Norwegian School of Economics
Bergen, Fall, 2020

A Comparison of Wind Forecasting Methods for Norwegian On-shore wind

A Perspective into the Nuances in Wind Speed to Power Conversion and the Economic Costs Associated with Wind Forecast Accuracy

Jae Meng Chong & Chris Kristiansen

Supervisor: Lars Jonas Andersson

Master Thesis, MSc in Economics and Business Administration,
Energy, Natural Resources, and the Environment &
Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Abstract

This paper examines various short-term forecasting methods to forecast hourly wind energy production in Norway. Performance of forecasting methods were compared across different months, through different evaluation metrics, to analyze the uniformity and dependability of methods. More than a decade of hourly wind speed data spanning across 69 locations along the Norwegian coast were utilized in the study.

Given the upcoming integration of EU and Nordic intraday electricity markets into the Cross-Border Intraday market (XBID), the study focuses on one-hour-ahead forecasting to be in alignment with the operations of intraday electricity market. With the final objective of predicting power production, a customized loss function, Power Curve Conversion Error with penalty, which takes into account both wind speed to power conversion and economic cost associated with over and under forecast, is included as part of the evaluation metrics to capture the true value of each model's predictions.

Forecasting methods undertaken consist of a mix of Statistical and Machine-Learning methods, with Naive forecasting used as the overall benchmark model. Other Statistical methods are ARIMA, and ARIMAX which includes the use of seasonal and time of day dummies. In terms of ML methods, Gradient Boosted Trees, Extremely Randomized Forest, and Neural Network are selected. Finally, a hybrid model of ARIMAX and Extremely Randomized Forest is also formulated. These methods are then evaluated on multiple evaluation metrics, namely: RMSE, MAE, Classification Accuracy, and the Power Curve Conversion Error with penalty.

The general implication of the study reveals that accuracy of models are consistent with their required computational intensity, with ML models outperforming statistical methods in most situations. The findings also suggest the Hybrid model to be the most suitable forecasting method for one-hour-ahead forecast under almost all evaluation metrics employed. This conclusion holds true for wind power forecasting under different seasons of the year as well.

Preface

This thesis is a culmination of our Masters of Science in Economics and Business Administration programme at the Norwegian School of Economics. The choice of topic, which integrates both our majors: *Energy, Natural Resources and the Environment and Business Analytics;* originated from our shared interest surrounding the field of energy data analysis, specifically relating to the integration of renewable energy into power markets.

To build or not to build wind farms in Norway has quickly become one of the most controversial and prevailing topics in the country. The Norwegian government's initiative to develop wind farms in Norway was met with significant resistance at ground level, with numerous projects forced to an halt in the past years (Adomaitis, By, & -, 2020). Given the timeliness of it all, working on this project has been riveting for the both of us. Questions such as how to balance the need for renewable energy and the protection of natural landscape; as well as how to quantify the economic benefits of these wind farms have been widely discussed. Hence, we decided to take on a different approach and focus on integration effects of this highly intermittent source of energy in the Cross-Border Intraday electricity market.

We are extremely grateful to our supervisor, Professor Lars Jonas Andersson, for his prolific feedback and insights. Jonas has been an exceptional mentor to us, generously offering his precious time for us despite how disruptive the volatile covid-19 situation in Bergen has been throughout the semester. A special shoutout to Professors Mette Bjørndal and Endre Bjørndal, who were not only with us at the beginning of our brainstorming process, but also pointed us to the right direction based on our interests, and gave us their continued assistance through clarifications on the datasets. We would also like to extend our gratitude to Professor Gunnar S. Eskeland, for his invaluable sharing and prompts that were especially critical for the formulation of our research question for the thesis. Last but not least, none of this would have been possible without the datasets from Kjeller Vindteknikk.

Contents

List of Figures

List of Tables

Acronyms

ANN - Artificial Neural Network

ARIMA - Auto Regressive Integrated Moving Average

ARIMAX - Autoregressive Integrated Moving Average with Explanatory Variable

ETR - Extra Tree Regressor

EU - European Union

GBT - Gradient Boosting Tree

GHG - greenhouse gases

EU - European Union

HAWT - Horizontal axis wind turbine

MAE - Mean Absolute Error

ML - Machine Learning

MLP - Multi-layer Perceptron

MSE - Mean Square Error

PCCEp - Power Curve Conversion Error with penalization

RES - Renewable Energy Source

RF - Random Forest

XBID - Cross-Border Intraday electricity market (EU and Nordic)

1. Introduction

The following section sets the stage for the rest of the paper with a brief account on the motivation behind the choice of topic, the general objective, as well as the relevance of the thesis. Besides that, it also aims to equip the reader with an overview of what to expect from the paper with a road map outlining the overall structure of the paper.

1.1 Motivation and Relevance

The adverse effects of anthropogenic climate change has become one of the most dire issues faced by modern society. Visible repercussions from climate change such as record high ambient temperatures, the rise of sea level, and the melting of ice caps, are already affecting the most vulnerable species in regions of Earth. Plants and animals are driven into extinction, with not just animals being displaced from their natural habitats, but also people from their homes. Just earlier this year at the World Economic Forum in Davos, Filippo Grandi, the UN High Commissioner, even proclaimed that the world needs to prepare for millions of people being driven from their homes by climate change ("UN," 2020), underscoring the monumental disruption to be expected from climate refugees and migrants in the near future.

Energy, albeit the key driver of development and economic growth, is also the largest contributor of greenhouse gas (GHG) emission. More than one quarter of GHG emissions in the European Union stems solely from electricity production ("Greenhouse Gas Emissions by Aggregated Sector," 2019). The constant growth in the human population, especially so in developing countries, coupled with their need for improved living conditions, ensues continuous development and year on year increase in energy demand. According to the International Energy Agency (IEA), CO2 emissions have been growing exponentially since the 1870s, spotting significant variances across countries and regions based on their level of economic development (Birol, 2017, p. 10).

British economist, Nicholas Stern, once described climate change as "the greatest market failure the world has ever seen.". Economically speaking, climate is both non-excludable and non-rivalrous, making it a public good; while greenhouse gases (GHG) emitted from economic activities represent negative externalities in their capacity in trapping heat, accentuating the need for government intervention.

With increasing awareness on the negative impacts of GHG as a global pollutant, global treaties such as the United Nations Framework Convention on Climate Change (UNFCCC) and its Kyoto Protocol and Paris Agreement were put in place to promote the global efforts on GHG reductions. A plethora of national policy measures, such as carbon trading, pollution taxes, emission quotas, as well as grants targeted at promoting sustainable energy, have also been introduced around the world by respective governments in light of the abatement potential in the sector. That said, the equitability associated with these environmental regulations are highly contented, given the heterogeneity and nuances surrounding the topic, ranging from income inequality, to international economic development inequality, and to varying degrees of decline in environmental quality in which people are exposed to.

Adoption of renewable energy, seems to be the only agreeable and executable solution in reducing emission intensity of our economies, considering how improbable the alternative solution of a worldwide collective reduction on the level of economic activity is. Having said that, one should not discount the fact that renewable energy sources (RES) also carry their own baggage of environmental impacts such as noise and sight pollution, ecosystem disruption, and water pollution.

More than 25% of the world's electricity demand today is being met by renewable sources, with growth driven mostly by China and the United States (Petrova, 2018). The cost of renewables have hence decreased significantly following the rising interests in investing in the sector. This remarkable adoption of renewable energy which has been impelled predominantly from strong policies support, however, also accentuates the greatest challenge relating to the growth of renewable en-

ergy supply in the overall energy mix – the intermittency of these energy supplies which relies heavily on weather and environmental conditions and the lack of cost-effective storage solution.

This uncertainty faced by wind power producers has been compared to by many as the newsvendor problem (DeMarle, 2019). In the electricity market, supply and demand must always be in balance. Hence, the variability surrounding the supply of wind poses a big challenge to the reliability and stability of the electrical system, which inevitably translates into monetary penalties for under/over electricity production, leading to high price volatility in the affected markets. In the absence of storage solutions, accurate models for predicting power output will be critical to establish wind energy as a reliable source of energy for practical applications such as generation scheduling, maintenance scheduling, and security analysis and energy transactions (Du, Wang, & Niu, 2019).

1.2 Research Question

Considering the aforementioned induction, the aim of the thesis is to answer the following research question:

"What is the most suitable forecasting method to predict hourly wind power production (1-hour-ahead) in Norway under different accuracy evaluation tools?

Do the methods' performance vary for different seasons of the year?

The paper contributes to the existing literature by focusing on short-term wind power forecasting in Norway. One could argue that the flexibility of Norwegian hydropower reservoirs could, theoretically, act as a storage capacity to combat the intermittency supply of wind energy within Norway. However, since Norwegian hydropower is mostly state-owned, it will be optimized to maximize the welfare of Norwegians specifically (Energy, 2016). The integration of EU and Nordic intraday electricity markets forces these supply dynamics to extend outward to the rest of the EU. Hence, more likely than not, these will be translated into economic costs elsewhere.

On the premise of this being a hypothetical study, it is not possible to investigate the market effects and actual economic cost associated with poor forecasts of the Norwegian wind power. Therefore, the solution adopted was to instead, model possible economic costs faced by system operators from the over and under forecasting of power. Therefore, a tailored loss function specific in answering the research questions and in capturing the nuances surrounding the research question is developed. It penalizes inaccuracies in forecasts based on each location's wind speed to power conversion curve, as well as taking into account the economic cost faced by system operators associated with the over and under forecasting of power through a penalty value.

Wind speed forecasting, can be categorized by time-scale into very short-term (seconds/ minutes), short-term (half-hour/hours), medium-term (hours/days), and long-term (weeks/months), and selected based on the objective of the study as well as wind speed data available (Farzaneh & Majid, 2017). The time duration of one-hour-ahead was decided based on the objective of focusing on the integration of wind energy into the power system (Lei, Shiyan, Chuanwen, Hongling, & Yan, 2009). Therefore, even though different forecasts can bring about its own economic significance and advantages, this paper does not aim to investigate wind speed forecasting methods of other time scales.

1.3 Thesis Overview

The remainder of this thesis is structured as followed:

Section 2 lays the foundation with the background on wind energy generation, features of the power curve, the value of forecasting short term wind speed, as well as the relevant literature upon which this study is based on.

Section 3 presents an overview of all data employed in the study, and delves into the key data set used in the analysis. It highlights relevant features in the wind speed data set with descriptive

statistics illustrating the variability of wind speed, followed by an examination into their stationarity. This section then moves on to reveal details on the programming softwares and libraries utilized for the study, and the steps taken for data preprocessing.

Section 4 outlines the empirical approaches of the analysis, elaborating upon all the forecasting methods deployed and the evaluation tools used to compare model performances.

Section 5 presents the findings from the analysis, while Section 6 discusses the results and deliberates over the limitations and shortcomings on both the methods prescribed, as well as of the study.

Finally, the paper concludes with Section 7, with the results obtained to answer our research question.

2. Background

The objective of this section is to arm the reader with some fundamentals on the related topics before diving into data and methodologies. This section will start off with the rudiments of wind energy and its power conversion, followed by a rundown on the value surrounding wind speed forecasting, and finally wrapping up with a review on existing literature.

2.1 Wind Energy Formation and Transformation

The exploitation of wind energy is not a new concept to people and has been around for over thousands of years, with usage ranging from transportation in the form of sailboats, to water pumps and windmills for food production.

Focusing on the concept of harnessing wind energy specifically for electrical power, the basic energy transformation process behind the creation of wind energy as well as the conversion of wind energy into electrical energy can be simplified to the following:

$$Solar \Rightarrow Wind \Rightarrow Mechanical \Rightarrow Electrical$$

2.1.1 Solar to Wind Energy

Wind energy can be categorized as a secondary generation of solar energy, and is formed by the absorption of radiation energy by the atmosphere. It is caused by a combination of the three concurrent events that creates differences in air pressure:

1. The sun's uneven heating of Earth's atmosphere (temperature gradients between poles and equator);
2. The irregularities of Earth's surface (land-sea distribution);

3. The rotation of Earth (Coriolis effect).

Simply stated, wind turbines operate opposite to a fan. Instead of using electricity to create wind, they use wind to generate electricity. Although a myriad of wind turbines are being designed and manufactured, this paper focuses on the most commonly used technology - lift-type Horizontal axis wind turbine (HAWT). (Rehman, Alam, Alhems, & Rafique, 2018)

2.1.2 Wind to Mechanical Energy

In the presence of wind, a wind turbine converts the kinetic energy from wind into electrical energy using the aerodynamic forces created by the turbine's rotor blades. When air flows over any surface, lift (perpendicular) and drag (parallel) motions are created concurrently, both of which can be captured to rotate a turbine's blade. That said, the majority of wind turbines deployed are lift-type rather than drag-type due to their better energy conversion efficiency. In a lift-based turbine, the airfoil shape of each turbine's blade creates differences in air pressure on the top and bottom side of the blade, leading to a stronger lift than drag, causing the rotor to spin. Illustrated in figure 1. *Source:* ("Lift and Drag," n.d.)

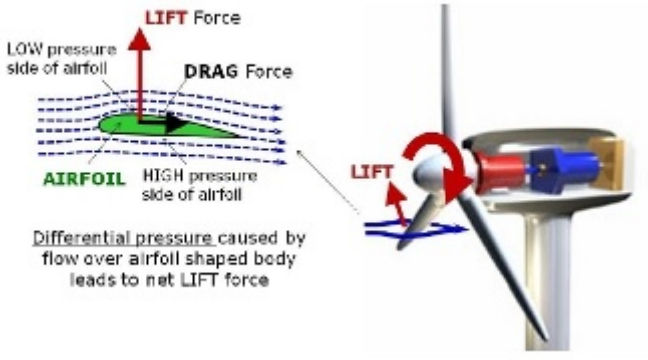

Figure 1: Mechanics of lift-type wind turbine

Kinetic Energy of Wind Parcel can be calculated as follows:

$$E(kinetic) = \frac{1}{2} \cdot M \cdot V^2 \tag{1}$$

$$M = \rho \cdot V \;\Rightarrow\; Mass\ of\ wind\ parcel = Density\ of\ air \cdot Volume\ of\ air\ parcel \tag{2}$$

$$V = A \cdot 1 = \pi r^2 \cdot vt \;\Rightarrow\; Volume = Cross\ sectional\ area \cdot arbitrary\ length \tag{3}$$

2.1.3 Mechanical to Electrical Energy

To create electricity from wind, the shaft of the turbine must be connected to a generator. The generator uses the turning motion of the shaft to rotate a rotor which has charged magnets and is surrounded by copper wire loops. Electromagnetic induction is created by the rotor spinning around the inside of the core, generating electricity. This is illustrated in figure 2 below, retrieved from: (Elsayed, 2017). Electricity is transmitted through a transformer in order to increase its voltage and make it successfully transfer across long distances. Power stations and fuse boxes receive the current and then transform it to a lower voltage that can be safely used by business and homes.

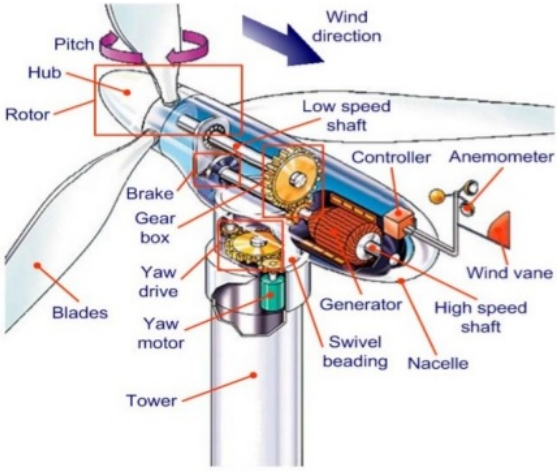

Figure 2: Major components of wind turbine

2.1.4 Betz' Law

According to the first law of thermodynamics, the law of conservation of energy states that energy cannot be created or destroyed, and can only be converted from one form to another. Betz' limit, which was first conceived by German Physicist Albert Betz in 1919, builds upon that theory and states that any wind turbine can only convert up to a maximum of 59% of kinetic energy from the wind into mechanical energy.

As illustrated in the fig. 3 below, if 100% of wind energy is converted, $P_{out} = 0$, giving us $V_{out} = 0$. Under that circumstance, there will be no movement of air past the turbine, which is not only unrealistic, but also results in no further movement of air into the wind turbine to generate more electricity. The only way for that scenario to hold true is if the turbine rotor blades somehow covers the full area, capturing all possible flow of air, while at the same time leaving room for receiving new air parcels.

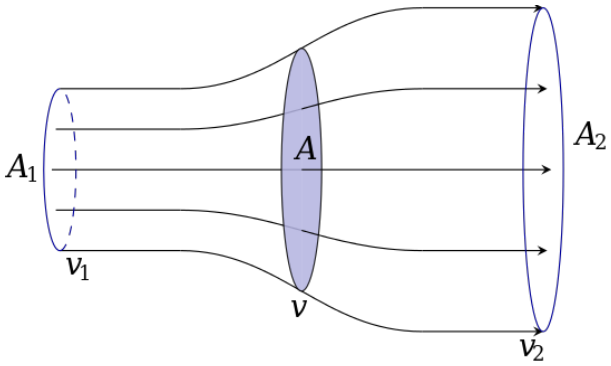

Figure 3: Air stream tube

Therefore, the total incoming wind energy that a wind turbine can produce is the energy difference between the energy of incoming wind into the turbine, and energy of outgoing wind from the turbine with efficiency measured by the ratio of extracted power and potential power as followed:

$$
\begin{aligned}
Efficiency(\eta) = \frac{P_{out}}{P_{in}} &= \frac{1}{2} \cdot \frac{(V_{in} + V_{out})(V_{in}^2 - V_{out}^2)}{V_{in}^3} \\
&= \frac{1}{2} \cdot (1 + x - x^2 - x^3) \ , \\
&\textit{where } x = \frac{V_{out}}{v_{in}}.
\end{aligned} \tag{4}
$$

The theoretical maximum value of $x = 1/3$ will result in the upper theoretical efficiency of 59%, where the best distance between each wind turbine is at 5 times the diameter of turbine rotor blade length apart.

Assuming a 100% energy conversion from kinetic energy of wind to mechanical, limited by Betz limit power coefficient of 59% Cp_{max}, power capacity P_{max} of each wind turbine can be calculated as follows:

$$
\begin{aligned}
P_{max} &= Cp_{max} \cdot \frac{E(kinetic)}{t} \\
&= Cp_{max} \cdot \frac{1}{2} \cdot M \cdot \frac{V^2}{t} \\
&= 0.59 \cdot \frac{1}{2} \cdot \rho \cdot \pi r^2 \cdot v^3
\end{aligned} \tag{5}
$$

It is evident that the final power production of a wind turbine is dependent upon a number of parameters such as air density - which is conditional on temperature, pressure, and humidity; turbine parameters - radius covered by the turbine blades; and last but most importantly, the velocity and direction of wind.

The cubic relationship of wind power output from a wind turbine and velocity of wind speed v^3, as represented in the final equation above, further accentuates the importance of predicting accurate

wind speed as it can lead to substantial differences in predicting wind power. Based on this relation, the doubling of wind speed can increase theoretical wind power eight folds.

Besides the Betz limit power coefficient, the capacity factor of a wind turbine, which indicates the ratio between a turbine's average and peak output, also affects the final energy production on site. The capacity factor depends not only on turbine characteristics, but also site specific wind characteristics with equation as followed:

$$Cf = \frac{P_{avg}}{P_{peak}} \tag{6}$$

2.2 Power Curve

Building upon aforementioned energy conversions principles, the evaluation of a wind turbine's actual power output can become rather complex when one attempts to take all influencing factors into consideration. To simplify this, each manufactured wind turbine today comes with a power curve modeling their optimal power performance. These power curves are derived from field measurements, with an anemometer placed close to the wind turbine to record and plot wind speed measurements against actual power output from the turbine ("The Power Curve of a Wind Turbine," 2003). The power curve, which depicts the relationship between the expected output power and wind speed at hub height, is especially useful in terms of energy assessment and turbine performance monitoring. Three different models of wind turbines (Effect 1, 2, and 3) from Kjeller Vindteknikk will be utilized in this paper as the conceptual wind speed to power output conversion of the sites. Their characteristics will be expanded upon in Chapter 3.

Figure 4 below illustrates a typical wind turbine power curve, retrieved from: (Xiao, Zhao, Yang, & Zhu, 2020). A power curve can be defined into four different zones based on the turbine's operational features, corresponding to a turbine-specific cut-in wind speed V_{in}, rated wind speed V_{rated}, and cut-out wind speed V_{out}. The cut-in wind speed represents the minimum wind speed

required for turbine blades to start rotating. Conversely, the cut-out wind speed represents the wind speed at which a turbine is forced to cease operation to protect its system from damage. Rated wind speed on the other hand refers to the lowest wind speed necessary for the turbine to generate maximum power.

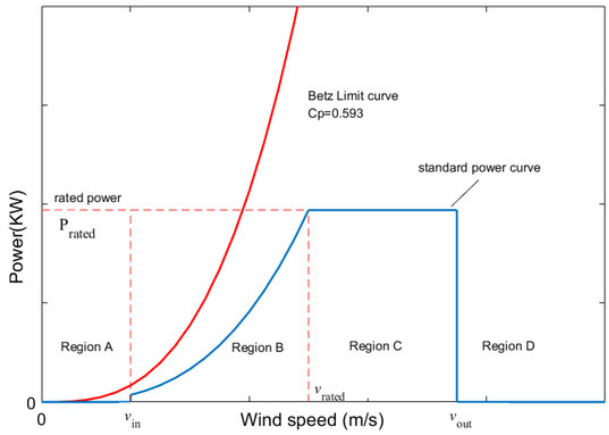

Figure 4: Theoreticaly power curve

Ergo no power is generated in Region A where actual wind speed is below turbine cut-in speed. Moving on to Region B, which is bounded by cut-in and rated wind speed, is the only region where the cubic relation between wind speed and power output from eq. 5 holds. As for Region C, which is bounded by the rated wind speed and cut-out wind speed, power output remains constant at rated power regardless of increment in wind speed within the region. Finally, when wind speed is greater than cut-out speed, as illustrated in Region D, turbine blades stop rotating, and power output drops drastically from maximum power to zero.

These characteristics of a wind turbine power curve represent an intriguing problem in the wind forecasting paradigm as the relationship between wind speed to power production differs greatly from having no effect to a cubic effect in the event of an inaccurate forecast.

When wind speed falls below the cut-in and above the cut-out as represented by Region A and D, no electricity is generated by the wind turbine. Therefore, when comparing operationally, Regions

B and C are considered the main working regions of a wind turbine. However, since our interest lies in the accuracy (and the effects of said accuracy) of wind speed forecasts, Region B is the key sector of our focus on inaccuracy penalization. Inaccurate forecasts occurring within Region A, C, and D will not be penalized as these will not translate into any forecast and actual power differences. This rationale will form the basis of our tailored evaluation method - Power Curve Conversion Error, which will be further elaborated upon in Chapter 4.

In spite of the utility of power curves, there exists uncertainty around the measurement of power curves, with studies indicating conversion errors of up to +/- 10% in certified power curves ("The Power Curve of a Wind Turbine," 2003). This is due to the strong fluctuations in site/time-specific wind speed, direction, and air density, adding to the difficulty in exact measurement of the column of wind gust passing through the rotor at any one point. The raw power curve, pre-processed for the public, encompasses cluttered points and requires averaging of all collected measurements to obtain the spruced power curves shown in fig. 4.

In order to combat the additional uncertainties that could be introduced by power curve conversions, rather than taking forecast accuracy measurements post speed to power curve conversion, this study incorporated the conversion effects into the evaluation tools instead. This will be further elaborated upon in Chapter 4.4.

2.3 Value of Forecasting Short-term Wind Speed

In a competitive electricity market such as the Nordic and EU, accurate forecasting of wind speed can bring in great value as market prices are determined based on the cost of energy imbalances (Farzaneh & Majid, 2017). Besides that, it is also a critical aspect in developing a robust and well-functioning hour-ahead and day-ahead markets (Wu & Hong, 2007).

2.3.1 Effects of Intermittent RES in Electricity market

Economically, renewable energy sources (RES) with intermittent supply (such as solar and wind) differ greatly from traditional sources of energy (such as coal, oil, and gas) in the fundamental fact that they do not require raw material for electricity generation. Hence, the cost of these intermittent RES are mostly fixed and predictable - during the initial setup, and the subsequent maintenance. This advantage of low operating costs grants them the ability to drive down electricity prices in the presence of strong supply.

The German market is a good example of an electricity market spotting a large intermittent RES mix. Gas power plants, which are cheaper to turn on and off compared to coal plants, are typically put on standby to make up for the shortage in energy production from these intermittent RES. Hence, during the time of low sun / wind, electricity prices will surge up to that of gas prices; while prices can even be driven down to negative during the periods of high renewable power output (Ådland, 2018).

The presence of negative prices stems not only from low operating cost of RES as mentioned previously, but also due to government subsidies, auctions, and other fixed price agreements such as power purchase agreements (PPA), which guarantees RES producers to make a profit despite the negative price biddings. On top of that, the significant cost associated with starting and shutting power plants also forces some plants to stay online during RES peak despite making a loss.

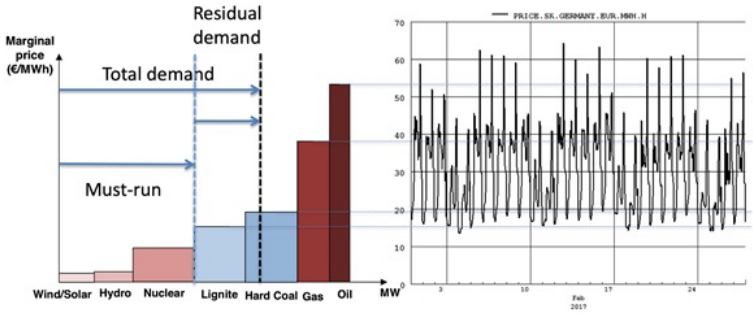

Figure 5: Energy supply stack and electricity prices

Referring to the example of energy supply stack and electricity prices in fig. 5 on the previous page, the price for any given hour reflects the marginal costs of the most expensive technology that needs to be activated in order to meet demand (Ådland, 2018).

This increase in supply-side intermittency, without sufficient flexible supply/ storage options, will therefore result in an increase in price gap between hours supplied by RES and the residual hours of demand. Coupled with the complexity wind farms face in predicting accurate power output (Papaefthymiou, 2009), precise forecasting of wind speed can hence be particularly beneficial in mitigating the fluctuations and risk electrical grid faces upon accepting wind energy (Smith et al., 2007).

2.3.2 Value of Adding Wind into Norway Energy Mix

Studies have identified the Norwegian coast as one of the most suitable locations on Earth for the harnessing of wind energy (Borsche, 2019), yet little research has been done on the potential economic cost from the integration effects of introducing this additional source of energy into the Cross-Border Intraday electricity market (XBID).

Norway's power generation is currently highly characterized by the significant presence of Hydropower - a RES distinguished for its flexibility, and most studies have been made relating to the optimization of the Norwegian hydropower capacity to regulate the influx of intermittent RES from European Union(Graabak, Korpås, Jaehnert, & Belsnes, 2019), none focusing on the Nowerigian wind potential - which is expected to double in capacity and meet electricity demand for up to 10% of the Norwegian population by 2021 (Reuters, 2020).

As much as the flexibility of Norwegian hydropower reservoirs could theoretically act as a storage capacity to combat the intermittency supply of wind energy, considerable dynamics, especially in terms of monetary impacts, will still be at play with the introduction of this additional source of energy given the integration of EU and Nordic electricity markets. Since traditional sources of

energy such as coal, oil, and gas drive power prices, demand and the supply of renewable energy drives intraday prices (Ådland, 2018). This thesis hence aims to bridge the gap by finding the most suitable method to predict hourly wind energy production one hour ahead for the intraday electricity market.

Besides that, given that Norway's hydropower is currently providing for almost all energy consumption within the country, the integration of the Nordic and EU electricity markets give rise to the opportunity for Norway to export clean energy to neighbouring countries, in alignment with the EU 2030 renewable energy target. Not only does this facilitate the phasing out of traditional and non-renewable forms of electricity generation such as coal and oil in other countries where harassing renewable energy is not physically or economically feasible, it also serves as a potential source of export for Norway.

2.4 Literature Review

The following section introduces the literature reviewed related to the topic, and gives credit to their contribution in the selection of methodologies adopted in this study. It will finally conclude with the differentiating points of this study to all existing literature.

One of the key aspects that set human beings apart from all other living things is understanding the concept of time. This has translated into a fixation for anticipating and planning for the future. Honing the art of predicting has henceforth been of priority since the age of time, with the first probability theory dating back to the late 17th centuries ("Foresightr," 2016).

Delving into the topic of interest, a multitude of research has been carried out in the realm of wind speed forecasting, incorporating various methods ranging from statistical to machine learning, and even the formulation of wind-specific methods to take into account the diurnal, nonnegative, and volatile nature of wind speed (Gneiting, Larson, Westrick, Genton, & Aldrich, 2006). This phenomenon is expectedly so given the global potential of wind power, the challenges to grid integrity

(Ayodele, Jimoh, Munda, & Agee, 2012), and economic costs associated with poor predictions elaborated upon in the previous subsection.

Wind speed forecasting exercises can have different time-scales based on the objective of the study. Journal (Farzaneh & Majid, 2017) investigates the defining aspects of each time-scale, starting from very short-term wind speed forecasting, which focuses on the forecast range of seconds to half-hour, all the way to long-term wind speed forecasting, forecasting between weeks to months. Besides that, (Farzaneh & Majid, 2017) also shed light on the objectives tied to each respective time scale. Based on the study, and further substantiated by (Lei et al., 2009), short term wind speed forecasting (half to six hours), aligns with the objective of this study, and is ideal for tackling problems related to power scheduling and grid integration of wind power. With the interest of the intraday market in mind, one-hour-ahead forecasting was ultimately decided upon.

Following that, (Yang, Zhang, Cui, Yang, & Huang, 2019) investigates into the effects of spatial heterogeneity in wind speeds and wind turbine power curves on wind power predictions errors. (Yang et al., 2019) proposed two possible approaches to wind speed and power forecasting, the first was to forecast wind speed before converting it into respective wind power forecast based on turbine power curve conversion tables, while the second was to formulate models directly to forecast power by mapping the relationship of wind speed data with power output. (Yang et al., 2019) reveal the effects of seasons and installed wind farm capacity on final wind power forecast errors, and established that errors in short-term wind power prediction can often be attributed to the inhomogeneity of the power curves (Yang et al., 2019). Hereinafter, this study will approach the research questions with the first method, taking special considerations in seasonal variations, as well as wind speed to power conversions.

Moving on to the method selection for the study, (Jiang, Yang, & Heng, 2019) and (Wang, Luo, Grunder, & Lin, 2017) categorizes short-term forecasting methods into four main types: physical, statistical, machine learning and hybrid models. Physical methods such as numerical weather

prediction (NWP) models although highly accurate, requires extensive environmental and atmospheric information along with high computational complexity as explained in (Wang, Wang, & Wei, 2015) and (Zhao et al., 2016). Following this, this study will focus on the comparison between statistical, ML, and hybrid methods.

ARMA/ARIMA models seem to be one of the more popular statistical methods adopted for short-term wind speed forecasting as seen in both (Torres, García, Blas, & Francisco, 2005) , and (Grigonytė & Butkevičiūtė, 2016), where hourly wind speed data from Navarre, Spain, and Riga, central Latvia, were used for forecasting. Both papers highlight the forecast superiority of the ARIMA model over the persistence (NAIVE) model based on RMSE comparison.

Building upon that, (Du et al., 2019) combines statistical methods with machine learning techniques for forecasting very short-term wind speed, one/two/three-step-ahead.. In (Du et al., 2019), wind speed data in the internals of 10-min from the Sotavento wind farm in Galicia, Northwestern Spain was used, while the hybrid model, known as Multi-objective moth-flame optimization (MOMFO), was a combination of Statistical method - Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) and ML method - Wavelet Neural Network (WNN). The performance of the MOMFO was next compared against a persistence model, a statistical model (ARIMA), and a machine learning model (Least Square Support Vector Machine, LSSVM), based on evaluation metrics MAE, RMSE, and MAPE. It concludes with MOMFO obtaining more accurate and stable wind power prediction amongst all other models.

Expanding the research into a time-scale relevant to this study, both (Farzaneh & Majid, 2017), and (Cadenas & Rivera, 2010) also performs short-term forecasting based on the idea of a hybrid model. Unlike (Du et al., 2019), both studies focus on one-hour-ahead hourly wind speed forecasts, and formulate hybrid models through the combination of ARIMA and NN. In both journals, ARIMA model is first employed to forecast the given wind speed time series, after which the errors were used in ANN to consider the nonlinear tendencies remaining in an attempt to reduce the final

error. (Farzaneh & Majid, 2017) utilizes hourly wind speed data from Binalood, Iran and evaluate the performance of 3 hybrid models - ARMA-MLP (Multi-layer Perceptrons), ARMA-RBF (Radial Basis Function), ARMA-ADALINE (Adaptive Linear Element) against ARMA model. Interestingly, (Farzaneh & Majid, 2017) concludes with the ARMA model outperforming the other three hybrid models based on MSE criteria, validating the competence of simple methods. In the other journal, (Cadenas & Rivera, 2010), hourly wind speed data from three different locations in Mexico were used to build three different models: ARIMA, ANN, and the proposed hybrid model ARIMA-ANN. Forecast performance of all three models were compared against statistical error measures, Mean Error (ME), MSE, and MAE, across all three sites, with results indicating the hybrid ARIMA-ANN model to outperform the other two models across all three sites and in all three accuracy measures. These two papers show the variation in forecast results on different wind data sets despite similar methods undertaken. It accentuates the fact that the 'best' forecasting method is often both data and evaluation metrics dependent, hence the necessity of this study to experiment with various forecasting and evaluation methods to pinpoint the most suitable model based on the norwegian wind speed data we have on hand.

Comparing accuracy evaluation methods, most of the aforementioned journals exploit traditional statistical error measures namely, ME, MAE, MSE, RMSE, and MAPE to compare models' performance. Research paper (Hering & Genton, 2010), however, scrutinizes these errors and asserts them as inappropriate in the wind forecasting paradigm for failing to capture properties of wind speed to power conversion into account. It instead introduces a new loss function, the Power Curve Error (PCE), as an alternative for a more realistic assessment of wind speed forecasting models. The PCE loss function in (Hering & Genton, 2010), is formulated with Generalized Piecewise Linear form to replicate a nondecreasing function, mapping speed to power conversion. In addition, PCE also included a penalty weight to penalize asymmetrical losses associated with under and over forecasting. It was the first evaluation metric that attempts to take into consideration the economic impacts of poor wind forecasting. In (Hering & Genton, 2010), a penalty weight of 0.73 was adopted to replicate a heavier penalization for under forecasts in model errors based on em-

pirical evidence from paper (Pinson, Chevallier, & Kariniotakis, 2007). This paper attested down regulation to be more expensive than up-regulation due to the higher cost incurred when system operators over-order from traditional sources. Finally, (Hering & Genton, 2010) failed to take into account the economic cost associated with different errors at different magnitudes of wind power. This paper therefore intends to build a loss function upon this idea from (Hering & Genton, 2010), but improve on it with the inclusion of the weight allocation based on magnitudes of wind power to capture the associated economic cost.

The research papers covered in this section demonstrate the popularity of wind speed forecasting in academia. They have imparted critical facts on the topic to help in the decision making for methodologies selection to be used in this paper. Unlike literature presented above, this study contributes to existing research with the focus on Nowergian wind speed forecasting and the economic cost associated with poor forecasts. Majority of current literature that are Norway and wind power specific revolves around the use of hydro-reservoirs as natural 'batteries' to smooth intermittency of RES as seen in (Graabak et al., 2019), (Førsund, Singh, Jensen, & Larsen, 2008), and (Destro, Korpås, & Sauterleute, 2016), underscoring a lack in literature surrounding the economic costs, which should be increasingly important given the integration into EU market. Finally, the reviewed literature have also served as a major inspiration on one of the key contributions of this paper - the development of a tailored loss function specific to this study.

3. Data

This section first familiarizes readers to all data sets utilized in the paper, followed by a deep-dive into the relevant features and compositions of the main data set - Norwegian Wind Speed data. It aims to illustrate the stochastic nature of the wind speed through visualization and interpretation. The section then concludes with stationarity tests, to establish a generic idea on linear assumptions in the data, and necessary steps taken to prepare the data for further analysis.

3.1 Overview of Dataset

Three data sets obtained from Kjeller Vindteknikk are used in this paper:

1. Norweigian Wind Speed Data (wind_data),
2. Wind Speed to Power Conversion Data (powerKVT),
3. Site Location Details Data (mill_data)

3.1.1 Nowegian Wind Speed Data

The main data set, Norwegian Wind Speed Data, is a complete data set containing hourly time series wind speed data across 69 locations along the Norwegian coast. The duration of data spans across 13 years, from 01.01.2000 04:00 to 01.03.2013 03:00, with wind measurement taken at a height of 100 meters above sea level (Blekastad & Landa, 2020).

3.1.2 Wind Speed to Power Conversion Data

The second data set, Wind Speed to Power Conversion data, three different models (Effect 1, 2, and 3) of wind turbines from Kjeller Vindteknikk will be utilized in this paper as the conceptual wind speed to power output conversion for our studies. Depending on the average wind speed recorded

at each site, the corresponding wind turbine model that maximizes power output will be allocated. Figure 6 below shows the power curve of each model along with their wind speed classification in table 1 below.

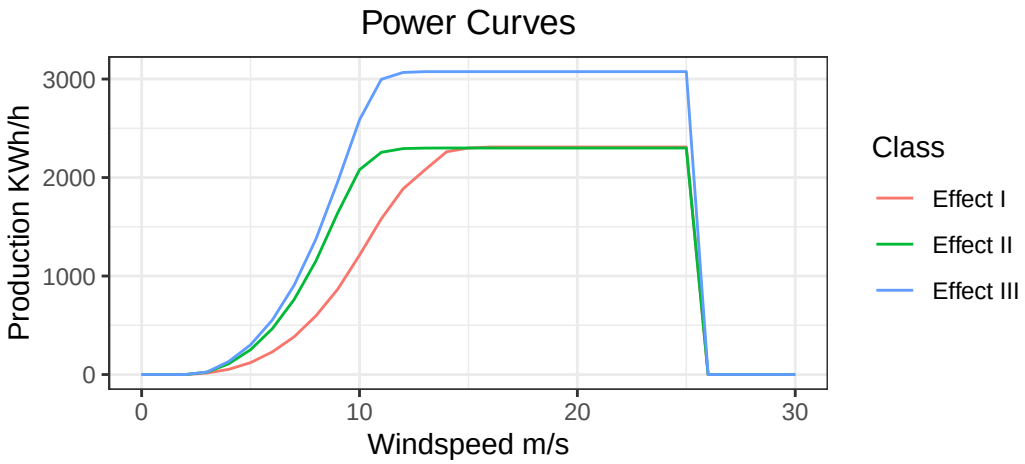

Figure 6: Plot of the Power curves

Table 1: Windspeed classification

	Effect I	Effect II	Effect III
Avg. wind speed	> 8.5m/s	7.5m/s - 8.5m/s	< 7.5m/s
Cut-in windspeed	3m/s	3m/s	3m/s
Cut-out windspeed	25m/s	25m/s	25m/s
Rated windspeed	17m/s	15m/s	14m/s

From fig. 6 above, we can infer that all three models have the same cut-in wind speed of 3m/s, and cut-out wind speed of 25m/s, indicating similar Region A and Region D. In terms of rated wind speed, Effect 3 spots the lowest at 14m/s, followed by Effect 2 at 15m/s, and finally Effect 1 of 17m/s.

Table 2: Amount of sites by class

	Effect I	Effect II	Effect III
Average Windspeed	> 8.5m/s	7.5m/s - 8.5m/s	< 7.5m/s
No. of sites	36	24	9

Table 2 on the previous page combines information from both the main dataset, and the power conversion dataset to reveal the distribution of sites based on the three wind turbine models. It can be inferred that more than half of the sites in our data set experiences high wind speed of above 8.5m/s and are categorized under Effect 1.

3.1.3 Sites Location Details Data

Lastly, the third data set contains location details of 68 of the 69 sites, with "Hog Jaren Trinn I" being the missing site. Location details include the coordinates and elevation of each site, along with the potential capacity and the respective electricity bidding zone they belong in.

Given the interest of this study, figure 7 below combines all three data sets to map out the location of each site, while highlighting the type of wind turbine allocated to each site.

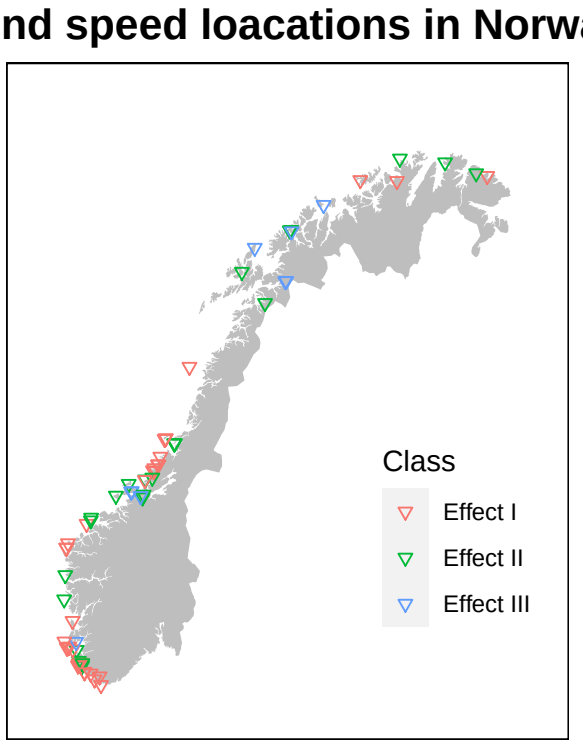

Figure 7: Map of all sites based on effect class

3.2 Descriptive Statistics

Given the data set consists of 12 years of hourly data across 70 locations, the sheer size of this data set presents complications not only in terms of time required for each process, but also compelled creative ways to aggregate data for the illustration of findings.

During preliminary checks of each individual site, it was discovered that both 'Lutelandet testanlegg' and 'Lutelandet' have identical time series of wind speed. The conscious decision to remove 'Lutelandet testanlegg' from the data set was made to avoid forecasting of duplicate sites, resulting in the total of 69 locations for the remainder analysis.

3.2.1 Summary Statistics

Referring to the summary statistics table 3 below, sites are aggregated based on their respective wind turbine model and compared to the combined of all sites. It can be observed that the mean wind speed is higher than median wind speed for all groups, indicating positive skewness in distribution. Besides that, all three groups spot similar characteristics in distribution, and the most conspicuous observation made from this table is the possibility of an incorrect input of maximum wind speed 72.11 m/s under the last column.

Table 3: Summary statistics

	All sites	Effect I	Effect II	Effect III
Min	0	0.01	0	0
1st Qu.	4.65	5.07	4.42	3.91
Median	7.64	8.23	7.28	6.47
Mean	8.444	8.998	8.069	7.224
3rd Qu.	11.43	12.17	10.91	9.73
Max	72.11	72.11	60.34	69.61

Investigating this unexpectedly high wind speed data point, table 4 depicts records of abnormally high wind speeds at "Sormakfjellet" spanning over an extended period of time in 20th January 2006. Diving further into the underlying reason, it was made clear that the dates coincide to the

period when storm Narve hits Norway (“"Narve" Feier Inn over Norge,” 2006), permitting us to overthrow the possibility of data input error.

Table 4: Abnormal windspeed

time	Sites	windspeed
2006-01-20 06:00:00	Sormarkfjellet	68.92
2006-01-20 07:00:00	Sormarkfjellet	70.66
2006-01-20 08:00:00	Sormarkfjellet	72.11
2006-01-20 09:00:00	Sormarkfjellet	70.92
2006-01-20 10:00:00	Sormarkfjellet	68.36
2006-01-20 11:00:00	Sormarkfjellet	67.36
2006-01-20 12:00:00	Sormarkfjellet	66.10
2006-01-20 14:00:00	Sormarkfjellet	65.99

3.2.2 Characteristics of Wind Speed

Following summary statistics which grouped sites by their wind turbine model, figure 8 below shows the kernel density plots of all 69 sites in 2008. The kernel density plot shows a probability density distribution, where the area under each curve adds up to exactly 1. The year 2008 was selected at random purely for illustration purposes to highlight the stochastic nature of wind speed in a given year.

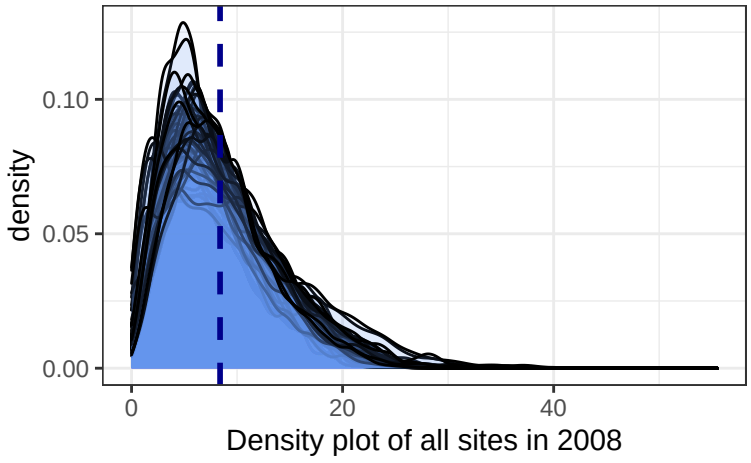

Figure 8: Density plot of all sites

A right skewed distribution is apparent amongst all sites, indicating the scarcity of strong wind, while accentuating prevalence of moderate wind. These kernel plots mimic that of Rayleigh distribution, a form of Weibull distribution with shape parameter of 2 ("Why the Weibull Distribution Is Always Welcome," 2013). Weibull distribution has often been used to describe wind speed variations, and even forms the basis in which wind turbine engineers design wind turbines on ("The Power Curve of a Wind Turbine," 2003).

3.2.3 Seasonality

Since wind energy is a form of solar energy, seasonality within the data set can be expected. In order to achieve a visual overview, the average wind speed at each location was compiled and plotted hourly and monthly as shown in fig. 9 below.

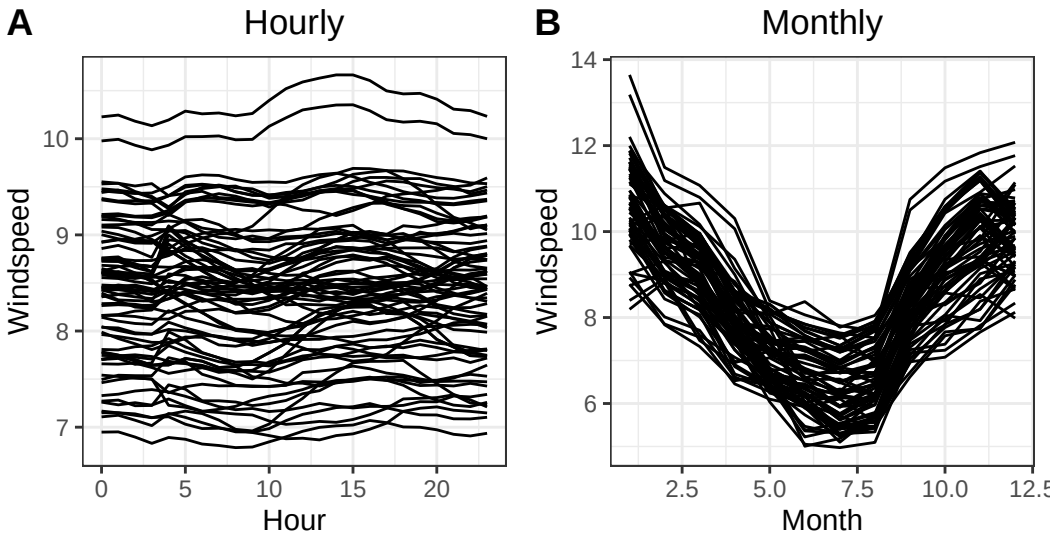

Figure 9: Average windspeed for all sites per hour and per month

In figure 9, each line represents one site. Looking at plot A of figure 9, no clear pattern can be observed, suggesting the absence or low daily seasonality. Although annual seasonality is more pronounced in plot B, illustrated by a clear dip in wind speed during summer months, it is pertinent to bear in mind that hourly wind speed data points were averaged out for the plot.

Having said that, for a better examination of the data set, time series will next be decomposed into their respective underlying components with Multiple Seasonal and Trend Decomposition using Loess (MSTL).

3.3 Decomposing

A time series y_t can be decomposed either additively or multiplicatively into three components: trend-cycle $\hat{T}_t$, seasonality $\hat{S}_t$, and remainder $\hat{R}_t$. While trend describes the long term increase or decrease in a time series, seasonality refers to the fixed change that occurs every year. (Hyndman & Athanasopoulos, 2018, ch. 2.3)

Additive decomposition is suitable for time series where overall fluctuations do not vary across time (Hyndman & Athanasopoulos, 2018, ch. 6.1), with formula as followed:

$$y_t = \hat{T}_t + \hat{S}_t + \hat{R}_t, \tag{7}$$

Conversely, a multiplicative decomposition is more suited for time series where fluctuations are observed to be increasing overtime (Hyndman & Athanasopoulos, 2018, ch. 6.1), with formula as followed:

$$y_t = \hat{T}_t \cdot \hat{S}_t \cdot \hat{R}_t \tag{8}$$

All 69 wind speed time series of interest do not possess varying fluctuations overtime, hinting at the additive behaviour of the time series. As outlined in eq. 7 above, all three underlying components of all time series are independent of each other.

The Multiple Seasonal and Trend Decomposition using Loess (MSTL) decomposition method is an extension of STL decomposition which allows the handling of complex seasonalities. The

STL is often deemed as the most versatile decomposition method, and was selected for this paper for its robustness (Hyndman & Athanasopoulos, 2018, ch. 11.1). The mstl() function provides room for user specified seasonal variation with s.window, smoothness of trend with t.window, robustness to outliers robust = TRUE/ FALSE, as well as estimation of trend-cycle and remainder of full data set (Hyndman & Athanasopoulos, 2018, ch. 11.1). In this study, s.window = 13 was selected in accordance with recommendations indicating the optimal balance between overfitting, and allowing slow changes over time the value can bring forth (Hyndman & Athanasopoulos, 2018, ch 6.6).

The sizable number of sites available in the data set made it unrealistic to plot out all 69 decompositions carried out. Therefore, site: Anstadblaheia, was singled out at random as an example plot for illustration purposes. Figure 10 below depicts the decomposition of wind speed time series in Anstadblaheia.

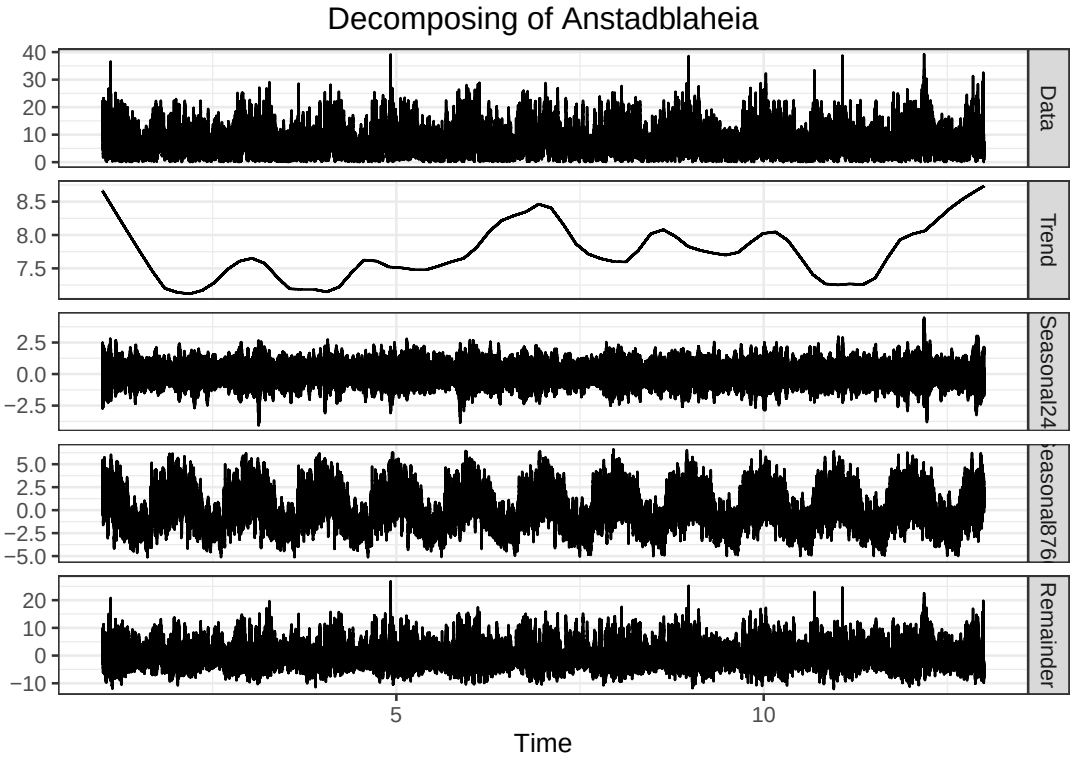

Figure 10: Decomposition example using Andstadblaheia

Similar to figure 9, obvious annual seasonality, as represented by the fourth plot - seasonality8766, can be observed in figure 10. In contrast, a pattern is harder to discern from the daily seasonality plot depicted by the third plot - seasonality24. Eyeballing the plots, both annual and daily seasonalities appear to be close to being additive in form, while no distinct long-term trend can be observed.

Moving forward, the decomposed series will be used to measure the strength and significance of seasonalities for all 69 wind speed time series (Hyndman & Athanasopoulos, 2018, ch 6.7). The importance of the seasonality within a time series can be calculated with the following equation:

$$F_s = max\left(0, 1 - \frac{Var(R_t)}{Var(S_t + R_t)}\right) \tag{9}$$

In eq. 9, F_s value close to 0 symbolises the absence of seasonality, while a value close to 1 reveals the presence of strong seasonality. Since both annual and daily seasonalities are of interest, S_t of daily and yearly components were computed independently. The results can be seen in table 5 on the next page.

Referring to table 5 on the next page, daily seasonality of all 69 time series are very weak, spotting the highest value of merely 0.03 at Tysvar, and lowest value of 0.014 at Falesrassa. Annual seasonality, albeit stronger than daily seasonality, is also considered low with all sites falling within the range of 0.13 - 0.26. The reason for the weak seasonality apparent in all time series of interest could perhaps be postulated to the high frequency of the data.

Following these discoveries, the covariance stationarity of the data will be investigated upon. Covariance stationary is a critical condition for forecasting time series with the statistical regression methods, ARIMA and ARIMAX, adopted in this study. Taking into account the importance in validating the covariance stationarity of all time series, formal assessment will next be carried out.

Table 5: Seasonality test result

Sites	Daily	Yearly	Sites	Daily	Yearly
Storheia	0.024	0.246	Mehuken I	0.014	0.171
Roan	0.023	0.236	Ytre Vikna trinn I	0.021	0.211
Skinansfjellet	0.017	0.149	Hundhammerfjellet demo II	0.023	0.219
Friestad	0.018	0.149	Havoygavlen	0.016	0.172
Royrmyra	0.017	0.15	Bessakerfjellet II	0.022	0.218
Nygardsfjellet trinn II	0.021	0.139	Sway Karmoy	0.017	0.171
Eikeland Steinsland	0.021	0.136	Hywind	0.017	0.175
Havsul I	0.019	0.153	Testomrade Stadt	0.019	0.134
Mehuken II	0.015	0.173	Vardoya	0.019	0.186
Kvenndalsfjellet	0.023	0.238	Asen II	0.019	0.155
Valsneset	0.025	0.238	Sandvesanden	0.019	0.164
Midtfjellet	0.017	0.176	Nordbo	0.022	0.164
Sormarkfjellet	0.02	0.206	Askjesundet	0.028	0.158
Fakken	0.024	0.162	Raudfjell	0.021	0.253
Andmyran	0.017	0.162	Hamnefjell	0.016	0.195
Gravdal	0.018	0.143	Donnesfjord	0.022	0.161
Rakkocearro	0.016	0.196	Hitra II	0.019	0.143
Hitra I	0.019	0.141	Froya	0.017	0.165
Sandhaugen	0.017	0.22	Geitfjellet	0.021	0.184
Hog Jaren trinn I	0.017	0.151	Svarthammaren	0.021	0.157
Nygardsfjellet trinn I	0.024	0.132	Remmafjellet	0.022	0.167
Lista	0.028	0.135	Svaheia	0.02	0.15
Tysvar	0.03	0.15	Tellenes	0.021	0.164
Utsira	0.017	0.178	Stigafjellet	0.019	0.141
Haramsfjellet	0.017	0.156	Makaknuten	0.018	0.142
Hundhammerfjellet demo I	0.023	0.22	Eigersund	0.02	0.147
Bessakerfjellet I	0.021	0.216	Kvinesheia	0.024	0.154
Harbakfjellet	0.023	0.24	Lutelandet	0.027	0.166
Kjollefjord	0.017	0.206	SWAY Kollsnes	0.02	0.166
Hundhammerfjellet	0.023	0.222	Ytre Vikna trinn II	0.021	0.206
Valsneset testsenter	0.026	0.246	Anstadblaheia	0.016	0.156
Fjeldskar	0.027	0.132	Sorfjorden	0.019	0.17
Smola	0.018	0.16	Hog Jaren trinn II	0.016	0.149
Kvitfjell	0.02	0.222	Falesrassa	0.014	0.17
Haroy	0.019	0.156			

3.4 Covariance stationarity

For data to be a covariance stationary process, three principal requirements must be met:

$$E(y_t) = \mu \tag{10}$$

$$Var(y_t) = \sigma^2 \tag{11}$$

$$cov(y_t, y_{t-s}) = \gamma_s \tag{12}$$

Eq. 10 and 11 demonstrates two of the three aforementioned requirements. It can be seen from the equation that the expected value and variance must be constant throughout all periods. The third requirement, as established by eq. 12, shows that the covariance of the series against itself must also be constant in all periods for a fixed number of periods s. These three requirements are pivotal for our result to convey economical significance as regression estimations from non-stationary data will yield spurious results, introducing bias to estimated parameters. (DeFusco, McLeavey, Pinto, & Runkle, 2015, pp. 472–473)

Unit Root Tests, such as the Augmented Dickey-Fuller (ADF) test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test are some of the options available in determining time series stationarity, as well as to identify the differencing term required to achieve stationarity. Taking into consideration the size of the data set, this paper adopts an ADF test based on the recommendations made by a research paper from Makèta Arlova and Darina Fedorova. The authors tested a plethora of stationarity tests and concluded the ADF test to be the most reliable option time series with large numbers of observations (Arlova & Fedorova, 2016, p. 63). In the presence of non-stationarity, time series will be differenced and retested again until covariance stationary process is achieved.

3.4.1 ADF test

The ADF test builds upon the original Dickey-Fuller test by augmenting lagged changes. This makes ADF preferable as provisions are made for serial correlations, increasing the robustness of the test.

The ADF method takes the following form:

$$\Delta y_t = \alpha + \theta y_{t-1} + \gamma_1 \Delta y_{t-1} + \varepsilon_t \tag{13}$$

with hypothesis test as follows:

$$H_0 : \theta = 0$$
$$H_1 : \theta < 0$$

Here $|\gamma_1| < 1$ ensures that under H_0 Δy_t follows a stable AR(1) model, while under H_1 it can be shown that y_t follows a stable AR(2) model. Lags of Δy_t can also be included to account for more dynamics present in the series. The regression of Δy_t on the lagged variables are then carried out, followed by a t-test on their coefficients. Based on the null hypothesis, when $\theta = 0$, we can objectively ascertain the non-stationarity of the time series. (Wooldridge, 2018, p. 612).

The ADF test was performed for all 69 sites, with results presented in table 6 on the next page.

Looking at table 6 we can discern test statistics for all 69 sites to be lower than critical value by a strong margin. Hence, the null hypothesis that the time series is non-stationary can be confidently rejected, implying stationarity in all series. In other words, no differencing will be required at all.

Table 6: Stationary tests results

Sites	ADF ts	ADF cv	Sites	ADF ts	ADF cv
Storheia	-25.102	-1.95	Mehuken I	-24.514	-1.95
Roan	-25.647	-1.95	Ytre Vikna trinn I	-23.132	-1.95
Skinansfjellet	-21.349	-1.95	Hundhammerfjellet demo II	-26.157	-1.95
Friestad	-22.532	-1.95	Havoygavlen	-23.68	-1.95
Royrmyra	-21.554	-1.95	Bessakerfjellet II	-24.446	-1.95
Nygardsfjellet trinn II	-30.545	-1.95	Sway Karmoy	-23.382	-1.95
Eikeland Steinsland	-24.21	-1.95	Hywind	-23.016	-1.95
Havsul I	-26.807	-1.95	Testomrade Stadt	-27.64	-1.95
Mehuken II	-24.584	-1.95	Vardoya	-24.497	-1.95
Kvenndalsfjellet	-24.15	-1.95	Asen II	-23.489	-1.95
Valsneset	-26.229	-1.95	Sandvesanden	-24.191	-1.95
Midtfjellet	-25.419	-1.95	Nordbo	-24.808	-1.95
Sormarkfjellet	-26.073	-1.95	Askjesundet	-26.023	-1.95
Fakken	-29.181	-1.95	Raudfjell	-24.556	-1.95
Andmyran	-24.19	-1.95	Hamnefjell	-22.805	-1.95
Gravdal	-22.532	-1.95	Donnesfjord	-28.308	-1.95
Rakkocearro	-22.9	-1.95	Hitra II	-27.287	-1.95
Hitra I	-26.654	-1.95	Froya	-24.729	-1.95
Sandhaugen	-24.2	-1.95	Geitfjellet	-28.801	-1.95
Hog Jaren trinn I	-22.048	-1.95	Svarthammaren	-29.281	-1.95
Nygardsfjellet trinn I	-34.608	-1.95	Remmafjellet	-31.385	-1.95
Lista	-24.754	-1.95	Svaheia	-23.879	-1.95
Tysvar	-28.019	-1.95	Tellenes	-23.07	-1.95
Utsira	-22.543	-1.95	Stigafjellet	-22.814	-1.95
Haramsfjellet	-26.081	-1.95	Makaknuten	-22.468	-1.95
Hundhammerfjellet demo I	-25.531	-1.95	Eigersund	-23.817	-1.95
Bessakerfjellet I	-25.152	-1.95	Kvinesheia	-23.144	-1.95
Harbakfjellet	-25.047	-1.95	Lutelandet	-27.237	-1.95
Kjollefjord	-22.439	-1.95	SWAY Kollsnes	-25.243	-1.95
Hundhammerfjellet	-25.461	-1.95	Ytre Vikna trinn II	-22.755	-1.95
Valsneset testsenter	-26.661	-1.95	Anstadblaheia	-25.696	-1.95
Fjeldskar	-23.2	-1.95	Sorfjorden	-27.866	-1.95
Smola	-25.871	-1.95	Hog Jaren trinn II	-21.766	-1.95
Kvitfjell	-24.904	-1.95	Falesrassa	-22.55	-1.95
Haroy	-26.464	-1.95			

3.5 Software used

Two programming languages, R and Python, were used for all the data analysis carried out in this paper, and the report is written fully in R markdown. Data preprocessing, visualisation, as well as the statistical models and evaluation methods were implemented in R with the following packages: *ffp2*, *lubridate*, *xts*, *urca*, *tidyverse*, *kableExtra*, *maps*, *mapdata*, *mapproj*, *grid*, *gridExtra*, *ggpubr*, *knitr*, *jpeg*, *png* and *janitor*.

Meanwhile, machine learning models were implemented in Python with the Scikit-learn library. Other libraries such as *pandas*, *numpy*, and *scipy*, were also used for data manipulation, calculation, and analysis.

3.6 Data Preparation

To formulate and evaluate forecasting models, all 69 univariate time series have to first be split into train set and test set. Since 2012 is the last complete year of data, all 606096 observations of data points from 01.01.2012 00:00:00 to 31.12.2012 23:00:00 were held out as test data. The reason for forecasting one-step-ahead for a full year was also to compare models' accuracies across different seasons. Therefore, besides the removal of duplicate site "Lutelandet testanlegg" as mentioned previously, data in the year of 2013, a total of 1420 observations in each time series, were also removed.

Train data differs slightly for Statistical Methods and Machine Learning Methods. For Statistical Methods, all data points prior to 2012 were used to train a model for each site. Besides that, data statistics presented prior were all based on the aforementioned train data.

However, for Machine Learning methods, one model is trained for all sites due to computational intensity required for not just training of models, but also for the tuning of each ML models'

hyperparameters. Therefore, data from all 69 sites were compiled into 1 data set, of 69 observations at each time frame.

Furthermore, all 69 univariate time series have to be framed into a supervised learning problem for ML model training - namely split into input and output components. For this paper, input components for all 3 ML models consisted of the following:

1. 6 lag observations ($t-1$ to $t-6$),
2. 4 seasonality splits (spring: 01 March to 31 May, summer: 01 June to 31 August, autumn: 1 September to 30 November, winter: 1 December to 28/29 February), and
3. 3 time of day splits (dawn: 00:00 to 07:00; day: 08:00 to 15:00 dusk: 16:00 to 23:00)

While the models output component, naturally, refers to the forecast wind speed value at $t = 0$. Annual and daily seasonality components are also one-hot-encoded (OHE) into input.

Finally, while ML models are trained with the full set of training data, due to the lengthy computational time, ML models' hyperparameters were iteratively tuned with only 10% randomly selected subset from compiled data of all sites from 01/01/2000 10:00 - 31/12/2011 23:00.

4. Methodology

This section dives into the crux of the report with theoretical explanations on all the methods employed to forecast the 69 univariate wind speed time series. Statistical methods utilized include the benchmark Naive method, ARIMA and ARIMAX. Machine Learning (ML) methods include two ensemble methods - Gradient Boosted Tree (GBR) and Extremely Randomized Tree (ETR), and a Multi-layer Perceptron (MLP) supervised Neural Network method. Finally, a hybrid model of ARIMAX and ETR is formulated as a combination of Statistical and ML. Subsequently, all models are trained and assessed through four different evaluation methods: MSE and RMSE, MAE, Classification Accuracy, and Power Curve Conversion Error with penalization, in an attempt at singling out the best model.

4.1 Statistical Forecasting Methods

4.1.1 NAIVE (Benchmark)

NAIVE forecast, the archetypal benchmarking model, is one of the simplest forecasting methods available (Hyndman & Athanasopoulos, 2018, ch. 3.1). Not only is it fast and easy to implement, the method also requires no assumption on the characteristics of the time series. Therefore, for the purpose of this study, the NAIVE model will also serve as the baseline performance by which other more computationally intensive forecasting methods are compared against. Having a benchmark model allows a clear indication on the gain/loss in accuracy performance with respect to the additional computational complexity from all other models. Besides that, it can also be used to provide a quantitative idea of how difficult the forecast problem at hand - forecasting of wind speed is.

The NAIVE takes on the straightforward assumption that the next value is the same as the current value, based on the persistence algorithm with equation as follows:

$$\hat{Y_{T+h|T}} = Y_T \tag{14}$$

As illustrated in eq. 14 above, the algorithm assigns a forecast value at h time step ahead with the same value as that at time T. Given that this study focuses on hourly one-step-ahead forecasting, the NAIVE model will simply be h = 1, and the forecast value at each time step is equal to the observed value 1 hour before. On that premise, it is reasonable to expect the NAIVE model to perform rather satisfactory since one can expect high correlation of wind speed with the past hour.

4.1.2. ARIMA and ARIMAX

ARIMA, which stands for "Auto-Regressive Integrated Moving Averages", is a popular method to forecast time series. Meanwhile, ARIMAX, short for ARIMA with Explanatory Variable(s), is simply an adaption of ARIMA to include exogenous variables. Similar to NAIVE forecast, ARIMA(X) also makes use of past values to predict future values, although with more complexity following linear assumptions and hence the requirement of stationarity and invertibility conditions. ARIMA model is a combination of the autoregression model, moving average model, and differencing; and forecasts based on the respective components:

p: order of the autoregressive (AR) term

q: order of the moving average (MA) term

d: degree of first differencing (I) required

The AR term represents the autoregressive component of an ARIMA model, and forecasts the variable of interest (yt) based on a linear combination of p number of past observations (Hyndman & Athanasopoulos, 2018, ch. 8.3). It can be mathematically expressed as follows:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \ , \tag{15}$$

where the value at time t is predicted by regressing p number of lagged values.

Conversely, the MA term represents the moving average component of an ARIMA model, and forecasts the same variable of interest (yt) based on a regression-like function on q number of past forecast errors(Hyndman & Athanasopoulos, 2018, ch. 8.3).

It can be expressed as follows:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \ , \tag{16}$$

where the value at time t is predicted by regressing q number of lagged errors. In the case of ARIMAX, each additional explanatory variable included will simply be represented by its own p and q term.

Autocorrelation (ACF) and Partial autocorrelation (PACF) plots are typically used to determine the values of p and q in a model manually. However with 69 models to forumate for each method, the 'auto.arima' function from the forecast package was used instead to retrieve the optimal parameter with the lowest Akaike's Information Criterion with correction (AICc) of each model.

Lastly, the integrated (I) term in ARIMA refers to the number of first differencing required for a time series to achieve covariance stationarity. Relevant stationarity tests, such as ADF and KPSS, as mentioned in the previous chapter, can be performed to determine the stationarity of time series as well as the number of differencing required (Hyndman & Athanasopoulos, 2018, ch. 8.1).

Therefore, in the presence of a non-stationary time series, differencing is carried out and the first order of differencing can be achieved by subtracting current value with a lagged value as follows:

$$y'_t = y_t - y_{t-1} \tag{17}$$

And second order of differencing can be achieved as follows:

$$y'_t = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \tag{18}$$

The ARIMA and ARIMAX models orders for each site can be seen in appendix 15. The ARIMAX method in this study simply builds upon the ARIMA method, with the inclusion of four dummy variables to represent the four seasons, and three dummy variables corresponding to the time of day, similar to the seven OHE dummies employed in ML methods.

4.2 Machine Learning Forecasting Methods

Expanding upon the statistical methods previously described, the paper moves on to more sophisticated forecasting methods using Machine Learning (ML) algorithms. ML methods, unlike ARIMA(X), do not require linear relationships to hold true. Since the true relationship of most wind speed time series are not necessarily linear, nonparametric regression techniques as these can often be better at handling and forecasting complex time series, albeit subjected to higher computational intensity (James, Witten, Hastie, & Tibshirani, 2017, p. 314). For the purpose of this study, two ensemble methods - Gradient Boosted Tree (GBR) and Extremely Randomized Tree (ETR), and a third method using Multi-layer Perceptron (MLP) supervised Neural Network are employed and compared against traditional statistical forecasting methods.

4.2.1 Regression Tree-Based Ensemble methods

Ensemble method, as the name suggests, combines a collection of forecasts to improve upon the base method(s) (Pedregosa et al., 2011, ch. 1.11). The Ensemble methods from Scikit-learn library specifically combine forecasts from multiple base estimators built by the same learning algorithm. They can be branched off into two main categories: averaging or boosting, based on how base estimators are built as well as how forecasts are combined. Under the averaging method, multiple

estimators are built independently, and forecasts are averaged out to reduce variance compared to forecasts from a single estimator (Pedregosa et al., 2011, ch. 1.11). Conversely, the boosting method builds estimators in succession, with the goal of minimizing bias when combining all estimators. The two ensemble methods picked for this study, ETR and GBR, fall into each of these categories respectively.

Furthermore, both ETR and GBR are also Regression Tree-Based Methods, thereby following the basic principles of decision trees, albeit with their independent modifications. When growing a regression tree, all predictors and cut points are first considered and selected based on minimum RSS, and the feature space is recursively partitions into subsets until a stopping criterion is reached (James et al., 2017, p. 307). To avoid overfitting of trained models, pruning can be introduced for decision trees. It is a strategy that requires one to first grow a large tree, then trim it down into a smaller subtree with lesser terminal nodes. Cost complexity pruning has the following equation:

$$minimize \sum_{m=1}^{|T|} \sum_{i:x_i \varepsilon R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T|) \tag{19}$$

In eq. 19 above, $|T|$ reveals the number of regions in a tree T, and R_m represent each respective region, while $\hat{y}_{R_m}$ corresponds to the mean response of each region. An optimal pruning parameter α can be determined based on cross validation to minimize the eq. 19 (James et al., 2017, pp. 308–310).

4.2.1.1 Extremely Randomized tree The first method, which takes on the averaging ensemble method, is the Extremely Randomized Tree (ETR) method. Similar to Random Forest (RF), ETR also averages forecasts based on randomized decision trees, but improves upon RF with the introduction of an additional randomization process. In Random Forest, two levels of randomness are introduced to reduce variance. Firstly by training trees on bootstrapped samples, and secondly, with the random subspace method to find the best split for each node with either all input variables or a random subset ("Machine Learning in Python," n.d., ch. 1.11.2.2).

Unlike RF where bootstrapped samples are used, ETR uses the whole data by default. Besides that, ETR also differs from RF in that splitting thresholds are selected randomly. The best split threshold is identified from all the randomly-drawn thresholds rather than based on the most discriminative one as in RF (Aznar, 2020). This characteristic gives ETR a computational advantage over traditional RF along with further reduction in variance, but at the expense of greater bias ("Machine Learning in Python," n.d., ch. 1.11.2.2).

The customizability of ETR parameters in sklearn allows the user to determine the computational cost v.s. accuracy of each model. The two key parameters for ETR are 'n_estimators', which determines the number of trees in the forest, and 'max_features' which selects the size of random subsets of features for splitting of nodes ("Machine Learning in Python," n.d., ch. 1.11.2.3). After parameters tuning and cross-validation the ETR parameters for model training in this study is illustrated in 11 below.

Model Training

```
params = {'n_estimators': 900,
          'criterion':'mse',
          'max_depth': 20,
          'min_samples_split': 100,
          'min_samples_leaf': 1,
          'max_features':'auto',
          'random_state' : 0,
          'n_jobs': -1}

etr_model = ExtraTreesRegressor(**params)

model = etr_model.fit(xtrain, ytrain)
```

Figure 11: ETR training parameters

4.2.1.2 Gradient Boosting tree Gradient Boosting Tree (GBT) can be characterized by the following three elements: 1) Optimization of a defined loss function; 2) Building of models iteratively from individual estimators (aka 'weak learners'); and 3) An additive model to add all weak learners together while minimizing loss function at every step.

The boosting component of the GBT method refers to how models are built sequentially with individual regression trees. These estimators are also known as 'weak learners' as higher weights are assigned to wrong predictions with higher errors (Elsinghorst, 2018). It stems from the idea of learning from mistakes, and focuses on improving through learning from the more 'difficult' cases it has encountered.

GBT can be mathematically represented by the following equation:

$$\hat{y}_i = F_M(x_i) = \sum_{m=1}^{M} h_m(x_i) \tag{20}$$

Where M represents the number of booting stages, and h_m represents the "weak learners".

The gradient component of this method refers to the gradient descent utilized to minimize the specified loss function of the model ("Machine Learning in Python," n.d., ch. 1.11.4), and is depicted in eqn. 21 below. At each training iteration, error rates are calculated and combined to derive the most updated gradient (i.e. the partial derivation of the loss function). The gradient indicates the steepness of the loss function, and is subsequently used to determine the direction to best steer the next estimator for minimizing the overall loss function ("Machine Learning in Python," n.d., ch. 1.11.4.5.1). Eqn. 21 below mathematically illustrates how h_m is fitted to minimize a sum of losses L_m given the previous ensemble F_{m-1}.

$$h_m = arg\min_h L_m = arg\min_h \sum_{i=1}^{n} l(y_i, F_{m-1}(x_i) + h(x_i)) \tag{21}$$

Considering the application of the boosting algorithm and the gradient descent in this method, the key parameters for GBT are 'n_estimators', which determines the number of boosting stages (trees in the forest); 'max_depth', which limits the number of nodes in the tree; 'learning rate', which determines the shrinkage of each tree's contribute; and 'loss', which decides the loss function to

optimize ("Machine Learning in Python," n.d., ch. 1.11.4.5.1). After parameters tuning and cross-validation the GBT parameters for model training in this study is illustrated in fig. 12 below.

Model Training

```
params = {'n_estimators': 500,
          'max_depth': 4,
          'min_samples_split': 750,
          'learning_rate': 0.1,
          'loss': 'ls',
          'min_samples_leaf' : 1,
          'subsample' : 1,
          'max_features':'auto',
          'random_state' : 0}

gtb_reg = GradientBoostingRegressor(**params)

model = gtb_reg.fit(xtrain, ytrain)
```

Figure 12: GBT training parameters

4.2.2 Neural Network

Artificial Neural Networks (ANN), as insinuated by its name, are computational algorithms used to process data by replicating the way a biological brain analyzes, processes, and transfers information with neurons and synapses through electrical and chemical signals.

Billions of neurons in the brain receive external information from their dendrites, process and convert these inputs within their cells, then transmit them as outputs through their axons to the next neuron. Depending on the signal strength within the neurotransmitter at the synaptic terminal, information is either accepted or rejected by the next neuron.

Analogous to a brain, a basic ANN model structure as illustrated in 13 below, can be simplified into three main types of layer: Input layer - to receive training data for learning; Hidden layer(s) - non-linear layers to perform computations on received information and process them into meaningful output; and Output layer - to determine and display forecasted output.

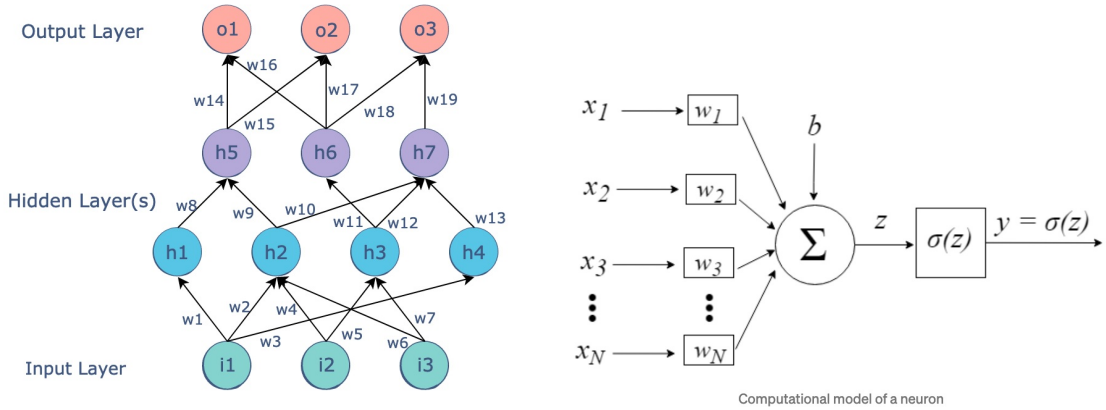

Figure 13: ANN model structure

Just like neurons in the brain, ANN consists of artificial neurons, known as nodes, which are responsible for information processing. These neurons/nodes are interconnected by arcs, which correspond to the dendrites-synapses-axons connections in the brain. Arcs are also assigned with a weight at each node to represent its relative importance as depicted in fig 13 above. Finally, a value is compiled from the weighted linear summation of all arcs plus bias as shown in eqn. 22 below, and fed into a predefined non-linear activation function along the incoming arcs, in an attempt to simulate the signal strength at each synaptic terminal in which the acceptance or rejection of incoming information is decided upon.

$$z = \sum_{i=1}^{N} x_i w_i + b \,, \tag{22}$$

There exists a plethora of neural networks with varying levels of complexity targeted for different uses. Under the realm of wind speed forecasting, Multilayer Perceptron (MLP), Radial Basis Function (RBF), and Recurrent Neural Networks (RNN) appear to be amongst the most popular methods (Farzaneh & Majid, 2017). This paper thus employs the most simple ANN method, the MLP as a basis of comparison with all other methods. The MLP method employed in this study, akin to the two previous ML methods, is also formulated from an off-the-shelf solution, 'MLPRegressor', available in the scikit-learn library.

Diving into the characteristics of 'MLPRegressor', instead of training the output layer with an activation function as in conventional ANN, MLPRegressor trains models using backward propagation ("Machine Learning in Python," n.d., ch. 1.17). Backpropagation refers to the iterative process of minimizing errors through weights optimization based on stochastic gradient descent similar to GBT. In the case of 'MLPRegressor', the loss function for backpropagation is defined as square error, and model's error outputs at each iteration are fed as inputs into the subsequent model, until an optimal output result is reached.

Having said that, one of the key disadvantages of ANN is the high time complexity associated with backpropagation and hyperparameters tuning on the number of hidden neurons, layers, and iterations; as well as the model's sensitivity to feature scaling ("Machine Learning in Python," n.d., ch. 1.17). After parameters tuning and cross-validation the ANN parameters for model training in this study is illustrated in fig. 14 below.

Model Training

```
params = {'hidden_layer_sizes':(100, ),
          'learning_rate_init': 0.00094,
          'max_iter': 385,
          'beta_1': 0.916,
          'beta_2': 0.995,
          'alpha':0.0386,
          'activation':'relu',
          'early_stopping': False,
          'validation_fraction': 0.1,
          'learning_rate': 'adaptive',
          'random_state' : None,
          'solver': 'adam',
          'batch_size': 'auto',
          'shuffle' : True,
          'n_iter_no_change':10}

ann_model = MLPRegressor(**params)

model = ann_model.fit(xtrain, ytrain)
```

Figure 14: ANN training parameters

4.3 Hybrid forecasting Method

Following recommendations from journals (Du et al., 2019), (Farzaneh & Majid, 2017), and (Cadenas & Rivera, 2010), as discussed in Chapter 2.4, the last forecasting method is a hybrid model of Statistical method - ARIMAX and Machine Learning method - ERT. Considering the limitation of ARIMA method which is based on linear assumption, this hybrid model aims to attain the best of both methods, employing the statistical method to predict linear components within the data, and machine learning method to capture the non-linear aspects. (Jingxing, Abdullah, Mingdi, Eunshin, & Romesh, 2020) The mathematical expression for this hybrid can be expressed as follows:

$$\hat{Y}_t = LP_t + NLP_t \tag{23}$$

Referring to eq. 23, LP_t represents the linear element to be forecasted by the statistical method, while NLP_t depicts the non-linear element to be forecasted by the machine learning method.

This approach is carried out by first training the model with ARIMAX, to capture the linear aspect. The residuals, which describe the non-linear dynamics remaining in the data, are next retrieved from the model. These residuals are next fed and trained by ERT. Having trained and built both models, values from both forecasts are next summed up to give one hybrid forecast. (Jingxing et al., 2020) The parameters of the final Hybrid-ERT model are illustrated in fig. 15 below.

Model Training

```
params = {'n_estimators': 1200,
          'criterion':'mse',
          'max_depth': 27,
          'min_samples_split': 120,
          'min_samples_leaf': 1,
          'max_features':'auto',
          'random_state' : 0,
          'n_jobs': -1}

etr_model = ExtraTreesRegressor(**params)

model = etr_model.fit(xtrain, ytrain)
```

Figure 15: Hybrid training parameters

4.4 Model Evaluation Tools

Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) are common scale-dependent forecasting accuracy measures (James et al., 2017, p. 33). Since the time series used in this study are all of the same scale - wind speed in m/s, scale-dependent accuracy measures are used instead of scale independent measures such as Relative Mean Square Error (relMSE) and Mean Relative Absolute Error (MRAE).

Having said that, (Hering & Genton, 2010), questions their suitability when the main concern is on the accuracy of power output prediction rather than the forecasted object wind speed. Hence, two additional methods, Classification Accuracy, and Power Curve Error with penalty will be employed in an attempt to take wind speed and power conversion into account to give the true value of each model's prediction. The classification accuracy compares the performance of each model in the lower (below 3m/s) and upper (>25 m/s) boundaries, classifying both actual and forecasted time series such that it takes the value of 1 when within the region of interest, and a value of 0 outside. Power Curve Error with penalty on the other hand, is a tailored loss function implemented with the purpose of assessing the quality of predictions in relation to wind speed and power output conversion, as well as the cost difference in the electricity market for under and/or over forecasting wind speed.

4.4.1 RMSE

Forecast errors refers to the difference between actual observed value and forecasted value and can be calculated with the following equation:

$$e_t = Y_t - \hat{Y}_t \tag{24}$$

Mean Square Error (MSE), as the name suggests, simply squares forecast errors at every period to make them all positive, and calculates the average squared error for each forecasting model. MSE

can be calculated by eq. 25 below:

$$RMSE = \sqrt{MSE}\ ,\ \ where\ \ MSE = \frac{\sum_{t=1}^{N}(Y_t - \hat{Y}_t)^2}{N} \tag{25}$$

Building upon MSE, Root Mean Square Error (RMSE) takes the equation a step further by taking the square-root of MSE as shown in eq. 25 above. This makes RMSE value more interpretable compared to MSE as it takes the value back down to the same magnitude as that of the forecast errors. Although RMSE value does not always increase with the variance of error, RMSE increases with the frequency distribution of error magnitude (Jj, 2016). Therefore, the key characteristic of both MSE and RMSE is the heavier penalization placed on bigger errors as opposed to smaller errors, which could be appreciated in specific cases.

4.4.2 MAE

Another metric routinely used to calculate forecast accuracy is the Mean Absolute Error (MAE). Similar to MSE, MAE removes negative forecast errors by taking the absolute value of all errors, summing them up, and calculates the average absolute error value by dividing the summed values against the total number of observations .(Hyndman & Athanasopoulos, 2018, ch. 3.4). MAE can be represented by eq. 26 below:

$$MAE = \frac{\sum_{t=1}^{N}|Y_t - \hat{Y}_t|}{N} \tag{26}$$

Since errors are not squared in MAE, MAE is usually preferred over MSE and RMSE as a metric for accuracy measurement if even penalization across error magnitude is of priority.

4.4.3 Classification accuracy

All three aforementioned measures, MSE, RMSE, and MAE take and compare the average accuracy of each forecast model. However, noting the characteristics of a wind turbine power curve

mentioned in Chapter 2.2 where Region A and D indicates the difference between operation and termination of a wind turbine, another interest of this study is in the accuracy surrounding the tails of the power curve.

Therefore, the classification accuracy, an accuracy metric typically used for classification problems, was employed as another measure to calculate the accuracy of forecasts falling within these two specific regions. Under this accuracy measure, forecasts were assigned to either of two variables taking the value of 1 or 0, representing the asymmetric cost of power production in wrong forecasting.

As illustrated in eq. 27 and eq. 28, for all observed values below the cut-in wind speed of 3m/s (Region A) and above the cut-out wind speed of 25m/s (Region D), the value of 1 will be assigned to their respective forecast dummy variable, while the value of 0 is assigned to all other observed values. The same procedure is done on all forecasted values across all models and all sites.

$$\delta_t^{left} = \begin{cases} 1\ , & if\ WS_t < 3 \\ 0\ , & Otherwise \end{cases} \tag{27}$$

$$\delta_t^{right} = \begin{cases} 1\ , & if\ WS_t > 25 \\ 0\ , & Otherwise \end{cases} \tag{28}$$

Subsequently, eq. 27 from observed value is compared against eq. 27 from forecast; while eq. 28 from observed value is compared against eq. 28 from forecast to get the number of correct predictions in Region A and Region D respectively.

Following that, the accuracy on the left (below 3m/s) and the right side (above 25m/s) of the power curve is calculated with the following equation with left tail as an example:

$$Accuracy\ LT = \frac{Number\ of\ correct\ prediction\ in\ region\ A}{Total\ number\ of\ observations\ in\ region\ A} \tag{29}$$

On the numerator, the number of correction predictions in the region corresponds to when both dummy variables of actual and forecast take the value of 1. At the denominator, the total number of observations in the region is calculated by summing up all instances of 1-1, 1-0, and 0-1 taken by dummy variables of actual and forecast.

4.4.4 Power Curve Conversion Error

Improving on the Classification Accuracy measure above, a tailored loss function, Power Curve Conversion Error with penalization (PCCEp), is developed to take into consideration both the position of observed values and forecast values in all four regions of the power curve, as well as to allocate a penalization value associated with the economic cost faced by system operators in the event of under and over forecasting electricity supply.

The first step in developing this loss function was to determine a method of assigning weights to each wind speed that signify the importance of a correct forecast at said wind speed. Since actual power production gains or losses corresponding to a wrong wind speed forecast is of interest here, the first step was to take the absolute value of the difference between the output power at a specific wind speed and the power value one wind speed above. Absolute value is taken to ensure positive values in the calculation since both losses and gains are of interest.

Next, the sum of all power output differences at each wind speed step are calculated for each power curve, and the weights of each wind speed is determined by dividing each of their absolute values against the sum of all power differences in each curve. The equation is shown as eq. 30, and the weightage table can be found in table 7 below, with final weights highlighted in blue.

Table 7: Weightage calculation table

	Effect I			Effect II			Effect III		
Windspeed	Power (kW)	Diff	Weight	Power (kW)	Diff	Weight	Power (kW)	Diff	Weight
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	16	16	0.0035	24	24	0.0052	26	26	0.0042
4	53	37	0.008	110	86	0.0187	133	107	0.0174
5	121	68	0.0147	252	142	0.0309	302	169	0.0275
6	230	109	0.0236	466	214	0.0465	554	252	0.041
7	383	153	0.0331	762	296	0.0643	907	353	0.0574
8	596	213	0.0461	1153	391	0.085	1375	468	0.0761
9	866	270	0.0584	1641	488	0.1061	1958	583	0.0948
10	1212	346	0.0749	2079	438	0.0952	2585	627	0.102
11	1580	368	0.0797	2256	177	0.0385	2997	412	0.067
12	1885	305	0.066	2294	38	0.0083	3067	70	0.0114
13	2077	192	0.0416	2299	5	0.0011	3075	8	0.0013
14	2262	185	0.04	2300	1	0.0002	3075	0	0
15	2300	38	0.0082	2300	0	0	3075	0	0
16	2310	10	0.0022	2300	0	0	3075	0	0
17	2310	0	0	2300	0	0	3075	0	0
18	2310	0	0	2300	0	0	3075	0	0
19	2310	0	0	2300	0	0	3075	0	0
20	2310	0	0	2300	0	0	3075	0	0
21	2310	0	0	2300	0	0	3075	0	0
22	2310	0	0	2300	0	0	3075	0	0
23	2310	0	0	2300	0	0	3075	0	0
24	2310	0	0	2300	0	0	3075	0	0
25	2310	0	0	2300	0	0	3075	0	0
26	0	2310	0.5	0	2300	0.5	0	3075	0.5
27	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0
30 and av	0	0	0	0	0	0	0	0	0
	Total	4620	1		4600	1		6150	1

$$weight_{ws} = \frac{|P(WS)_{ws} - P(WS)_{ws-1}|}{\sum_{ws=0}^{k} |P(WS)_{ws} - P(WS)_{ws-1}|} \tag{30}$$

Referring to eq. 30, variable k will first be set to the maximum wind speed present in all sites to ensure all observations are accounted for. Subsequently, penalization of both under and over prediction of wind speed is performed based on the difference between observed and forecasted wind and their position on the power curve.

$$PCCE = \begin{cases} (WS_t - \hat{WS}_t) \cdot \begin{cases} \left(1 - \sum_{ws=\hat{ws}_t}^{ws_t} weight_{ws}\right), & \sum_{ws=\hat{ws}_t}^{ws_t} weight_{ws} > 0.5 \\ \sum_{ws=\hat{ws}_t}^{ws_t} weight_{ws}, & \sum_{ws=\hat{ws}_t}^{ws_t} weight_{ws} \leq 0.5 \end{cases}, & WS_t \geq \hat{WS}_t \\ (\hat{WS}_t - WS_t) \cdot \begin{cases} \left(1 - \sum_{ws=ws_t}^{\hat{ws}_t} weight_{ws}\right), & \sum_{ws=ws_t}^{\hat{ws}_t} weight_{ws} > 0.5 \\ \sum_{ws=\hat{ws}_t}^{ws_t} weight_{ws}, & \sum_{ws=\hat{ws}_t}^{ws_t} weight_{ws} \leq 0.5 \end{cases}, & WS_t < \hat{WS}_t \end{cases} \tag{31}$$

Eq. 31 reveals that in both cases of under or over forecast, each forecast error is multiplied by the sum of the weights assigned between the observed and forecasted wind speed values. This approach gives the value of 0 when both forecast and observed value falls under region A, C, D of the power curve (with the exception of wrong wind speed post cut-out speed), and focuses mostly on the penalization in region B.

Furthermore, since the maximum possible change in power is from turbine peak power to 0, the weight of forecast error should also be capped at 0.5 to reflect that. Therefore, as indicated in Eq. 31, the PCCE value of summed weights above and below the value of 0.5 are calculated independently. Weights with values larger than 0.5 post-summation will be subtracted by 1 instead to convey the actual power difference stemming from the forecast error.

As mentioned in Chapter 2.4, empirical evidence from the Dutch electricity market promulgated it to be more costly for a system operator when wind speed is underestimated than overestimated, and postulated a penalty weight of 0.73 to reflect the economic cost associated with this (Pinson et al., 2007). Therefore, the previously formulated method PCCE, is expanded into the Power Curve Conversion Error with penalization (PCCEp) to take that into account.

$$PCCEp = \begin{cases} \lambda \cdot (WS_t - \hat{WS}_t) \cdot \begin{cases} \left(1 - \sum_{ws=\hat{ws}_t}^{ws_t} weight_{ws}\right), & \sum_{ws=\hat{ws}_t}^{ws_t} weight_{ws} > 0.5 \\ \sum_{ws=\hat{ws}_t}^{ws_t} weight_{ws}, & \sum_{ws=\hat{ws}_t}^{ws_t} weight_{ws} \leq 0.5 \end{cases}, & WS_t \geq \hat{WS}_t \\ (1-\lambda) \cdot (\hat{WS}_t - WS_t) \cdot \begin{cases} \left(1 - \sum_{ws=ws_t}^{\hat{ws}_t} weight_{ws}\right), & \sum_{ws=ws_t}^{\hat{ws}_t} weight_{ws} > 0.5 \\ \sum_{ws=\hat{ws}_t}^{ws_t} weight_{ws}, & \sum_{ws=\hat{ws}_t}^{ws_t} weight_{ws} \leq 0.5 \end{cases}, & WS_t < \hat{WS}_t \end{cases} \quad (32)$$

This study follows the penalization weight of 0.73 recommended in, taking into consideration that the Dutch electricity market in which the study was based on, operates within the EU power market just like Norway today (“Norway and the European Power Market,” 2016).

5. Results and Analysis

In the two preceding sections, characteristics of the data set as well as the methodologies of both forecasting models and evaluation tools were introduced and discussed. Just like the sizable input hourly time series for 69 sites and over 10 years of training data, our forecasting results consist of 69 hourly time series, forecasted over a period of 1 year, from 7 different forecasting models. Therefore, delving into every site/model forecasts discussions will be impractical. To convert the results into a more approachable format for visualization and discussion, models are evaluated largely based on averaged errors across all sites.

5.1 RMSE

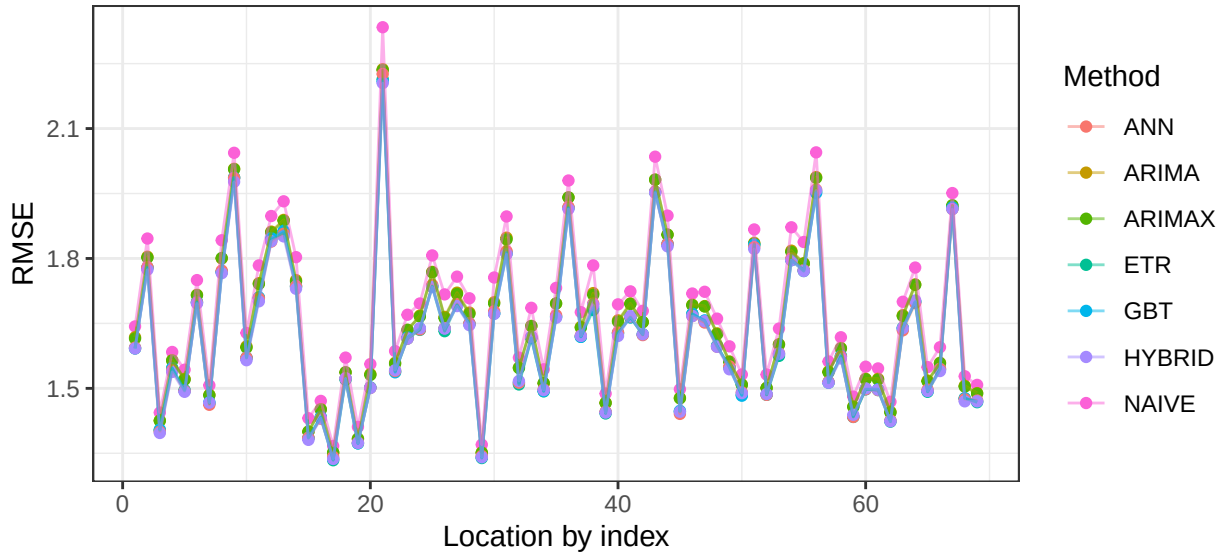

Figure 16: Location RMSE for different methods

Fig.16 above depicts the RMSE value of all models across all 69 sites. The results appear to be rather consistent across all models, with the best accuracy of RMSE around 1.1 achieved at site 17 and 29, corresponding to “Rakkocearro” and “Kjollefjord” , and the distinctively worst accuracy at site 21 - “Nygardsfjellet trinn I”.

Table 8: Forecast result using RMSE

	NAIVE	ARIMA	ARIMAX	GTB	ETR	ANN	HYBRID
RMSE	1.6841	1.6501	1.6487	1.6267	1.6257	1.627	1.6248

The first accuracy measure, RMSE, can be used to reflect the degree of differences between the actual and forecast values. The lower a model's RMSE value, the more accurate it is. In the one-step-ahead hourly forecast of wind speed exercise, RMSE of all models fall in the range of 1.62 - 1.69. The performance results appear to be consistent with the computational intensity required of each model.

The NAIVE benchmark model performs the worst, with the highest RMSE value of 1.684 while the Hybrid model performs the best with RMSE of 1.6248 slightly outperforming the three other ML models of similar accuracy performance within the 1.62 range.

5.2 MAE

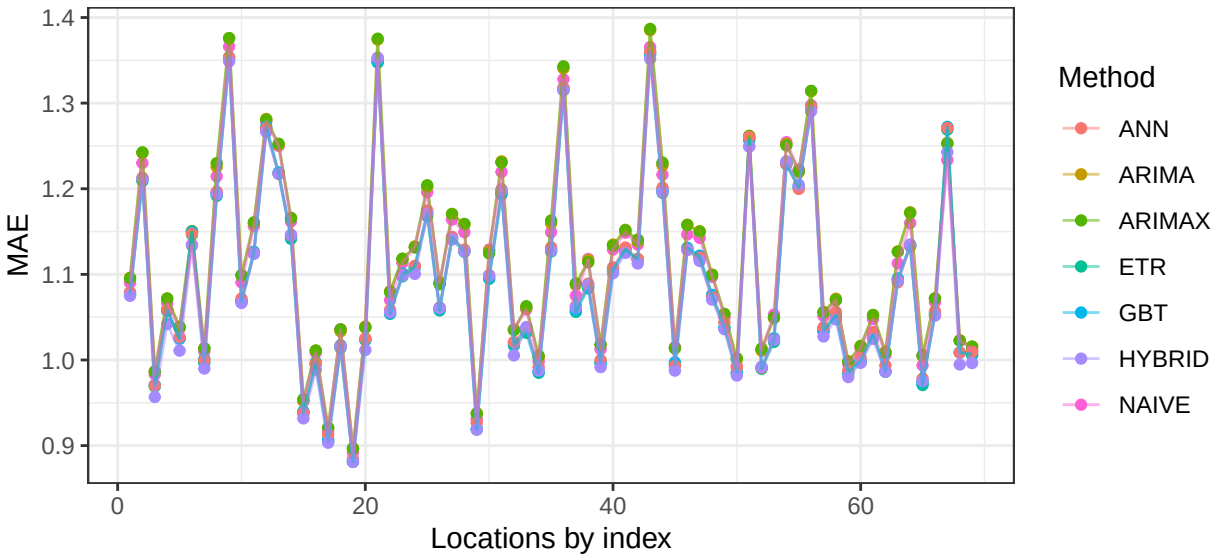

Figure 17: Location MAE for different methods

Fig. 17 above depicts the MAE value of all models across all 69 sites. Similar to RMSE, the results

also appear to be rather consistent across all models, with the best accuracies of MAE around 0.9 achieved at site 17, and 19, corresponding to Rakkocearro" "- and "Sandaugen", and the worst accuracies of MAE close to 1.4 at sites: 9 - "Mehuken II", 21 - "Nygardsfjellet trinn I", and 43 - "Testområde Stadt".

Table 9: Forecast result using MAE

	NAIVE	ARIMA	ARIMAX	GTB	ETR	ANN	HYBRID
MAE	1.1093	1.1153	1.1154	1.0942	1.0923	1.0953	1.0899

Unlike RMSE which penalizes bigger errors more than smaller errors, the next measure, MAE, reflects the level of forecasting error, with even penalization across all error magnitudes. Hence, compared to MAE, RMSE is more sensitive towards outliers and large errors. This presents a drawback where the response variable is right skewed, similar to this wind speed data set, as the RMSE error can be sharply punished by just a few data points on the right end of the distribution tail. Therefore, MAE could be an alternative measure to impede this hazard.

The most distinctive difference in accuracy performance of models under MAE is the fact that both ARIMA and ARIMAX models perform worse than the benchmark NAIVE model with MAE values of 1.153, 1.154 and 1.1093 respectively. On the other hand, similar to RMSE, Hybrid model is still the best model with the lowest MAE of 1.089, with the three ML models close behind with MAE in the range of 1.092-1.095. On that premise, it is reasonable to draw the conclusion that the Hybrid and ML models perform better than the other three models when it comes to traditional accuracy measures.

5.3 Classification accuracy

In table 10 below, Acc. LT refers to the accuracy on the left tail of the power curve with wind speed below 3 m/s, while Acc. RT refers to that of the right tail with wind speed above 25 m/s. Referring

to table 10, it is discernible that all models perform around 5% points better in LT, with accuracy in the range of 55-60%, than in the RT, with accuracy of 50-56%. It was surprising to find the persistence model outperforming all models, hinting, yet again, at the strength in the correlation of current wind speed with the past hour.

Table 10: Forecast result using classification accuracy

	NAIVE	ARIMA	ARIMAX	ETR	GTB	ANN	HYBRID
Acc. LT	0.591	0.5547	0.5553	0.569	0.5694	0.5757	0.5748
Acc. RT	0.559	0.5196	0.5211	0.5142	0.5105	0.5051	0.5423

In the right tail of a power curve, inaccurate forecast corresponds to a power production difference from peak power (2310kW/2300kW/3075kW) to zero, while in the left tail, the power production difference is only from the turbine's corresponding cut-in power (16kW/24kW/26kW) and zero. Therefore, the right tail accuracy can be deemed as having higher economic importance for model selection compared to the left. Based on this reasoning, in terms of overall ranking of models, the Hybrid model comes in second after the NAIVE model with Acc. RT of 0.543, and Acc. LT of 0.5748.

5.4 Power Curve Conversion Error

Much like the plots of RMSE and MAE, fig. 18 on next page depicts the PCCEp value of all models across all 69 sites also displaying consistency across all models. This similarity in location results comparison across all evaluation tools point to similarities in forecasts results and accentuates the possibility of outliers in our test sets time series.

Since PCCEp does not penalize wrong forecasts as evenly as RMSE or MAR, the accuracy value variation across all locations is the smallest amongst the three metrics. Moreover, most locations also have rather good accuracy values falling within the range of 0.06 - 0.08. The best accuracy of

PCCEp below 0.05 is observed at site 17 - “Rakkocearro”, while a very pronounced worst accuracy of PCCEp around 0.11 is illustrated by the spike in site 21 - “Nygardsfjellet trinn I”.

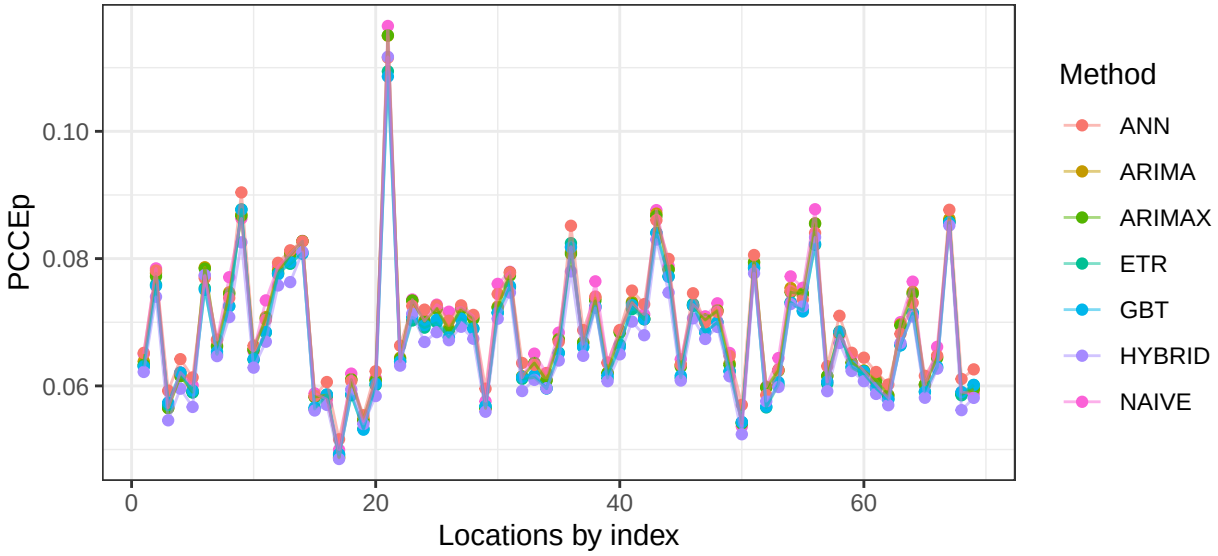

Figure 18: Location PCCEp for different methods

Table 11: Forecast result using PCCEp

	NAIVE	ARIMA	ARIMAX	GTB	ETR	ANN	HYBRID
PCCEp	0.0695	0.069	0.0688	0.0677	0.0676	0.0698	0.0666

Moving on to compare the averaged PCCEp values across all models, it is interesting to observe the benchmark NAIVE model performing marginally better than the worst model - ANN, with a PCCEp value of 0.698 v.s. 0.695. Taking into account the nuances incorporated into this metric, one can postulate that the ANN model might be under forecasting on a higher frequency than other models, hence reflecting the poor performance in PCCEp when under forecasts are more heavily penalized, but not under RMSE or MAE.

Once again, the Hybrid model is the most accurate model under this evaluation method, with a PCCEp value of 0.666.

5.5 Distribution of the forecasts

Further investigations into the distributions of forecasts reveals the benchmark NAIVE model to achieve the highest number of correct forecasts, congruous with the assumption of strong correlation present between the last and next hour of wind speed. Besides that, all models formulated with linear assumptions made no correct forecast, while the ML models managed to predict less than half the number of correct observations compared to NAIVE.

Table 12: Forecast distributions

	Total UF	Total OF	Correct	UF/OF	UF/Incorrect	OF/incorrect
NAIVE	295902	302961	7233	0.977	0.494	0.506
ARIMA	296316	309780	0	0.957	0.489	0.511
ARIMAX	297025	309071	0	0.961	0.490	0.510
ETR	300347	303020	2729	0.991	0.498	0.502
GBT	300341	303066	2689	0.991	0.498	0.502
ANN	321337	282367	2392	1.138	0.532	0.468
HYBRID	300574	305522	0	0.984	0.496	0.504

Diving into the percentage split of under forecasting and over forecasting, only ANN has more under than over forecast errors amongst all the models, confirming the hypothesis above why ANN was the worst model under PCCEp. Both statistical methods, ARIMA and ARIMAX models, perform better in this arena compared to the ML methods with 48.9% and 49% of wrong forecasts stemming from under forecasting. The remaining two ML methods, ETR and GBT, on the other hand, have a more even split of under forecast and over forecast, both achieving 49.8% and 50.2% for each sector respectively.

5.6 Seasonal performance

Comparing models performance across different months and based on the different evaluation methods, varying results of the 'best' models can be inferred. The proposed Hybrid model, by and large, gives the best results throughout the year whilst never being the worst model. It also

overthrows the benchmark NAIVE model across all months except the summer months of June, July, and August, albeit just mildly and specifically under the PCCEp metrics.

Table 13: seasonal performance

	Model	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
RMSE	NAIVE	2.0156	2.0698	1.8634	1.6274	1.4981	1.2272	1.2718	1.2645	1.7242	1.9109	1.7432	1.5984
	ARIMA	1.9688	2.0209	1.8316	1.5842	1.4718	1.2177	1.2609	1.2601	1.6808	1.8548	1.7076	1.583
	ARIMAX	1.9661	2.0153	1.8346	1.5826	1.4717	1.2142	1.2584	1.2543	1.6824	1.855	1.7063	1.5836
	GTB	1.9318	1.9956	1.8084	1.5686	1.4544	1.19	1.2354	1.2342	1.6623	1.829	1.6915	1.5643
	ETR	1.9288	1.9963	1.8074	1.5683	1.4532	1.1887	1.2343	1.2325	1.6623	1.8263	1.6927	1.5614
	ANN	1.9349	1.9959	1.8142	1.5609	1.4539	1.1901	1.2366	1.2333	1.6638	1.8248	1.6951	1.5615
	HYBRID	1.9278	1.9926	1.8102	1.5607	1.4496	1.1902	1.238	1.2345	1.6629	1.8259	1.6913	1.5562
MAE	NAIVE	1.282	1.4086	1.2577	1.1022	1.042	0.8513	0.8932	0.8714	1.2056	1.2324	1.1647	1.0167
	ARIMA	1.2913	1.4014	1.2658	1.0985	1.0468	0.8691	0.904	0.8916	1.1956	1.2228	1.1651	1.0453
	ARIMAX	1.2928	1.3973	1.2711	1.0986	1.0476	0.865	0.9007	0.8823	1.1991	1.2266	1.1661	1.052
	GTB	1.2665	1.3819	1.2459	1.0818	1.0262	0.8361	0.8762	0.8607	1.1831	1.2101	1.1519	1.025
	ETR	1.2632	1.3814	1.2426	1.0801	1.0245	0.8338	0.8742	0.8595	1.1825	1.2084	1.1517	1.0202
	ANN	1.2715	1.3838	1.2544	1.0766	1.0245	0.8372	0.8779	0.8597	1.1836	1.2066	1.1592	1.0236
	HYBRID	1.2597	1.3773	1.2458	1.0736	1.0181	0.8342	0.8754	0.857	1.1818	1.2048	1.1513	1.0142
PCCEp	NAIVE	0.0917	0.0986	0.0791	0.0714	0.0591	0.0413	0.0453	0.0404	0.0829	0.0876	0.0764	0.0621
	ARIMA	0.0916	0.0979	0.0789	0.0693	0.0586	0.0414	0.045	0.0403	0.0815	0.0853	0.076	0.0634
	ARIMAX	0.0893	0.095	0.0797	0.0697	0.0591	0.0436	0.0472	0.0419	0.0805	0.0844	0.0752	0.0614
	GTB	0.0871	0.0938	0.078	0.0686	0.0582	0.0424	0.0463	0.0413	0.0792	0.0828	0.0742	0.0613
	ETR	0.0873	0.0943	0.0777	0.0684	0.058	0.042	0.0459	0.0411	0.0795	0.0827	0.0741	0.0617
	ANN	0.0897	0.0965	0.0807	0.0701	0.06	0.0437	0.0477	0.0425	0.0825	0.0851	0.0774	0.0632
	HYBRID	0.0856	0.0922	0.0771	0.0673	0.057	0.0417	0.0457	0.0407	0.0786	0.0818	0.0731	0.0598

6. Discussion

Building upon the previous chapter which underlined the forecasting results gathered from the study, the following section dives into a discussion of these results and their economic implications. Subsequently, potential shortcomings imposed by both the data sets and methodologies were explored and developed into suggestions on ways to enhance the current research topic.

6.1 Discussion of Results and Implication of Study

6.1.1 Forecasting Results

Summarizing the results of the study to answer the research question, forecast accuracy results suggest the Hybrid model to be the most suitable forecasting method for this one-hour-ahead hourly Norwegian wind power production forecast. This conclusion holds true under almost all of the evaluation metrics employed in the study, as well as for wind power forecasting under different seasons of the year.

Besides the Hybrid model, two other ML models - GTB and ETR, also demonstrated better forecast results over the benchmark persistence model consistently across all accuracy measures as depicted in 14 below.

Table 14: Method comparision

Method	RMSE	MAE	Acc. LT	Acc. RT	PCCEp
NAIVE	1.6841	1.1093	0.591	0.559	0.0695
ARIMA	1.6501	1.1153	0.5547	0.5196	0.069
ARIMAX	1.6487	1.1154	0.5553	0.5211	0.0688
ETR	1.6257	1.0923	0.569	0.5142	0.0676
GTB	1.6267	1.0942	0.5694	0.5105	0.0677
ANN	1.627	1.0953	0.5757	0.5051	0.0698
HYBRID	1.6248	1.0899	0.5748	0.5423	0.0666

In general, models' accuracies appear to be consistent with the computational intensity required of each model, with minor improvements in terms of accuracies. Although the overall accuracy

differences might seem minute, mostly in the scale of 2-3 decimal points, considering the characteristics in wind speed to power conversion as discussed in the previous section, small differences in wind speed forecasting error could easily translate into significant differences in actual power forecast.

In terms of seasonal forecast accuracy, it is also interesting to note the significantly different results indicated by PCCEp from that RMSE and MAE during the summer months. For the months of June, July, and August, the benchmark NAIVE model is the worst performing model under RMSE, while the ARIMA model was singled out as the worst performing for MAE. Surprisingly, they were both the best performing models for PCCEp for the same months. This could be attributed to the fact that Norwegian wind speeds in summer months have the pronouncedly smaller standard deviations compared to all other seasons, coupled with the fact that PCCEp focuses predominantly on wrong forecasts within Region B of power curve, allowing both models - which are usually insensitive to sporadically large or small values, to do remarkable well in these months under PCCEp.

Finally, as depicted in fig.16, fig. 17, and fig. 18, site 21 - "Nygardsfjellet trinn I" can easily be singled out as the location with the worst forecast accuracy, and site - "Rakkocearro" the best consistently across all three aforementioned metrics. Looking at the errors between the best and the worst sites reveals significant differences as depicted in fig. 19 **A** at the next page. While errors from "Rakkocearro" stay close to the center, errors from "Nygardsfjellet trinn I" experiences huge fluctuations with spikes above 20.

Moreover, further investigation into their differences based on test data wind speed values exposes strikingly large amounts of outliers in "Nygardsfjellet trinn I", along with a highly right-skewed and double peaked distribution depicted in fig. 19 **B** one next page. This reveals the complexity and difficulty in generalizing forecasts for several sites, where the preferable method could be to build site-specific forecast models.

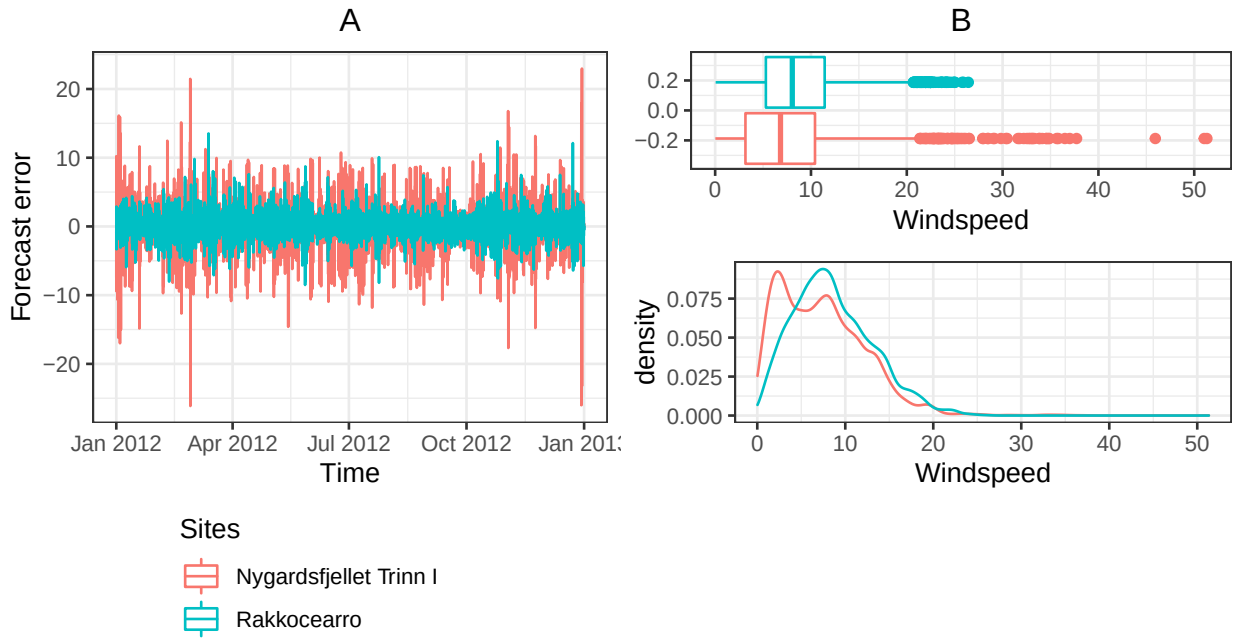

Figure 19: Comparison plot of the worst and the best

6.1.2 Economic Implications

When it comes to comparing models, PCCEp not only takes into account forecast errors made in different magnitudes of wind power based on each site's power curve, but also errors made in difference between under/over forecasting. Therefore, PCCEp can easily be adapted for different wind turbine power curves and different penalty values to reflect different electricity markets.

Since the study is based on hypothetical wind sites in Norway, it was not possible to perform empirical studies on actual cost incurred by various parties from poor wind forecasts due to the absence of actual data. Economic implications of this study hence focuses on economic costs faced by system operators in the event of under/over forecasting. As mentioned earlier, the under/over forecasting of wind speed leads to the under/over forecasting in wind power, and results in the necessary up-regulation (increasing generation) and down-regulation (reducing generation) of electricity carried out by the system operators. Typically, when wind speed is under forecasted, system operators will end up over-ordering surpluses of electricity from traditional sources, in-

evitably driving prices up (Hering & Genton, 2010).

That said, in different electricity markets, and at different time of the year, there will also be different costs associated with individual up-regulation and down-regulation since incurred costs are highly dependent on local conditions, such as the cost of the backup electricity source to cover for RES supply variation.

Since the λ (penalty) value of 0.73 was quoted from a 2002 Dutch electricity market data, we found it prudent to investigate into the different ranges for the applicability to reflect the reality of different electricity markets as well as different seasonalities. Following the logic of 0.73 reflecting a heavier penalization for under forecasting, the smaller a penalty value, the more skewed the metric will be towards heavier penalization of over forecasting. A range of different λ value, from 0.25 to 0.85 is employed and the results are illustrated in fig. 20 below.

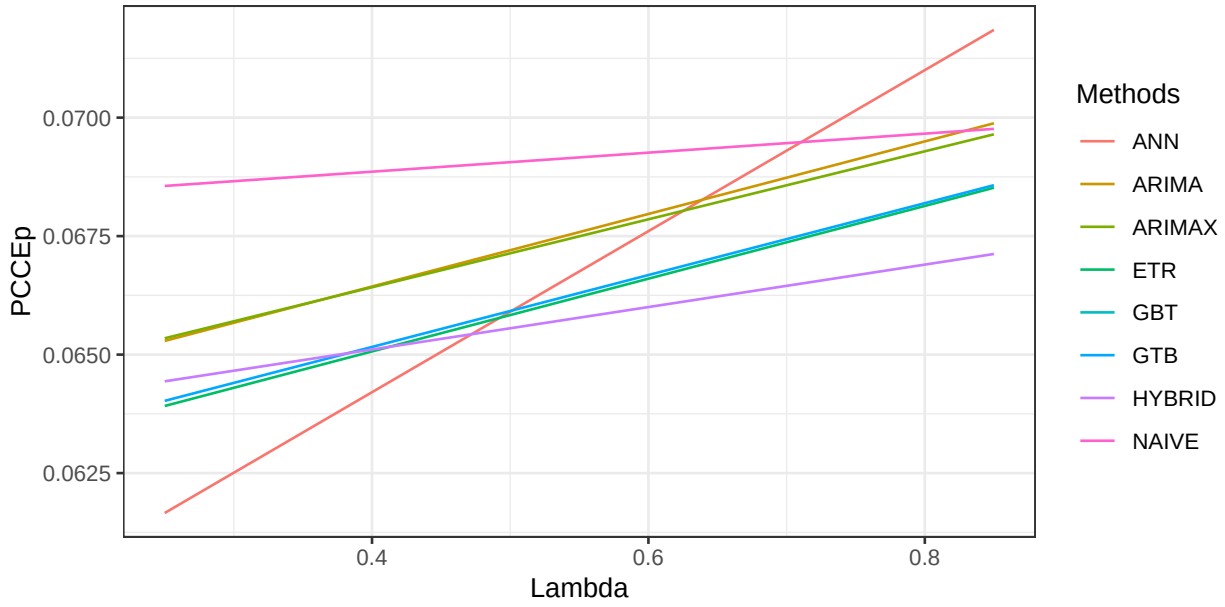

Figure 20: Result using different Lambda values

Based on the results above, there do not seem to be a preferred model that outperforms the rest when comparing across different λ values. That said, for $\lambda < 0.48$, where over forecasting is heavier penalized than under, ANN model seems to be the best performing. In contrast, for λ

> 0.48, Hybrid model seems to be preferred. Judging from the steepness of all the plots, the Hybrid model also spots the lowest gradient, suggesting that a combination of linear and non-linear models could make the forecast more robust. Furthermore, the ETR and GBT models both have very similar accuracy performance across all lambda values, just like the ARIMA and ARIMAX models.

Finally, the positive slope coefficient across all models also suggests that all the models have higher tendencies to under forecast than over forecast. However, keeping in mind that the PCCEp values were aggregated over all 69 sites, there is also the possibility of poor under forecasts occurring only at specific sites.

6.2 Limitations

Even though the final aim of the study was focused on the production of wind power, wind speed accuracy was adopted instead as a variable to emulation forecasted power in an attempt to combat the additional uncertainties that could be introduced during wind speed to power conversion. Most of the limitations surrounding this study, as will be explained in the following paragraphs, can be attributed to this problem.

6.2.1 Data limitations

One of the most apparent limitations of this study stems from the variation surrounding expected v.s. actual wind speed to power conversions. As explained in chapter 2.2, theoretical wind turbine power curves were built upon measurements under standard test conditions of direct front flow wind, standard air pressure and temperature, all of which are unlikely to be stable in a real site.

With only univariate wind speed data available, we were not able to take into account site-specific attributes such as the local turbulence, complexity of terrain, actual direction of wind flow, and air density at each hour. All these idiosyncrasies, coupled with the fact that wind speed data was

more than a decade old and does not take into consideration the possible and potential impacts of climate change on wind speed, might also result in significant differences on the final predicted wind power output compared to that from the expected power output derived from a theoretical power curve.

Besides that, in this study, the types of wind turbines that could be allocated to each location were limited to three choices based on Kjeller Vindteknikk power curves. Therefore, simply categorizing the sites into either of the three options based on their historical average wind speed introduces another layer of imprecision into the study. Firstly from the limited options, and more importantly, on the use of the site's average wind speed as a means of classification. Considering the right-skewed distribution of wind speed at each site, taking the average wind speed albeit easy, might not be the most befitting method. This sweeping assumption made on the choice of turbine naturally translates into the subsequent wind speed to power curve conversion assumption, which could result in significantly different expected v.s. actual power output. Although this is adequate as a preliminary wind potential study of Norway, however, if site specific wind power potential is of significance to the study, the optimal power curve for each site should be established by a more extensive research and categorization of their respective power curves.

6.2.2 Methodology limitations

When it comes to methodologies, one of the limitations comes from the sheer size of the main wind speed data set, as iterated previously. Not only does it result in significant computational intensity constraints, forcing us to aggregate ML models instead of formulating site-specific ML models for better accuracy, data size used for ML parameters tuning were also forced down to random 10% subsample instead of the whole data set because of the time required to run each iteration.

Aside from ML model formulations, forecasts results must also be interpreted based on averaged values. These aggregations, albeit necessary, could have easily introduced biases, or concealed certain trends that should otherwise have been discernible.

Another point is the additional source of error introduced due to the discrete nature of power curve conversion data set. Most forecasted wind speeds from the models are not in whole numbers, resulting in a rounding effect during the calculation of the PCCEp loss function. Forecasted wind speed values have to be rounded off to the nearest whole number for the summation of weights as seen in eq. 30 implementation of the accuracy evaluation measure.

The final limitation surrounds the interpretation of the loss function PCCEp. The way it was formulated focuses predominantly on the purpose of purely detecting the most appropriate forecast method based on penalization. Therefore, the loss function was only able to rank forecast models, but not provide any level values. Hence, if quantifying economical cost was the fundamental purpose, it could be of interest to attempt at converting the loss function in level to represent costs.

6.3 Suggestion for Future Research

Based on the limitations highlighted in the previous subsection, one of the extensions that could possibly improve forecast accuracy would be the incorporation of additional covariate information that will have time-specific impact on either wind speed (weather forecasts) or wind power (air temperature, air pressure, wind direction) into the forecasting models.

A more sophisticated approach such as the clustering of wind sites prior to forecasting could definitely help make the data sets a lot more manageable, serving as an alternative to all the averaging and aggregating currently employed. Sites could be clustered based on their wind speed characteristics, power curves, and perhaps even by sites' location as deduced in a theoretical study by (Blekastad & Landa, 2020), there is a high correlation between power and location distance, suggesting a highly correlated wind speed and location distance relationship as well. Proper clustering could contribute greatly to this study, possibly giving better and more insights into the data.

Further research on this topic could revolve around the refining of the tailored loss function PCCEp. Diving into the weightage component of PCCEp, the turbine power curve could be converted into

continuous values before incorporation into the loss function. This could have prevented errors from being introduced due to the rounding effect of forecasted wind speed. Besides that, other methods of assigning weights to the loss function could also be explored.

In terms of the penalty component of PCCEp, more research into Norway specific power market conditions such as the cost related to down-regulation and up-regulation of electricity could be carried out to find the value that best represents the norwegian electricity market. Building upon that, month-specific penalty value could also be adopted to reflect the possible changes in up/down regulation costs with changing seasonalities, given that the hydroreservoir levels are highly influenced by seasons too.

7. Conclusion

Given the stochastic nature of wind speed and the promise of harnessing wind energy, wind forecasting is a widely researched topic. A plethora of studies and immeasurable models recommendations have been made. However, to determine the best model is largely data specific. In other words, what works for one data set might not always be the best for another. Moreover, the method of justifying the 'best' model is often contentious as well with so countless accuracy measures available on top of problem-specific nuances to consider.

The purpose of this study was therefore targeted at answering the research questions: *"What is the most suitable forecasting method to predict hourly wind power production (1-hour-ahead) in Norway under various accuracy evaluation tools? And, do the methods' performance vary at different times of the year?"* Besides the implementation of various forecasting methods, the analysis also focuses on factors relating towards formulating an accuracy evaluation tool suitable in capturing both the nuances surrounding wind speed to wind power conversion, and the economic costs related to wrong forecasting of wind power faced by system operators.

Eleven years (2000 - 2011) of univariate hourly wind speed time series across 69 sites along the Norwegian coast were used to train 7 different forecasting models, and one year (2012) of wind speed data for each location was hold-out as test data for accuracy comparison of these models. To simulate the conversion of wind speed to wind power production, each location was also categorized based on their average wind speed into their respective turbine power curves based on the recommendations from Kjeller Vindteknikk.

The results from the study reveal forecast accuracies to be quite similar across all models. This highlights the strong correlation between current and previous hour's wind speed since the benchmark NAIVE model was able to achieve accuracies just slightly worse than all other models. That said, the computational intensity required of each model is also rather aligned with their respective accuracies, with ML models performing better than statistical methods in most situations albeit

only to a small extent. In general, when forecasting accuracy is prioritized, the Hybrid model is the best choice of model regardless of seasons and evaluation metric.

Focusing on the tailored solution, PCCEp, can easily be adapted for different turbines, markets and seasons. Therefore, depending on the economic costs faced by system operators, two different model recommendations will be made. Keeping sites locations and their associated turbine's power curves constant, results from PCCEp, suggest that when the cost bears by system operators are higher for over forecasting , the ANN model is preferred. Conversely, when the cost of under forecasting is higher, the Hybrid model of ARIMAX and ETR gives better accuracy. Therefore, it is recommended for detailed research to be developed in this field to more precisely allocate the correct penalty at different months considering Norwegian hydropower optimization is largely dependent on water level and dams capacity, hence highly correlated to the month of year.

8. References

Adomaitis, N., By, & -. (2020). Norway will slow down onshore wind power developments. ArcticToday. Retrieved from https://www.arctictoday.com/norway-will-slow-down-onshore-wind-power-developments/

Arlova, M., & Fedorova, D. (2016). Selection of unit root test on the basis of length of the time series and value of ar(1) parameter. *Statistika*, 47–64.

Ayodele, T. R., Jimoh, A. A., Munda, J. L., & Agee, J. T. (2012). Challenges of grid integration of wind power on power system grid integrity: A review. *INTERNATIONAL JOURNAL of RENEWABLE ENERGY RESEARCH*, *2*(4), 618–626.

Aznar, P. (2020). *What is the difference between extra trees and random forest?* Retrieved from https://quantdare.com/what-is-the-difference-between-extra-trees-and-random-forest/

Ådland, H. M. (2018). Power markets : Strategy and regulatory framework. NHH.

Birol, D. F. (2017). CO2 emissions from fuel combustion. INTERNATIONAL ENERGY AGENCY. Retrieved from https://euagenda.eu/upload/publications/untitled-110953-ea.pdf

Blekastad, M., & Landa, K. J. (2020). *The power of wind – a portfolio approach* (PhD thesis).

Borsche, T. (2019). Handling the risk of renewable investment - nordic lessons for the european market. THEMA Consulting Group. Retrieved from https://www.strommarkttreffen.org/2019-01_Borsche_Erfahrung_mit_PPAs_in_norwegischen_Windprojekten.pdf

Cadenas, E., & Rivera, W. (2010). Wind speed forecasting in three different regions of mexico, using a hybrid arima–ann model. *Renewable Energy*, *35*(12), 2732–2738. https://doi.org/10.1016/j.renene.2010.04.022

DeFusco, R. A., McLeavey, D. W., Pinto, J. E., & Runkle, D. E. (2015). *Quantitative investment analysis* (Third). John Wiley & Sons.

DeMarle, P. (2019). The basics of the newsvendor model. Medium. Retrieved from https://

medium.com/@pdemarle/the-basics-of-the-newsvendor-model-ef756f203433

Destro, N., Korpås, M., & Sauterleute, J. F. (2016). Smoothing of offshore wind power variations with norwegian pumped hydro: Case study. *Energy Procedia*, *87*, 61–68. https://doi.org/10.1016/j.egypro.2015.12.358

Du, P., Wang, J., & Niu, T. (2019). A novel hybrid model for short-term wind power forecasting. *Applied Soft Computing*, *80*, 93–106. https://doi.org/https://doi.org/10.1016/j.asoc.2019.03.035

Elsayed, M. M. (2017). Wind energy overview 170908. Engineering. Retrieved from https://www.slideshare.net/MoustafaMElsayed/wind-energy-overview-170908

Elsinghorst, D. S. (2018). Machine learning basics - gradient boosting & xgboost. Shirin's playgRound. Retrieved from https://www.shirin-glander.de/2018/11/ml_basics_gbm/

Energy, M. of P. and. (2016). The history of norwegian hydropower in 5 minutes. Government.no. Retrieved from https://www.regjeringen.no/en/topics/energy/renewable-energy/the-history-of-norwegian-hydropower-in-5-minutes/

Farzaneh, T., & Majid, M. (2017). Hourly wind speed prediction using arma model and artificial neural networks. *International Journal of Smart Electrical Engineering*, *6*(3), 101–107. https://doi.org/http://ijsee.iauctb.ac.ir/article_537457_c1f3b4a13a66f9297c4e05cf7a9a398e.pdf

Foresightr. (2016). foresightR. Retrieved from https://foresightr.com/2016/05/06/a-brief-history-of-forecasting/

Førsund, F. R., Singh, B., Jensen, T., & Larsen, C. (2008). Phasing in wind-power in norway: Network congestion and crowding-out of hydropower. *Energy Policy*, *36*(9), 3514–3520. https://doi.org/10.1016/j.enpol.2008.06.005

Gneiting, T., Larson, K., Westrick, K., Genton, M. G., & Aldrich, E. (2006). Calibrated probabilistic forecasting at the stateline wind energy center. *Journal of the American Statistical Association*, *101*(475), 968–979. https://doi.org/10.1198/016214506000000456

Graabak, I., Korpås, M., Jaehnert, S., & Belsnes, M. (2019). Balancing future variable wind and

solar power production in central-west europe with norwegian hydropower. *Energy*, *168*, 870–882. https://doi.org/10.1016/j.energy.2018.11.068

Greenhouse gas emissions by aggregated sector. (2019). Retrieved from https://www.eea.europa.eu/data-and-maps/daviz/ghg-emissions-by-aggregated-sector-5

Grigonytė, E., & Butkevičiūtė, E. (2016). Short-term wind speed forecasting using arima model. *Energetika*, *62*(1-2). https://doi.org/10.6001/energetika.v62i1-2.3313

Hering, A. S., & Genton, M. G. (2010). Powering up with space-time wind forecasting. *Journal of the American Statistical Association*, *105*(489), 92–104. https://doi.org/10.1198/jasa.2009.ap08117

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Forecasting: Principles and practice. OTexts. Retrieved from https://otexts.com/fpp2/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning with applications in r* (Eight). Springer.

Jiang, P., Yang, H., & Heng, J. (2019). A hybrid forecasting system based on fuzzy time series and multi-objective optimization for wind speed forecasting. *Applied Energy*, *235*, 786–801. https://doi.org/10.1016/j.apenergy.2018.11.012

Jingxing, W., Abdullah, A., Mingdi, Y., Eunshin, B., & Romesh, S. (2020). *Integrative density forecast and uncertainty quantification of wind power generation*. 1–8. https://doi.org/https://arxiv.org/pdf/1808.07347.pdf

Jj. (2016). MAE and rmse - which metric is better? Human in a Machine World. Retrieved from https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

Lei, M., Shiyan, L., Chuanwen, J., Hongling, L., & Yan, Z. (2009). A review on the forecasting of wind speed and generated power. *Renewable and Sustainable Energy Reviews*, *13*(4), 915–920. https://doi.org/https://www.sciencedirect.com/science/article/abs/pii/S1364032108000282

Lift and drag. (n.d.). Boston University. Retrieved from http://people.bu.edu/dew11/liftanddrag.html

Machine learning in python. (n.d.). Retrieved from https://scikit-learn.org/stable/getting_started.html

"Narve" feier inn over norge. (2006). Retrieved from https://www.aftenposten.no/norge/i/4ozeG/narve-feier-inn-over-norge

Norway and the european power market. (2016). NVE. Retrieved from https://www.nve.no/norwegian-energy-regulatory-authority/wholesale-market/norway-and-the-european-power-market/

Papaefthymiou, G. (2009). Using copulas for modeling stochastic dependence in power system uncertainty analysis. *IEEE TRANSACTIONS ON POWER SYSTEMS*, *24*(1), 40–49. Retrieved from https://www.researchgate.net/publication/224358098_Using_Copulas_for_Modeling_Stochastic_Dependence_in_Power_System_Uncertainty_Analysis

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Petrova, V. (2018). Renewables supply 25. Renewables Now. Retrieved from https://renewablesnow.com/news/renewables-supply-25-of-global-power-in-2017-iea-606070/

Pinson, P., Chevallier, C., & Kariniotakis, G. N. (2007). Trading wind generation from short-term probabilistic forecasts of wind power. *IEEE Transactions on Power Systems*, *22*(3), 1148–1156. https://doi.org/10.1109/tpwrs.2007.901117

Rehman, S., Alam, M. M., Alhems, L. M., & Rafique, M. M. (2018). Horizontal axis wind turbine blade design methodologies for efficiency enhancement—a review. *Energies*, *11*(3). https://doi.org/https://www.mdpi.com/1996-1073/11/3/506

Reuters. (2020). Norway to slow down onshore wind power developments - et energyworld.

ETEnergyworld.com. Retrieved from https://energy.economictimes.indiatimes.com/news/renewable/norway-to-slow-down-onshore-wind-power-developments/76475010

Smith, J. C., Parsons, B., Milligan, M. R., Acker, T., Zavadil, R., Schuerger, M., & Demeo, E. (2007). European wind energy conference and exhibition 2007, ewec 2007. In *Best practices in grid integration of variable wind power: Summary of recent us case study results and mitigation measures* (Vol. 1, pp. 500–509). EWEC 2007.

The power curve of a wind turbine. (2003). Danish Wind Industry Association. Retrieved from http://xn--drmstrre-64ad.dk/wp-content/wind/miller/windpower%20web/en/tour/wres/pwr.htm

Torres, J., García, A., Blas, M. D., & Francisco, A. D. (2005). Forecast of hourly average wind speed with arma models in navarre (spain). *Solar Energy*, *79*(1), 65–77. https://doi.org/10.1016/j.solener.2004.09.013

UN: Prepare for 'millions' more climate displaced. (2020). Climate Centre. Retrieved from https://www.climatecentre.org/news/1242/un-prepare-for-a-millionsa-more-climate-refugees

Wang, D., Luo, H., Grunder, O., & Lin, Y. (2017). Multi-step ahead wind speed forecasting using an improved wavelet neural network combining variational mode decomposition and phase space reconstruction. *Renewable Energy*, *113*, 1345–1358. https://doi.org/10.1016/j.renene.2017.06.095

Wang, Y., Wang, J., & Wei, X. (2015). A hybrid wind speed forecasting model based on phase space reconstruction theory and markov model: A case study of wind farms in northwest china. *Energy*, *91*, 556–572. https://doi.org/10.1016/j.energy.2015.08.039

Why the weibull distribution is always welcome. (2013). The Minitab Blog. Retrieved from https://blog.minitab.com/blog/understanding-statistics/why-the-weibull-distribution-is-always-welcome

Wooldridge, J. M. (2018). *Introductory econometrics, a modern approach* (Seven). Cengage.

Wu, Y.-K., & Hong, J.-S. (2007). Power tech. In *A literature review of wind forecasting technology in the world*. IEEE Lausanne Power Tech.

Xiao, Z., Zhao, Q., Yang, X., & Zhu, A. (2020). A power performance online assessment method of a wind turbine based on the probabilistic area metric. Multidisciplinary Digital Publishing Institute. Retrieved from https://www.mdpi.com/2076-3417/10/9/3268/htm

Yang, M., Zhang, L., Cui, Y., Yang, Q., & Huang, B. (2019). The impact of wind field spatial heterogeneity and variability on short-term wind power forecast errors. *Journal of Renewable and Sustainable Energy*, *11*(3), 033304. https://doi.org/10.1063/1.5064438

Zhao, J., Guo, Z.-H., Su, Z.-Y., Zhao, Z.-Y., Xiao, X., & Liu, F. (2016). An improved multi-step forecasting model based on wrf ensembles and creative fuzzy systems for wind speed. *Applied Energy*, *162*, 808–826. https://doi.org/10.1016/j.apenergy.2015.10.145

Appendix

Table 15: Model order for ARIMA and ARIMAX

Windfarm	ARIMA	ARIMAX	Windfarm	ARIMA	ARIMAX
Storheia	4,0,0	1,0,3	Mehuken I	2,0,3	2,0,0
Roan	2,0,0	1,0,3	Ytre Vikna trinn I	5,0,0	2,0,0
Skinansfjellet	3,0,1	2,0,5	Hundhammerfjellet demo II	1,0,2	1,0,3
Friestad	2,0,2	2,0,2	Havoygavlen	1,0,2	1,0,3
Royrmyra	2,0,2	1,0,5	Bessakerfjellet II	1,0,2	5,0,0
Nygardsfjellet trinn II	2,0,1	2,0,2	Sway Karmoy	5,0,2	3,0,2
Eikeland Steinsland	5,0,3	2,0,3	Hywind	3,0,3	3,0,2
Havsul I	3,0,3	4,0,0	Testomrade Stadt	3,0,3	2,0,5
Mehuken II	2,0,3	2,0,5	Vardoya	1,0,2	1,0,4
Kvenndalsfjellet	4,0,0	1,0,3	Asen II	2,0,3	3,0,2
Valsneset	3,0,0	1,0,1	Sandvesanden	2,0,2	2,0,5
Midtfjellet	1,0,4	1,0,3	Nordbo	1,0,2	3,0,1
Sormarkfjellet	2,0,0	2,0,0	Askjesundet	2,0,2	4,0,2
Fakken	5,0,0	2,0,2	Raudfjell	1,0,5	1,0,3
Andmyran	2,0,0	2,0,5	Hamnefjell	1,0,3	1,0,1
Gravdal	2,0,4	3,0,2	Donnesfjord	1,0,5	2,0,2
Rakkocearro	1,0,1	2,0,0	Hitra II	1,0,2	1,0,2
Hitra I	4,0,0	3,0,0	Froya	1,0,3	2,0,2
Sandhaugen	2,0,2	1,0,4	Geitfjellet	5,0,0	2,0,3
Hog Jaren trinn I	2,0,2	4,0,1	Svarthammaren	5,0,0	1,0,5
Nygardsfjellet trinn I	2,0,2	2,0,2	Remmafjellet	5,0,0	5,0,0
Lista	3,0,1	2,0,2	Svaheia	2,0,2	4,0,4
Tysvar	3,0,0	1,0,3	Tellenes	2,0,4	2,0,3
Utsira	3,0,1	3,0,2	Stigafjellet	3,0,2	2,0,4
Haramsfjellet	3,0,0	3,0,1	Makaknuten	3,0,2	2,0,3
Hundhammerfjellet demo I	1,0,3	2,0,3	Eigersund	2,0,2	2,0,3
Bessakerfjellet I	4,0,0	4,0,1	Kvinesheia	5,0,1	5,0,0
Harbakfjellet	2,0,0	1,0,4	Lutelandet	4,0,0	3,0,3
Kjollefjord	1,0,1	2,0,0	SWAY Kollsnes	2,0,2	2,0,3
Hundhammerfjellet	2,0,2	2,0,2	Ytre Vikna trinn II	1,0,3	4,0,1
Valsneset testsenter	1,0,1	3,0,3	Anstadblaheia	4,0,0	4,0,1
Fjeldskar	2,0,2	2,0,4	Sorfjorden	1,0,5	2,0,1
Smola	1,0,3	2,0,2	Hog Jaren trinn II	2,0,2	2,0,2
Kvitfjell	1,0,1	1,0,2	Falesrassa	5,0,0	2,0,2
Haroy	3,0,3	5,0,0			