Norwegian University
of Life Sciences

Master's Thesis 2023 30 ECTS
Faculty of Science and Technology

Indirect Forecasting of Power Generation From a PV Power Plant Through Data-Driven and Physical Methods

Sigrid Vøllo
M.Sc. Environmental Physics and Renewable Energy

Preface

In this thesis, I have made forecasts of solar power production for a large-scale power plant with the use of machine learning. This has been done using what I have learned in the past five years as a student in Environmental Physics and Renewable Energy. The process has included many weeks of preprocessing of the data to understand the underlying connections in the data, several weeks to make a method for forecasting using various techniques, and finally much time now in the end to analyze and understand the outputs of the forecasts.

All of this would have been impossible without the help of my wonderful supervisors. I would therefore like to thank Jo Gjessing from Scatec for providing me with data as well as improving my understanding of how utility-scale PV work, Magnus Moe Nygård from IFE for helping me with building my methods and understanding substantially more of how academic work is done, and Heidi S. Nygård for helping me make it all come together as a scientific work and for all good advice along the way. I would also like to thank Leonardo Rydin Gorjão for helping me with all the minor and major issues I faced with my code and writing along the way.

As I am writing this, I am slowly starting to realize that my five years as a student here at NMBU are coming to an end. There have been so many people here at the university along the way helping us through difficult times and concepts. My family has also been a great support and motivator in these years. I also feel the need to thank my teachers at Numedal Videregående Skole, who inspired me to make the choice I did five years ago and are part of the reason I have had these great years. And lastly, my amazing fellow students, at the university as a whole and at our study program, but especially the amazing people in GRUS. You have made these years wonderfully memorable both as great study buddies and amazing friends.

Sigrid Vøllo
Ås, May 2023

Abstract

Most people are familiar with the fact that the weather conditions are neither constant nor controllable. The power generation from photovoltaic (PV) technologies are highly dependent on the weather, thus accurate power generation forecasts are necessary if one wants knowledge about power generation hours and days ahead for PV power plants. Understanding and accurately forecasting PV power generation is also central to ensuring the stability of the electrical power system.

In this thesis, two methods, Method 1 and 2, were developed to forecast power generation with a 24-hour horizon and a 1-hour resolution for a PV power plant. In the development of the methods, three years of measurement data from a utility-scale PV power plant with an installed capacity above 100 MW were utilized. Weather forecasts from the weather forecasting service *Yr* and empirical data generated with the *Python* library *pvlib* were used. These data were concatenated and preprocessed with outlier detection, missing values imputation, and min-max scaling. After this, the forecasting methods were developed. These were compared to forecasts by the commercial power generation forecast provider Solargis.

Both methods in this thesis used an indirect approach where the initial step of forecasting Global Horizontal Irradiance (GHI) was equal in both methods. The GHI was forecasted with a machine learning method using Random Forest Regression (RFR). For this forecast, the input features were forecasted ambient temperature, clouds, low clouds, medium-height clouds, and precipitation, local measurements of GHI from the previous day, and historic measurements of ambient temperature and wind direction. To forecast the power generation Method 1 used the same RFR method that was used to forecast the GHI. This time, the input features were forecasts for medium clouds, wind direction, and humidity, clear sky irradiance, the GHI measurements from the previous day, and the forecasts for the GHI made in the initial step. Method 2 used a series of physical and empirical operations to calculate a forecast for power generation based on the GHI forecasts.

From the evaluation of the results, Method 1 produced the forecast with the highest skill score of 0.200. Solargis received a score of 0.122 and Method 2 0.004. In general, it was observed that Method 1 had a tendency to underestimate power generation, and Method 2 and Solargis overestimated it. It was generally seen that Method 1 had difficulties forecasting the peak generation hours. Some of the bias in the results might partially be a result of power curtailment. Because of conditions in the electrical power grid, the power generation from the PV power plant was reduced with curtailment, this was not evident in the weather data used to make the forecast in this thesis.

It was concluded that Method 1 is a viable method for power generation forecasts for PV power plants and that it has an accuracy enabling it to compete with commercial solutions. With more time to refine the method, it could become a precise and reliable tool. An accurate power generation forecast is beneficial for both the Transmission System Operator (TSO) and power plant operators for them to get better control of their operations, which can potentially result in more efficient operations for both parties.

Sammendrag

At værforhold verken er konstante eller kontrollerbare er noe de fleste er kjent med. Effekten fra PV-teknologi er svært avhengig av været, og nøyaktige effektprognoser er derfor nødvendig dersom man ønsker kunnskap om generert effekt fra et PV-kraftverk. En evne til å kunne nøyaktig predikere den genererte effekten fra et PV-kraftverk er også sentralt for å sikre stabilitet i det elektriske kraftsystemet.

I denne oppgaven ble to metoder, metode 1 og 2, utviklet for å predikere effekten til et PV-kraftverk med 24 timers horisont og én times oppløsning. I utviklingen av metodene ble det benyttet tre år med måledata fra et storskala PV-kraftverk med installert effekt over 100 MW. Værmeldinger fra værvarslingstjenesten *Yr* og empiriske data generert med *Python*-biblioteket *pvlib* ble brukt. Disse dataene ble satt sammen og preprosessert med avviksdeteksjon, imputering av manglende verdier og min-maks-skalering. Etter dette ble prediksjonsmetodene utviklet. Disse ble sammenlignet med prediksjoner fra den kommersielle leverandøren av effektprediksjoner, Solargis.

Begge metodene i denne oppgaven brukte en indirekte fremgangsmåte hvor fremgangsmåten for å predikere global horisontal innstråling (GHI) var lik i begge metodene. GHI ble predikert med en maskinlæringsmetode ved bruk av "Random Forest Regression" (RFR). I denne modelen ble værmeldingsverdier omgivelsestemperatur, skyer, lave skyer, middels høye skyer og nedbør, sammen med lokale målinger av GHI fra forrige dag, og historiske målinger av omgivelsestemperatur og vindretning bruk som variabler. For å forutsi den genererte effekten brukte metode 1 den samme RFR-metoden som ble brukt til å predikere GHI. Her ble værmeldingsverdier for middels høye skyer, vindretning og luft fuktighet, sammen med irradians ved klar himmel, GHI-målingene fra forrige dag og prediksjonene for GHI brukt som variabler. Metode 2 brukte en rekke fysiske og empiriske operasjoner for å lage en prediksjon for den genererte effekten basert på GHI-prediksjonene.

Fra evalueringen av resultatene produserte metode 1 prediksjoner med høyest "skill score" på $0,200$. Solargis fikk en verdi på $0,122$ og metode 2 fikk verdien $0,004$. Generelt ble det observert at metode 1 hadde en tendens til å underestimere den genererte effekten, mens metode 2 og Solargis overestimerte den. Det ble generelt observert at metode 1 hadde problemer med å predikere i timene hvor effektgenereringen fra PV-kraftverket var på det høyeste. Deler av feilene i resultatene kan delvis være et resultat av "curtailment". På grunn av forhold i kraftnettet kan kraftgenereringen fra PV-kraftverket bli redusert med "curtailment". Når dette forekom ble den genererte effekten redusert mens værdataene som er brukt for å lage prediksjonene i denne oppgaven derimot ikke ble påvirket av dette.

Denne oppgaven konkluderte med at metode 1 er en brukbar metode for prediksjon av effekt fra PV-kraftverk og at den har en nøyaktighet som gjør at den er i stand til å konkurrere med kommersielle løsninger. Med mer tid til å forbedre metoden kan metoden bli et presist og pålitelig verktøy. Nøyaktige prediksjoner av effekt er gunstig for både operatøren av transmissjonssytemet (TSO) og kraftverksoperatørene for at de skal få bedre kontroll over driften, noe som potensielt kan resultere i mer effektiv drift for begge parter.

Nomenclature

AC Alternating Current

aFRR automatic Frequency Restoration Reserve

AM Air Mass Ratio

ANN Artificial Neural Network

DC Direct Current

DHI Direct Horizontal Irraiance

DSO Distribution System Operator

ECMWF European Centre for Medium-Range Weather Forecasts

EU European Union

FCR Frequency Containment Reserves

FFR Fast Frequency Reserve

GHI Global Horizontal Irradiance

GTI Global Tilted Irradiance

IEA International Energy Agency

IFE Institute for Energy Technology

MBE Mean Bias Error

mFRR manual Frequency Restoration Reserve

MPP Maximum Power Point

MPPT Maximum Power Point Tracker

MSE Mean Squared Error

NMBU Norwegian University of Life Sciences

NWP Numerical Weather Prediction

PAPE Peak Absolute Percentage Error

POA Plane of Array

PV photovoltaic

RFR Random Forest Regression

RMSE Root Mean Squared Error

STC Standard Testing Conditions

TSO Transmission System Operator

Contents

Chapter 1

Introduction

1.1 Motivation

150 years of human CO_2 emissions have resulted in a new set of challenges that needs to be solved by the population of the entire world [1, 2]. The emission of CO_2 and other greenhouse gasses has led to an increase in the greenhouse effect, causing global warming with generally higher temperatures of the earth's surface [2]. It is not only temperature rise in itself that is troubling, changes in local and global weather patterns in the earth's climate patterns are also becoming evident [2]. The changes in climate will eventually affect the entire world when problems with growing food, rising waters, and natural catastrophes become more frequent [1].

The rise of global temperatures started with CO_2 emission, but it can also be stopped through a reduction in CO_2 emissions according to the United Nations [2, 1]. As 84% of today's energy production stems from CO_2-intensive fossil fuels, a large part of the solution is to switch from fossil coal, oil, and gas, to cleaner renewable energy sources like wind-, solar-, and hydropower [3].

In addition to having to replace much of today's energy production, the world's total energy consumption is increasing, which means even more new renewable energy production is needed in the future [3]. On the positive side, renewable energy can be generated almost anywhere [4]. Since sun, wind, and water is available to some degree in most places around the world, there is no dependency on costly fuels from other countries [4]. The backside, however, is that natural resources are dependent on factors outside human control. Solar power needs sunlight, wind power needs wind and hydropower is dependent on enough water in the river or dam. Without knowing how power production is affected by these uncontrollable conditions, it is difficult to know how much power is being sent out on the electrical power grid.

The electrical power grid is an extremely complex system, with loads and generators varying in size, location, and operation [5]. For the power grid to function optimally for all its consumers and generators, it is important that there is a balance between generation and consumption [5]. This balance is carefully controlled by the transmission system operators [5]. With the increasing penetration of renewable energy in

the power grid, it is becoming increasingly important to be able to have accurate forecasts of the power generation from these renewable energy resources to report to the TSO [6].

1.2 Objective

Many publications have had good results in forecasting power generation and solar irradiance. Beanli et al., Babar et al., and El-Baz et al. all achieved good results with their method and beat their set reference model [7, 8, 9]. This thesis will be a part of this academic work to make accurate power forecasts for power plants using PV technology. For this thesis, a utility-scale PV power plant has been chosen with three full years of data. This site was chosen because of its varying cloud condition that makes power and irradiance forecasting challenging.

The objective of the thesis is to make 24-hour forecasts with a 1-hour resolution for the power generation at the chosen site. In this thesis, this will be done with an indirect approach in two steps, where the first step is to forecast the GHI with a machine learning model using RFR. In the second step, the power generation will be forecasted using the GHI from the first step.

For the forecast of power generation, two methods will be used and compared with each other. The first method uses the same approach as was used to forecast the solar irradiance with RFR. In the second model, the generated power is calculated from the forecasted GHI through physical and empirical methods in several steps. These methods will be compared with a forecast done by the commercial forecast provider Solargis and a smart persistence model, as well as some results from the literature. Several metrics are used to compare the forecasts, however, the skill is chosen as the main metric of evaluation.

Chapter 2

Theory

The concepts and theory presented in the following sections will give a good fundament before the method and results are presented later. First, some factors affecting solar irradiance on Earth are presented in Section 2.1, then the basics of power generation from PV technology are described in 2.2. Next, an introduction to the electrical power system is given in Section 2.3. In Section 2.4, machine learning and some important concepts surrounding the field of machine learning are presented before some preprocessing steps are presented in Section 2.5.

2.1 Solar irradiance

The Sun is the largest energy resource we have, every second the Sun emits $3.8 \cdot 10^{26}$J of energy, of which the Earth receives $1.7 \cdot 10^{18}$J [10]. To compare, the entire world consumed $6.35 \cdot 10^{18}$J in all of 2021 according to "Our World in Data" [3]. Even though these astonishingly high numbers, when estimating the solar energy that might be used for electric energy consumption, they might not be too accurate. The irradiance on ground level on Earth is dependent on several aspects, some of these will be explained in the following sections. The effects of the orbit and tilt of the Earth will be explained in Section 2.1.1, the effects of the atmosphere in Section 2.1.2, and a brief review of the effects of clouds will be given in Section 2.1.3.

2.1.1 Earth tilt and rotation

It is common knowledge today that the Earth both revolves around the Sun and its own axis. The Sun is the main source of energy on Earth, and the Earth's movements will have an effect on the energy the Earth receives [10]. Firstly, the orbit in which the Earth revolves around the Sun is slightly elliptical. However, this has such a small effect on the received energy that it can be neglected for the purpose of this thesis. The effect of the tilt of the Earth in relation to the Sun is a big contributor to giving the Earth seasons. The tilt of the Earth is 23.5°, which means that the Earth's rotation axis is deviating 23.5° from the plane in which the Earth orbits the Sun. This is shown with the diagonal red axis in Figure 2.1. This axis is fixed in one direction, and because of this, the Earth experiences the seasons. The tilt both affect the length of the day and the peak irradiance [10].

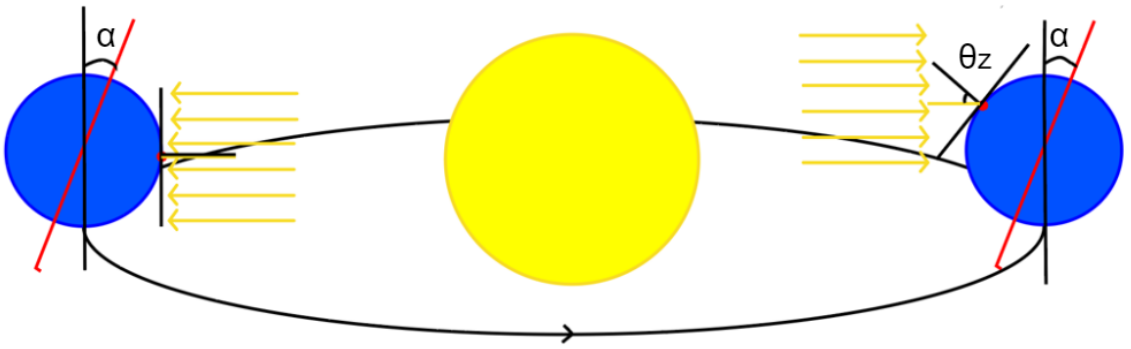

Figure 2.1: *The Figure shows Earth in two positions in relation to the sun and the tilt angle, α, of the Earth's axis (diagonal red lines) in relation to the vertical (vertical black lines). On the red point on the northern hemisphere in the right Earth position, it is winter, and the zenith angle between the solar irradiance and the zenith is θ_z. The left position shows the irradiance on the same point but in summer, here the zenith angle is 0. Figure adapted from Camacho [11].*

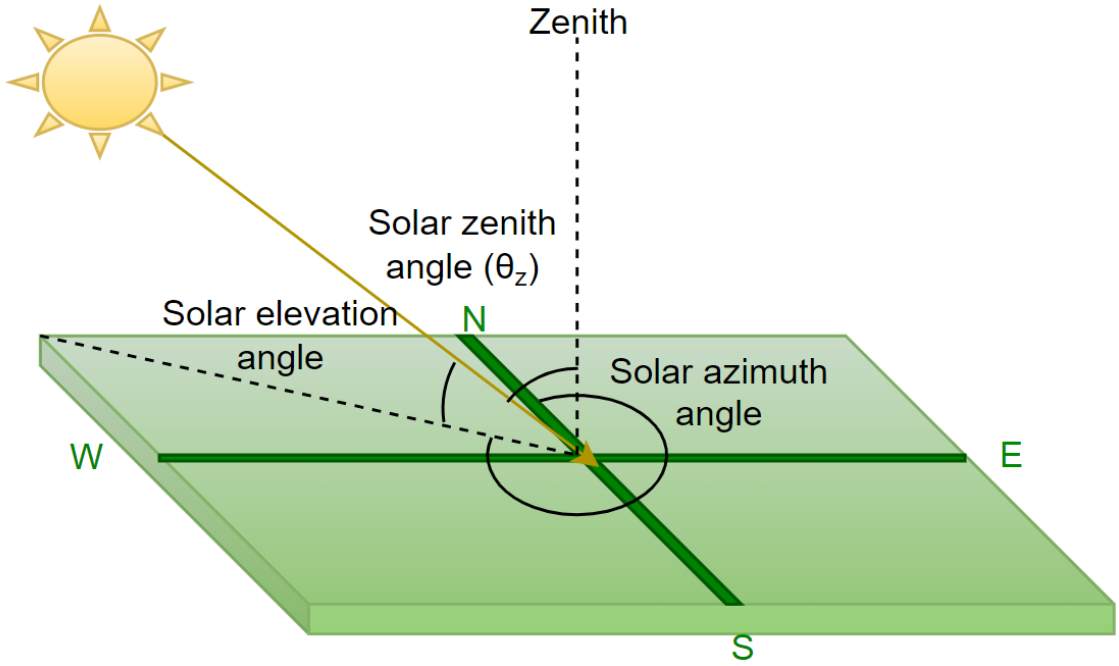

Figure 2.2: *Representation of the relevant angles used. The solar zenith angle is the angle between the solar ray and the zenith. The solar elevation angle is the angle between the solar ray and the plane. Lastly, the solar azimuth angle is the angle of orientation from North in a clockwise direction, i.e. 0° is north, 90° is east, 180° is south and 270° is west.*

The reason for the larger peak irradiance is also evident from Figure 2.1. The peak irradiance is dependent on the zenith angle and the solar constant [11]. The zenith angle (θ_z) is the angle between the Sun's rays and the zenith, e.i. the normal of the plane. This is shown in Figure 2.2, together with the solar azimuth and solar elevation angle. The solar constant (G) is the total solar irradiance on a plane perpendicular to the Sun rays outside the atmosphere with an average distance between the Sun and Earth of 149.6 $\cdot 10^9$ m. The value varies slightly, but a constant of 1361 $^{W}/_{m^2}$ is often used.

The irradiance (P) outside the atmosphere can be estimated with the formula

$$P = G \cos \theta_z \tag{2.1}$$

where G is the solar constant and θ_z is the solar zenith angle. This is the maximum irradiance possible to obtain on a horizontal plane on the Earth's surface with the same solar zenith angle. From Figure 2.1, the right earth experiences winter in the northern hemisphere. There, the solar zenith angle is large thus making the irradiance in Equation 1 small. On the left Earth, there is summer at that point and the solar zenith angle is 0, the irradiance outside the atmosphere is, therefore, equal to the solar constant [11].

2.1.2 The atmosphere

The rotation and tilt of the Earth decide the seasons and the irradiance on the Earth [12]. However, the Earth is covered by an atmosphere that also interacts with the radiation from the Sun, and is a substantial obstacle the Sun's radiation meets on its way to the surface of the Earth. A lot of the radiation is either absorbed or reflected here, as can be seen from Figure 2.3. The first, smooth, spectrum shows the black body irradiance of a body with a temperature of 5900 K. The Sun's mean surface temperature is approximately 6000 K [10]. The next spectrum represents the radiation at the top of the atmosphere, and the lower spectrum is the radiation that reaches Earth. Some atmospheric gasses like H_2O, CO_2, O_2 and O_3 absorb certain wavelengths causing the dips in the lower spectrum in Figure 2.3 [12].

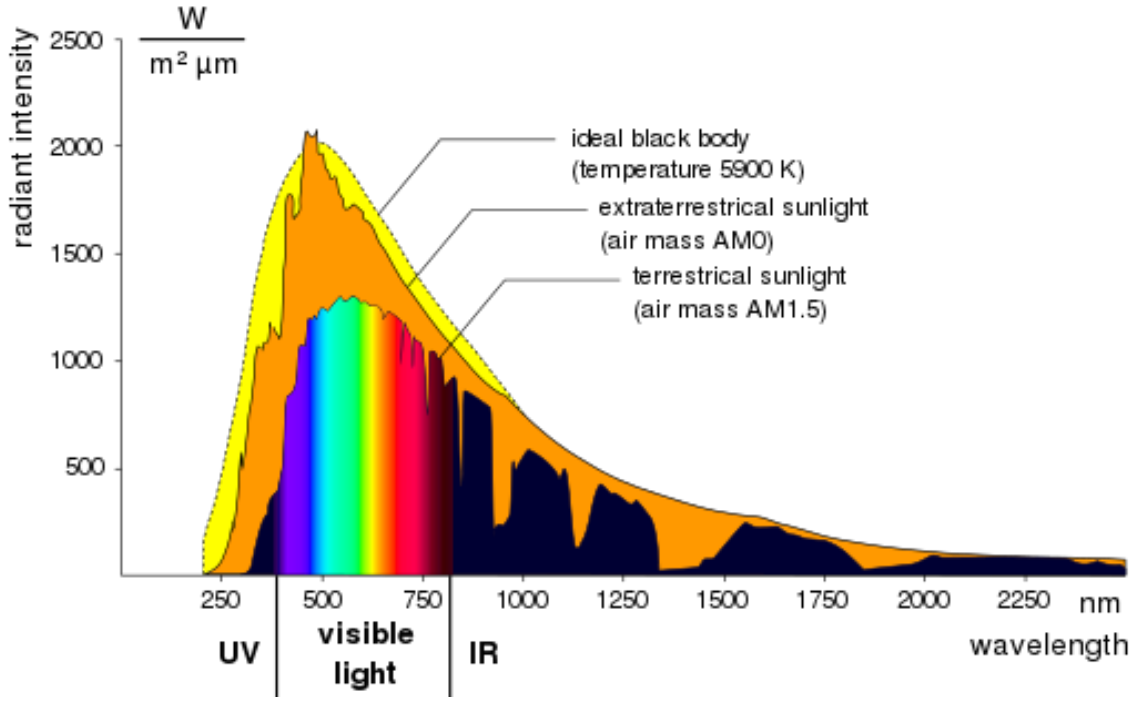

Figure 2.3: *Solar spectrum at ground level (AM1.5), outside the atmosphere (AM0) and the black body radiation of a mass with a temperature of 5900 K. Figure by Degreen under CC BY-SA 2.0 DE license [13].*

Solar radiation is also subject to scattering on its way through the atmosphere [14]. This happens when an interaction with a particle in the atmosphere causes solar radiation to change direction. Rayleigh and Mie scattering are two kinds of

scattering that often are used to describe the interaction. Rayleigh scattering occurs when particles the radiation are small, i.e. when interacting with the atmospheric gases N_2, O_2, and CO_2, and will scatter shorter wavelengths much more than longer wavelengths. That is, blue light is much more likely to be scattered than red light [14]. Mie scattering does not differentiate between the wavelengths like Rayleigh scattering does [15]. This form of scattering occurs with larger particles like water droplets, and scatter all wavelengths equally [15].

The intensity of radiation of the wavelengths that reach Earth in Figure 2.3 is for a given Air Mass Ratio (AM) of 1.5. AM is a measure of how much atmosphere the radiation is passing through [10]. At the top of the atmosphere, the radiation has not yet passed through any atmosphere and has a AM0 spectrum as shown in the middle orange plot in Figure 2.3. The irradiance at AM0 will be equal to the solar constant of 1361 $^{W}/_{m^2}$, this means that the integral over all the wavelengths of the middle orange curve in Figure 2.3 is equal to the solar constant. On the surface of the Earth, the AM will vary throughout the day, depending on the solar zenith angle(θ_z), and is given by the equation

$$AM = \frac{1}{\cos(\theta_z)} \tag{2.2}$$

where θ_z is the solar zenith angle [10]. This formula states that the AM will be lowest under clear sky conditions when the solar zenith angle is 0°, e.i., when the Sun is directly overhead [10].

The radiation that is scattered but reaches the ground, is called diffuse radiation [10]. Together with the direct unscattered radiation from the Sun, this forms the GHI. This is the total irradiance on a horizontal surface on Earth [10].

2.1.3 Clouds

Many of the effects listed above are somewhat possible to calculate just by knowing the local time and location on Earth. The weather, on the other hand, is much more unreliable, but will also affect the irradiance on the surface greatly. More specifically, the weather phenomena of clouds can have a large impact on solar radiation [12].

Clouds have an average albedo of 0.6, e.i. 60% of the incident irradiance is reflected back to space [12]. The cloud albedo will vary with varying cloud thickness, where thin clouds will reflect less of the irradiance than thick clouds. The darker the underside of the cloud is, the more of the Sun's irradiance is reflected by the cloud [12]. The irradiance that is not reflected is scattered with Mie scattering by the cloud droplets, thus making the cloud look white [15].

It is normal to separate the clouds depending on how high in the atmosphere the base of the clouds are [12]. For the mid-latitude regions, low clouds are typically below 2000 m, medium-height clouds are between 2000 m and 5000 m, and high clouds are above 5000 meters. Some clouds, like larger thunderclouds, will cover all heights. Depending on where the clouds are located, they will also have different properties, Figure 2.4 show how clouds at different heights can look [12].

The clouds in the highest region in Figure 2.4 are thin and look white or sometimes transparent [12]. These clouds form in a part of the atmosphere where the air is relatively cold and dry and they consist mainly of ice crystals. The medium-height clouds are white or gray and consist of water droplets, however, when the temperature is low, there can also be some ice crystals. Precipitation can form in the altostratus clouds from Figure 2.4 if they are sufficiently thick. When this happens the base of the cloud usually lowers and if the precipitation reaches all the way down to the ground, the cloud is reclassified as a nimbostratus cloud in the lower region. The clouds in the lower region are almost always consisting of droplets, but also these clouds can consist of ice crystals if it is cold enough. The precipitation that reaches the ground is mainly from these clouds which can range from white to dark grey in color depending on cloud type and thickness. Some clouds can also cover all the height regions as cumulonimbus in Figure 2.4. These are dark thunderclouds that can cause heavy precipitation [12].

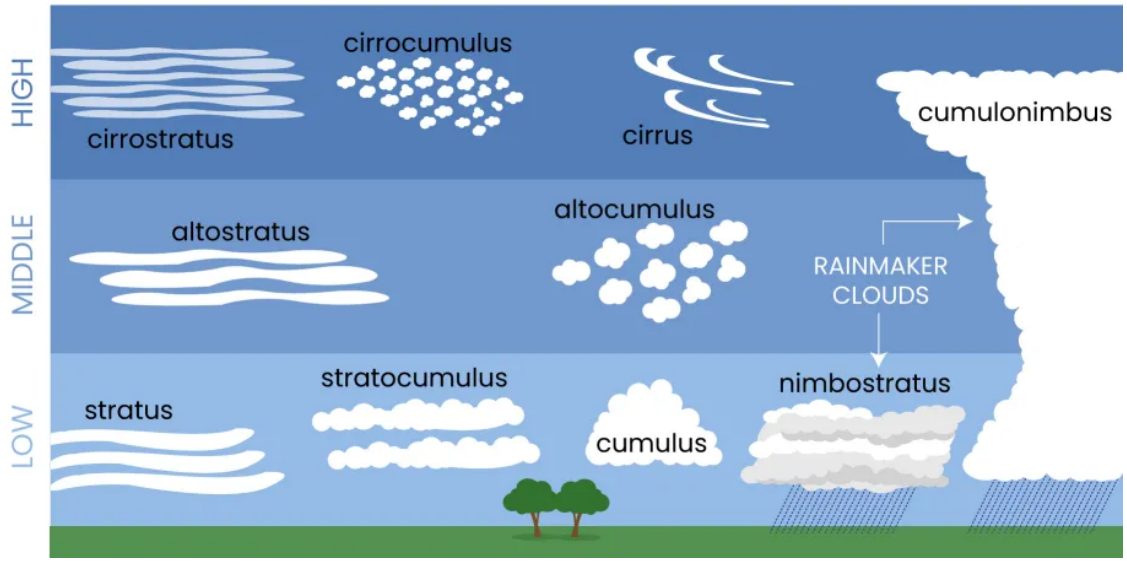

Figure 2.4: *Clouds at low, medium, and high altitudes in the atmosphere. Figure from UCAR - center for science education by L.S. Gardiner under CC Attribution 4.0 International license [16]*

Different cloud types will affect the GHI, however, so will the time of day and location on the globe. Only knowing the irradiance is therefore not enough to describe how much the clouds or other atmospheric effects affect the irradiance. The clearness index (K_t) is a measure of global horizontal transmittance through the atmosphere and is given by

$$K_t = \frac{GHI}{G\cos(\theta_z)} \tag{2.3}$$

where G is the solar constant outside the atmosphere and θ_z is the solar zenith angle [17]. Since the equation is dependent on the solar zenith angle, the clearness index will be relatively independent of the time of day, thus making it possible to compare cloud conditions throughout a day based on this index. A clearness index of 0 indicates no atmospheric transmittance, whereas 1 represents 100% transmittance [17].

2.2 Solar energy generation with photovoltaic (PV)

Today, solar energy is the fastest-growing source of energy in the European Union (EU) [18]. PV technology is a way of directly utilizing solar energy as electric energy, and with a price decrease for PV power from 5.55 $/w in the year 2000 to 0.27 $/w in 2021, it is a highly competitive source of energy in the EU [18, 19]. In 2022, 6.2% of the world's electricity demand was covered by power from PV power plants according to International Energy Agency (IEA) [20]. This might seem small, but in 2011, solar energy accounted for 0.3% of the total electricity generation [3].

It is evident that solar energy and more specifically, PV is an important part of future energy generation. It is therefore important to understand the technology and infrastructure that is necessary. In the following section, the fundamental physics in PV technology is presented in Section 2.2.1, basic concepts of the PV module will be introduced in Section 2.2.2, the function of solar tracking will be explained in Section 2.2.3, and finally inverters as a connection to the power grid in Section 2.2.4. The theory in this Section is based on the textbook "Solar energy - The physics and engineering of photovoltaic conversion, technologies and systems" by Smets et al. unless stated otherwise [10].

2.2.1 The fundaments of PV thecnoology

The fundaments of PV technology date back to Alexandre-Edmund Becquerel's discovery of the photovoltaic effect in 1839. The photovoltaic effect is described as a generation of an electric field at the junction between two materials when they are exposed to light. This happens because the energy in the light excites electrons in the illuminated material from the valence band up to the conductive band. The free electrons will move to the other material, thus making a negative charge in that material and a positive charge in the illuminated material which causes an electrical field [21]. In solar cells, these two materials are two layers of semiconductors where one layer is doped with an element with more valence electrons than the semiconductor, and the other layer is doped with an element with fewer electrons. When this junction, called a p-n junction, is made into a closed circuit, it can generate electrical energy when illuminated with solar radiation.

Solar radiation consists of energy quanta called photons, with energy (E) given by

$$E = h\frac{c}{\lambda} \tag{2.4}$$

where h is the Planck constant, c is the speed of light and λ is the wavelength of the photon. From this equation, one can see that the lower the wavelength, the higher the energy of the photon is.

The energy difference between the valence band and the conductive band is called the bandgap, and the electrons need that exact energy to be exited. If there is too little energy, the electron will not be exited, if there is too much, the difference between the received energy and the bandgap will be deposited in the material as thermal energy. For crystalline silicon, the most common semiconductor material used in PV technology, the bandgap is 1.12 eV which, through equation 2.4 translates to a wavelength of 1107 nm.

From Figure 2.3 one can see that all the visible light has a smaller wavelength than 1100 nm. The energy from photons with a wavelength larger than 1100 nm is lost, as well as the difference in energy between the bandgap energy and photons with a wavelength smaller than 1100 nm. These are the two major loss mechanisms taken into account when calculating the Shockley-Queisser limit which is a theoretical upper efficiency limit for single junction solar cells. For crystalline silicon, the Shockley-Queisser limit is not directly applicable because of its indirect bandgap (which this thesis will explore in further detail), however, a limit of 29.43% has been derived by Richer et al [22].

2.2.2 The PV module

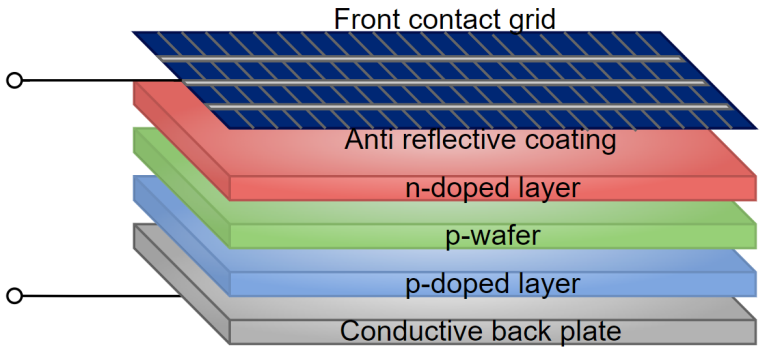

Figure 2.5: *Example of how the layers in a solar cell can be arranged and connected to an external circuit [10].*

One semiconductor junction makes the power-generating component of a solar cell. Figure 2.5 shows the layers of a crystalline silicon solar cell, which is the most used semiconductor material in PV modules. The middle portions of the Figure represent the p-n junction. When sunlight is shining on this junction, the photovoltaic effect sets up an electric field making the electrons move into the front contact grid, through the circuit, and back through the back contact. The amount of current flowing will depend on the intensity of the irradiance. PV modules are typically capable of converting between 16% and 24% of the irradiance to electric power [23]. The Figure also shows an anti-reflective coating, this is to reduce the reflection from the cell, thus increasing the amount of absorbed energy by the solar cell. Note that the current will only move in one direction through the solar cell, thus producing Direct Current (DC).

The relation between the current (I) and voltage (V) is shown as the full red curve in Figure 2.6. The power (P) is related to the voltage through the equation

$$P = IV. \tag{2.5}$$

The Maximum Power Point (MPP) is marked with a circle, the product of the current and voltage at this point generates the highest power. These curves are often made with values derived from the solar cell under Standard Testing Conditions (STC) when the irradiance on the solar cell is 1000 $^{W}/_{m^2}$ with an AM1.5 spectrum and a cell temperature of 25°C. However, under normal operation, these conditions will not be met most of the time because of the Sun's movement and varying weather.

Lower irradiance will cause the maximum current in the solar cell to fall, thus decreasing power generation as seen from the dotted green line in Figure 2.6. In Figure 2.6, the curve for increased cell temperature is also shown with a dashed blue line. Increased temperature causes the maximum voltage to fall and the maximum current to increase slightly which leads to an overall fall in maximum power. It can therefore be beneficial to use PV technology in areas with some wind and low temperatures to cool down the solar cells.

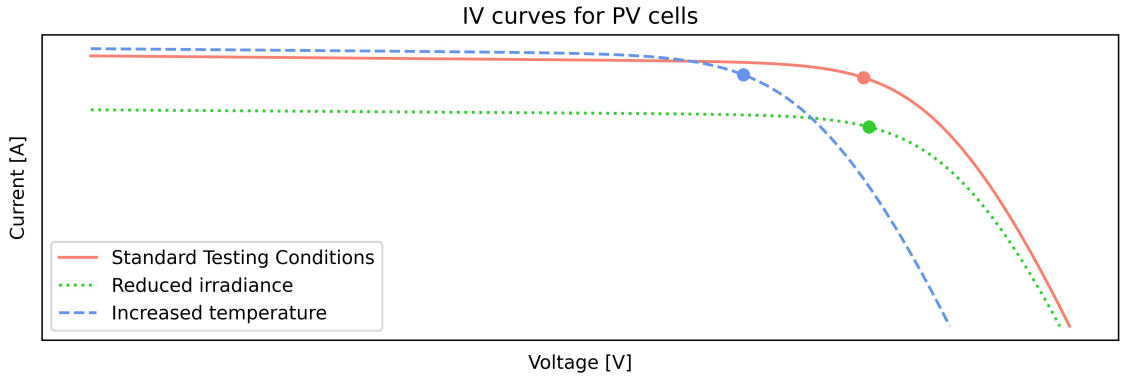

Figure 2.6: *Relationship between current and voltage from a PV cell with STC, reduced irradiance, and increased temperature. The MPP for each curve is marked with a circle. Figure plotted with functions and data from pvlib [24]*

When used in a module, many solar cells are connected in series as in Figure 2.7. At the top of the Figure, there are three bypass diodes, one for each of the three substrings in the series. These are in place to reduce the effects of shading of single strings or cells. When a single cell in a substring does not receive solar energy, there will be no energy generation in the cell. As the solar cells are connected in series, the current in the rest of the string drops to the same level. The bypass diode ensures that the shading of one solar cell will only affect one string and not the entire module, thus reducing generation loss.

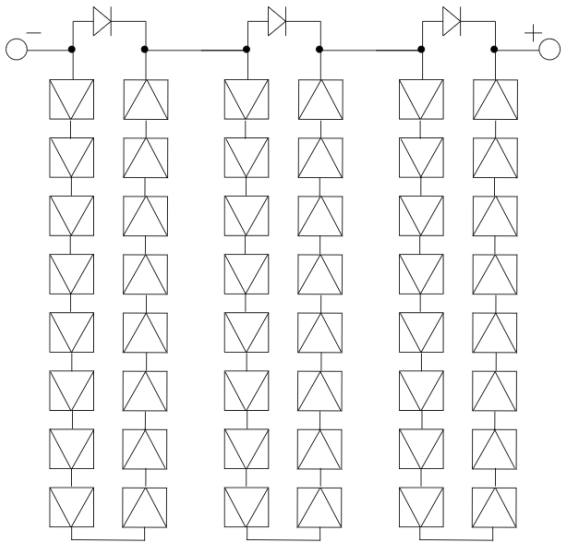

Figure 2.7: *Solar module with three substrings with solar cells connected in series with bypass diodes connected in parallel over each substring [10].*

2.2.3 Tracking

As noted earlier, the Sun's place in the sky varies, both with the time of day and seasons of the year. Figure 2.1 and Equation 2.1 also demonstrate how a larger angle of incidence gives a lower irradiance on a surface. Looking at a horizontal surface on Earth, the surface will get a high irradiance in the middle of the day when the solar zenith angle is at its lowest, i.e. when the Sun is at its peak position in the sky, and a lower irradiance in the mornings and evenings when the solar zenith angle is larger.

A way of minimizing the angle of incidence on the PV modules' surfaces is by using a mounting system with tracking for the modules [11]. This can be done with single or dual-axis tracking, where dual-axis tracking regulates both the module's azimuth angle and the module's tilt angle, and the single-axis only regulates the tilt angle. The tilt angle is the angle of the module plane in relation to the horizontal plane. This kind of modification to the mounting system can greatly enhance generation in the morning and evening, thus increasing the overall efficiency of the power plant and making it possible to generate more energy[11].

When the modules are placed in several rows, there is a possibility of module-on-module shading in the mornings and evenings [25]. This is solved by backtracking, which decreases the tilt angle at times with low solar elevation angles Even though this makes the irradiance on the modules lower, this is often less than the losses one would have had from partial shading of a module. Figure 2.8 demonstrates how the tracker uses backtracking as well as maximum tilt angle to make sure there is no module-on-module shadowing [25]. The dotted green line shows how backtracking makes the tilt angle gradually increase up to the maximum tilt angle in the morning, compared to how the full, red line without any regulation starts the day with a high tilt angle. The dashed blue line show how tracking looks with only a maximum tracking angle without backtracking.

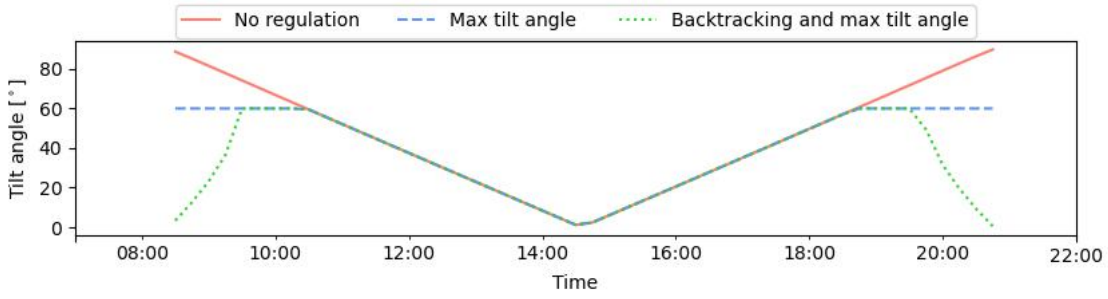

Figure 2.8: *Tracking angle for solar modules with only single-axis tracking, tracking and maximum tilt angle, and tracking, maximum tilt angle, and backtracking. The plot is generated with pvlib [26].*

2.2.4 Inverting the power

Since the energy generated by the solar power plant is in DC, the energy needs to be converted to Alternating Current (AC) before it is transferred to the power grid. This is done with an inverter. The inverter is equipped with an Maximum Power Point Tracker (MPPT) that will adjust the current drawn from the solar modules so that it operates at the MPP. The relation between current and voltage can be seen

in Figure 2.6. This relationship will depend on the power the modules receive from the Sun, therefore the MPPT must adapt throughout the day.

There are several system architectures to choose from when implementing an inverter to a PV system. Three possible architectures are central, string, and micro-inverter. With a micro-inverter, one or several modules are connected to one inverter. In a system with string inverters, the modules are connected in series to form strings. Each string is connected to a string inverter. When a central inverter is used, there is one large inverter for several parallel connected strings with PV modules. Which architecture is used, depends on the park's design as well as costs. Generally, the more inverters are used, the less one will be affected by shading or loss of generation in some parts of the architecture. Since each inverter has its own MPPT, only the inverter with the shadowed modules will experience a loss in power. However, since inverters can be expensive, this tradeoff must be considered for each architecture.

2.3 The electrical power system

The electrical power system is a vast network of generators, loads, and power lines often spanning over several countries [5]. The system is traditionally built up roughly as shown in Figure 2.9. The power from generation units is transformed up to a high voltage to be transmitted over large areas on the transmission system to the distribution system [5, 27]. In the distribution system power is distributed from the transmission system to consumers like industries, residential housing, and institutions on a lower voltage [27]. The distribution system is controlled by a local Distribution System Operator (DSO) and the transmission system by a regional or national TSO [28, 29]. In this thesis, the main focus is on the TSO which is responsible for transporting the generated power from the power plants to the distribution system as well as maintaining the balance between power generation and consumption in the transmission system [29]. In this section, the importance of power balance will be introduced in Section 2.3.1 followed by the workings of the power market in Section 2.3.2, Section 2.3.3 introduces the solution the balancing market provides with curtailment as an example in Section 2.3.4.

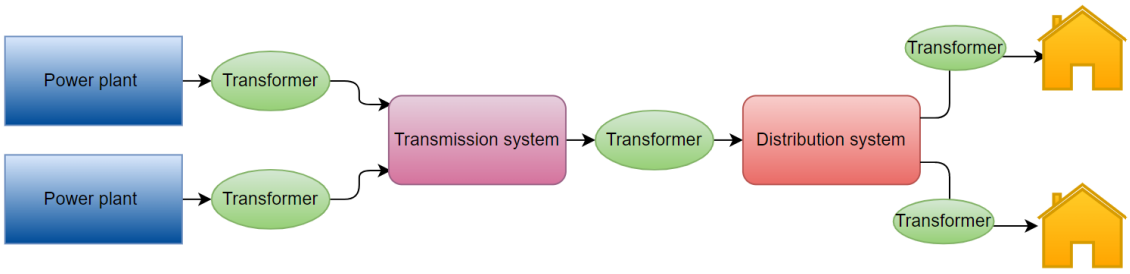

Figure 2.9: *Simple schematic representation of the electrical power system [27].*

2.3.1 Power balance

When discussing power systems, the balance between generation and consumption is very important [30]. The power grid connecting the loads and generators in the power system does not have any storage capacity. This means that the consumption (and losses) must be equal to the generation at any given time [5].

However, the power system consists of many rotating generators in i.e. hydro, nuclear, and gas power plants [5]. These provide the system with some degree of slack because of their rotating property, and they rotate with the same synchronous frequency of 50 Hz (in Europe). When the system load exceeds the system generation, energy from the rotational kinetic energy of the generators will be used to balance out this difference. This leads to the generators losing rotational energy, thus slowing down the frequency. The opposite happens when the generation exceeds the loads, leading to an increase in frequency [5].

The frequency of the power grid at normal conditions fluctuates between 49.9 and 50.1Hz without it causing any harm to the system components [31]. However, larger deviations could lead to a potential breakdown of the grid. If the frequency drop for a generator is sufficient, this will eventually lead to the generator powering down, thus leading to a further drop in frequency. This kind of event could lead to the entire system collapsing and cause a power outage [31].

2.3.2 The power marked

The power generation is scheduled every day to match up with the anticipated load as well as possible [5]. The scheduling is on a daily and hourly basis and tries to choose the mix of generation units that gives the lowest cost. Power generation units can be split into three categories based on their role in the generation mix; baseload, load-following, and peaking units. The baseload units are the cheapest power generators that can give a constant and steady flow of power, like coal or nuclear power plants. The next unit follows the load, and when demand increases, the generation unit can ramp up generation to follow, like in a hydroelectric power plant. Finally, the peak units are used for demand peaks and are relatively expensive. For these cases, gas turbines can be used [5]. This is the traditional generation mix, however, with a higher penetration of renewable energy like solar and wind, a fourth unit is added to the mix; must-take units [27]. These units only generate when the weather conditions are right and can not be stored easily and must therefore be deployed when available [27].

The scheduling is based on the prices of the power market [32]. The participants in the power market are the power generators, brokers, energy companies, large industrial customers, and power suppliers trading on behalf of small and medium size consumers and industries. Bids are given on generation and the prices are set based on demand in three organized markets. These are the day-ahead, intra-day, and balancing markets. On the day-ahead market, power for each hour the following day is traded. In the Nordic market Nord Pool, all trading must be done between 0800 and 1200. On the day of generation, from the time of clearing the day-ahead market until one hour before the hour of operation, trading can be done at the intra-day market. This is to account for changes in anticipated load and generation due to for example changes in the weather forecast which can affect the power generation

from renewable energy sources. The balancing markets are used within the hour of generation and will be described in more detail in the following section [32].

2.3.3 The balancing market

To prevent frequency-caused power outages to happen, the TSO uses the balancing market [31]. The balancing market is a marketplace that enables both generators and large consumers to get paid by the TSO to alter their generation or consumption. The types of balancing products can be categorized by their response time and duration, i.e., how quickly they can be connected and how long they can be connected. In Norway these categories as defined by the Norwegian TSO, Statnett, are [31]:

- Fast Frequency Reserve (FFR)
  - 0.7-1.3 seconds response time; 5-30 seconds duration.
  - Slows down the change in frequency.
- Primary reserve - Frequency Containment Reserves (FCR)
  - 30 seconds response time; minimum 15 minutes duration.
  - Stops the change in frequency and stabilizes the frequency at a new level.
- Secondary reserve - automatic Frequency Restoration Reserve (aFRR)
  - Full response within 2 minutes; duration as long as the bid period lasts.
  - Brings the frequency back to the normal frequencies (49,9-50,1 Hz).
- Tertiary reserve - manual Frequency Restoration Reserve (mFRR)
  - Full response within 12.5 minutes; duration as long as the bid period lasts.
  - Releases aFRR and maintains balance until a new balance is reached in the energy market.

2.3.4 Curtailment

One way of regulating power generation is by using curtailment [33]. Curtailment can be used both on power consumption and generation to achieve balance in the system by reducing consumption or generation depending on whether the grid frequency is too low or too high. It is however much more common to curtail power generation than consumption. In particular, renewable energy sources are often the subject of curtailment. The curtailment can be used to solve congestion problems caused by bottlenecks in the grid or to control the frequency in cases of overgeneration in the system [33].

2.4 Machine learning

The following Section about machine learning is mainly based on the textbook "Python Machine Learning" by Raschka unless otherwise specified [34]. Machine learning is a way humans have made computers learn connections and patterns in data through different algorithms in a similar fashion as humans learn. The algorithm runs many times with increasing accuracy [35]. As humans, machine learning is able

to extract knowledge from data. Through machine learning the world has gotten technologies like reliable spam filters, voice recognition software, search engines as well as cancer detection software. There are endless possibilities yet unexplored for the application of machine learning.

There are many different kinds of machine learning. One common way of grouping different algorithms is supervised, unsupervised, or reinforced learning. When the algorithm is supervised, it is trained with labeled data. That means that it is trained towards a specific predetermined solution and evaluated on how well it performs compared to this solution. In unsupervised machine learning the algorithm is trained with unlabeled data. This is often used to find hidden connections within the data. Lastly, reinforcement learning is implemented by rewarding desired behavior, thus making the model learn the rules itself. In this section, the focus will be on supervised learning since the GHI and power generation that are the targets of the predictions are known in the data set that is used. First, an introduction to some basic machine learning models will be given in sections 1 and 2, before moving on to the ensemble model in Section 3, and at the end one of the most common problems in machine learning, the problem with over and under fitting will be explained in Section 4.

2.4.1 Artificial Neural Network (ANN)

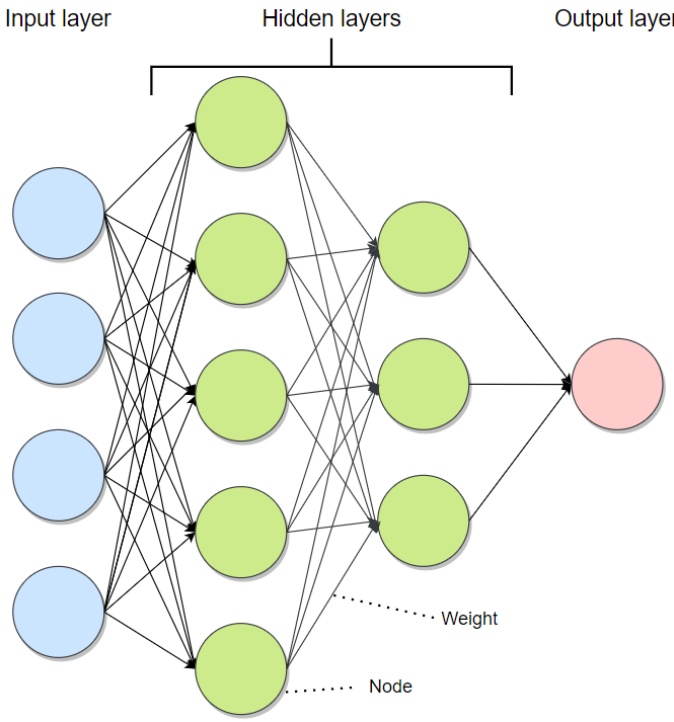

Figure 2.10: *Schematical representation of an ANN network with an input layer, two hidden layers, and an output layer. Nodes are represented by circles, the arrows between show how the output from one node is multiplied with a weight before entering the next node [34].*

One subfield of machine learning is ANN. As the name implies, these algorithms mimic the way neurons signal each other in the brain's neural network. An ANN is built up of layers of nodes, one input layer, one or more hidden layers, and one output layer. The nodes are connected to the other nodes in the previous and next

layers as shown in Figure 2.10. The number of nodes in the input layer is the number of features in the dataset. If the problem is a classification problem, the number of output nodes is the number of classes that are to be classified, and if it is a regression problem as in Figure 2.10, there is only one output node that can take on continuous values.

All the nodes from one layer are multiplied with weights before entering the next node, the arrows in Figure 2.10 show where the weights are added. In the nodes in the hidden layers and output layer, the node values multiplied with the weights are summarized and sent through an activation function. The weights decide how much to emphasize the information from the previous nodes, and the activation function adds nonlinearity to the result by sending the result through a nonlinear function like a logarithm, hyperbolic tangent function, or other functions. With each iteration, the weights are updated based on a given optimization criteria so that the output is as close as possible to the true value.

There are many kinds of ANN models from small single-layer networks to large ANN networks that have intricate connections. However, no matter what kind of model, hyperparameters are used to make the model generate the best possible results. For ANNs hyperparameters decide things like how much to update the weights after each iteration, the regularization (this will be explained in Section 2.4.4), the type of activation function, and much more.

2.4.2 Decision tree

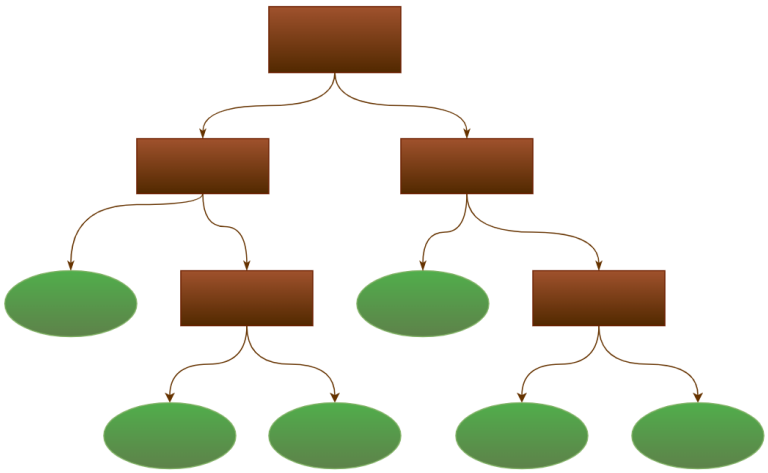

Figure 2.11: *Illustration of a decision tree. The brown boxes represent internal nodes with thresholds or questions, and the green circles represent leaf nodes [34].*

Another learning algorithm is the decision tree. The decision tree is a supervised learning algorithm that, instead of weights, uses questions or threshold nodes to separate the data into smaller and smaller nodes. As the name implies, this model architecture resembles that of a tree. Figure 2.11 shows how the data start at the stem of the tree and is divided by decisions made in the internal nodes of the tree. The internal nodes decide where the data go by a question or threshold in the node, these are the brown squares in Figure 2.11. If there are no more nodes coming out

of a node, it is called a leaf node as seen as the green ovals in Figure 2.11, and the structure now looks like an upside-down tree. Since this tree can grow very deep with a large training set, it is normal to prune it. This is done by altering the hyperparameters that decide the depth of the tree, the deeper the tree, the more complex the algorithm is.

Decision trees can be used for both classification and regression and will, therefore, have different measures of success depending on the purpose. The regression model uses the within-node variance, which looks at how much the data deviates from the mean of the data in each node. The sequence of splits that minimize this variance will be chosen as the best model.

2.4.3 Ensemble learning

Decision trees are so-called weak learners. That means simple models that often are only slightly better than random guessing. A commonly used practice is to use an ensemble of weak learners to create a strong learner. A random forest is an example of an ensemble model using bagging. The model uses the mean of an ensemble of different deep decision trees. In each decision tree in the ensemble, a random set of data is selected from the full data to train the tree, and at each node in the tree, a random set of features are selected to make the split. Since the model uses the mean of many trees it is not much affected by noise from one single tree, therefore, it is often not necessary to prune the trees. This means that the hyperparameters that prune the decision trees are less important. One hyperparameter that is important for RFR is the number of trees in the random forest. More trees often lead to better performance and more computation time.

2.4.4 Over-and underfitting

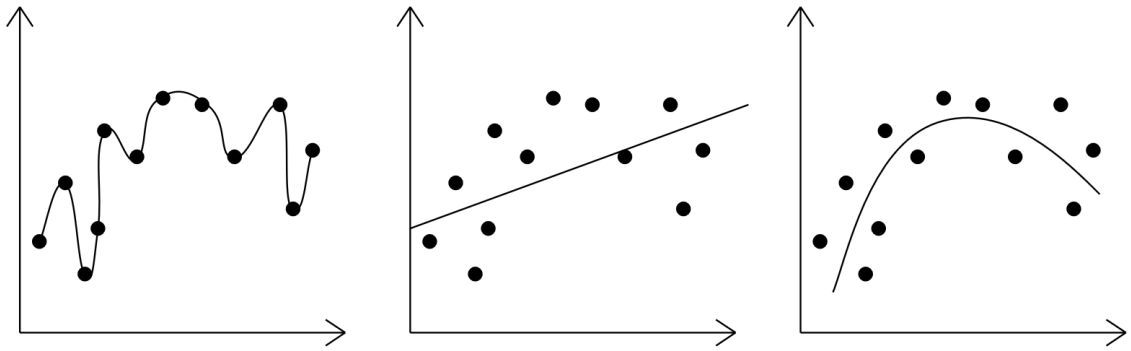

Figure 2.12: *Graphs showing overfitting, underfitting, and a good tradeoff to the plotted points [34].*

A common problem with designing machine learning models is that they might adapt too much from the training data. By doing so, the algorithm will predict or classify perfectly or really well within the training data, but it loses its ability to adapt to new unseen data, this is called overfitting. In order to manage this, regularization is used. This is a method to generalize the algorithm so that it is able to adapt to new data. However, one must be careful to not regularize too much, so that it will lose its ability to find patterns in both the training and the test data. The trade-off between overfitting and underfitting is visualized in Figure 2.12 where the left graph

shows overfitting to the plotted points, the center shows underfitting and the right plot shows a good tradeoff between over and underfitting to the points.

2.5 Preprosessing

No matter what one does in the model-building step, the result will always depend on the data that is put into the model [34]. In the following section, a few important aspects of the prepossessing of the data will be explained starting with missing data in Section 2.5.1, outliers in Section 2.5.2, the use of feature selection in Section 2.5.3, and finally, some evaluation metrics in Section 2.5.4.

2.5.1 Missing data

Missing data is a normal problem in data analysis [34]. There can be different sources for the missing data, whether it is equipment error, computer error, or human error. As with the sources of the missing data, there are also several different ways to deal with them. One method is to delete the row or column with the missing values. This is a simple technique that does not add estimates to the data, however, it can lead to large amounts of missing data and loss of valuable information. Another technique is imputation, this can potentially give a complete dataset which can be beneficial for several machine learning techniques. Some commonly used imputation methods are mean imputation, rolling mean, interpolation, modeled data, and using data from other sources [34][36]. Imputing missing data can lead to a bias in the data and change in variance and should only be used if necessary [37]

2.5.2 Outliers

Another normal problem one can face when investigating the data is outliers. An outlier is a data point that deviates significantly from its expected behavior [38]. Replacing or removing abnormal values without understanding the reason for their abnormality might remove important information from the data. Therefore, it is important to understand why these values deviate from the rest of the data. If it is reasonable that the deviation is because of errors in the data, it is considered safe to take action against the outliers [38].

Handling outliers can be done by replacing the outliers with missing values or a value that follow the data more closely [38]. If one chooses to replace an outlier with a value, the missing data methods in the previous section can be applied.

2.5.3 Feature selection

With a large number of features, it can sometimes be beneficial to extract the most relevant ones [39]. This can reduce both noise in the data and the computational time of the machine learning model [39]. This can be done with a variety of different feature selection algorithms that use different techniques [40].

One simple feature selection method is using the Pearson correlation matrix [41]. The correlation is a measure of the linear relationship between two features where -1 is a total negative correlation, 1 i is a total positive correlation and 0 is no correlation

at all. When using this method, a cutoff value is set so that the features with a correlation to the target with an absolute value lower than the cutoff, are not seen as a relevant feature [41].

Another method is the Boruta method [39]. Boruta is a RFR based algorithm that removes features that are proven to be less relevant than random noise [39]. This is done by replacing one of the features with a shadow attribute containing the shuffled data for the feature and then doing a test to see if the predictive capabilities of the RFR model change. If it does not change, the feature does not give any relevant information and is therefore removed. This is done for all the features in the data [39].

2.5.4 Evaluation metrics

There are many metrics for evaluating the accuracy of a model. In table 2.1 five metrics will be explained.

Table 2.1: *Evaluation criteria with their symbols, formulas, and explanations.*

Symbol	Formula	Explanation
MBE	$\frac{1}{n}\sum_{i=0}^{n} y_i - y_i'$	**Mean bias error** gives an estimate on the average bias of the model, positive MBE means the model on average underestimates the truth, negative values mean the model overestimate on average [42]. It is not usually a measure of fit since large over and underestimating can cancel each other out and also give the desired 0 value. MBE is in the same unit as the results. In the equation, y_i is the truth, y_i' is the model formula, and n is the number of samples [42].
MSE	$\frac{1}{n}\sum_{i=0}^{n}(y_i - y_i')^2$	**Mean squared error** is the mean of the squared difference between the model estimate and truth [42]. Because of the squaring of the errors, large errors are penalized more and small errors less. In the formula, y_i is the truth, y_i' is the model estimate, and n is the number of samples [42].
RMSE	$\frac{1}{n}\sum_{i=0}^{n}\sqrt{(y_i - y_i')^2}$	**Root mean squared error** is the squared rood of Mean Squared Error (MSE), which leaves this metrics with the same qualities as the Mean Squared Error (MSE) [42]. Because of the squared root, it penalizes large errors less. This metric will be in the same unit as the results. In the formula, y_i is the truth, y_i' is the model estimate, and n is the number of samples [42].
PAPE	$\frac{100\%}{N}\sum_{i=0}^{N}\sqrt{(y_i - y_i')^2}$	**Peak absolute percentage error** is a measure of how well the peak value in each period is estimated. The output is the percentage difference between the model's peak value and the true peak value [43]. This metric is scale independent. In the formula, y_i is the period peak of the truth, y_i' is the period truth of the model estimate, and N is the number of periods [43].
SS	$1 - \frac{RMSE_{forecast}}{RMSE_{reference}}$	**Skill score** is a measure of how well a models estimate is compared to a reference model [44, 45]. If the output is less than 0, the model is worse than the reference, the output is 0, the models are equal. Finally, if the output is between 0 and 1, the model estimate is better than the reference model [45].

Chapter 3

Method

In this chapter, the method used to forecast power generation is explained. The proposed method is summarized in the flow chart in Figure 3.1. Before starting on the description of the method, a short literature review stating how others have approached similar problems will be given in Section 3.1. Next, the case for this thesis will be presented in Section 3.2, and an introduction to the software used is given in Section 3.2.1. Then the presentation of the proposed method starts with presenting the available data in Section 3.3, followed by the steps to preprocess the data in Section 3.4. This includes missing data and outlier handling, scaling, and data splitting. The process of feature selection is then explained in Section 3.5. After this, the two different forecast methods are introduced in Section 3.6. Finally, the evaluation is presented in Section 3.7.

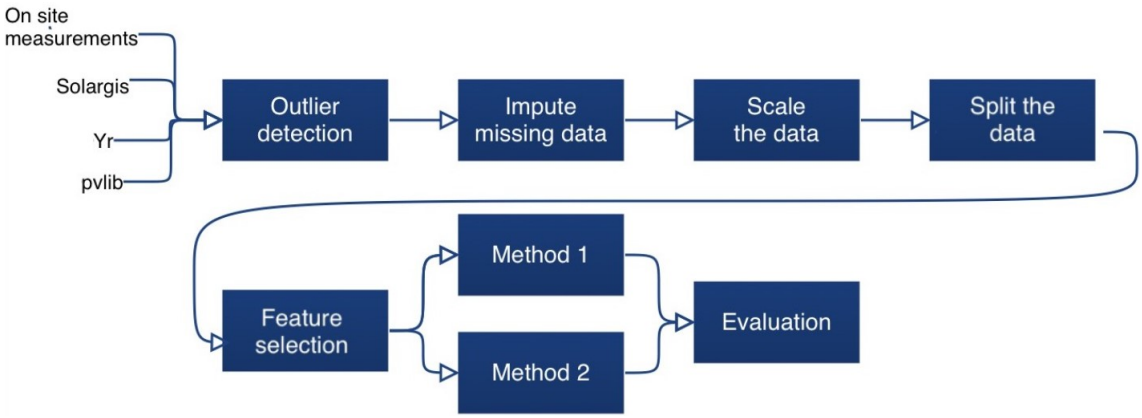

Figure 3.1: *Flow chart of the method used to preprocess the data, make the models and evaluate the results.*

3.1 State of the art

With the increase in renewable power generation from PV technology, there has also been an increase in scientific work concerning PV power forecasting [46]. Gupta et al. and Antonanzas et al. collected and summarized various works in their review publications in 2021 and 2016 [46, 45]. In these reviews, forecasting of solar power is divided into direct and indirect forecasting methods. Direct forecasting methods forecast the power generation directly whereas indirect forecasting methods forecast

solar irradiance first, then power generation is forecasted based on the solar irradiance forecast. Many works focus solely on solar irradiance, as this in itself is the most difficult element to forecast. The works done about irradiance forecast are also relevant for power forecasts, as many of the same methods are used [45]. Gupta et al. also divides the forecast methods into two groups: data-driven and physical models. The data-driven methods extract useful information from the features in the given training data and use that to make a forecast model, whereas the physical model uses numerical methods and physical formulas to calculate the forecasts [46]. In this thesis, both these methods will be examined. However, first, a short review of solutions for similar problems will be presented.

Babar et al. used RFR in their publication from 2020 to forecast solar irradiance at high latitudes in Norway and Sweeden [8]. There was a need for better irradiance forecasts than the ones provided by a satellite-driven method and reanalysis data. The satellite-driven forecasts had a large number of missing values as well as a tendency to underestimate the irradiance, and the reanalysis data had a tendency to overestimate for high latitudes. A RFR model with the satellite-driven data and reanalysis as input was tested, and proved to be better than both the satellite-driven data and the reanalysis data as well as outperforming a linear regression model [8].

Another solar radiation forecast was performed by Benali et al. in 2018 for a location in France [7]. They compared the performance of smart persistence, ANN and RFR on forecasts of GHI, beam normal irradiance, and diffuse horizontal irradiance. The forecasts from these models had a 1-hour resolution and 6 different time horizons, with time horizons from 1 to 6 hours ahead. RFR had the lowest Root Mean Squared Error (RMSE) for all time horizons, with the lowest RMSE being for the shortest horizon and the highest for the longest horizon. Smart persistence scored better than the ANN model for the shortest time horizon but scored worst on all longer horizons [7].

The oldest article reviewed in this thesis is by Bacher et al. from 2009 [47]. They forecasted PV power generation from rooftop systems in a village in Denmark with a typical peak capacity between 1 and 4 kW. Before forecasting future power generation, the power generation values were normalized with a clear sky model. The power was then forecasted using a statistical autoregression model with three different variations. One with only historical power generation as input, one with only Numerical Weather Prediction (NWP) for solar irradiance, and one with both. It was found that for longer horizons (day ahead forecasts), the two last methods were superior two the first [47].

Larson et al. also used NWP to forecast power generation. In their case, they used fixed-axis rooftop PV on two American schools with a peak installed capacity of 1 MW [48]. They used an indirect approach with physical methods. For the calculation of GHI, two methods were used. The first used the GHI provided by one of two NWP providers directly. The other used the cloud coverage index provided by the NWP providers and clear sky irradiance to calculate GHI through a deterministic model. Power generation was found through a linear relationship between power and GHI for both methods. The authors found that their method could reduce errors in forecasts of PV generation in the day-ahead market [48].

In the article by El-Baz et al. from 2018 in Germany, power generation from a rooftop PV system with a fixed axis and an installed capacity of 3 kW was forecasted [9]. An indirect approach was used where the first step entailed tuning a clear-sky model for power generation which took shading from nearby buildings into account. In the second step a data-driven method with bagging regression trees with data from NWP. Bagging regression trees is closely related to RFR, where both use an ensemble of decision trees for regression [49]. The input features provided by the NWP were temperatures, wind direction, wind speed, cloudiness, and humidity [9].

Riise et al. researched forecasting of solar irradiance on two sites in Norway using RFR with data from the weather forecasting service *Yr* in 2023 [50]. The data that were found to be most influential were lagged GHI measurements, clear sky irradiance, low, medium, and total cloud cover, and relative humidity. These features were used to forecast GHI with 1-hour resolution and horizons between 1 hour and 48 hours. The models were tested on data from the sites it was trained on, in addition, one of the models was tested on the other site. All skill scores for the tests gave positive values, meaning all models performed better than smart persistence [50].

3.2 Case

For this thesis, the site chosen as the case is a utility-scale PV plant using crystalline silicon modules [51]. The modules are installed on mounting systems with single-axis tracking using backtracking and maximum tilt angle. The site has an installed generation capacity above 100 MW and uses several central inverters to convert the power from DC to AC. These inverters are connected to smaller transformers which transform up the voltage before sending it to a central transformer station. In the central transformer station, the voltage is transformed to a high level to be connected to the transmission grid [51].

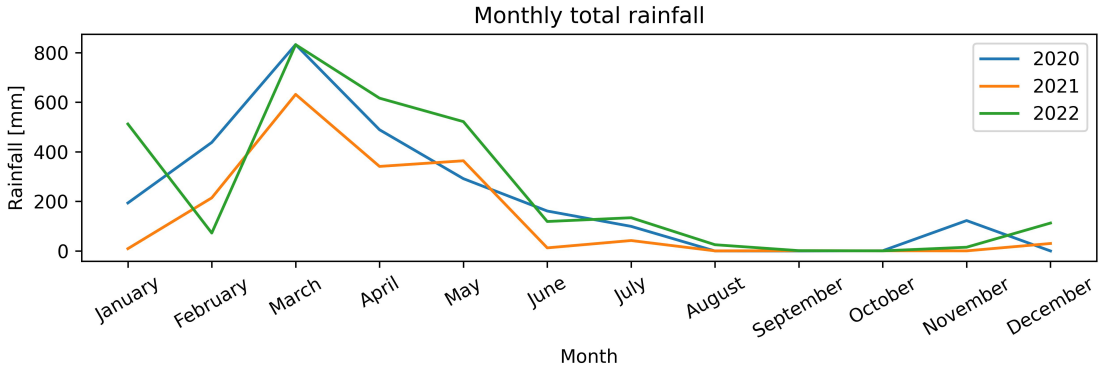

Figure 3.2: *Total measured rainfall for each month of the period 2020 to 2022 at the site.*

The site is also equipped with various sensors monitoring the weather conditions spread over the site. These measurements will be presented in Section 3.3. The site lies close to the equator and experiences a rainy season and a dry season. This can be seen in Figure 3.2, which shows the monthly precipitation on the site for the three years of data used in the thesis. Most of the precipitation falls in the period between January and July. Any forecasting model for PV generation needs to be developed to fit the conditions of the PV plant location. Riise et al. demonstrate this in their

article concerning PV forecasting models [50]. Models trained with data from the site got skill scores close to 0.3, whereas models trained on data from a different site got a skill score of less than 0.05 [50]. Therefore, some general knowledge about the weather on the site is beneficial.

3.2.1 Software

When dealing with large data sets it is important to have good tools to do so. *Python* is one of those tools. *Python* is a programming language with a large collection of libraries, enabling it to solve a wide spectrum of problems [52]. In this thesis, focusing on solar energy and machine learning, there are especially three important libraries that are used:

- *pvlib* is a library containing functions and methods enabling the user to simulate PV systems and solar trajectories [26].
- *scikit-learn* provides a variety of machine learning algorithms and data analysis tools [53].
- *Darts* is a library specialized in time series, focusing on anomaly detection and forecasting [54].

Most of the programming has been done with *Python* on a regular laptop, however, the grid search described in Section 3.6 was too computationally demanding and time-consuming for a regular computer. Therefore the Orion High Performance Computing system at the Norwegian University of Life Sciences (NMBU) was used [55]. This made it possible to run several computations in parallel, thus significantly decreasing the total computation time [55].

3.3 The data

The data used in the forecast process came from four sources:

- On-site measurements of weather parameters
- Meteorological forecasts from the weather forecasting service *Yr*
- Matemathical features from *pvlib*
- Estimates from the commercial company Solargis

Together, the first three sources formed the basis for the forecasts. The data from Solargis was used as a reference as well as for imputations of some features. In the following section, an overview of the data and its features will be given. A full overview of the data is given in Figure 3.1

Table 3.1: *Overview of data used from the different sources. Data that were not used in any steps of the process are not listed in the table.*

On site	Solargis	Yr	pvlib
Module temperature	Ambient temperature	Humidity	Clear sky irradiance
Ambient temperature	GHI forecasted	Ambient temperature	Solar elevation angle
Wind direction	GHI estimated	Wind direction	Module tracking angle
Wind speed	Generation forecasted	Wind speed	Estimated module temperature
Precipitation	Generation estimated	Precipitation	
Pressure		Pressure	
Plane of Array irradiance		Total cloud cover	
GHI		High clouds	
Power generation		Medium clouds	
		Low clouds	

3.3.1 On-site measurements

The measurements on site were made by various sensors placed across the site. The median value was taken across all measuring sensors for each feature and was used as data in this thesis. Table 3.2 lists the measurements taken on the site, and the total number of measurement points. All the sensors register measurements every 15 minutes. These measurements were resampled to 1 measurement per hour by averaging.

Smart persistence for both power generation and GHI is made based on the measurement data and added as a feature to the data. This feature was used to produce the skill scores for the results and was made by shifting the GHI and generation features 24 hours forward [45]. Since one of the simplest forecast models for forecasts with a 24-hour horizon is stating that yesterday's events will be repeated today, this is a good reference to compare the results to [45].

Table 3.2: *Number of sensors N for on-site measurements*

Data measured	N
Module temperature	12
Ambient temperature	13
Wind direction	13
Wind speed	13
Precipitation	1
Pressure	1
Plane of Array irradiance	23
GHI	25
Power generation	1

3.3.2 Yr

As Riise et al., this thesis also uses meteorological data from the publicly available weather forecast service *Yr* in the forecasts [50]. *Yr* provides weather forecasts and other meteorological information from NWPs [56]. From their webpage yr.no, one can access forecasts 48 hours into the future with a one-hour resolution. *Yr* has its main focus on the Nordics but also provides weather forecasts for the rest of the world by using the weather forecasting model from the European Centre for Medium-Range Weather Forecasts (ECMWF) [56]. This model has a spatial resolution of approximately 9 km^2 and is updated every 4 hours [56]. To use forecasts made in the past, Institute for Energy Technology (IFE) has logged 48-hour forecasts from *Yr*'s webpage every hour from 2018 to 2022. This has given a data set of 48 hours with forecasts for each hour between 2018 and 2022.

For the purpose of 24-hour irradiance forecasts, only the 24-hour weather forecasts issued at midnight were used for the years 2020 to 2022. This made a continuous forecast that equals checking the weather forecast every midnight. This is beneficial for the purpose of this thesis, where the forecasts for the following day were issued at midnight.

3.3.3 pvlib

pvlib is a library in *Python* that provides tools for simulations of PV systems [26]. For this thesis, version 0.9.4 was used. Tools form *pvlib* make it possible to for instance simulate the steps between clear sky irradiance and AC power output. These steps will be explained in further detail in Section 5. As well as providing functions to calculate result time series based on input time series, the library also provides mathematical simulations of useful features based on a few input values. Given the latitude, longitude, altitude, and time of day, a *pvlib* function will return the Sun's elevation angle or clear sky irradiance. With some more input, one can also get the solar panel's tracker angle. These functions were being used as potential features for the machine learning process as well as in the prepossessing of the measurement data [26].

3.3.4 Solargis

Solargis provides historic, recent, and forecast data for solar power generation [57]. These data consist of Direct Horizontal Irraiance (DHI), Global Tilted Irradiance (GTI), GHI, power generation data, as well as weather data for a given site. Satellite data is used in a semi-empirical solar radiation model for historic and recent radiation data. The forecasts use NWP models that are dynamically improved with data from the satellite model [58]. For the meteorological data, NWP models are used. To get better accuracy and homogeneity for temperature, the temperature data is post-processed to make its spacial resolution smaller [59]. As with the on-site measurements, the Solargis data also provide one measurement every 15 minutes, these data were therefore also resampled.

3.4 Data prepossessing

3.4.1 Outliers

For most of the features, no large outliers or deviations that needed to be removed were found. The pressure data did contain a lot of outliers where the pressure was significantly lower than in the rest of the distribution, however, because of the generally low quality of this feature, this feature was not imputed. Also in the generation data, there were occasional deviations like the ones shown in Figure 3.3. The value of the spike on the first day in Figure 3.3 is about 80% of the median daily generation for the entire three-year period. Therefore, it is reasonable to consider that the events are the cumulated generation logged into the system after some time without logging. As well as large positive outliers, there were also periods with curtailment where the generation was much lower than the measured irradiance indicates because of excess power in the electrical grid [33]. Since the curtailment losses happen due to factors outside the power plant, the most extreme generation reductions due to curtailment were also removed.

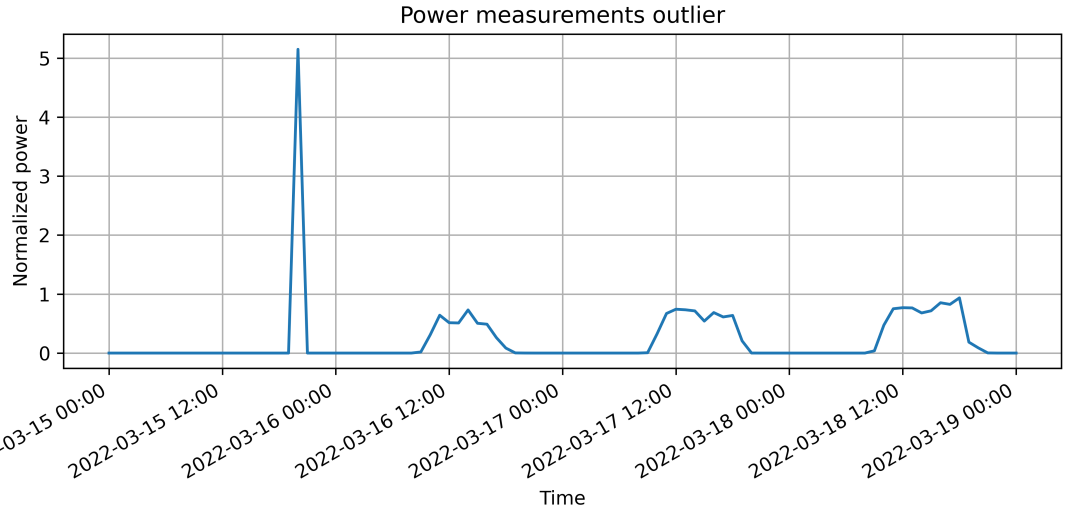

Figure 3.3: *Plot showing the normalized measured power generation for four days in March 2022. The plot shows an anomaly on the first day where the measurement is significantly larger than for the following days.*

Table 3.3: *Percentage of missing and removed values for the target features where the full data set consists of 26256 hourly measurements.*

	GHI	**Power generation**
Missing	0.29%	0.23%
Removed	-	0.11%
Total missing	0.29%	0.34%

As a means to efficiently remove the most extreme outliers, a z-score was used with a cut-off value of 4. To be able to take the daily periodicity into account, the data were grouped by the solar elevation angle and the z-score was calculated for every single group. The recommended cut-off value when using z-scores is normally 3 [60]. However, since the z-scores for each elevation group were not perfectly normally

distributed, a higher z-score made sure no normal values were cut off. The portion of the data removed with this process was 0.11% as shown in Table 3.3.

3.4.2 Missing data

As with most datasets, all the datasets used in this thesis also have some degree of missing data. For the target features, i.e., GHI and power generation, no imputations were made except for filling out 0 values when the solar elevation angle was less than 0. At these times there will never be any irradiance or power generation. Some different methods of imputation were used for the explanatory features, this is explained in the following paragraphs and summarized in Table 3.4

No strong seasonality was found in the wind speed measurement. However, the rest of the measurement data have a strong daily seasonality. Since the wind speed does not depend directly on any of the other features, a simple linear interpolation was used. For wind direction and ambient temperature, a rolling seasonal mean of the last five days was used to impute missing data. The amount of missing data for precipitation and pressure was so large that these features were dropped instead of imputed. Even though no missing values were imputed in irradiance, they were filled in the smart persistence model. This was done to have a complete set of explanatory features with few missing values, thus giving the model more to train on. The missing smart persistence values were filled with the estimated GHI from Solargis.

Table 3.4: *Percentage of imputed values for measurement data and technique used. The measurement data had a total of 105120 15-minute measurements, whereas the Yr data had 26256 hourly measurements*

Measurement	**Method**	**Missing**	**Imputed**
Ambient temperature	Solargis. Rolling mean	0.64%	0.52% 0.12%
Module temperature	Ambient temperature transformed with *pvlib*	1.78%	1.78%
Wind direction	Rolling mean	0.42%	0.42%
Wind speed	Interpolation	0.64%	0.64%
Pressure		3.6%	
Precipitation		1.6%	
Smart persistence	Solargis	0.52 %	0.42%
Yr data	Filled with forecasts done 25 to 48 hours ahead for the same timestamps. Mean value if elevation < 0	23%	2.6% 6.6%

The *Yr* data had several long segments of missing data spanning several days as well as occasional missing values. As only the first 24 hours of the 48-hour forecasts were used as features, the 25 to 48 hours ahead forecasts were used to impute if there were missing values. If there were missing values in the 24-hour forecast for one day, the 48-hour forecasts made the day before were used to impute. This technique had relatively good accuracy with a RMSE of 0.7°C for the temperature forecasts and 0.6 mm for the precipitation. The nighttime values when the Sun's elevation is less than 0° were imputed with mean values. This will not contribute to the GHI or power

forecasts, however, since the model needed days with no missing values, this will could increase the number of days used for forecasting.

The design of the method does not allow any missing data in either the features or targets for the 24-hour forecast horizon. That means that was is not possible to do forecasts on days with one or more missing values. Therefore, the true portion of missing data would be the percentage of days with at least one missing value. The distribution of the days with no missing values is shown in Figure 3.4. In the total preprocessed dataset, 26.3% of the days had one or more missing values for the GHI forecasts and 32.1% for the power forecasts.

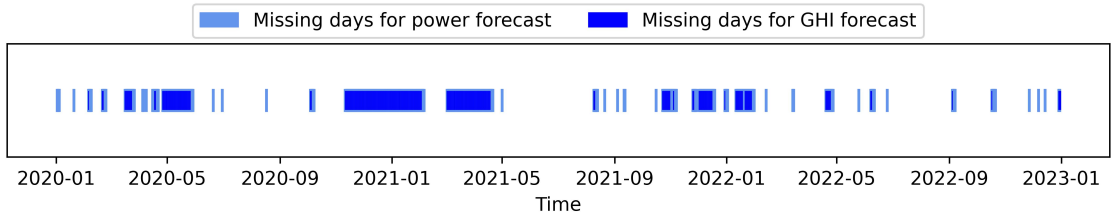

Figure 3.4: *Missing days for the GHI forecast in dark blue and for the power generation forecast in light blue.*

3.4.3 Scaling

There are many ways to scale the data before initiating the learning process. For these data min-max normalization was chosen. The equation for the scaling is given by

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3.1}$$

where x is the feature, and x_{scaled} is the scaled feature [61]. This scaling sets the range of the data to [0,1] with 0 being the smallest and 1 the largest value for each feature in the training set [61]. For the targets, this scaling would set the maximum values to 1 and no irradiance or power generation to 0. This makes the results easy to read even in their scaled form. However, this method can be highly influenced by outliers, and should not be used without preprocessing or good knowledge of the data [61]. Of the features used in this thesis, the power generation values were the ones with outliers, and these were removed before scaling.

3.4.4 Splitting the data

The data were split into three smaller datasets for training, validation, and testing. There is no set truth for what the best data split is, but for this thesis, 1/3 splits were used, to get one full year of data with all seasons represented for testing [62]. 1/3 of the data was set aside for the final test of the data, and 1/3 of the remaining data was used to validate the models made with the training data. This gave the split 44% of the data for training, 22% for validation, and 33% for testing. With this split the training data was raging from 01/01/2020 to 06/05/2021, the validation data from 07/05/2021 to 23/12/2021, and the test data from 24/12/2021 to 31/12/2022.

When splitting the data for the GHI forecasts, the splits were made to the full dataset, meaning all of 2022 would be used for testing. For the forecasts of the power

generation, a new approach was used due to a large amount of missing data in the first two year as is evident from Figure 3.4. Therefore the fractions were based on the days without any missing values, leading to the test set starting at the 7^{th} of May 2022. This resulted in less than a full year of testing. However, this approach did leave more data for training and validating the model properly. With this new split, the interval for the training data was ranging from 01/01/2020 to 28/11/2021, the validation data from 29/11/2022 to 06/05/2022, and the test data from 07/05/2022 to 31/12/2022.

3.5 Feature selection

When features from all sources were added together, there were 20 features that could be used to forecast GHI and 21 that could forecast power generation. As it would take too long to check all the feature combinations with this number of features, a feature selection using the Boruta method was carried out. This method was chosen because of its speed and ease of implementation [39]. As with the forecasting model used in this thesis, Boruta is also based on RFR [39]. In Degenhart et al.'s comparison of feature selection methods for RFR, the Boruta model is seen as the most powerful feature selection method with good stability in the feature selection [40]. As Boruta is part of the library BorutaPy that is based on *scikit-learn*, it was also easy to implement to a *scikit-learn* RFR [39].

To avoid being overly affected by the randomness of the RFR model and to make sure all potentially influential features are included, the model was run repeatedly 10 times. The results of these iterations are shown in the appendix Figure 5.1 and Figure 5.2 in the Appendix. All features that passed the test at least once, i.e. gets ranked 1 at least once, were selected for further processing.

3.6 The forecasting methods

For this thesis, RFR was chosen as the forecasting model. RFRs have been used with good results by Benali et al. for intra-day forecasts (1-6 hours) in France and Babar et al. for locations at high latitudes in Norway and Sweeden [7, 8]. In the literature review by Gupta et al., the ensemble learning techniques (as RFR is a part of) also outperforms ANN and other machine learning techniques [46].

The main goal of this thesis was to forecast power generation from a PV power plant. This was done through two different methods, one data-driven, and one physical method. The flow chart in Figure 3.5 demonstrates the two different ways the power generation was forecasted. The top chart is the data-driven model hereby referred to as Method 1, and the physical method in the bottom chart is Method 2. A review of both methods will be given after the introduction of the general RFR model that is used initially in both models.

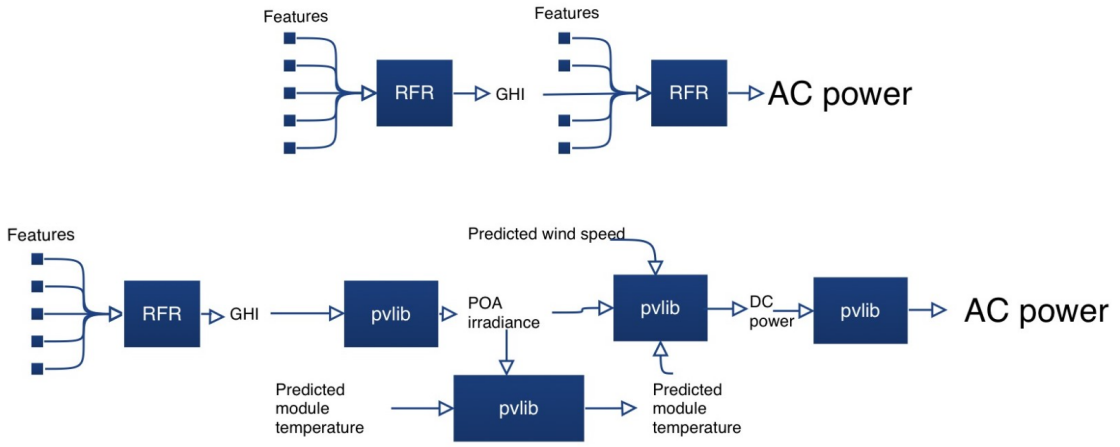

Figure 3.5: *Flow chart showing the proposed methods, with Method 1 in the top chart and Method 2 in the bottom chart.*

3.6.1 General Random Forest Regression (RFR) method

The method for optimizing the RFR models is shown in Figure 3.6. First, a grid search over a large number of different combinations of features was performed. Here the feature combinations were the only hyperparameter that was changed in each iteration. This would give the best combination of features as well as the number of features necessary.

Because the RFR model is a random model, even with the same parameters, it will produce a slightly different model each time it is called. Therefore, the 100 best feature combinations were run once more with 10 reiterations of each model. The feature combination with the best average Mean Squared Error (MSE) after this process was considered the best and selected for further optimization.

The next step was to find the optimal combination of hyperparameters. This was done by searching over all combinations of different values for the selected hyperparameters combined with the 5 best feature combinations. As with the feature combination search, also here the 100 models with the lowest was MSE run 10 times to get the model with the best average MSE value after 10 iterations.

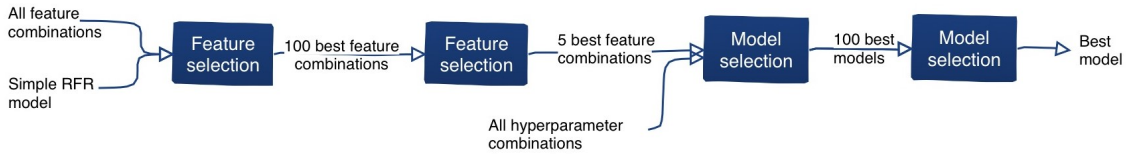

Figure 3.6: *Flowchart showing the proposed RFR optimization method.*

3.6.2 Method 1

Method 1 is a completely data-driven method that uses the RFR method described above first to estimate the GHI. The RFR method was used again on the estimated GHI together with all the other features selected by the Boruta feature selection to directly output a forecast for power generation.

3.6.3 Method 2

Method 2 is a partial physical method that consists of a series of operations to get from GHI to power generation. The first step towards a power forecast was to forecast the GHI. This was done with the data-driven method using RFR described above. The data produced from these forecasts were put through several *pvlib* functions as shown in Figure 3.5 (the *pvlib* functions used for each operation are listed in footnotes). The forecasted GHI data was transformed to Plane of Array (POA) irradiance[1], which, together with the forecast module temperature [2] and wind speed was used to calculate DC power [3] which was transformed to AC power [4]. POA and forecast ambient temperature were used to get an estimate of forecast module temperature.

3.7 Evaluation criteria

To evaluate the performance of the models, the evaluation metrics listed in Table 2.1 in Chapter 2.5.4 were used. MSE was used as the objective function, e.i. the value that was minimized when evaluating which model performs the best in the feature and hyperparameter combination selection in the RFR models.

As well as using MSE to select the best performing RFR model, the rest of the metrics in Table 2.1 were used to compare the results from Methods 1 and 2 with each other as well as with the model from Solargis. Skill score was used as the main metric for comparison and this metric was also used to compare the results with results from literature.

[1] pvlib.irradiance.get_total_irradiance(surface_tilt,surface_azimuth, solar_zenith, solar_azimuth, dni, ghi, dhi)

[2] pvlib.temperature.faiman(poa_global, temp_air, wind_speed, u0, u1)

[3] pvlib.pvsystem.pvwatts_dc(g_poa_effective, temp_cell, pdc0, gamma_pdc, temp_ref

[4] pvlib.pvsystem.PVSystem(inverter_parameters).get_ac(model, p_dc)

Chapter 4

Results and discussion

In this section, the results will be presented and discussed. First, the measured data from the site and the forecast from *Yr* are examined to show the quality and correlation in Section 4.1.1. Next, the results from the GHI forecast are evaluated in Section 4.1.2 and the results from the power forecast are evaluated and discussed in Section 4.1.3. In the following three sections parts of the method are discussed with feature selection in Section 4.1.4, missing data in 4.1.5, and the train test split in Section 4.1.6. Then the effects curtailment might have had on the results is discussed in Section 4.1.7. After this, the results are compared to similar methods from the literature in Section 4.1.8. Lastly, in Section 4.2, the implications and applications of the results are discussed briefly.

4.1 Results

4.1.1 The data

To make any model or forecast, one first needs to understand the data that is being utilized in the model. In Figure 4.1 some of the data from the on-site measurements and weather forecasts for the site for the first week of August 2022 is plotted. Figure 4.2 show the correlation between all features during daytime for the entire time period.

From the on-site measurements, one can see a correlation between the GHI and the ambient and module temperatures. On days 6 and 7, when the irradiance is consistently high throughout the day, there are also consistent smooth curves for the temperatures with a high maximum. The same relationship can be seen for the days with low irradiance. Those days have more inconsistent temperature curves with a much lower maximum temperature, e.i. on days 3 and 4. One can also see a connection between the days with precipitation and GHI, where the irradiance is reduced on days with precipitation. However, not all days with reduced irradiance have any precipitation as can be seen from day two. For the other measured features, it is difficult to see any clear correlation in both Figure 4.1 and 4.2, which underlines the importance of deeper analysis with machine learning to see the connections that might be hidden from the human eye.

Looking at the weather forecast in Figure 4.1 there are relationships between the

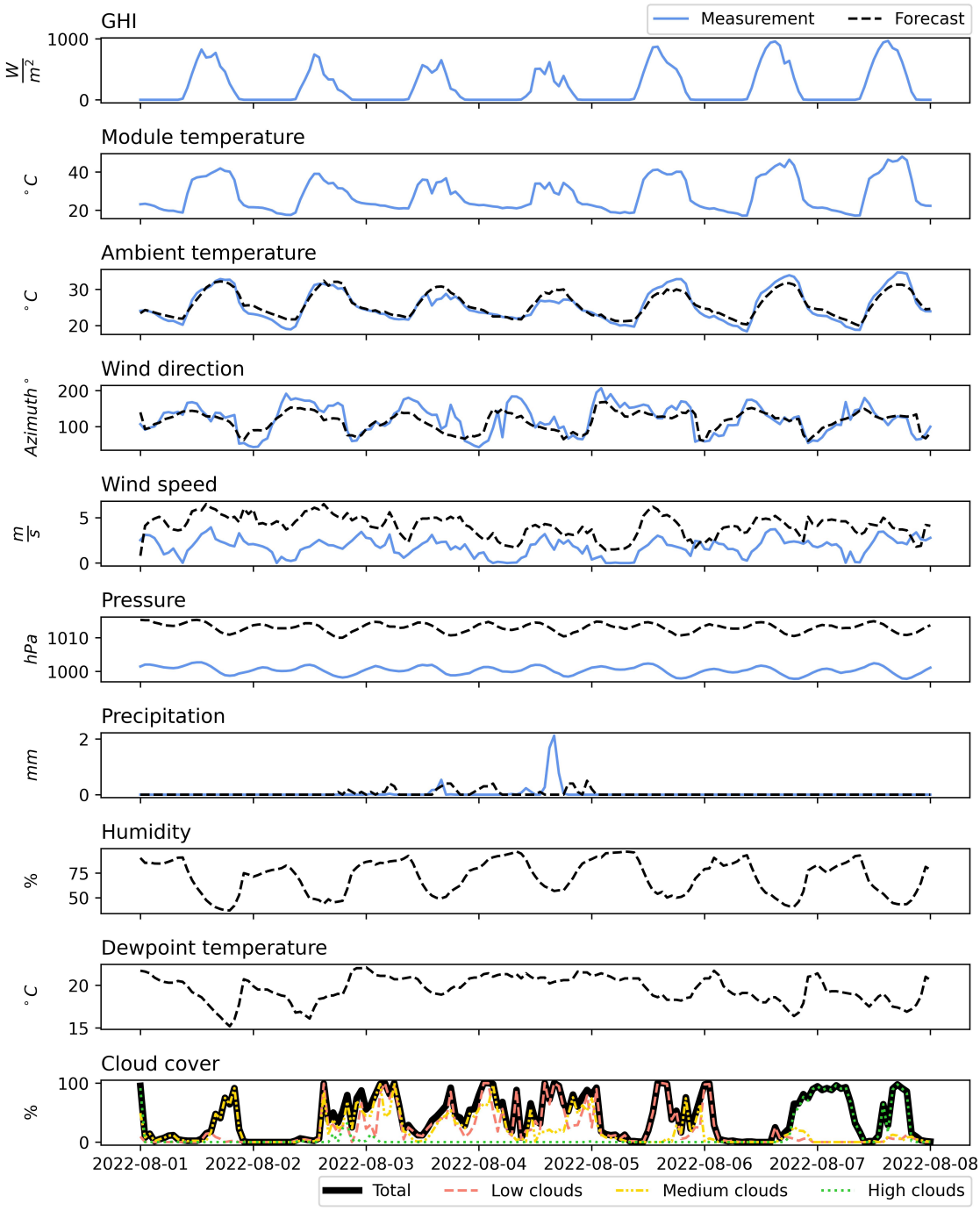

Figure 4.1: *Measurement data (solid blue lines) and weather forecasts (dashed black lines and multicolored lines) from the first week in August 2022 for the site.*

forecasted and measured values. The forecasted ambient temperature has values that highly resemble the measurement values with only slight deviations around midday from day 3 onward, the correlation between these features for the entire dataset is also high at 90%. For the precipitation, the weather forecast is spread out over a larger time period of two days, whereas the measured precipitation is more intense over two short time periods of a few hours. This is reflected in the low correlation of 16% between the forecasted and measured precipitation. Wind direction, with a correlation of 67%, shows a good forecast where the general trend in wind direction is well covered by the forecast.

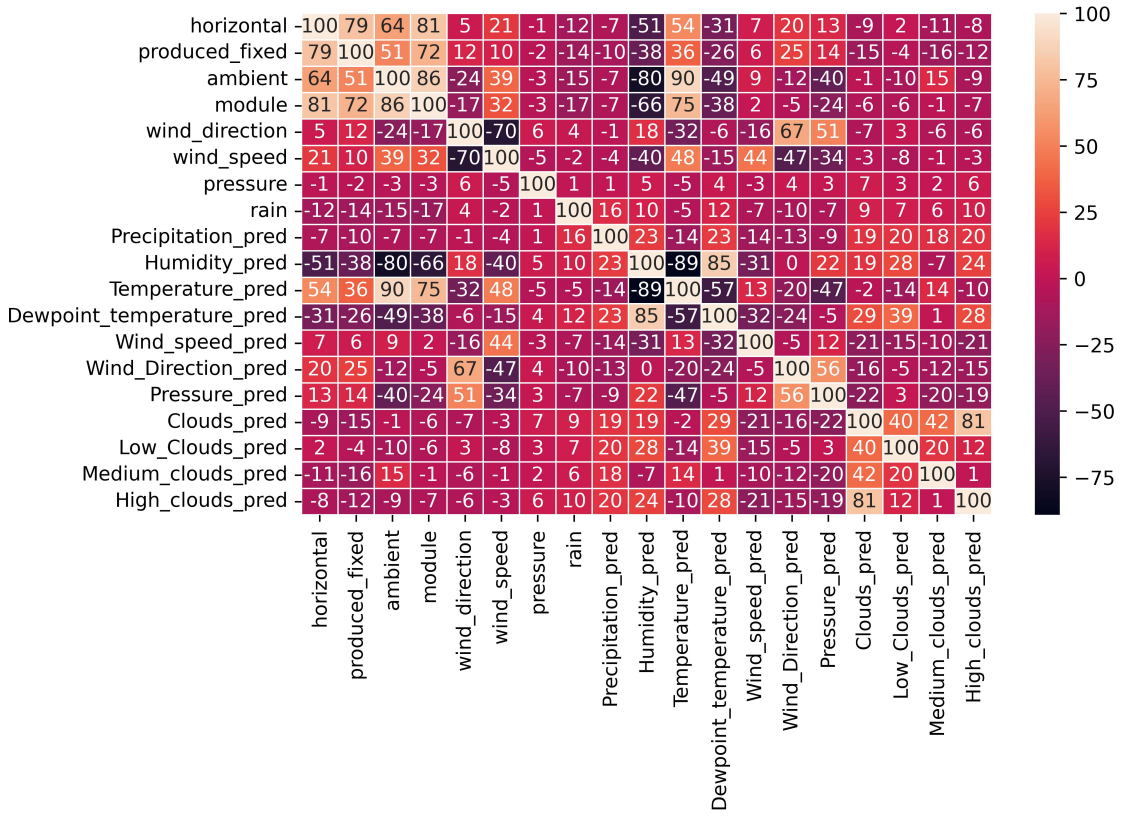

Figure 4.2: *Correlation plot for all measured and forecasted features at daytime hours when the solar elevation angle is larger than 0. All feature names ending with _pred are forecasts from yr. The feature "horizontal" is the GHI and the "produced_fixed" is the power generation.*

The forecast for wind speed and pressure is consistently higher than the measured values. For the pressure, the forecast follows the measurement trend quite well, the wind speed forecast also follows the measurement trend well for the first four days. On the last two clear sky days, there is more difficult to see similarities between the trend in the forecasts and measurements. The correlation between the measured and forecasted wind speed and pressure are 44% and 3%. The low correlation for measured and forecasted pressure is likely related to the outliers in the measured data. Finally, the forecasted cloud cover for low and medium-height clouds seems to correlate well with the reduced irradiance on days 1 to 5. Days 6 and 7 also have some cloud cover, but mainly by high clouds.

From the matrix in figure 4.2, the highest correlation to the GHI is from temperature features, where ambient, module, and forecasted ambient temperature had correlations of 64%, 81%, and 54% respectively. It should however be noted that correlation is a measure of linear relationships and will not give a good score for non-linear relationships like sinusoidal or quadratic relationships [63].

4.1.2 The Global Horizontal Irradiance (GHI) forecast

Table 4.1 shows the evaluation metrics for the forecast of GHI for Solargis, the data-driven RFR model, and smart persistence for the same days as in Figure 4.1. The data-driven model has a better score on the MSE based metrics MSE, RMSE, and

Table 4.1: *Summary statistics for the forecasts of normalized GHI from Solargis and a data-driven RFR method for the test set spanning from 24/12/2021 to 31/12/2022.The best score for each metric is marked in bold. The scores for the data-driven method are the mean after 10 iterations.*

Metric	Solargis	Data-driven	Smart persistence
MBE	-0.0174	0.00631	**-0.0008**
MSE	0.00823	**0.00791**	0.01337
RMSE	0.0907	**0.0896**	0.1156
PAPE	**18.4**%	22.2%	22.8%
SS	0.216	**0.231**	-

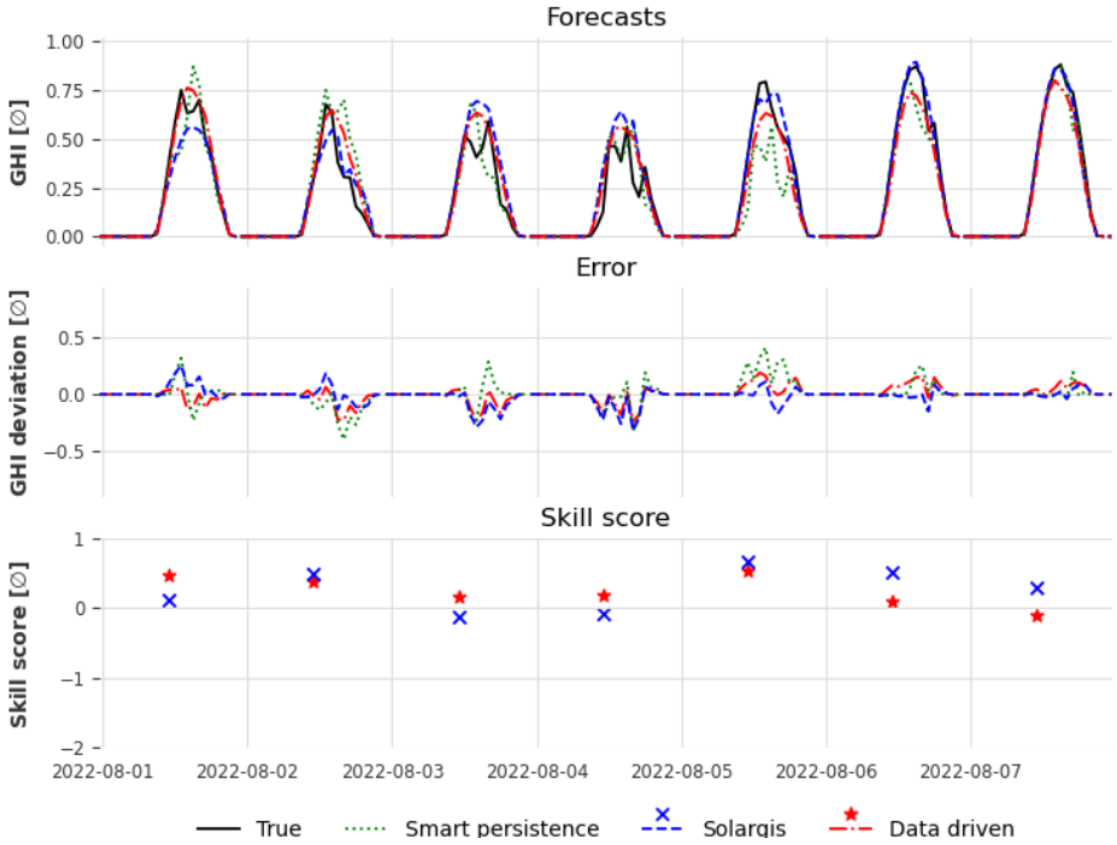

Figure 4.3: *In the top plot, the true normalized GHI is plotted together with normalized forecasts from smart persistence, Solargis, and the data-driven method. The middle plot shows the bias between the forecasts and the true values. In the bottom plot, the daily skill score is plotted for Solargis and the data-driven method.*

skill score. The solargis model does however have a lower Peak Absolute Percentage Error (PAPE), and the Mean Bias Error (MBE) for smart persistence is the one closest to zero. The MBE also shows that the Solargis model has a tendency to overestimate the results, whereas the data-driven model tends to underestimate them. Smart persistence seems to overestimate and underestimate roughly equally. All three models also have a comparable high PAPE, suggesting that all models have problems with forecasting the days' peak irradiance.

On the clear sky days on days 6 and 7 in figure 4.3, the skill score for the data-driven model is relatively low compared to the total skill score in Table 4.1. For day 6, the model forecasts a lower peak GHI, resulting in a low skill score. On the next, the data-driven model returns a good forecast with a bias close to zero, however, here the smart persistence is very good because of the two consecutive clear-sky days, resulting in a low skill score for the data-driven model. However, over the full test set, the data-driven model and Solargis score better than the smart persistence. The skill scores in table 4.1 indicate that in terms of RMSE, the forecast from Solargis is 21.6% better than the smart persistence, and the data-driven model produces a forecast that is 23.1% better.

4.1.3 The power forecast

In Table 4.2 the evaluation metrics for the power forecasts are presented. This show that Method 1 scores best on the MSE-based metrics, e.i., MSE, RMSE, and skill score, and the smart persistence model has the lowest PAPE and the MBE closest to 0.

Table 4.2: *Summary statistics for the forecasts of normalized power generation from Solargis, Method 1 and Method 2 for the test set spanning from 07/05/2022 to 31/12/2022. The best score for each metric is marked in boldface. The scores for Method 1 are the mean after 10 iterations.*

Metric	Solargis	Method 1	Method 2	Smart persistence
MBE	-0.0525	0.0156	-0.0182	**0.0006**
MSE	0.0187	**0.0155**	0.0240	0.0242
RMSE	0.137	**0.124**	0.155	0.156
PAPE	11.4%	15.2%	12.7%	**10.9%**
SS	0.122	**0.200**	0.004	-

The MBE values for Solargis and Method 2 are negative while Method 1 and smart persistence have positive MBE values. This means that Solargis and Method 2 have a tendency to overestimate the power, whereas Method 1 and smart persistence tend to underestimate it. However, as the MBE is almost zero for smart persistence, its underestimation and overestimation are roughly equal. The MBE scores correlate well with the results from the GHI forecast. The underestimating tendency did increase for Method 1 and decreased for Method 2. Thus, the underestimation is increased by the RFR and decreased by the physical connections.

From the plotted results in Figure 4.4, one can see that Method 1 is not able to forecast the clear-sky days at the end of the plotted week well because of underestimations. It is also evident that Solargis generally has negative forecast errors from looking at the second plot. This is even more evident in Figure 4.5 where the majority of the power forecasts made by Solargis are larger than the actual power measurements. For Method 1 it is evident that the forecasts underestimate the higher power generation values in the upper left corner of the third plot in Figure 4.5. In the bottom of the right plot in Figure 4.5, there is a linear correlation between Method 2 and the true values, indicating some pattern in the underestimations from Method 2. This could be an interesting connection causing this, however, this is not investigated further in this thesis.

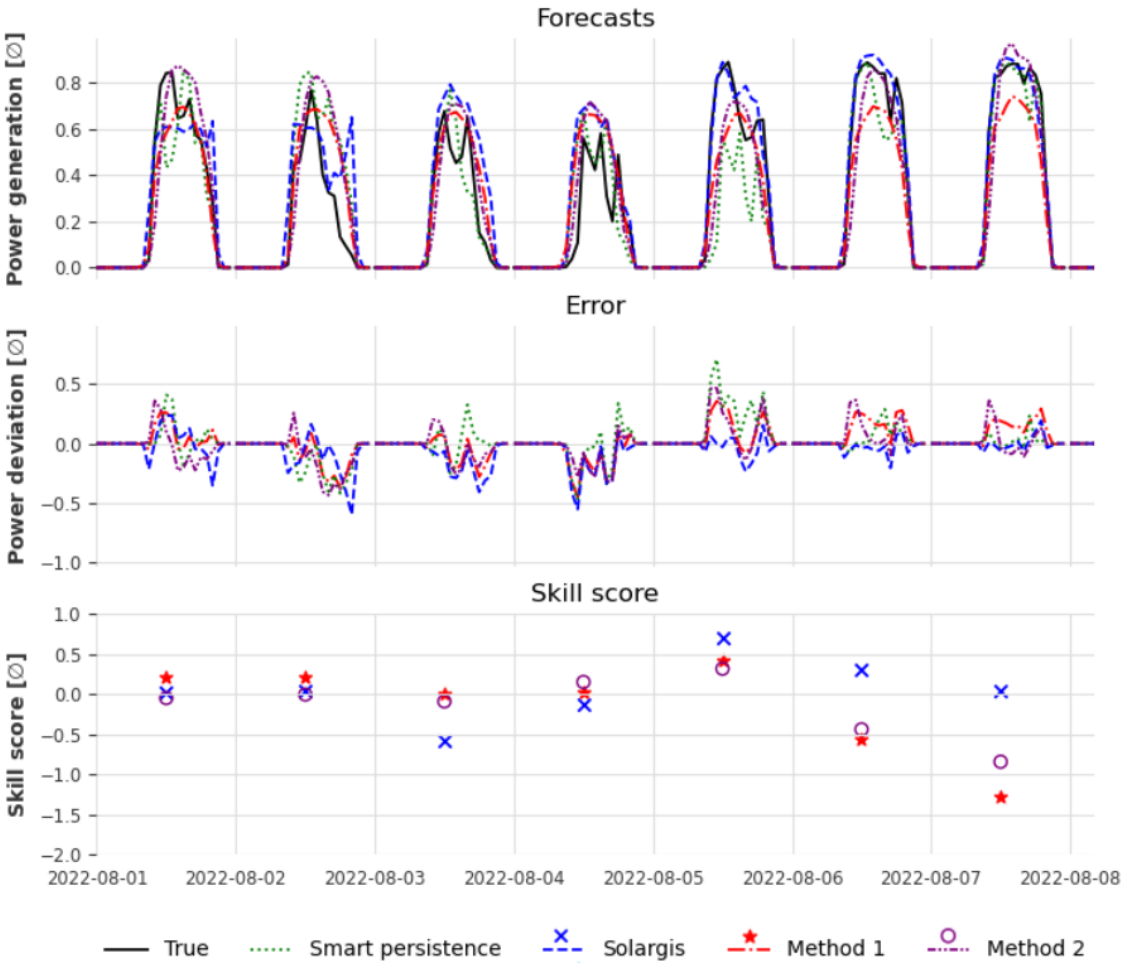

Figure 4.4: *In the top plot, the true normalized power generation is plotted together with normalized forecasts from smart persistence, Solargis, Method 1, and Method 2. The middle plot shows the bias between the forecasts and the true values. In the bottom plot the daily skill score is plotted for Solargis, Method 1, and Method 2.*

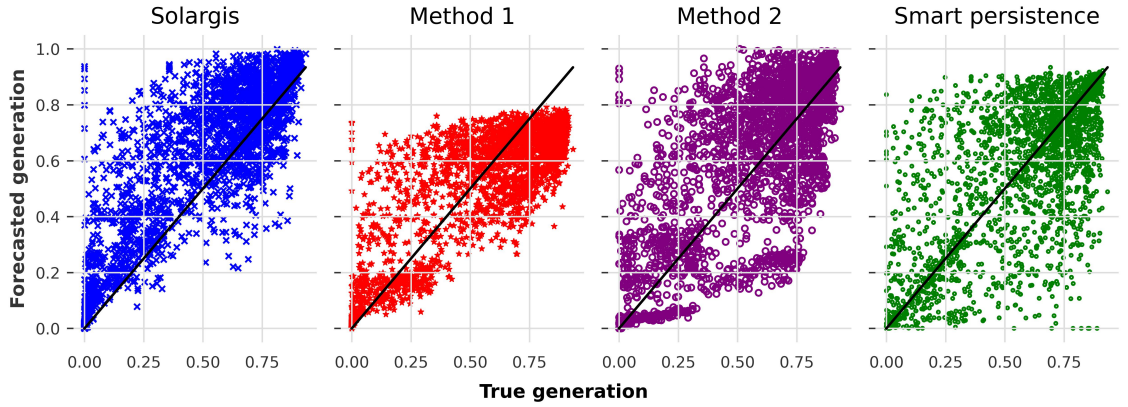

Figure 4.5: *Forecasted power generation plotted against the actual power generation for the three methods. The black diagonal in each plot represents where the forecast is equal to the true value.*

The PAPE values are generally lower for the power forecasts than for the GHI forecasts, indicating that the models are better at forecasting peak generation than peak GHI. The difference in PAPE between Method 1 and Method 2 shows that

Method 2 might have a higher ability to correct the peak estimation errors from the GHI forecast than Method 1. Alternatively, Method 2 might be inherently optimistic in its forecasting, thus making up for the errors in the GHI forecasts with opposite errors in the power forecasts. The upper right corners in the two middle plots in Figure 4.5 show well how Method 2 is better at forecasting the peak power generation than Method 1.

From the positive skill scores, it is evident that all of the methods are better than smart persistence in terms of RMSE. Looking at Figure 4.4, the skill scores of Method 1 are best on the three first overcast days and the Solargis model scores best when it is a higher irradiance in the three last days. This also correlates well with Solargis being prone to overestimate and Method 1 to underestimate.

Figure 4.6 shows a density plot for the daily skill scores in the test period. Method 1 has the highest density above 0 with a relatively sharp peak around 0.25. To the right in the plot, one can see that Solargis is the one with the highest skill scores, however, the total skill score is dragged down by the higher density of skill scores below 0. Method 2 has its peak just below 0 and a relatively even distribution on both sides of 0. Looking at the minimum values, both Method 1 and 2 have a minimum skill score of about -11 whereas the minimum skill score for the Solargis model is at -4.4. These minima typically happen when two perfect clear-sky days follow each other. On these days smart persistence will get a very low RMSE value, resulting in a highly negative skill score for small errors in the modeled forecast.

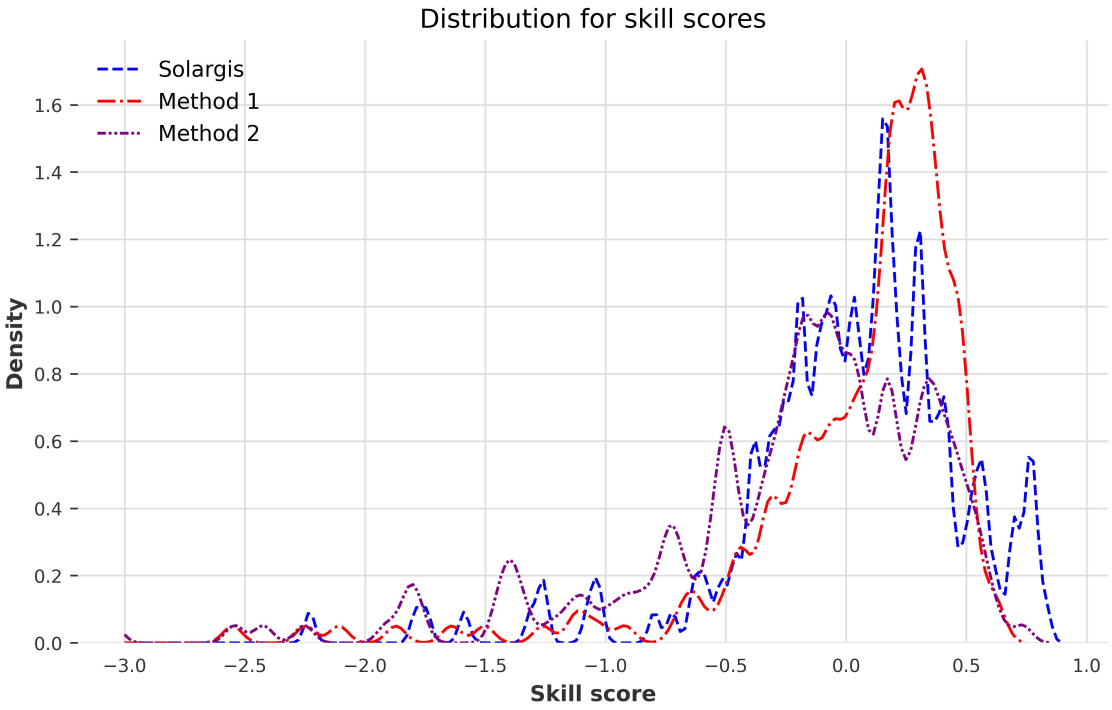

Figure 4.6: *The Figure shows the distribution of daily skill scores for the three models in the range* -3 *to* 1*. Outside the bounds of the plot, the Solargis, Method 1, and Method 2 have 2, 4, and 4 days outside the index of the plot respectively. Their minimum skill score are* -4.4*,* -11.6*, and* -11.3*, respectively.*

Looking at the general shape of the forecasts, the data-driven models for both the power and GHI produce smooth curves and are not able to capture much of the hour-to-hour variability. This is also transferred to Method 2 which is built on the data-driven GHI model. Looking at the Solargis model, this has more hour-to-hour variability, however, it does not necessarily improve the forecasts for the days plotted in Figures 4.3 and 4.4. In the same Figures, it is also evident that the daily power generation curves from Method 1 have a relatively small variation compared to both the other two power generation models and the data-driven GHI model. It does seem like the data-driven model smooths out the variations in the data, applying this twice, as is done in Method 1, might enhance this smoothing effect on the results.

4.1.4 Feature selection

For the first RFR model that forecasts the GHI, eight features were chosen as forecaster features. Those were *Yr* forecasts for ambient temperature, clouds, low clouds, medium height clouds, and precipitation, local measurements of GHI for the previous day, and historic measurements of ambient temperature and wind direction.

It seems like the model chose features that directly impact the irradiance. From Figure 4.1 and 4.2, the temperatures highly resemble the irradiance curves and have a high correlation. The days with significantly high coverage for low and medium height clouds and days with forecast precipitation were also the days with reduced irradiance, as seen from the first four to five days in Figure 4.1. For the entire dataset these do, however, have low correlations with the GHI in the correlation matrix in figure 4.2.

For the power forecast with Method 1, seven features were chosen. Those were medium clouds, wind direction, humidity, clear sky irradiance, the GHI for the previous day, and the forecasts for the GHI. Of these, the three first are forecasted values from *Yr*, the clear sky irradiance is generated with *pvlib*, the GHI from the previous day's measurement data from the site, and finally, the forecasted GHI which is the forecasts made in the previous step.

This model has some overlap in the feature choice with the model for irradiance as both models use medium height clouds and GHI from the previous day. A new feature is the wind direction. This feature could have been chosen since the wind has a cooling effect on the effect leading to an increase in power generation [10]. The overlapping features with the irradiance model, as well as other irradiance features like clear sky irradiance and humidity, could indicate that the irradiance forecast does not capture enough of the variations in the measured irradiance.

4.1.5 Missing data

As stated in Section 3.4.2, there is a large amount of missing data, with up to 32% for the power generation forecasts. This is in part because of the missing data in the *Yr* forecast and targets, but some of it also comes due to the need to have a full day of data to be able to do the forecasts for that day. Since days with only one missing value were discarded, a lot of useful information is lost. An alternative to this would be to build the RFR model differently so that it could do day-ahead forecasts one hour at a time and make it able to skip hours with missing values. This

should be possible when basing the model on *scikit-learn's* RFR model and might also be possible with the *Darts* RFR model given more time.

4.1.6 The train-test split

In the method, two different splits of the data were used in the method for forecasting of GHI and power generation. Because of the uneven distribution of missing data, the split securing 33% of the data for testing of the GHI forecasts lead to 42% of the data being used for testing. This was altered in the forecasts of power generation so that 33% of the actual data was used for testing. However, this led to there no longer being a full year of data for testing. That means that all the seasons were not represented when testing the model.

To check this, the clearness index was plotted for the training, validation, and testing set as well as for the entire data set in Figure 4.7. This shows that the test set has some more hours with a higher clearness index, except for that, the test set has a roughly equal clearness index distribution as the full data set. However, the validation set differs from the other data sets by having a lower density at higher clearness indexes and a higher density at lower indexes. Looking at the yearly precipitation graph in figure 3.2 and the train test splits presented in section 3.4.4, this does make sense. The test set, raging from May to December of 2022 does not cover much of the rainy season. However, the validation set, raging from December 2021 to May 2022, mainly covers months with large amounts of rainfall. As the validation data was used when selecting the features and hyperparameters, this could have had a notable impact on the results. This should therefore be taken into account in later attempts working with these data by using techniques like k-fold validation where the model is validated on different validation sets [62].

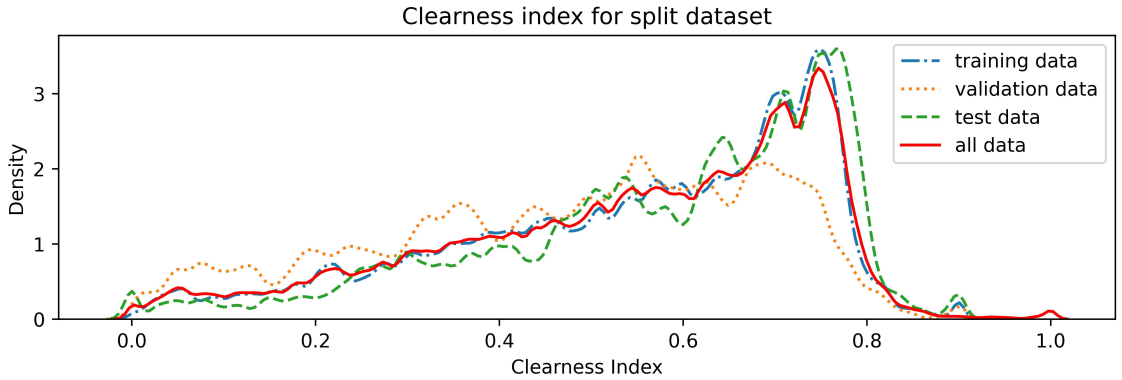

Figure 4.7: *Density plot of the clearness index for the train, validation, and test set as well as for the entire full data set.*

4.1.7 Curtailment

One of the factors making power forecasts difficult for the site chosen for this thesis is the amount of curtailment. The site experiences generation curtailment in 11.4% of the power plant's generation hours. The curtailment of the power is based on the state of the power grid [33]. Therefore, when the forecasts are made based on mainly the weather conditions on the site, the models have difficulties catching the

sudden drop in power generation. This problem will affect the results of all the models presented here, but not in the same way.

Method 1 has input data with what seems like random drops in the generation data. The RFR model in Method 1 will try to learn the pattern in these data based on the input features. Since the selected hyperparameters and features are the ones that gave the lowest MSE on the validation data, the selected model could be the one that safely underestimates so that the random curtailment losses give minimal effect on the MSE. That would be a possible explanation of why Method 1 has a positive MBE and a large PAPE.

On the other hand, Method 2 does not see any of the curtailment losses when the model is being built. It is purely based on the forecasted irradiance as well as physical relations to transform the irradiance forecasts into power forecasts. On days with curtailment losses, the model will therefore forecast as if the power plant was unaffected by curtailment. This is a possible explanation for why the model tends to overestimate, and it might also be an explanation for why is generally better at forecasting the daily peak generation. If the curtailment losses are sufficiently large, this could also be an explanation for the low skill score found for Method 2.

4.1.8 Literature

In the publications by Larson et al., Bacher et al., and El-Baz et al., the results were also evaluated with skill scores [48, 47, 9]. The skill scores from selected results from the publications are listed together with the results from the methods from this thesis in Table 4.3. In the publication from Larson et al. two methods were used to obtain the GHI used to calculate the power generation: using cloud cover from a NWP to calculate GHI (CC) and using GHI directly from a NWP (GHI) [48]. The best results from each method are shown in the table. Bacher et al. used different input features in their statistical model [47]. They made one model using historic power generation (P) as input, one using forecasted GHI from a NWP (GHI), and one using both as input (P + GHI) [47]. El-Baz et al. only used one method in their bagged decision tree method [9].

Table 4.3: *Overview of the skill scores for the methods in this thesis and from methods used in literature. For skill scores, a larger value is better and the highest possible value is 1 [45].*

Method	**Skill score (RMSE)**
Solargis	0.12
Method 1	0.20
Method 2	0.00
Larson et al. [48] (2009) CC	0.23
Larson et al. [48] (2009) GHI	0.25
Bacher et al. [47] (2016) P	0.17
Bacher et al. [47] (2016) GHI	0.36
Bacher et al. [47] (2016) GHI + P	0.36
El-Baz et al. [9] (2018)	0.49

Table 4.3 shows that the latest method by El-Baz et al. provides the highest skill score. Compared to the results in this thesis, all but one of the methods presented here from other publications are superior. However, the case for these methods is not exactly equal, as they are all for small-scale rooftop PV systems and the case in this thesis is a utility-scale power plant. The large PV system used in this thesis is more complex with more inverters, tracking, and curtailment losses. Because of the size of the rooftop systems, they are likely not subject to power curtailment, making the power only dependent on weather conditions. This could partially explain why the skill scores from this thesis are lower than the methods presented here.

Riise et al. made forecasts for GHI at locations in Norway [50]. They also used skill score, however, in their publication, they based the score on MSE instead of RMSE as in this thesis. This highlights the difficulty in comparing forecasting methods with other publications as there is no set way of evaluating the results from the forecasts. Riise et al. obtained a MSE skill score of 0.31 when forecasting with a 24-hour horizon. From Table 4.1 a MSE skill score of 0.41 can be calculated for Method 1. This shows that the method used in this thesis is comparable to other results using a similar approach.

4.2 Implications and potential applications

The results presented in this thesis demonstrate that it is possible to make power generation forecasts with the same level of accuracy as commercial solutions with relatively simple methods and freely available weather forecasts. With much focus on the field of power generation, one can start to look at how good power forecasts can potentially change how solar power is used as a more flexible resource. This discussion is presented with no consideration of the various regulations of different nations' power grids.

As mentioned in Section 2.3, the power grid needs stability, that is, the generated power must at all times equal the consumed power in the grid [5]. This has proven more challenging with the increasing penetration of renewable energy resources in the generation mix [6, 64]. The ability to generate reliable power generation forecasts is therefore important. In Richter et al.'s 2021 publication, they highlight the importance of reliable power generation forecasts for the TSO for several of their functions of planning, maintaining, and running the power grid [6]. The contribution of this thesis can potentially help with day-ahead forecasts that the TSO can use to better plan the power generation mix for the coming day. The method used in this thesis is also easy to implement on small-scale power plants as the main input data is taken from the open weather forecast service *Yr*. With some improvements to the method to increase the accuracy, this method could be useful for small and large PV generators as well as the TSO.

Another useful application would be to reduce the use of curtailment. As stated in Section 2.3.4, curtailment of power generation is used when there is too much power generation in a system [33]. A better knowledge of when power is generated from sources that are dependent on uncontrollable weather conditions is therefore beneficial to avoid discarding large amounts of renewable energy. In an article by Kraiczy et al., the benefits of forecasts are discussed in regard to congestion and reactive power management at the distribution level [64]. They found that accurate

power forecasts would give good congestion forecasts that in turn could lead to a reduction in the use of curtailment [64].

A last possible application with good knowledge about tomorrow's power generation could be to sell flexibility to the grid. Combining this with battery storage could mean the PV power plant could provide flexibility when there is too much generation by charging the batteries instead of discarding the power. Power from the batteries could be delivered to the grid when there is a need for more power generation in the system. For a power plant to be able to sell the generation, it must have reliable forecasts both to report the forecasted generation and the available flexibility.

Chapter 5

Conclusion and further work

The main objective of this thesis was to forecast power generation from a PV power plant. The forecast should forecast the generation 24 hours ahead and have a 1-hour resolution. In this thesis, this was solved with two methods, Method 1 and Method 2, which both used an indirect approach. The first step in the indirect approach for both methods entailed forecasting the GHI with a RFR model using on-site measurements and forecasts from *yr* as input. In method 1 the power generation was obtained using the RFR model again with the forecasted GHI together with on-site measurements and *yr* forecasts as input. Method 2 used the forecasted GHI and a series of physical and empirical operations to obtain the generated power. Both methods were compared to forecasts obtained by the commercial forecast provider Solargis for the same site.

Skill score was chosen as the main metric of evaluation in this thesis. Based on this metric, Method 1 performed best with a skill score of 0.200. Solargis and Method 2 received a score of 0.122 and 0.004 respectively. The positive skill scores indicate that all methods produce forecasts with better RMSE than the smart persistence method.

The results also show that Method 1 had a tendency to underestimate power generation and was unable to correctly forecast the peak production hours. This was reflected in the method's positive MBE of 0.0156 as well as the PAPE score of 15.2% which is the highest for methods used in this thesis. It was found that curtailment of power generation might be the cause of the positive MBE in the forecast from Method 1. This is because the target feature, power generation, has several drops caused by curtailment that can not be explained by the input features. Method 2 got a PAPE of 12.7% and MBE of -0.0182, for Solargis, these scores were 11.4% and -0.0525 respectively. However, on these two metrics, none of the more advanced methods above were able to beat smart persistence which got a PAPE of 10.9% and MSE of 0.0006.

Given the good skill score for Method 1, it is possible to conclude that this is a viable method for power forecasting, especially given more time to refine the method to make it more precise and reliable. With better forecasts, it will be easier for the TSO to operate the electrical power grid with renewable energy on a day-to-day basis. It can also be beneficial for the power plant operators since better grid operation can

lead to less power curtailment. Lastly, it can potentially be a tool that can help the power plant operators sell flexibility to the grid. However, there is much room for further work in this field.

5.1 Further work

There are countless angles to approach the topic of power production, and there are also a large variety of paths to take the work in this thesis onwards. Some interesting paths that will be discussed here are:

- Different time horizons
- Different climate zones
- Account for curtailment

5.1.1 Different time horizons

Different time horizons on the forecast will have different accuracies, as is evident from the report of Banali et al. [7]. They show that shorter horizons (1 hour) have better accuracy than longer (6 hours) [7]. Richter et al. demonstrate how forecasts of different horizons serve different processes that the TSO is responsible for, they look at horizons of 25 minutes up to 1 week ahead [6]. It could, therefore, be interesting to test how the methods used in this thesis would respond to shorter time horizons. With shorter time horizons, one could also expect better weather forecasts, thus better input data to the model. These shorter, more accurate forecasts could then be used to update the power production to the TSO at the intra-day market, giving them a better overview of the production in their grid. There is also possible to test longer horizons. *Yr* only issues 1-hour forecasts up to 48 hours ahead, however, they do have up to 9 days ahead with a 6-hour resolution. This could give a rough estimate of next week's power production.

In relation to the forecast horizons, there are also varying practices for when the forecasts are being issued. In this thesis, the forecasts are issued at midnight and for the following day. If the forecasts made in this forecast are going to be reported to a TSO, it needs to be made in time for their deadline. So if the TSO has set the deadline at 18:00, the forecast for the following day must be made by 18:00.

5.1.2 Different climate zones

Another aspect is to look at how the model would work in different climate zones. As Riise et al. demonstrated in their report, there are differences in accuracy if a model is trained in one region of Norway and applied to another region within Norway [50]. Thus, changing climate zone will likely also make an impact on the result. Training the model in Norway and applying it in the deserts in Egypt, might be challenging because of the large difference in the climate. However, finding the best model parameters and features for the site in this thesis and doing the training on another site might be possible. If this is viable, it could greatly reduce the time needed to make optimize the model for every single site. Alternatively, one could

train the model on several time series from different climate zones initially, thus making the model familiar to all relevant climate zones.

5.1.3 Account for curtailment

Lastly, in future work, it would also be natural to look at the problems following curtailment. The methods in this thesis are not able to forecast the curtailment or differentiate between days with and without curtailment. Three possible approaches to reduce the effects of curtailment could be:

- only train the model on days without any curtailment losses,
- add a boolean variable stating when there were losses in the training data,
- add an estimate of the lost power generation to the actual generation on times with curtailment.

However, these approaches do not come without problems of their own. The first approach could end up cutting out a lot of the data. 30% of the data is already lost because of missing values, so this would reduce the data available to build the model even more. This would not be a problem with the second approach. Catching the effect of curtailment with a boolean variable, could, however, be difficult. With a boolean variable indicating when the curtailment losses are, there is no way of knowing how much power production is lost. The model might be adaptable enough to figure it out, however, that must be tested in a future project. The last approach would give a time series of generation data that seems unaffected by curtailment losses and has a good potential of giving a good basis for training the model. However, with 11.4% curtailment loss, a significant portion of the data the model is trained on would have to be estimated instead of using true measurement values. This could lead to replacing one bias in the data caused by curtailment losses with a new bias caused by estimated values. These three options demonstrate that accounting for curtailment in future methods is attainable, however, all these approaches will have some drawbacks that must be taken into account.

It is evident that there are many ways of making the power system more efficient through the use of machine learning models.

Bibliography

[1] United Nations. *What Is Climate Change? [Web]*. Accessed on: 19/04/2022. URL: https://www.un.org/en/climatechange/what-is-climate-change.

[2] National Aeronautics and Space Administration (NASA). *Global Warming vs. Climate Change [Web]*. Accessed on: 01/05/2022. URL: https://climate.nasa.gov/global-warming-vs-climate-change/.

[3] H. Ritchie, M. Roser, and P. Rosado. "Energy [Web]". In: *Our World in Data* (2022). Accessed on 17/04/2023. URL: https://ourworldindata.org/energy.

[4] United Nations. *Renewable energy – powering a safer future [Web]*. Accessed on: 19/04/2022. URL: https://www.un.org/en/climatechange/raising-ambition/renewable-energy.

[5] A. V. Meier. *Electric Power Systems: a Conceptual Introduction*. Wiley Interscience, 2006. ISBN: 978-0-471-17859-0.

[6] A. Richter, T. Schroeter, and M. Wolter. "Importance of TSO Forecast in Power System Processes – Challenges in Load, Generation, Storage and Sector Coupling Forecast". In: *ETG Congress 2021* (2021), pp. 402–407.

[7] L. Benali, G. Notton, A. Fouilloy, C. Voyant, and R. Dizene. "Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components". In: *Renewable Energy* 132 (2019), pp. 871–884. DOI: 10.1016/j.renene.2018.08.044.

[8] B. Babar, L. T. Luppino, T. Boström, and S. N. Anfinsen. "Random forest regression for improved mapping of solar irradiance at high latitudes". In: *Solar Energy* 198 (2020), pp. 81–92. DOI: 10.1016/j.solener.2020.01.034.

[9] W. El-Baz, P. Tzscheutschler, and U. Wagner. "Day-ahead probabilistic PV generation forecast for buildings energy management systems". In: *Solar Energy* 171 (2021), pp. 478–490. DOI: 10.1016/j.solener.2018.06.100.

[10] A. Smets, K. Jäger, O. I. R. V. Swaaij, and M. Zeman. *Solar Energy - The physics and engineering of photovoltaic conversion technologies and systems*. UIT Cambridge Ltd, 2016. ISBN: 9781906860325.

[11] E. F. Camacho, M. Berenguel, F. R. Rubio, and D. Martínez. *Control of Solar Energy Systems*. Springer, 2012.

[12] C. D. Ahrens and R. Henson. *Essentials of Meteorology - An Invitation to the Atmosphere*. Cenage Learning, 2017. ISBN: 978-1-305-62845-8.

[13] Degreen. *File:Sonne Strahlungsintensitaet.svg [Web]*. Accessed on 15/04/2023. Wikimedia commons. URL: https://commons.wikimedia.org/wiki/File:Sonne_Strahlungsintensitaet.svg.

[14] A. Knudby. *Remote Sensing*. Pressbooks, 2021.

[15] National Oceanic and Atmospheric Administration. *The Color of Clouds [Web]*. Accessed on: 28/04/2022. URL: https://www.noaa.gov/jetstream/clouds/color-of-clouds.

[16] L. Gardiner. *Cloud Types [Web]*. Accessed on 02/05/2023. UCAR Center for Science Education. URL: https://scied.ucar.edu/learning-zone/clouds/cloud-types.

[17] E. L. Maxwell. *A quasi-physical model for converting hourly global horizontal to direct normal insolation*. 1987. URL: https://www.nrel.gov/docs/legosti/old/3087.pdf.

[18] European Commission. *Solar energy [Web]*. Accessed on: 17/04/2022. URL: https://energy.ec.europa.eu/topics/renewable-energy/solar-energy_en.

[19] Our World in Data. *Solar (photovoltaic) panel prices*. Accessed on 17/04/2023. 2022. URL: https://ourworldindata.org/grapher/solar-pv-prices.

[20] International Energy Agency (IEA). "Snapshot of Global PV Markets 2023". In: *Photovoltaic Power System Programme* (2023). Accessed on 28/04/2023. URL: https://iea-pvps.org/wp-content/uploads/2023/04/IEA_PVPS_Snapshot_2023.pdf.

[21] W. L. Hosch. *Photovoltaic effect [Web]*. Accessed on 07/05/2023. Britannica. URL: https://www.britannica.com/science/photovoltaic-effect.

[22] A. Richter, M. Hermle, and S. W. Glunz. "Reassessment of the Limiting Efficiency for Crystalline Silicon Solar Cells". In: *IEEE Journal of Photovoltaics* 3 (2013), pp. 1184–1191. DOI: 10.1109/JPHOTOV.2013.2270351.

[23] L. Mæhlum and K. A. Rosvold. *Solceller [Web]*. Accessed on 13/05/2023. Store Norske Leksikon. URL: https://snl.no/solceller.

[24] pvlib. *Calculating a module's IV curves [Web]*. Accessed on: 28/04/2023. URL: https://pvlib-python.readthedocs.io/en/v0.9.0/auto%5C_examples/plot%5C_singlediode.html.

[25] J. Smalley. *How does solar backtracking make projects more productive? [Web]*. Accessed on 2023/15/04. Solar Power World. URL: https://www.solarpowerworldonline.com/2015/07/how-does-solar-backtracking-make-projects-more-productive/.

[26] W. F. Holmgren and C. W. H. andand Mark A. Mikofski. "pvlib python: a python package for modeling solar energy systems". In: *Journal of Open Source Software* 3.29 (2018), p. 884. DOI: 10.21105/joss.00884.

[27] National Renewable Energy Laboratory (NREL). *Solar Power and the Electric Grid*. 2010. URL: https://www.nrel.gov/docs/fy10osti/45653.pdf.

[28] Europead Distribution System Operators. *What Is A DSO? [Web]*. Accessed on 13/05/2023. URL: https://www.edsoforsmartgrids.eu/about-dsos/what-is-a-dso.

[29] K. Hofstad. *Transmission System Operators [Web]*. Accessed on 22/04/2023. Store Norske Leksikon. URL: https://snl.no/Transmission_System_Operators.

[30] J. D. Glover, T. J. Overbye, and M. S. Sarma. *Power System Analysis & Design*. Cengage Learning, 2016. ISBN: 978-1-305-63618-7.

[31] Statnett SF. *Introduksjon til Statnett sine reservemarkeder*. Accessed on 23/04/2023. URL: https://www.statnett.no/for-aktorer-i-kraftbransjen/systemansvaret/kraftmarkedet/reservemarkeder/introduksjon-til-reserver/.

[32] Olje- og energidepartementet. *Kraftmarkedet*. Accessed on 24/04/2023. URL: https://energifaktanorge.no/norsk-energiforsyning/kraftmarkedet/.

[33] Next Kraftwerke. *What is Curtailment of Electricity? [Web]*. Accessed on 24/04/2023. URL: https://www.next-kraftwerke.com/knowledge/curtailment-electricity.

[34] S. Raschka and V. Mirjalili. *Python Machine Learning*. Pact Publishing Ltd., 2019. ISBN: 978-1-78995-575-0.

[35] IBM. *What is machine learning?* Accessed on 11/05/2023. URL: https://www.ibm.com/topics/machine-learning.

[36] Colombia Mailman School of Public Health. *Missing Data and Multiple Imputation [Web]*. Accessed on 03/05/2023. URL: https://www.publichealth.columbia.edu/research/population-health-methods/missing-data-and-multiple-imputation.

[37] N. Mittag. *Imputations: Benefits, Risks and a Method for Missing Data*. 2013. URL: https://home.cerge-ei.cz/mittag/papers/Imputations.pdf.

[38] R. J. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2021. ISBN: 978-0987507136.

[39] M. B. Kursa and W. R. Rudnicki. *Feature Selection with the Boruta Package*. 2010. DOI: 10.18637/jss.v036.i11.

[40] F. Degenhardt, S. Seifert, and S. Szymczak. "Evaluation of variable selection methods for random forests and omics data sets". In: *Briefings in Bioinformatics* 20 (2019), pp. 492–503. DOI: 10.1093/bib/bbx124.

[41] S. Manoj. *Discovering the shades of Feature Selection Methods [Web]*. Accessed on 30/04/2023. Analytics Vidhya. URL: https://www.analyticsvidhya.com/blog/2021/04/discovering-the-shades-of-feature-selection-methods/.

[42] M. Padhma. *End-to-End Introduction to Evaluating Regression Models [Web]*. Accessed on 15/04/2023. Analytics Vidhya. URL: https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/.

[43] S. Dai, F. Meng, H. Dai, Q. Wang, and X. Chen. "Electrical peak demand forecasting- A review". In: *(unpublished)* (2021). DOI: 10.48550/arXiv.2108.01393.

[44] E. Wheatcroft. "Interpreting the skill score form of forecast performance metrics". In: *International Journal of Forecasting* 35 (2019), pp. 573–579. DOI: 10.1016/j.ijforecast.2018.11.010.

[45] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Pison, and F. Antonanzas-Torres. "Review of photovoltaic power forecasting". In: *Solar Energy* 136 (2016), pp. 78–111. DOI: 10.1016/j.solener.2016.06.069.

[46] P. Gupta and R. Singh. "PV power forecasting based on data-driven models: a review". In: *International Journal of Sustainable Engineering* 14 (2021), pp. 1733–1755. DOI: 10.1080/19397038.2021.1986590.

[47] P. Bacher, H. Madsen, and H. A. Nielsen. "Online short-term solar power forecasting". In: *Solar Energy* 83 (2009), pp. 1772–1783. DOI: 10.1016/j.solener.2009.05.016.

[48] D. P. Larson, L. Nonnenmacher, and C. F. Coimbra. "Day-ahead forecasting of solar power output from photovoltaic plants in the American Southwest". In: *Renewable Energy* 91 (2016), pp. 11–20. DOI: 10.1016/j.renene.2016.01.039.

[49] S. Khillar. *Difference Between Bagging and Random Forest [Web]*. Accessed on 10/05/2023. Difference Between. URL: http://www.differencebetween.net/technology/difference-between-bagging-and-random-forest/.

[50] H. N. Riise, Ø. S. Klyve, A. Dobler, and M. M. Nygård. *Irradiance predictions from global and public weather forecasts*.

[51] J. Gjessing. Personal comunication. 11/05/2023.

[52] G. van Rossum and F. L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

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

[54] J. Herzen, F. Lässig, S. G. Piazzetta, T. Neuer, L. Tafti, G. Raille, T. V. Pottelbergh, M. Pasieka, A. Skrodzki, N. Huguenin, M. Dumonal, J. Kościsz, D. Bader, F. Gusset, M. Benheddi, C. Williamson, M. Kosinski, M. Petrik, and G. Grosch. "Darts: User-Friendly Modern Machine Learning for Time Series". In: *Journal of Machine Learning Research* 23.124 (2022), pp. 1–6. URL: http://jmlr.org/papers/v23/21-1177.html.

[55] Norwegian University of Life Sciences. *Orion [Web]*. Accessed on 12/05/2023. URL: https://orion.nmbu.no/.

[56] Yr. *Location forecast data model [Web]*. Accessed on 13/03/2022. URL: https://developer.yr.no/doc/locationforecast/datamodel/.

[57] Solargis. *Solargis*. Accessed on: 27/02/2022. URL: https://solargis.com/.

[58] Solargis. *Methodology - Solar radiation modeling[Web]*. Accessed on: 27/02/2023. URL: https://solargis.com/docs/methodology/solar-radiation-modeling.

[59] Solargis. *Methodology - Meteorological models and post-processing[Web]*. Accessed on: 27/02/2022. URL: https://solargis.com/docs/methodology/meteo-data.

[60] A. P. Gulati. *Dealing with outliers using the Z-Score method [Web]*. Accessed on 12/05/2023. Analytics Vidhya. URL: https://www.analyticsvidhya.com/blog/2022/08/dealing-with-outliers-using-the-z-score-method/.

[61] Codecademy Team. *Normalization [Web]*. Accessed on 17/04/2023. Codecademy. URL: https://www.codecademy.com/article/normalization.

[62] J. Korstanje. *Advanced Forecasting with Python With State-of-the-Art-Models Including LSTMs, Facebook's Prophet, and Amazon's DeepAR*. Apress, 2021. ISBN: 978-1-4842-7150-6.

[63] K. F. Frøslie. *Korrelasjon [Web]*. Accessed on 10/05/2023. Store Norske Leksikon. URL: `https://snl.no/korrelasjon`.

[64] M. Kraiczy, A. Altayara, F. Wenderoth, K. Winter, P. Hofbauer, S. Meilinger, and M. Braun. "Benefits of advanced PV power forecasts for congestion management and reactive power management at the distribution level". In: *ETG Congress 2021* (2021), pp. 77–82. ISSN: 978-3-8007-5549-3.

Appendix

The code used in this thesis can be accessed on GitLab on:

https://gitlab.com/sigridvo/master_thesis

Figure 5.1: *The results after 10 iterations with the Boruta algorithm for the GHI forecasts. All features receiving rank 1 at least once were used in the forecasts.*

	Temperature forecast	Wind speed forecast	Wind Direction forecast	Humidity forecast	Pressure forecast	Clouds forecast	Fog forecast	Low Clouds forecast	Medium clouds forecast	High clouds forecast	Dewpoint temperature forecast	Precipitation forecast	clear sky	Smart persistence
0	1	3	1	1	1	4	6	1	1	5	1	2	1	1
1	1	2	1	1	1	1	5	1	1	3	1	3	1	1
2	1	4	1	1	1	4	7	1	1	6	2	3	1	1
3	1	3	1	1	1	4	6	1	1	5	1	2	1	1
4	1	3	1	1	1	5	7	1	1	4	2	6	1	1
5	1	3	1	1	1	4	6	1	1	5	1	2	1	1
6	1	5	1	1	1	3	7	1	1	6	3	2	1	1
7	1	3	1	1	1	6	7	1	1	4	2	4	1	1
8	1	3	1	1	1	1	5	1	1	4	2	1	1	1

Figure 5.2: *The results after 10 iterations with the Boruta algorithm for the power generation forecasts. All features receiving rank 1 at least once were used in the forecasts.*

	Temperature forcast	Wind speed forcast	Wind Direction forcast	Humidity forcast	Pressure forcast	Clouds forcast	Fog forcast	Low Clouds forcast	Medium clouds forcast	High clouds forcast	Dewpoint temperature forcast	Precipitation forcast	clear sky	tracker	GHI smart persistence	GHI forcast
0	3	1	1	1	2	1	8	6	1	4	5	7	1	1	1	1
1	4	1	1	1	1	2	8	6	1	3	5	7	1	1	1	1
2	4	1	1	1	1	1	8	6	1	3	5	7	1	2	1	1
3	3	1	1	1	1	1	7	5	1	2	4	6	1	1	1	1
4	4	1	1	1	1	3	8	6	1	2	5	7	1	1	1	1
5	2	1	1	1	1	1	6	4	1	1	3	5	1	1	1	1
6	4	1	1	1	1	1	7	5	1	2	3	6	1	1	1	1
7	3	1	1	1	1	1	7	5	1	2	3	6	1	1	1	1
8	4	1	1	1	3	1	1	9	7	2	5	5	8	1	1	1
9	4	1	1	1	1	2	1	8	5	1	3	6	7	1	1	1

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway