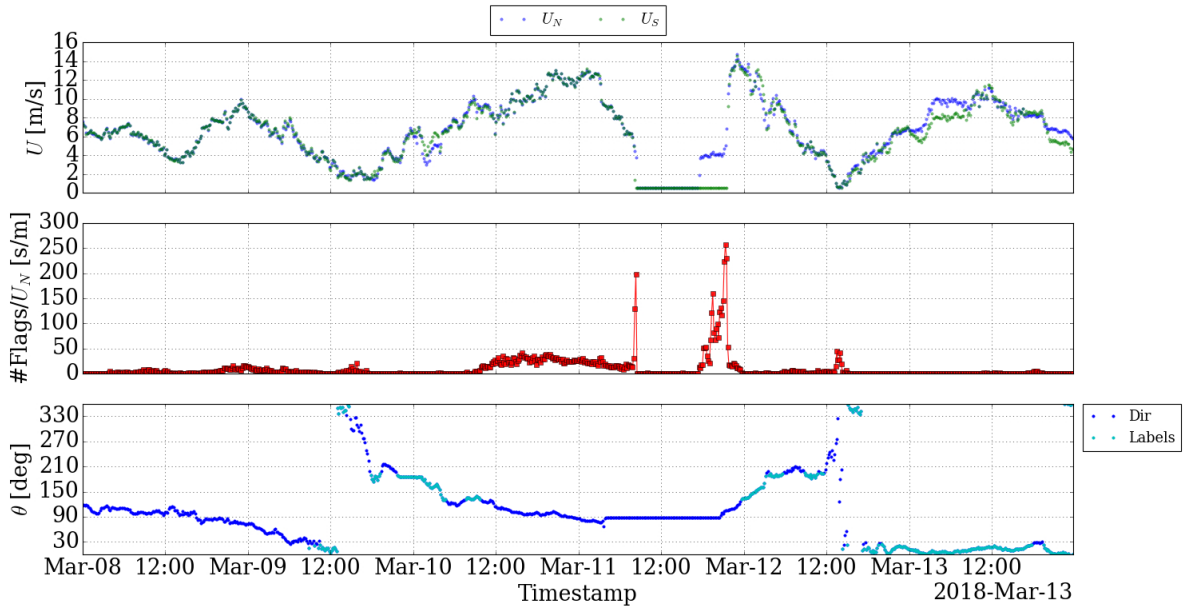


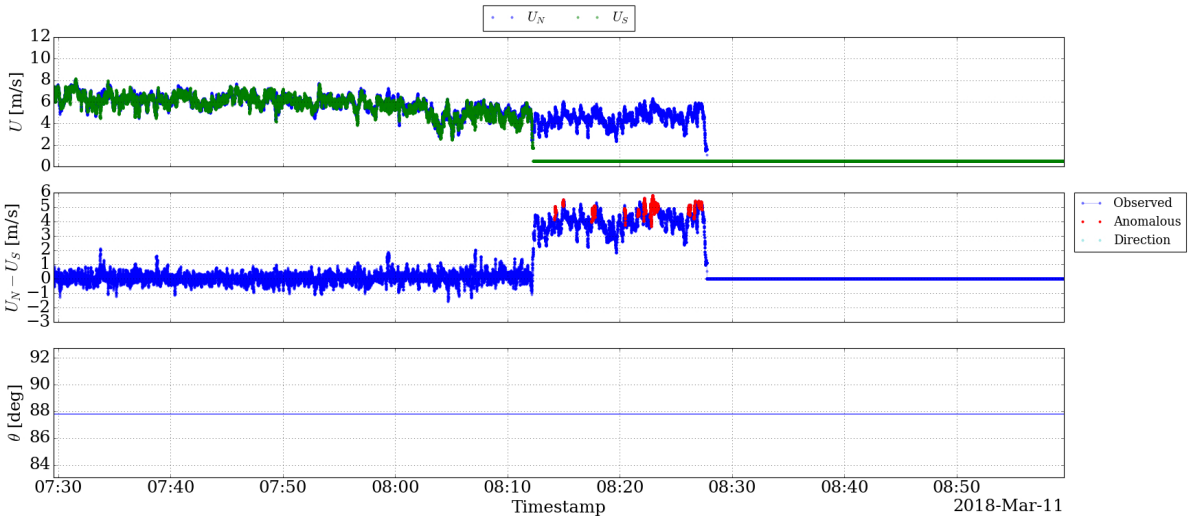
Figure 8.18: 10-minute averaged time series of the wind speed and direction during which all instruments freeze, together with the total number of flags per 10-minute period which are normalized by the northern wind speed.

and 8:30, but not nearly as many when compared to the periods where the northern cup was damaged by lightning as shown in Figure 8.15.

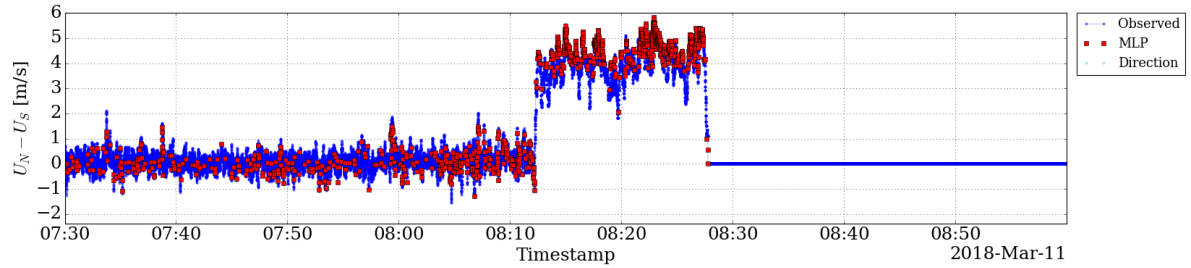
A similar pattern in predictions is noted in Figure 8.19c where the model tries and fails to model the faulty bias. The model consistently underestimates the difference, but this does not lead to high detection rate that was observed when the northern cup stopped due to the broken arm. The major difference between both events lies in the magnitude of the wind speed. Lower wind speeds are observed during icing events, because the anemometers slow down due to gradual accumulation of ice on the cups, before they come a complete standstill. It can be seen that the confidence intervals are relatively wide and that the absolute fluctuations in wind speed are too small to result in flagged measurements.

A similar conclusion can be drawn when looking at the time series when both cups unfreeze. In Figure 8.20, it can be seen that the northern cup unfreezes partially before 18.00, after which no anomalous densities are observed. The sonic anemometer appears to unfreeze a bit after around 20.45, whilst the southern cup remains frozen. Around 21.15, a small incremental increase can be observed in the northern wind speed signal, which gives rise to more flags at a higher density. Then later at 21.30, when the northern cup seems to completely unfreeze whilst the southern anemometer is standing still, almost all measurements are flagged and all windows during that period are marked.

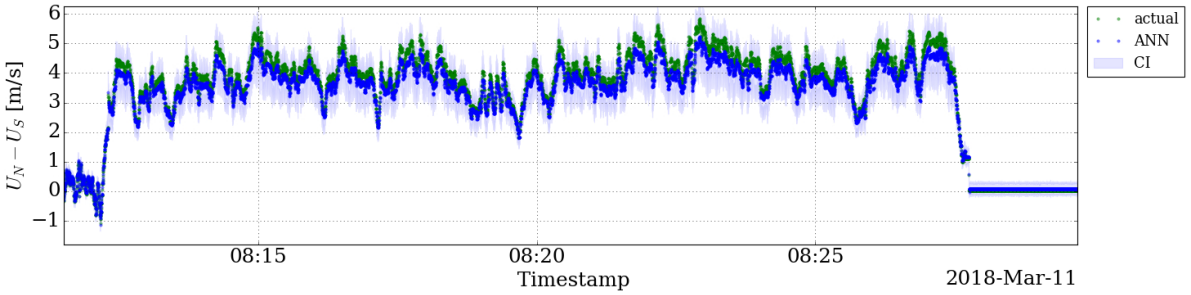
Although the confidence intervals increase during these periods of higher wind speeds, as seen in Figure 8.20b, the bias between the predicted residual and the observed residual seems to become larger when one of the cups is failing for increasing wind speeds, indicating that the model is able to detect such anomalous sequences. It is also observed



(a) Anomalous densities



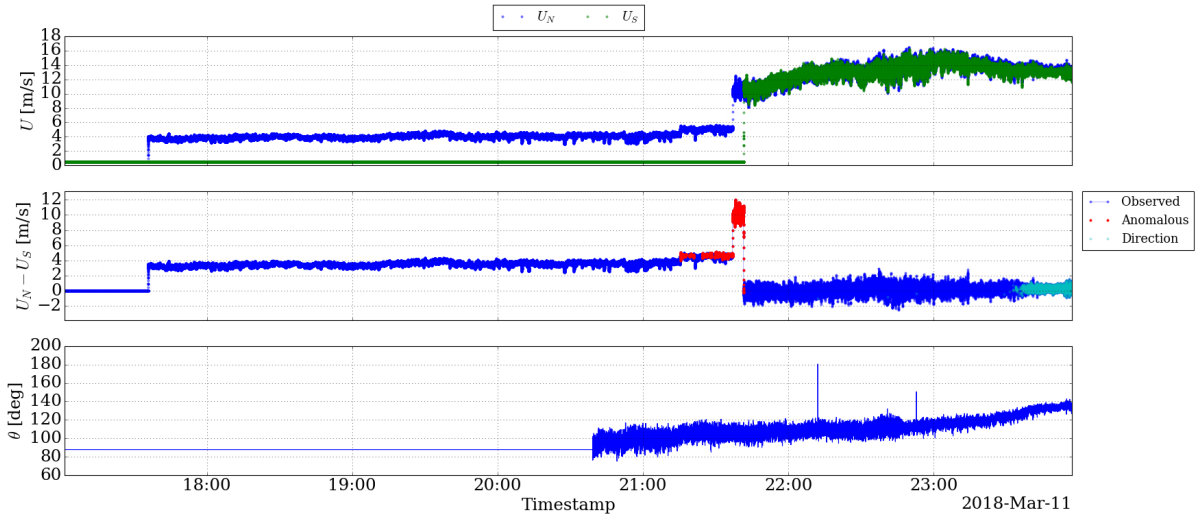
(b) Flagged measurements



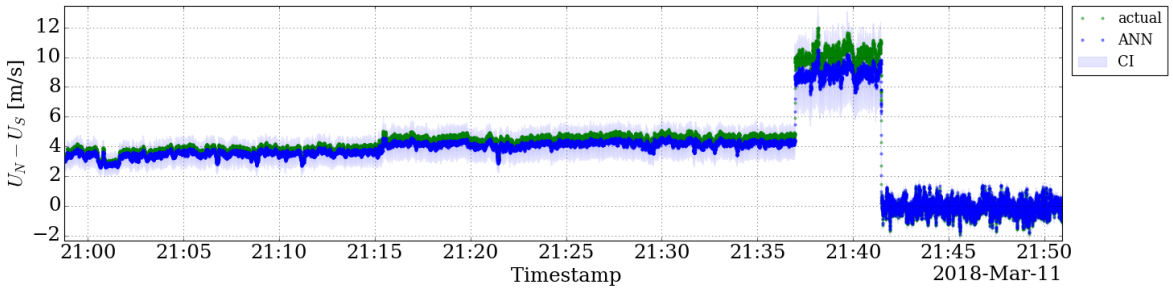
(c) Zoomed-in predictions

Figure 8.19: Model performance during the start of an icing event with measurements taken at 10Hz.

that failure of one cup significantly increases the variability of predictions, giving rise to larger confidence limits, which both contribute to limitations at lower wind speeds. When both cups are not frozen, the confidence limits are significantly smaller for similar wind speeds compared to when the southern cup is still frozen.



(a) Flagged measurements



(b) Zoomed-in predictions

Figure 8.20: Model performance during the end of an icing event with measurements taken at 10Hz.

More icing events are inspected as well, from which similar conclusions can be drawn. Another icing event is shown, together with the number of labels in each 10-minute period, in Figure 8.21. During this period, a low number of flags is observed, and it appears the wind direction sensor is frozen, meaning that it can not certainly be said that the alleged presence of ice specifically cause these rises. In both discussed events, however, steep increases before and after both instruments freeze over are observed, meaning that the model correctly gives an indication that icing may be occurring. Labelled data with true negatives, for example from installed cameras, is required for quantitative validation.

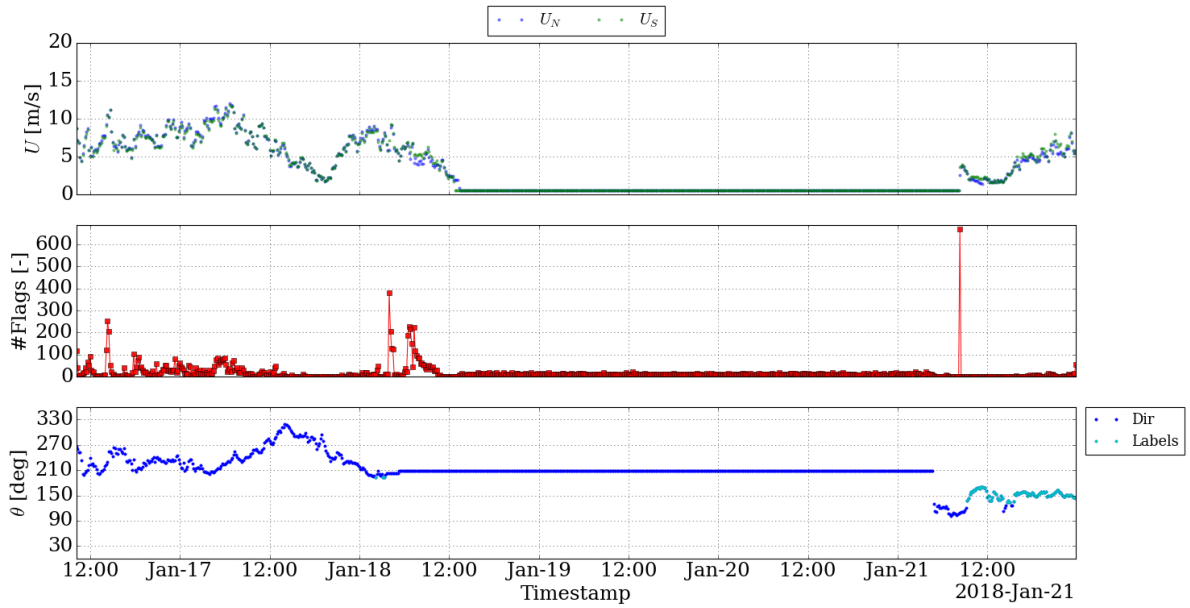


Figure 8.21: Second 10-minute averaged time series of the wind speed and direction during which all instruments freeze, together with the total number of flags per 10-minute period.

CHAPTER 9

Conclusion & Recommendations

Cup anemometers remain a popular choice for measuring wind speed, despite recent innovations within the area of measurement techniques. Once operational, however, three distinct error sources can cause cup anemometers to respond differently compared to their calibration and as such degrade the accuracy of their measurements. These three categories are: operational errors, mounting errors and mechanical errors. Where the first two error sources are quantified through uncertainty analysis, mechanical errors are only removed through conservative filtering and are thus defined as anomalous. This study has demonstrated the potential of a general detection model which exploits the spatial correlation between two collocated cup anemometers in order to flag these anomalies.

The spatial correlation of the anemometers becomes specifically apparent when filtering out mechanical errors according to logged timestamps of wrongful operation and extraneously measured variables, such as temperature and wind speed. This correlation is lower when wind direction sectors are included and either cup is in the wake of the mast they are installed on. Therefore, these sectors are removed as well, and a dataset that fully conforms to normal behaviour is created. Three different prediction models with increasing complexity (SGD, ARIMA and MLP) were trained using this data, resulting in regression methods that were able to predict the next residual between both cups' measurements in time. For each method, means were established to evaluate the predictions' confidence intervals. The ARIMA and MLP showed clear advantages over the SGD model, as they exhibited a much lower MSE over a 5-fold CV. The MLP model proved to be the most effective method of the three in terms of anomaly detection with a much lower FPR compared to the ARIMA model, because the adaptive confidence intervals perform better at higher wind speeds.

The MLP model proved to be most effective in modelling data conforming to normalcy, and thus this model was trained on the entire dataset to conduct a performance evaluation and validation. The model performance is evaluated by examining both the number of flags per 10 minutes to draw conclusions on a larger time scale as well as inspecting the 10Hz series in terms of anomalous densities and extremeness of anomalies.

The largest limitation of the model seems to its dependency on the magnitude of the wind speed. Though normalized inputs and adaptive confidence intervals are used, the number of flags showed similar trends when compared to the wind speeds on normal unseen data. The magnitude of fluctuations in wind speed are higher at higher wind speeds, meaning that more accurate predictions are made for lower wind speeds. Normalizing the number of flags by the wind speed may reduce these trends; the dependency on the results remain and need to be taken into account for future research

and interpretation of the results.

Periods during which one of the anemometers was operating with structural damage were clearly distinguishable, with significant increases in the number of of flags, the number of anomalous densities of flags and the extremeness of the found anomalies. Periods in which the cups freeze, however, were not as distinct. When one cup completely stops, e.g. in the case of icing, prediction uncertainties become very high. When higher wind speeds are present, the model is able to flag all measurements if either cup is failing. However, when a correctly working anemometer is slowing down, for example due to ice formation on the instrument, this large uncertainty combined with lower wind speeds causes measurements to not be flagged, making this period indistinguishable from normal behaviour when referring only to the flags. Just before the instruments stop, however, large spikes in flags are observed.

It can be concluded that using collocated cup anemometer measurements as means for anomaly detection can work, but much is still left unexplored. For example, different prediction models and statistical methods can be considered can be used to create an anomaly detection model. Changes in the set-up of the model can be considered as well, for example by predicting the ratio both instantaneous cup measurements. Furthermore, labelled data containing true positives and negatives to perform quantitative validation is required. Additional observational tools such as camera images or a heated reference instrument could be used to provide labels for icing events, which can then be used to validate the detection models capabilities under these circumstances. Other measurement set ups can be considered as well, such as a top-reference cup pairs or cups on the same direction but different altitudes. More machine-learning problems arise too. It can be valuable to employ clustering models on regions with relatively many anomalies to see if a certain error source results in distinct behaviour such that more specific labels can be assigned to measurements.

If the model in its developed state is to be used as a tool, several recommendations must be made as well. In particular, several simple filters of obviously faulty data points can be employed before the model results are used, for example when one of the cups is standing still. The number of flags on a 10-min scale can be used over long periods which may reveal periods in the time series where faulty data is gathered. Encountering a defined extreme has been shown to also be indicative of the start of an event that needs to be looked at. The most important conclusion at this conjecture is that the model can not be used yet as an independent filtering method. Visual inspection and sound considerations are still needed to ultimately determine whether regions with large flags are truly faulty or not. The model can aid in making this process easier and point out where these periods may be found, but expert judgement remains invaluable, even with machine learning tools.

APPENDIX A
Technical Drawings

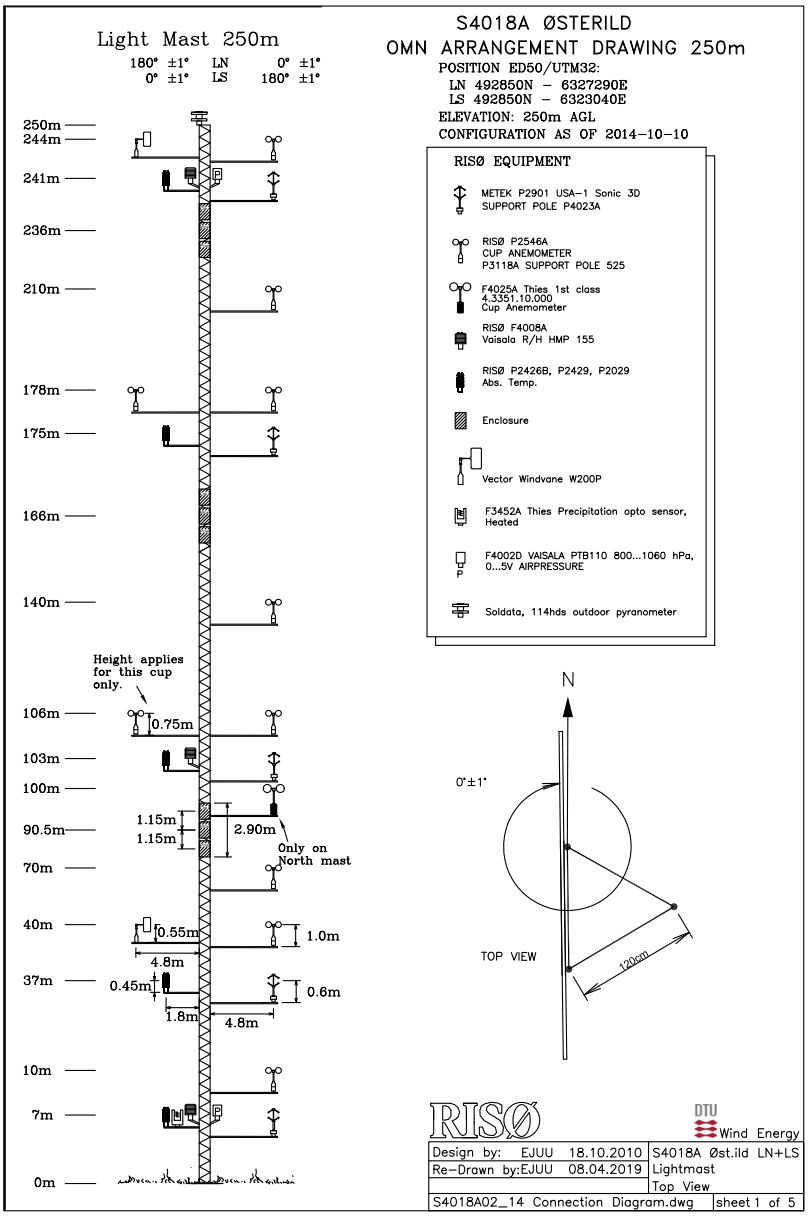


Figure A.1: Schematic overview of the LMN and LMS.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Azorin-Molina, C., Asin, J., McVicar, T. R., Minola, L., Lopez-Moreno, J. I., Vicente-Serrano, S. M., & Chen, D. (2018). Evaluating anemometer drift: A statistical approach to correct biases in wind speed measurement. *Atmospheric Research*, *203*, 175–188.

Bailey, B. H., McDonald, S. L., Bernadett, D. W., Markus, M. J., & Elsholz, K. V. (1997). *Wind resource assessment handbook: Fundamentals for conducting a successful monitoring program* (No. NREL/SR-440-22223; ON: DE97000250). National Renewable Energy Lab., Golden, CO (US); AWS Scientific, Inc., Albany, NY (US).

Baseer, M. A., Meyer, J. P., Rehman, S., Alam, M., Al-Hadhrami, L. M., & Lashin, A. (2016). Performance evaluation of cup-anemometers and wind speed characteristics analysis. *Renewable Energy*, *86*, 733–744.

Bégin-Drolet, A., Lemay, J., & Ruel, J. (2013). Time domain modeling of cup anemometers using artificial neural networks. *Flow Measurement and Instrumentation*, *33*, 10–27.

Bishop, C. M. (2005). *Neural networks for pattern recognition*. New York, Oxford University Press.

Bottou, L. (2012). Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade* (Pages 421–436). Berlin, Springer.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control - fifth edition*. John Wiley & Sons.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (Pages 93–104).

Busch, N. E., & Kristensen, L. (1976). Cup anemometer overspeeding. *Journal of Applied Meteorology*, *15*(12), 1328–1332.

Chan, T. F., Golub, G. H., & LeVeque, R. J. s. (1983). Algorithms for computing the sample variance: Analysis and recommendations, *37*(3), 242–247.

Chandola, V., Banerjee, A., & Kumar, V. (2010). Anomaly detection for discrete sequences: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *24*(5), 823–839.

Clark, S. H., Clay, O., Goglia, J. A., Hoopes, T. R., Jacobs, L. T., & Smith, R. P. (2009). Investigation of the NRG #40 anemometer slowdown. In *Proceedings of the AWEA Windpower Conference and Exhibition* (Pages 4–7).

Cortes-Ciriano, I., & Bender, A. (2019). Reliable prediction errors for deep neural networks using test-time dropout. *Journal of chemical information and modeling*, *59*(7), 3330–3339.

Dahlberg, J.-Å., Friis Pedersen, T., & Busche, P. (2006). *Accuwind - methods for classification of cup anemometers* (Risø-R-1555 (EN)). Denmark, Forskningscenter Risø.

Dahlberg, J.-Å., Gustavsson, J., Ronsten, G., Friis Pedersen, T., Schmidt Paulsen, U., & Westermann, D. (2001). *Development of a standardised cup anemometer suited to wind energy applications. publishable final report.* (Report: FOI-S-0108-SE, October 2001 EU contract JOR3-CT98-0263). [s.l.], Aeronautical Research Institute of Sweden.

Follrichs, U., & Glocker, S. (1997). Tilt angle sensitivity of different reference anemometers. In *State of the art on power performance assessment for wind energy conversion systems - 30th meeting of experts* (Pages 21–24).

Gal, Y., & Ghahramani, Z. (2014). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning* (Pages 1050–1059).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gupta, M., Gao, J., Aggarwal, C. C., & Han, J. (2013). Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *26*(9), 2250–2267.

GWEC. (2019). *Global wind report 2018* (Report). https://gwec.net/wp-content/uploads/2019/04/GWEC-Global-Wind-Report-2018.pdf

Heldman, D. R., & Moraru, C. I. (2010). *Encyclopedia of agricultural, food, and biological engineering*. Crc Press.

Herlau, T., Schmidt, M. N., & Mørup, M. (2019). *Introduction to machine learning and data mining*. Technical University of Denmark.

Hill, D. J., & Minsker, B. S. (2010). Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling & Software*, *25*(9), 1014–1022.

IEC 61400-12-1:2017. (2017). *Wind energy generation systems - part 12-1: Power performance measurements of electricity producing wind turbines. edition 2* (Standard). International Electrotechnical Commission.

Kimura, S., Abe, K., Tsuboi, K., Tammelin, B., & Suzuki, K. (2001). Aerodynamic characteristics of an iced cup-shaped body. *Cold Regions Science and Technology*, *33*(1), 45–58.

Kristensen, L. (1998). Cup anemometer behavior in turbulent environments. *Journal of Atmospheric and Oceanic Technology*, *15*(1), 5–17.

Lindelöw, P. J. P., Friis Pedersen, T., Gottschal, J., Vesth, A., Wagner, R., Schmidt, P., & Courtney, M. S. (2010). *Flow distortion on boom mounted cup anemometers* (Risø-R-Report-1738 (EN).). Denmark. Forskningscenter Risø.

Maccready, P. B. (1966). Mean wind speed measurements in turbulence. *Journal of Applied Meteorology*, *5*(2), 219–225.

Madsen, H. (2007). *Time series analysis*. CRC Press.

Manwell, J. F., McGowan, J. G., & Rogers, A. L. (2002). *Wind energy explained: Theory, design and application*. John Wiley & Sons Ltd.

Matsubara, Y., & contributors, G. (2018). *Pymysql documentation release 0.7.2* (Report). https://readthedocs.org/projects/trio-mysql/downloads/pdf/stable/

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference* (Pages 56–61).

MEASNET. (2009). *Anemometer calibration procedure. version 2 (october 2009)* (technical report). Madrid, Spain.

Metek. (No date). *Sonic Anemometer uSonic/uSonic-3 Scientific (former: USA-1)*. https://metek.de/product/usonic-3-scientific/

Papadopoulos, K. H., Stefantos, N. C., Schmidt Paulsen, U., & Morfiadakis, E. (2001). Effects of turbulence and flow inclination on the performance of cup anemometers in the field. *Boundary-Layer Meteorology*, *101*(1), 77–107.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (Pages 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Patterson, J. (1926). The cup anemometer. *Transactions of the Royal Society of Canada*, *20*, 1–53.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pindado, S., Barrero-Gil, A., & Sanz, A. (2012). Cup Anemometers' Loss of Performance Due to Ageing Processes, and Its Effect on Annual Energy Production (AEP) Estimates. *Energies*, *5*(5), 1664–1685.

Pindado, S., Cubas, J., & Sanz-Andrés, Á. (2013). Aerodynamic analysis of cup anemometers performance: The stationary harmonic response. *The scientific world journal*, *2013*, 1–11.

Pindado, S., Cubas, J., & Sorribes-Palmer, F. (2014). The Cup Anemometer, a Fundamental Meteorological Instrument for the Wind Energy Industry. Research at the IDR/UPM Institute. *Sensors*, *14*(11), 21418–21452.

Pindado, S., Cubas, J., & Sorribes-Palmer, F. (2015). On the harmonic analysis of cup anemometer rotation speed: A principle to monitor performance and maintenance status of rotating meteorological sensors. *Measurement*, *73*, 401–418.

Pindado, S., Sanz, A., & Wery, A. (2012). Deviation of cup and propeller anemometer calibration results with air density. *Energies*, *5*(3), 683–701.

Pindado, S., Vega, E., Martínez, A., Meseguer, E., Franchini, S., & Sarasola, I. P. (2011). Analysis of calibration results from cup and propeller anemometers. Influence on wind turbine Annual Energy Production (AEP) calculations. *Wind Energy*, *14*(1), 119–132.

Ramos-Cenzano, A., Ogueta-Gutierrez, M., & Pindado, S. (2019). On the output frequency measurement within cup anemometer calibrations. *Measurement*, *136*, 718–723.

Ramos-Cenzano, A., Ogueta-Gutiérrez, M., & Pindado, S. (2019). Cup anemometer measurement errors due to problems in the output signal generator system. *Flow Measurement and Instrumentation*, *69*, 101621.

Robinson, T. R. (1850). On a new anemometer. *Proceedings of the Royal Irish Academy (1836-1869)*, *4*, 566–572.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python, In *9th python in science conference*.

Siegel, D., & Lee, J. (2011). An auto-associative residual processing and k-means clustering approach for anemometer health assessment. *International Journal of Prognostics and Health Management*, *2*, 50–61.

Sim, S.-K., Maass, P., & Lind, P. G. (2019). Wind speed modeling by nested ARIMA processes. *Energies*, *12*(1), 69.

Solov'Ev, Y. P., Korovushkin, A. I., & Toloknov, Y. N. (2004). Characteristics of a cup anemometer and a procedure of measuring the wind velocity. *Physical Oceanography*, *14*(3), 173–186.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Sun, L., Chen, C., & Cheng, Q. (2012). Feature extraction and pattern identification for anemometer condition diagnosis. *International Journal of Prognostics and Health Management*, *3*, 8–18.

Swytink-Binnema, N., Godreau, C., & Arbez, C. (2019). Detecting instrumental icing using automated double anemometry. *Wind Energy*, *22*(1), 80–88.

Vega, E., Pindado, S., Martínez, A., Meseguer, E., & García, L. (2014). Anomaly detection on cup anemometers. *Measurement Science and Technology*, *25*(12), 127002.

Wadham-Gagnon, M., Swytink-Binnema, N., Bolduc, D., Tété, K., & Arbez, C. (2015). Ice detection methods and measurement of atmospheric icing. In *16th International Workshop on Atmospheric Icing of Structures (IWAIS)*.

Wind Sensor. (No date). *WindSensor P2546A-OPR Cup Anemometer*. https://www.windsensor.com/en/products/p2546a-opr-cup-anemometer

Wyngaard, J. C. (1981). Cup, propeller, vane, and sonic anemometers in turbulence research. *Annual Review of Fluid Mechanics*, *13*(1), 399–423.